LREC-COLING 2024

ParlaCLARIN IV
Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora

Proceedings

Editors
Darja Fišer, Maria Eskevich, David Bordon

May 20, 2024
Torino, Italia

Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora

ISBN 978-2-493814-24-1
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association
and the International Committee on Computational Linguistics

ParlaCLARIN IV @ LREC-COLING2024: Introduction

Parliamentary data is an important source of scholarly and socially relevant content, serving as a verified communication channel between the elected political representatives and members of the society. The development of accessible, comprehensive and well-annotated parliamentary corpora is therefore crucial for the information society, as such corpora help scientists and investigative journalists to ascertain the accuracy of socio-politically relevant information, and to inform the citizens about the trends and insights on the basis of such data explorations. Research-wise, parliamentary corpora are a quintessential resource for a number of disciplines in digital humanities and social sciences, such as political science, sociology, history, and (socio)linguistics.

The distinguishing characteristic of parliamentary data is that it is spoken language produced in controlled circumstances. Such data has traditionally been transcribed in a formal way but is now also increasingly transcribed with speech-to-text software as well as released in the original audio and video formats, which encourages resource and software development and provides research opportunities related to structuring, synchronisation, visualisation, querying and analysis of parliamentary corpora. Therefore, a harmonised approach to data curation practices for this type of data can support the advancement of the field significantly. One of the ways in which the research community is supported in this line of work is through the conversion of existing corpora and further development of new cross-national parliamentary corpora into a highly comparable, harmonised set of multilingual resources. These allow researchers to share comparative perspectives and to perform multidisciplinary research on parliamentary data. We envision that the ParlaCLARIN IV workshop, as a venue for knowledge and experience exchange on the topic, will contribute to the development and growth of the field of digital parliamentary science.

This fourth ParlaCLARIN workshop is a continuation of the 2018[1], 2020[2] and 2022[3] editions held at the respective LREC conferences, see references below. On the one hand, it continues to bring together developers, curators and researchers of regional, national and international parliamentary debates from across diverse disciplines in the Humanities and Social Sciences. On the other hand, we envisage the appearance of new discussion threads, tasks, and challenges that are partially inspired by or related to the new data releases such as ParlaMint[4] and data formats such as ParlaCLARIN[5] .

The Call for Papers has invited original, overview and position papers with the focus on one of the following topics:

- Compilation, annotation, visualisation and utilisation of historical or contemporary parliamentary written or audio records
- Harmonisation of existing multilingual parliamentary resources, containing either synchronic or diachronic data or both
- Linking or comparing of parliamentary records with other datasets of political discourse such as party manifestos, political speeches, political campaign debates, and social media posts, and to other sources of structured knowledge, such as formal ontologies and LOD datasets (in particular for the description of speakers, political parties, etc.)

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

[2] https://www.clarin.eu/ParlaCLARIN-II

[3] https://www.clarin.eu/ParlaCLARIN-III

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

[5] https://github.com/clarin-eric/parla-clarin

In 2024 the following special themes were also brought for discussion at the workshop:

- Enrichment of parliamentary proceedings (with e.g. sentiment annotation, political profiling of speakers etc.) and research using such data
- Machine translation of parliamentary proceedings and research using such data
- Argument mining of parliamentary debates

The workshop programme is composed of a keynote talk by Ines Rehbein from the Universität Mannheim and 24 peer-reviewed papers (of which 8 are presented as posters and 5 as demos) by 69 authors from 15 countries (the three most represented: Germany (5), Slovenia (4) and Czech Republic (3)). Two papers report on the work that was carried out by the co-authors representing the institutions in more than one country.

We would like to thank the reviewers for their careful and constructive reviews which have contributed to the quality of the event.

The ParlaCLARIN IV workshop was held in person with the a possibility of hybrid attendance in Turin (Italy), as part of the 2024 Joint International Confrence on Computational Linguistics, Language Resources and Evaluation (LREC-COLING2024).

D. Fišer, M. Eskevich, D. Bordon May 2024

Organizing Committee

- Darja Fišer, Institute of Contemporary History & CLARIN
- Maria Eskevich, Huygens Institute, KNAW
- David Bordon, University of Ljubljana

Program Committee

- Kaspar Beelen, The Alan Turing Institute, GB
- Siddharth Bhargava, Fondazione Bruno Kessler, IT
- Andreas Blaette, University of Duisburg-Essen, DE
- Hajo Boomgaarden, University of Vienna, AT
- Robert Borges, Uppsala University, SE
- Çağrı Çöltekin, University of Tübingen, DE
- Tomaž Erjavec, Dept. of Knowledge Technologies, Jožef Stefan Institute, SI
- Francesca Frontini, Istituto di Linguistica Computazionale "A. Zampolli" - ILC Consiglio Nazionale delle Ricerche - CNR, IT
- Maria Gavriilidou, ILSP / Athena RC, GR
- Turo Hiltunen, University of Helsinki, FI
- Pasi Ihalainen, University of Jyväskylä, FI
- Tatsuya Kawahara, Kyoto University, JP
- Haidee Kotze, Utrecht University, NL
- Anna Kryvenko, NISS (Ukraine); INZ (Slovenia), UA
- Cristina Lastres-López, University of Seville, ES
- Bente Maegaard, University of Copenhagen, DK
- Christian Mair, University of Freiburg, DE
- Maarten Marx, University of Amsterdam, NL
- Monica Monachini, Institute of Computational Linguistics "A. Zampolli" - CNR, IT
- Jan Odijk, Utrecht University, NL
- Maciej Ogrodniczuk, Institute of Computer Science, Polish Academy of Sciences, PL
- Petya Osenova, Sofia University "St. Kl. Ohridski" and IICT-BAS, BG
- Stelios Piperidis, Athena RC/ILSP, GR
- Maria Pontiki, Institute for Language and Speech Processing (ILSP), Athena R.C., Greece, GR

- Simone Paolo Ponzetto, University of Mannheim, DE
- Valeria Quochi, Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale "A. Zampolli", IT
- Hugo Sanjurjo-González, University of Deusto, ES
- Sara Tonelli, FBK, IT
- Turo Vartiainen, University of Helsinki, FI
- Tanja Wissik, Austrian Academy of Sciences, AT

Invited Speaker

- Ines Rehbein, University of Mannheim Data and Web Science Group, DE

Table of Contents

ParlaCLARIN IV Workshop Program

20 May 2024

9:00–9:10	**Welcome and Introduction**
9:10–10:30	**ParlaMint**
9:10–9:30	*Parliamentary Discourse Research in Political Science: Literature Review* Jure Skubic and Darja Fišer
9:30–9:50	*Compiling and Exploring a Portuguese Parliamentary Corpus: ParlaMint-PT* José Aires, Aida Cardoso, Rui Pereira and Amalia Mendes
9:50–10:10	*Gender, Speech, and Representation in the Galician Parliament: An Analysis Based on the ParlaMint-ES-GA Dataset* Adina I. Vladu, Elisa Fernández Rei, Carmen Magariños and Noelia García Díaz
10:10–10:30	*Bulgarian ParlaMint 4.0 corpus as a testset for Part-of-speech tagging and Named Entity Recognition* Petya Osenova and Kiril Simov
11:00–12:00	**Keynote**
11:00–12:00	*Resources and Methods for Analysing Political Rhetoric and Framing in Parliamentary Debates* Ines Rehbein

20 May 2024 (continued)

12:00–12:40	**Creation of Parliamentary Language Resources**
12:00–12:20	*PTPARL-V: Portuguese Parliamentary Debates for Voting Behaviour Study* Afonso Sousa and Henrique Lopes Cardoso
12:20–12:40	*Polish Round Table Corpus* Maciej Ogrodniczuk, Ryszard Tuora and Beata Wójtowicz
14:00–15:00	**Analysis of Parliamentary Discourse**
14:00–14:20	*Investigating Multilinguality in the Plenary Sessions of the Parliament of Finland with Automatic Language Identification* Tommi Jauhiainen, Jussi Piitulainen, Erik Axelson, Ute Dieckmann, Mietta Lennes, Jyrki Niemi, Jack Rueter and Krister Lindén
14:20–14:40	*Exploring Word Formation Trends in Written, Spoken, Translated and Interpreted European Parliament Data – A Case Study on Initialisms in English and German* Katrin Menzel
14:40–15:00	*Quantitative Analysis of Editing in Transcription Process in Japanese and European Parliaments and its Diachronic Changes* Tatsuya Kawahara
15:00–15:40	**Language Technology for Parliamentary Discourse**
15:00–15:20	*Automated Emotion Annotation of Finnish Parliamentary Speeches Using GPT-4* Otto Tarkka, Jaakko Koljonen, Markus Korhonen, Juuso Laine, Kristian Martiskainen, Kimmo Elo and Veronika Laippala
15:20–15:40	*Making Parliamentary Debates More Accessible: Aligning Video Recordings with Text Proceedings in Open Parliament TV* Olivier Aubert and Joscha Jäger

20 May 2024 (continued)

15:40–16:00 Poster pitches

16:30–17:45 Poster session

16:30–17:45 *Russia and Ukraine through the Eyes of ParlaMint 4.0: A Collocational CADS Profile of Spanish and British Parliamentary Discourses*
Maria Calzada Perez

16:30–17:45 *Multilingual Power and Ideology identification in the Parliament: a reference dataset and simple baselines*
Çağrı Çöltekin, Matyáš Kopp, Meden Katja, Vaidas Morkevicius, Nikola Ljubešić and Tomaž Erjavec

16:30–17:45 *IMPAQTS: a multimodal corpus of parliamentary and other political speeches in Italy (1946-2023), annotated with implicit strategies*
Federica Cominetti, Lorenzo Gregori, Edoardo Lombardi Vallauri and Alessandro Panunzi

16:30–17:45 *ParlaMint Ngram viewer: Multilingual Comparative Diachronic Search Across 26 Parliaments*
Asher de Jong, Taja Kuzman, Maik Larooij and Maarten Marx

16:30–17:45 *Investigating Political Ideologies through the Greek ParlaMint corpus*
Maria Gavriilidou, Dimitris Gkoumas, Stelios Piperidis and Prokopis Prokopidis

16:30–17:45 *ParlaMint in TEITOK*
Maarten Janssen and Matyáš Kopp

16:30–17:45 *Historical Parliamentary Corpora Viewer*
Alenka Kavčič, Martin Stojanoski and Matija Marolt

16:30–17:45 *The dbpedia R Package: An Integrated Workflow for Entity Linking (for ParlaMint Corpora)*
Christoph Leonhardt and Andreas Blaette

16:30–17:45 *Video Retrieval System Using Automatic Speech Recognition for the Japanese Diet*
Mikitaka Masuyama, Tatsuya Kawahara and Kenjiro Matsuda

16:30–17:45 *One Year of Continuous and Automatic Data Gathering from Parliaments of European Union Member States*
Ota Mikušek

20 May 2024 (continued)

16:30–17:45 *Government and Opposition in Danish Parliamentary Debates*
Costanza Navarretta and Dorte Haltrup Hansen

16:30–17:45 *A new Resource and Baselines for Opinion Role Labelling in German Parliamentary Debates*
Ines Rehbein and Simone Paolo Ponzetto

16:30–17:45 *ParlaMint Widened: a European Dataset of Freedom of Information Act Documents (Position Paper)*
Gerda Viira, Maarten Marx and Maik Larooij

17:45–18:00 Closing Remarks

Parliamentary Discourse Research in Political Science: Literature Review

Jure Skubic, Darja Fišer
Institute of Contemporary History
Privoz 11, 1000 Ljubljana
jure.skubic@inz.si, darja.fiser@inz.si

Abstract

One of the major research interests for political science has always been the study of political discourse and parliamentary debates. This literature review offers an overview of the most prominent research methods used in political science when studying political discourse. We identify the commonalities and the differences of the political science and corpus-driven approaches and show how parliamentary corpora and corpus-based approaches could be successfully integrated in political science research.

Keywords: parliamentary discourse, political science, parliamentary corpora

1. Introduction

Parliamentary debates are one of the best sources of information about political discourse, which is inherently valuable for research in the humanities and social sciences. Especially political science is particularly involved in the analysis of political power and authority exercised through parliamentary discourse.

This literature review is part of a series of literature reviews produced as part of the ParlaMint project (Erjavec et al., 2022). Similar reviews have been compiled for sociology (Skubic and Fišer, 2022) and history (Skubic and Fišer, 2022) and are important for better understanding how the humanities and social sciences use qualitative and quantitative research methods in analyzing parliamentary discourse. The ParlaMint project has developed comparable corpora of parliamentary transcripts for more than 20 European countries and offered literature reviews, showcases, and tutorials mentioned earlier to promote the use of the corpora in a wide range of scholarly communities interested in the study of parliamentary discourse and debate. In this paper, we review existing political science research focusing on written parliamentary records and the commonly used research methods. We view these approaches as complementary to other common political science research techniques and types of data sources such as surveys, records of election results, media content, etc.

This literature review is organized as follows. In the first part, we outline the selection process of relevant papers and explain the research methods they employ. In the second part, we summarize each of the selected papers in terms of 1) the research topic, 2) the data collection, 3) the research method, and 4) a brief discussion of possible improvements to the research. We conclude the review with a discussion of how this area of political science could benefit from the use of corpus data and the use of corpus-assisted research methods or other text mining methods.

2. Political Science Methods

Parliamentary discourse is an important focus of political science research at the (inter)national or local level. Like many other social sciences, it draws on and complements various methodological traditions in the study of politics and governance, legislation, and political discourse to increase the relevance and reliability of its research findings (Lauer, 2021). The methodological pluralism of political science allows it to address contemporary issues and problems that arise in the broad field of social sciences in general (Franco et al., 2021), and to focus on topics that might go unaddressed in other social science disciplines. Although political science has in recent years taken a quantitative turn (ibid.), qualitative methods and approaches are still widely used, prove to be highly effective, and provide meaningful insights into important research questions.

Blaxill (2022) notes that political scientists are interested in language and discourse as a means of studying political power, change, institutions, etc. Since political discourse is about the text and speech of professional politicians and political institutions (van Dijk, 1997), documentary sources are a valuable source of data for political science. Documents (texts, laws, etc.) are usually collected from official websites or archives of relevant organizations (parliaments, libraries, etc.) or by visiting archives, bureaus, and other organizations (Franco, 2021). In addition, political scientists often triangulate data and metadata collected from official parliamentary minutes and policy texts with data from other sources such as interviews, (social) media, newspapers, etc. This makes the ParlaMint corpora directly relevant for political science researchers.

3. Literature Selection

3.1 Selection of Papers

When selecting relevant papers for this literature review, the following criteria were followed. We used the following scholarly search engines to search for relevant papers:

- Elsevier (https://www.elsevier.com)
- Project MUSE (https://muse.jhu.edu)

- SAGE Journals (https://journals.sagepub.com)
- Springer Link (https://link.springer.com)
- Taylor and Francis Online (https://www.tandfonline.com)
- Wiley Online Library (https://onlinelibrary.wiley.com)

We searched for keywords such as "parliamentary discourse", "parliamentary debates" and "parliamentary proceedings" and applied the following filters to narrow down search options:

- Publication period: 2012 – 2022,
- Discipline: political science,
- Article ranking: "most relevant" or "most cited",
- Relevant journals: additional filters were sometimes needed to search papers in relevant journals.

Because the number of papers was still high, we performed an additional selection process, analyzing the abstract, topic, data collection, and methods used for each paper. At this stage, many papers were screened out either because of a lack of methodological explanation or because the research did not focus on parliamentary data. We focused only on papers that specifically addressed parliamentary and/or legislative documents or interwove them with other data sources. After completing the selection process, we selected 24 relevant papers from the following political science journals: Parliamentary Affairs, British Politics, French Politics, Comparative European Politics, Political Communication, Ethnicities, The British Journal of Politics and International Relations, Australian Journal of Political Science, Political Analysis Journal, and Journal of Contemporary European Studies.

3.2 Overview of Methods

All 24 papers relevant for this review are listed in the Google Spreadsheet.[1] We thoroughly analyzed all of them and paid special attention not only to the data selection process or the methods employed, but also to the thematic focus of each paper. In all the reviewed papers, methods and data selection process were clearly explained and all of them used parliamentary records as the main source of data. The research questions of the analyzed papers were highly heterogeneous, so we decided not to group the papers thematically. Out of 24 analyzed papers, 12 employed content analysis, 3 (critical) discourse analysis, 2 sentiment analysis, 2 thematic analysis, 2 papers employed a mixed methods approach and 3 papers employed one of the many text-as-data approaches (1 paper social network analysis and 2 papers quantitative text analysis with supervised machine learning techniques). Due to methodological and in a few cases thematic similarities of some papers, we decided not to include all 24 papers in this review but analyze no more than 2 representative papers for each methodological approach.

4. Reviewed Research and Methods

4.1 Content Analysis

Content analysis (CA) is one of the most widely used research techniques in social sciences and its main goal is to analyze data in a specific context and extract meaningful information from the analyzed documents (Krippendorff, 2018). According to Blassnig (2022), it is perhaps one of the most used methods in the field of political science, mainly due to the general influence of other disciplines such as sociology, history, philosophy, etc. It is used to interpret textual data through the process of coding and identifying themes or patterns (Lilja, 2021), and to analyze the self-representation of political actors through the analysis of political and parliamentary speeches, debates, party platforms, etc. In political science, researchers often decide for triangulation of content analysis with other either qualitative (e.g., CDA) or quantitative (often digital) methods. Due to advances in computational research approaches, content analysis is becoming more and more digitized with researchers using computer software to systematically import and analyze large volume of text documents without spending considerable amount of time reading or paying for expensive coding (Provalis research, 2019). Although qualitative content analysis still prevails in political science, quantitative content analysis is once again gaining recognition and is becoming increasingly more popular.

4.1.1 Emotions in EU Parliamentary Debates

Research problem: The aim of Sanchez Salgado's (2021) paper was to explore the verbal display and role of emotions in the European Parliament (EP). She analyzed how emotions are expressed inside the EP and how they reflect not only power but also status dynamics.

Data collection: The data for her research consisted of 25 plenary debates in English, French, Dutch and Spanish that took place in EP between 2009 and 2017. The author focused on two topics in EP in which she expected emotions to play a crucial role: the financial crisis (2009 – 2014, 14 debates) and the refugee crisis (2014 – 2017, 11 debates) as the two most challenging crises that the EU had faced before 2020. She accessed the debates on the website of the EP in September 2017. For the first she selected those debates which included "economic crisis" or "financial crisis" in their title as for the latter she used the debates the title of which included the words "migration" and "refugees". The automatic coding she employed only included keywords which correspond to basic primary and secondary emotions as defined by Parrot (2001). She was particularly interested in analysis of emotional patterns and structures which

[1] Those papers can be found in the first sheet in the Google Spreadsheet titled "*All papers*". The second sheet, titled "*Papers selected for report*" includes papers, which are in detail discussed below. Link to the spreadsheet: https://docs.google.com/spreadsheets/d/1dd9YCDs9G53NBxxg0Bxhfbjx3QxjWjgN4tPjCLw2WVg/edit#gid=0.

were identified through an in-depth reading of all the debates in which emotion markers were used within their discursive context.

Research method: The author opted for an in-depth (qualitative) content analysis of 25 debates in EP in which she observed explicit emotion keywords present in discourses. For qualitative content analysis of emotions, she used the Atlas.ti[2] data analysis software (ATLAS.ti Scientific Software Development GmbH, 1993), which contributed to the efficiency, consistency, and transparency of her analysis. In her Atlas.ti analysis she considered only emotion keywords, whereas the in-depth contextual analysis accounted for all types of implicit and explicit references to emotions.

Discussion: Sanchez Salgado's research is one in a few which focuses on international (EU) parliamentary debates. What could be seen as a potential shortcoming of the research is in author not elaborating on why she specifically chose debates in those languages and not any other. She points out that the transcriptions since 2012 are not available in English, which could be seen as a limitation, however it also shows how emotions are expressed in various languages.

4.1.2 Exploring Feminist Arguments in German Parliamentary Debates

Research problem: Och (2019) analyzed the parliamentary discourse around two instances of feminist policy adoption in two conservative German governments. She showed that in both analyzed governments feminist arguments dominated the debates.

Data collection: Och analyzed documents from the 16th (2006) and 18th (2015) legislative period of the German parliament. She identified suitable documents with the help of document and information system for parliamentary processes of the Bundestag. This system returned all parliamentary documents linked to respective bills, which included verbatim protocols of plenary debates in both chambers as well as verbatim protocols of the committee hearings and bill documents presented by the federal government to parliament for information purposes or in response to parliamentary questions. She also included documents published by Federal Ministries for Family Affairs, Senior Citizens, Women and Youth, Federal Ministry of Justice and Consumer Affairs as well as statements and speeches by the responsible ministers if they directly discussed the respective policy.

Research method: The author employed qualitative content analysis on a series of parliamentary documents of the German parliament by reading all the documents and coding them by hand to identify statements that contained arguments of either of the two broad coding categories: utility-driven arguments and feminist arguments. She coded arguments as utility-driven if the policy was justified as a means to a non-feminist end and as feminist if they showed feminist attitudes and behavior (referring to gender equality, sex-based discrimination, inequalities or challenging the elimination of traditional gender roles) as defined by Carroll (1984).

Discussion: Och was the only coder and coded all the texts by hand. This could be identified as a potential research problem which could be avoided if more coders were involved in coding process and if computer-assisted methods were used to avoid coding by hand.

4.2 Discourse Studies

Discourse Studies has been developing at the intersection of language and society. It combines various qualitative and quantitative research methods as well as different genres such as news reports and parliamentary debates (van Dijk, 2018). In this review, we identified two salient methods of Discourse Studies, namely discourse analysis (DA) and critical discourse analysis (CDA).

In political science, **discourse analysis (DA)** is most frequently used to study parliamentary debates and parliamentary discourse. It is frequently referred to as political discourse analysis (PDA) (Dunmire, 2012) and can sometimes be mistakenly equated to content analysis even though it does not focus on the analysis of content but rather on the analysis of language through specific text and context. One of the main foci of DA is to examine how political power, power abuse and domination manifest through discourse practices and structures (ibid.).

Critical discourse analysis (CDA), or critical-political discourse analysis, is one of the most visible categories of discourse studies frequently applied to parliamentary communication. It provides a critical context in which political debates occur and analyzes the relationship between power and the traditional ideology in implied discourse (van Dijk, 2018). A contribution that CDA can make to political studies is mostly in offering a general theoretical perspective on discourse which recognizes the constitutive potential of discourse within and across social practices without reducing social practices to their discursive aspect (Farrelly, 2010).

4.2.1 Parliamentary Discourse on Immigration

Research problem: May (2016) analyzed the parliamentary discourses on immigration in Canada and France and wanted to find out what arguments were introduced in parliamentary arenas to justify more restrictive immigration policies.

Data collection: May's analysis was stretched between January 2006 to December 2013. The two countries were chosen because of the very similar discussions about immigration and because they developed different models of integration and management of cultural diversity. He analyzed parliamentary debates following seven bills which included a high number of immigration indicators.

[2] https://atlasti.com

During the coding procedure he and another coder read through the debates and compiled a list of coding units which was inspired by the literature review. Then they identified the phrases and clusters of meaning which resulted in a hierarchical coding structure which included 32 nodes. They refined the coding procedure by introducing new nodes based on the themes they considered relevant, which resulted in the introduction of new nodes into the structure. After that the inter-coder reliability test was performed followed by the discursive analysis.

Research method: The author employed critical discourse analysis (CDA). After identifying the main 32 keywords (refugee, asylum seekers, Roma, financial cost, immigration, multiculturalism, etc.) in the chosen parliamentary debates, he opted for lexical analysis with the Nvivo software[3] (QSR International Pty Ltd., 2020) to code specific discursive constructions.

Discussion: May gave no specific account as to where the analyzed debates were downloaded from and what language they were in (relevant for Canada which is bilingual). The paper could also benefit from a more thorough description of the discourse analysis since it is mentioned as the primary method used.

4.2.2 Political Discourse about COVID-19

Research problem: Jarvis (2021) analyzed the conceptions of time during the COVID-19 pandemic within the UK parliamentary discourse. He showed that construction of temporality was important for social, political, and historical positioning of the virus and that such constructions had impact on UK government's response to the virus.

Data collection: Jarvis analyzed more than 120 texts including parliamentary speeches, newspaper articles, press releases, public letters, accouchements, and policy statements. The timeframe of his analysis was limited to the first six months of 2020 since this was the timeframe crucial to the government's communication of the crisis. He designed his own corpus by collecting the texts directly from the official website of the Prime Minister's office. All the texts were thoroughly read to determine their relevance for the research and all the texts that referred to the pandemic or its response were included in the corpus for future analysis. Jarvis organized coding material around various index categories (the virus, the UK government's response, the scientific response, the public, temporality) and reread all the texts through his framework. This allowed for the distribution and coding of the data according to different themes and their subcategories.

Research method: The author employed discourse analysis via the framework method as defined by Ritchie and Spencer (2002). He analyzed qualitative data through summarizing, sifting, and sorting research material and classifying large volumes of data in its own terms. Jarvis performed a detailed analysis which involved a thorough reading of the corpus in four stages: 1) familiarization with the documents, 2) coding via paraphrasing of short text sections, 3) developing an analytical framework from the coded material, and 4) applying this framework to the corpus.

Discussion: Jarvis' paper shows the importance of collecting data from various sources and strengthens the notion that political scientists often use different sources to gather relevant data for their analysis. It is also one of the few studies in political science where a corpus was created to analyze the data.

4.3 Sentiment Analysis

Sentiment analysis is a growing research method at the intersection of linguistics and computer-based automated approaches which attempts to automatically determine the sentiment contained in a certain text (Taboada, 2016). Automated sentiment analysis presents an innovative approach in social sciences, the main aim of which is to measure the polarity or tonality of texts by identifying and assessing expressions that people use to evaluate persons, events, or identities (Haselmayer and Jenny, 2017). Although it is becoming increasingly popular in political science mainly because the digitization of legislative transcripts has increased the potential application of established tools for analyses of emotion in text (Cochrane et al., 2021), many political scientists are still more comfortable using human-based content analysis to analyze emotions. The potential problem of analyzing sentiment in parliamentary debates is that unlike text, speeches consist of intonation, facial expressions and body language which are hard to determine just by looking at the transcripts. Hence coders frequently focus not only on reading the transcripts but also on watching video clips of the debates to grasp emotions in their entirety.

4.3.1 Gender Influence on Negativity in Parliament

Research problem: Haselmayer, Dingler and Jenny (2022) analyzed how the gender of the MPs and the context of debates influenced the level of negativity in parliamentary speeches and showed that female MPs used less negative language than male MPs mainly because of gender differences in socialization and stereotypical expectations.

Data collection: The authors focused their analysis on 52.132 speeches from plenary debates in the Austrian National Council. Those speeches were delivered by more than 500 different MPs from 7 Austrian parties (SPÖ, ÖVP, FPÖ, BZÖ, LiF, Greens and Team Stronach) throughout 24 years (from 1993 to 2013). Speeches from cabinet members (approximately 4.000) and short speeches with less than five sentences (around 500) were excluded from the analysis.

Research method: The authors applied sentiment analysis with word embeddings to plenary speeches in Austrian parliament. They researched negative

[3] https://lumivero.com/products/nvivo/

parliamentary speeches and relied their analysis on machine learning based on crowdcoded training set. The classifier used data and word embeddings from FastText library[4] (META, 2015). The authors calculated meaningful word vectors by using subwords and the Gensim library[5] (LGPL, 2009). Each sentence was represented as a sequence of word vectors which preserved information on word order and captured dependencies between words. They also used a recurrent neural network (The Gated Recurrent Unit – GRU) to deal with a sequential data input. In the stage of pre-processing the text, stop words and punctuation were included. They trained this procedure on around 20.000 sentences which contained a continuous negativity score ranging from neutral to very negative (0 – 4). The model was then trained 60 times with a dropout of 40 % over the entire network.

Discussion: Although this is a political science research, the data collection and analysis descriptions are highly computational and therefore require some computational knowledge to be fully understandable. Since one of the common goals is to familiarize other political and social scientists with automated sentiment analysis, a more simplified description of the methods would be useful.

4.3.2 Emotions in Political Speech

Research problem: Cochrane et al. (2021) analyzed a new dataset of annotated texts and videos form the Canadian House of Commons to examine whether transcripts capture the emotional content of speeches, to compare strategies for the automated sentiment analysis in text and test the robustness of the approach based on word embeddings.

Data collection: Their data collection consisted of official Hansard transcripts and video clips. To gather the latter, the authors recorded every third Question Period in the Canadian parliament between January 2015 and December 2017. This covered the last 10 months of Stephen Harper's conservative and the first 23 months of Justin Trudeau's liberal government. They trimmed the videos from the start of the first question to the end of the last answer which produced 102 videos of approximately 45 minutes in length and randomly selected ten time-points (*mm:ss*) in each of them. The sentence beginning just prior to the time-point was extracted as its own video clip. The average length of the extracted clip was approximately 9 seconds and it contained 23 words. These videos clips were added to a Qualtrics[6] survey instrument and randomly assigned to one of three independent, bilingual coders for manual coding. For all but one video clip the authors were also able to identify the corresponding official Hansard transcriptions. For speeches in French, the coders used the official English translations. The coders were asked to assign a sentiment score to each clip depending on eleven-point scale (0 – 10, negative – positive) as well as activation (subdued – aroused) of the speech fragment. Since the presentation of clips was randomized, same clips were often presented to the same coder at different times. The texts of the speech fragments were also randomly presented to three independent coders who were asked to indicate the sentiment and activation for each fragment on eleven-point scale. Throughout their analysis, the authors also tested five widely used sentiment dictionaries (Lexicoder 3.0, Sentiwordnet 3.0, Hu-Liu Lexicon, VADER, and Jockers-Rinker's Lexicon) to test their efficacy.

Research method: Researchers employed sentiment analysis with the help of automatically generated sentiment dictionaries. In addition, sentiment was manually coded by coders to improve reliability of the research results.

Discussion: This paper shows that when conducting sentiment analysis, political scientists can rely on video clips of the parliamentary debates and use them to triangulate data gathered from the analysis of official parliamentary transcriptions which improves the reliability of the research.

4.4 Mixed Methods Approach

Mixed methods approach draws on the strengths of qualitative and quantitative research methods which generates a more complete picture of the research problem (Shorten and Smith, 2017). It is a highly complementary approach where the results of one research method can be validated, elaborated, and clarified by the other. Such triangulation allows not only for more valid research results but also reduces research bias and unwarranted selectivity of source materials, which according to Thies (2002) are the two biggest problems of qualitative research. Mixed methods offer more in-depth findings and forces researchers to develop a broader set of research skills which produce valid research results (Tzagkarakis and Kritas, 2022).

Corpus-assisted discourse studies (CADS) could be understood as a special type of mixed methods approach as they combine qualitative discourse analysis with predominantly quantitative corpus-assisted research approach. Rubtcova et al. (2017) show that it is a useful research method for the study of political discourse and parliamentary data especially when the data has already been collected in a corpus (as in ParlaMint). This approach uses corpus techniques to examine a particular political discourse type and analyze certain patterns of language with one of the greatest strengths being minimization and reduction of the research bias (Partington, 2012).

4.4.1 Performance, Gender, and Affective Atmosphere in the time of Brexit

Research problem: Parry and Johnson (2021) examined the parliamentary discourse regarding threats to Members of the Parliament in the context of

[4] https://fasttext.cc

[5] https://radimrehurek.com/gensim/

[6] https://www.qualtrics.com/support/survey-platform/survey-module/survey-tools/survey-tools-overview/

broader discussions about emotionality, polarization, and toxicity in discourse in the UK.

Data collection: The primary source material were Hansard transcripts of the debate on September 25. The debate started with the PM's address at 6.30 in the afternoon and ended 3 hours later. They also used data provided by the UK Parliament's YouTube channel; this allowed them to watch relevant sections and capture gestures, use of space and affective atmosphere. In addition, they used the Nexis database[7] for the newspaper analysis. Here they searched for "Tracy Babin" and "Paula Sherriff" since the names of the two MPs were determined to provide the most relevant results regarding the research topic. They read the articles and retained those that focused on the abuse of the female MPs or those which called for the new standards in public life and language. This news sample comprised 97 articles, mostly from national news outlets.

Research method: The authors employed a mixed methods approach combining performance analysis of the Hansard transcripts and UK Parliament YouTube coverage of the debates and discourse analysis of national as well as local newspaper coverage of the parliamentary debates. Using the performance approach allows the researchers to conduct research beyond the linguistic content of political speech and to focus on style, form, gesture, and the use of physical space.

Discussion: This is not a typical use of the mixed methods approach since the authors did not combine quantitative and qualitative but rather two qualitative approaches. This research is significant also because it is the only one in our sample which employed performance analysis. This paper also shows how important it is to not only focus on one data source but rather combine various sources and different types of data.

4.4.2 Religious Freedom in Debates on Same-sex Marriage in Australia

Research problem: Poulos (2019) explored why and how the term "religious freedom" appeared in the title of the Australian bill to legalize same-sex marriage. He wanted to analyze how debates about same-sex marriages changed over time.

Data collection: Poulos analyzed 663 speeches made in Australian parliament during the marriage legislation debates between 2004 and 2017. This research was based on 20 bills proposing amendments to the Marriage Act allowing for same-sex marriage or recognizing same-sex marriages. Data was taken from the Australian Parliament House website using the homepages of the respective bills as well as the Hansard. Once the same-sex marriage bills were identified, PDFs of the Hansard files were collected for every speech and then converted to the text file using an online converter.[8] Poulos removed all the metadata (speakers' names, electorates, ministerial roles, time stamps, etc.), interjections and procedural statements included in the Hansard files. The speeches were chronologically grouped into three different sub- corpora (the first one from 2004, the second one between 2006 and 2016 and the third one from 2017). Then, two other sub-corpora were created, this time according to whether the speakers explicated a position in support or in opposition to the same-sex marriage and then chronologically sorted again according to support or the opposition.

Research method: The author opted for corpus-assisted discourse analysis. Poulos analyzed the text files with the help of two software packages, namely AntConc[9] (Anthony, 2018) and WordSmith Tools[10] (Oxford University Press, 1996). The first was used to generate word frequency lists, concordances, and identify collocates and the second one to identify keywords. This analysis was triangulated with manual discourse coding using the NVivo software. To examine whether the arguments were framed for or against the same-sex marriage, each sub-corpus of the supportive speeches was analyzed against corpus, which included the speeches which opposed same sex marriage and vice versa. The author examined the most frequent words and lexical keywords from each of the sub-corpora and performed the analysis of how the framing of the same-sex marriage "issue" changed over time.

Discussion: This is a rare example of research which deliberately discarded the available metadata. This is uncommon in social sciences which usually relies on metadata to provide additional information during analysis.

4.5 Thematic Analysis

Thematic analysis is a highly useful approach in qualitative research since it allows for the identification of prominent themes and provides several ways to interpret meaning from a certain dataset. Its focus is to find not only the major themes of analyzed data, but also to come up with various fine-grained subthemes that match the main themes and therefore make the interpretation of results much more straightforward (Gherghina, Tap and Soare, 2022). It is commonly understood as an umbrella term for various research approaches rather than a single method. In political studies, thematic analysis (sometimes referred to as qualitative document/content analysis) is particularly useful for the study of legislation and policy and is also becoming increasingly important in the study of parliamentary debates.

Sometimes thematic analysis is equated to content analysis and much of this confusion is because thematic analysis originated from content analysis before branching off to serve similar but distinct research goals (Joffe, 2012). The main difference between the two lies in the possibility of quantification

[7] https://www.lexisnexis.com/en-us/professional/nexis/nexis.page
[8] https://pdftotext.com
[9] https://www.laurenceanthony.net/software/antconc/
[10] https://www.lexically.net/wordsmith/

of data in content analysis by measuring the frequency of categories and themes, whereas thematic analysis is strictly qualitative. Consequently, content analysis has a wider selection of coding approaches, is more practical and straightforward whereas thematic analysis supports deeper immersion and is more intuitive.

4.5.1 Parliamentary Debates About Emigrants

Research problem: Gherghina, Tap and Soare (2022) analyzed the ways in which members of the Romanian parliament refer to emigrants; not only the ambivalent attitude but also the representation of emigrants and their needs.

Data collection: The authors focused on analyzing parliamentary speeches from the plenary sessions in the Chamber of Deputies (lower house of the Romanian parliament) in the two terms between 2012 – 2016 and 2016 – 2020 with an incomplete second term (data was available only until March 2020 whereas the term ended in November 2020). This yielded 239 parliamentary speeches which covered the developments after the financial crisis and important events (elections, anti-government protests) in which the diaspora actively participated. The speeches were split between the two terms as follows: 135 speeches with the average length of 530 words from the first term and 104 speeches with average length of 517 words from the second term. The speeches were publicly available on the official website of the Chamber of Deputies. Before the analysis, data was coded in three stages. First, coders independently read all relevant speeches and grouped them into predefined themes. Second, an inter-coder reliability test was used to identify borderline and missing themes. In the final phase, the list of main themes was enriched with the relevant sub-themes and applied to the speeches.

Research method: The authors employed deductive thematic analysis based on the pre-established themes which were derived from the literature. This allowed for the identification of comment themes as well as provided various ways to interpret meaning from the dataset of speeches selected for the analysis.

Discussion: One shortcoming that authors mention is an underrepresentation of Romanian emigrants in Romanian politics which could influence the content of speeches about the diaspora. In addition, not all the speeches were collected which could have some impact on reliability of the research results.

4.6 Text as Data and Computational Approaches

Computational methods have in the last couple of years gained in popularity which allowed for the development of new research approaches and new methods to analyze textual documents inside social sciences. One such is **text-as-data approach** which consists of a broad set of techniques and relies on automated or semi-automated analysis of text (Gilardi and Wüest, 2020). It allows researchers to analyze extensive amounts of textual data, significantly reduces the cost of analyzing large collections of text and allows researchers to deploy language-agnostic analytical tools. Text-as-data is a relatively new approach in political science in comparison to the more traditionally used content analysis and qualitative methods (Krippendorff, 2018). It combines new sources of data, machine-learning tools, and social science research design to develop and evaluate new insights (Grimmer, Roberts and Steward, 2022) and understands text as numerical data suitable for quantitative analysis. The aim of this approach is not to replace the insights of qualitative research but rather complement and extend it (Mochtak, personal communication, 2023).

Quantitative text analysis (QTA) is an example of the text-as-data approach and refers to the process of analyzing text data by using statistical procedures. It is an automated and systematic method for processing extensive amounts of text (e.g., parliamentary debates, policy documents, party manifestos, etc.) (Slapin, 2018) which most commonly occurs in three basic steps: 1) defining a corpus from the texts for analysis, 2) determining the unit of analysis, and 3) creating document feature matrix.

Social network analysis (network analysis) refers to the study of social structures by using networks and graph theory. It analyzes links between nodes, which in political science most commonly represent either persons, organizations, or states while links represent some form of connection between them (Ward, Stovel and Sacks, 2011). Social network analysis is becoming an increasingly used computational method in political science and is commonly used when researchers want to establish connections between political actors from an extensive dataset. As shown in Skubic et al. (2022), network analysis can be extremely useful for the comparative analysis of argumentative and structural power of parliamentarians in different European parliaments.

4.6.1 Populism and Parliamentary Polarization in German Parliament

Research problem: Lewandowski et al. (2021) examined how the German parliamentary discourse changed after two populist and two non-populist parties entered parliament and analyzed how populism shaped the behavior of new parties as well as how other parties respond when the new contesters arrive.

Data collection: The authors based their analysis on a GermaParl corpus[11] (Blätte and Blessing, 2018) which includes parliamentary debates from the German parliament. They analyzed legislative periods 9 to 19 (from 1980 to 2020) and focused on two populist (The Left, AfD) and two non-populist parties (Greens, PDS). Only speeches delivered by members of the parliament were analyzed and speakers not belonging to a parliamentary group were

[11] https://github.com/PolMine/GermaParlTEI

excluded. The analysis of populist language was based on all speeches from the period of interest (190.000) whereas the analysis of polarization was based on a subset of approximately 113.000 speeches. When measuring parliamentary polarization, the authors only included those speeches to which they could assign a substantial topic using a topic modelling approach which resulted in a lower number of analyzed speeches.

Research method: They applied qualitative text analysis of parliamentary speeches to measure populism and issue-based polarization. To measure populist speech, they used a dictionary-based approach. Firstly, they used a specific word list (suggested by Rooduijn and Pauwels, 2011) to create a lexicon of key terms which indicated the use of populist references. Then they calculated the frequency of those terms relative to the length of a speech as well as used keyword-in-context analysis to examine the context in which the identified words occurred. For measuring political polarization, the authors used the Wordfish algorithm[12] (Slapin and Proksch, 2008) for which they needed to subset all speeches along three dimensions: the parliamentary group, the primary topic of the speech and the legislative period in which the speech was made. All speeches of a single parliamentary group in each legislative period were clustered about a single topic.

Discussion: This is the only reviewed research which uses Wordfish algorithm, which is written in the programming language for statistical computing R. As shown later in the discussion chapter, R is especially important for political science since it is easy to understand and provides data in tabular format and is therefore most used programming language for extracting political positions from textual documents.

4.6.2 Analyzing the Politics of Brexit Debate Abroad

Research problem: Sierens and Brack (2020) examined to what extent the attention given to Brexit differs across different parliaments and if parties emphasized the same issues across different levels. They specifically analyzed how Brexit was framed and discussed in the Belgium parliament.

Data collection: Research relied on a unique database of parliamentary questions in three different Belgian parliamentary assemblies (Federal, Flemish, and Walloon). The authors gathered data from January 23, 2013 (when David Cameron announced his intention to hold a referendum about Brexit) until October 2017. Data for analysis were retrieved directly from the websites of all three assemblies. At the federal level, the authors analyzed parliamentary questions asked in the Chamber of Representatives and used the keyword "Brexit" to classify all questions that dealt with this specific topic. At the regional levels, they focused on questions that had the word "Brexit" in their titles. Altogether, they retrieved 146 parliamentary questions in the Federal parliament (94 oral and 52 written), 88 parliamentary questions in Flanders (57 oral and 31 written) and 37 parliamentary questions in Wallonia (12 oral and 25 written). For the purpose of comparative analysis, the authors categorized data into series of questions divided into "who" questions ("who asks who?", "who asks what?", etc.) and "what" questions. The former were classified according to the MPs party and presence/absence in the governmental coalition. In the latter, each parliamentary question was categorized according to the main issue emphasized in the parliamentary question. According to these criteria the data was coded into four most frequent categories (general information on Brexit, trade and economic consequences, negotiation strategy, specific issues).

Research method: In the first step, the authors conducted a descriptive comparative analysis of the gathered parliamentary questions. It relied on Social Network Analysis that allowed the authors to focus on the structural relationships between the different units of analysis. For each level of the government, they drew networks of parliamentary questions and computed various indicators of those networks (density, average degree, etc.). In the second step they employed loglinear modelling (a special case of generalized linear models for multivariate cross-classified categorical data (Sierens and Brack, 2020)) of the frequency of associations and interactions between categorical variables.

Discussion: This is the only reviewed paper that employs social network and loglinear model analysis. Although authors provide some explanation of the methods, there is no emphasis on a more detailed explanation (e.g., which software was used for network analysis, how to work with such software, etc.).

5. Discussion and Conclusion

In this literature review we showed the most common methods and approaches political scientists use when conducting research on parliamentary debates and discourse. One of the core interests of political science is to analyze power relations inside parliaments as well as a means through which the power is displayed. Parliamentary discourse not only reflects the power and authority of the parliamentarians, but also allows parliamentarians to present their interpretation of specific issues to different external audiences (Laver et al., 2003).

One of our main findings is the similarity between methods and approaches used in political science and sociology, as shown in Skubic and Fišer (2022). Our extensive research showed that more than half of the reviewed political science papers employed one of the research methods that are traditional in political science (either content or discourse analysis). We also find that political scientists often employ such methods to analyze data which is frequently manually collected and downloaded from various sources (e.g., parliamentary websites, repositories of relevant

[12] http://www.wordfish.org

organizations, etc.) rather than using more modern and less time-consuming and resource heavy computational techniques. This is confirmed by Mochtak (personal communication, 2023) who states that more than 90 % of political science research still employs traditional data collection and research methods (with content analysis being the most used). According to Mochtak, political science is slow when it comes to adjustment and modification of research methods and approaches to more modern, less time-consuming, and more technologically advanced methods. Our review shows that in some cases this transition has already been made but such research is scarce, hard to identify, and often lacks methodological explanation.

Even when political scientists use modern computational methods to conduct research and collect data, they are often reluctant to perform big-corpora and big-data analyses or employ methods which they find hard to comprehend. According to Mochtak, political scientists only rarely rely on complicated programming language or computational methods. Probably one of the most used in political science is R programming language mainly because it offers tidy data in tabular format and is relatively easy to use. When political scientists deal with large amounts of quantitative data, they want them to be organized, easily accessible and easy to use (one example of such data is V-DEM data[13]). Databases therefore need to be made approachable, accessible and offer functional API for political scientists to consider using them.

We find that despite the quantitative turn of political science in recent years, political scientists still predominantly use qualitative or mixed methods. In addition, software and tools for computational qualitative analyses (such as Nvivo, Atlas.ti or MAXQDA) have in recent years become more popular. This not only allows researchers to analyze data faster, more efficiently and in a more organized way but also attributes to more replicable and relevant research results and minimizes researcher bias which is otherwise common in solely qualitative research. Reliability and relevance of results is further enforced by data collection triangulation which is common in political science. Often researchers rely not only on parliamentary but also other sources such as newspapers, (social) media, interviews, etc., which assures higher quality of the conducted research.

If we want to encourage political scientists to start incorporating corpora such as ParlaMint in their research and use corpus-assisted methods more actively, we firstly need to make it highly approachable and accessible (Mochtak, personal communication, 2023). Datasets such as ParlaMint are very useful and offer an abundance of valuable data but are often too complex for political scientists to use. Our first aim should therefore be to make data available in a format which political scientists would be familiar with. In addition, tutorials, workshops, showcases, and user manuals should be offered to political scientists so they could familiarize themselves with the ParlaMint concept, workflow, and the variety of data it offers. We agree with Kytö (2011) that corpus compilers should also provide rich, useful, and user-friendly documentation as to how the corpus data is gathered, processed, and annotated and should clearly and in detail document their compilation decisions, offering user guides, corpus manuals and training materials which would accompany the release versions of corpora. This would enable political scientists to reuse corpora in a contextualized way, which would significantly ease their process of data collection and analysis.

In addition, the ParlaMint community should also focus on providing data with rich and useful metadata. Metadata such as gender, role, party affiliation, political orientation, etc. are useful, but other metadata such as sentiment score, emotions, policy areas of agenda points etc. would be an additional added value. Collecting and assigning such metadata is usually a time-consuming process which requires a lot of effort and human resource and is frequently very specific to the research question at hand. This is why corpora such as ParlaMint would be even more interesting for political scientists if it allowed them to directly add, edit and share additional metadata layers. Machine translations of parliamentary debates would also provide important additional possibilities for more international research and parliamentary discourse comparisons.

The argument that we want to put forward with this literature review is not that the current predominantly qualitative research methods in political science should be replaced with more quantitative corpus-assisted approaches in their entirety, but rather that corpus data and corpus-analytical techniques could effectively be used alongside the traditional qualitative approaches. We understand corpora as potentially powerful tools which would help political scientists not only to simplify data collection processes and help them generate relevant results much more effortlessly but would also contribute to more transparent, verifiable, and reproducible research.

6. Acknowledgements

The work described in this paper was funded by the Slovenian Research Agency research programme P6-0436: Digital Humanities: resources, tools, and methods (2022-202), the CLARIN ERIC ParlaMint project (https://www.clarin.eu/parlamint) and the DARIAH-SI research infrastructure (https://www.dariah.eu).

7. Bibliographical References

Anthony, L. (2018). AntConc (Version 4.2.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

[13] https://github.com/vdeminstitute/vdemdata

Blassnig, S. (2022). Content Analysis in the Research Field of Political Communication: The Self-Presentation of Political Actors. In Standardisierte Inhaltsanalyse in der Kommunikationswissenschaft–Standardized Content Analysis in Communication Research: Ein Handbuch-A Handbook (pp. 301-312). Wiesbaden: Springer Fachmedien Wiesbaden.

Blätte, A., & Blessing, A. (2018, May). The GermaParl corpus of parliamentary protocols. In proceedings of the eleventh international conference on language resources and evaluation (LREC 2018).

Blaxill, L. (2022, June). Parliamentary Corpora and Research in Political Science and Political History. In Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference (pp. 33-34).

Carroll, S. J. (1984). Woman candidates and support for feminist concerns: The closet feminist syndrome. Western Political Quarterly, 37(2), 307-323.

Cochrane, C., Rheault, L., Godbout, J. F., Whyte, T., Wong, M. W. C., & Borwein, S. (2022). The automatic analysis of emotion in political speech based on transcripts. Political Communication, 39(1), 98-121.

Dunmire, P. L. (2012). Political discourse analysis: Exploring the language of politics and the politics of language. Language and Linguistics Compass, 6(11), 735-751.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., ... & Fišer, D. (2023). The ParlaMint corpora of parliamentary proceedings. Language resources and evaluation, 57(1), 415-448.

Farrelly, M. (2010). Critical discourse analysis in political studies: An illustrative analysis of the ‘empowerment’agenda. Politics, 30(2), 98-104.

Cauchon, S. Introduction to Political Science Research Methods. eBook – Adobe PDF. 1st Edition. https://ipsrm.com/.

Gherghina, S., Tap, P., & Soare, S. (2022). More than voters: Parliamentary debates about emigrants in a new democracy. Ethnicities, 22(3), 487-506.

Gilardi, F., & Wüest, B. (2020). Using text-as-data methods in comparative policy analysis. In Handbook of research methods and applications in comparative policy analysis (pp. 203-217). Edward Elgar Publishing.

Haselmayer, M., & Jenny, M. (2017). Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding. Quality & quantity, 51, 2623-2646.

Haselmayer, M., Dingler, S. C., & Jenny, M. (2022). How Women Shape Negativity in Parliamentary Speeches—A Sentiment Analysis of Debates in the Austrian Parliament. Parliamentary Affairs, 75(4), 867-886.

Jarvis, L. (2022). Constructing the coronavirus crisis: Narratives of time in British political discourse on COVID-19. British Politics, 17(1), 24-43.

Joffe, H. (2012). Thematic analysis. Qualitative research methods in mental health and psychotherapy: A guide for students and practitioners, 209-223.

Krippendorff, K. (2018). Content analysis: An introduction to its methodology. Sage publications.

Kuckartz, U. (2019). Qualitative text analysis: A systematic approach. Compendium for early career researchers in mathematics education, 181-197.

Kytö, M. (2011). Corpora and historical linguistics. Revista Brasileira de Linguística Aplicada, 11, 417-457.

Lauer, J. (2021). Methodology and political science: the discipline needs three fundamentally different methodological traditions. SN Social Sciences, 1(1), 43.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. American political science review, 97(2), 311-331.

Lewandowsky, M., Schwanholz, J., Leonhardt, C., & Blätte, A. (2022). New parties, populism, and parliamentary polarization: evidence from plenary debates in the German Bundestag. The Palgrave handbook of populism, 611-627.

Lilja, M. (2021). Russian political discourse on illegal drugs: A thematic analysis of parliamentary debates. Substance Use & Misuse, 56(7), 1010-1017.

May, P. (2016). Ideological justifications for restrictive immigration policies: An analysis of parliamentary discourses on immigration in France and Canada (2006–2013). French Politics, 14, 287-310.

Och, M. (2019). Conservative feminists? An exploration of feminist arguments in parliamentary debates of the bundestag1. Parliamentary Affairs, 72(2), 353-378.

Parry, K., & Johnson, B. (2023). Humbug and outrage: A study of performance, gender and affective atmosphere in the mediation of a critical parliamentary moment. The British Journal of Politics and International Relations, 25(1), 3-20.

Partington, A. "Corpus analysis of political language." In: C.A. Chapelle (Ed.) The Encyclopedia of Applied Linguistics. Blackwell Publishing Ltd. (2013).

Poulos, E. (2020). The power of belief: religious freedom in Australian parliamentary debates on same-sex marriage. Australian Journal of Political Science, 55(1), 1-19.

Provalis Research. (2014). QDA Miner.

Sanchez Salgado, R. (2021). Emotions in European parliamentary debates: Passionate speakers or unemotional gentlemen?. Comparative European Politics, 19(4), 509-533.

Scott, Mike. “WordSmith Tools Manual.” (2018): https://lexically.net/downloads/version7/HTML/

Shorten, A., & Smith, J. (2017). Mixed methods research: expanding the evidence base. Evidence-based nursing, 20(3), 74-75.

Sierens, V., & Brack, N. (2021). The politics of the Brexit debate abroad: an analysis of parliamentary questions on Brexit in Belgian parliaments. Journal of Contemporary European Studies, 29(4), 519-534.

Skubic, J., & Fišer, D. (2022, June). Parliamentary discourse research in sociology: Literature review. In Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference (pp. 81-91).

Skubic, J., & Fišer, D. (2022). "Parliamentary Discourse Research in History: Literature Review." In Proceedings of the Conference on Language Technologies and Digital Humanities.

Skubic, J., Angermeier, J., Bruncrona, A., Evkoski, B., & Leiminger, L. (2022). Networks of power: Gender analysis in selected european parliaments. In 2nd Workshop on Computational Linguistics for Political Text Analysis (CPSS).

Slapin, J. (2020). Three Basic Steps of Quantitative Text Analysis.

Slapin, J. B., & Proksch, S. O. (2008). A scaling model for estimating time-series party positions from texts. American Journal of Political Science, 52(3), 705-722.

Taboada, M. (2016). Sentiment analysis: An overview from linguistics. Annual Review of Linguistics, 2, 325-347.

Thies, C. G. (2002). A pragmatic guide to qualitative historical analysis in the study of international relations. International Studies Perspectives, 3(4), 351-372.

Tzagkarakis, S. I., & Kritas, D. (2022). Mixed research methods in political science and governance: approaches and applications. Quality & Quantity, 1-15.

Van Dijk, T. A. (2018). Discourse and migration. Qualitative research in European migration studies, 227-245.

Van Dijk, T. A. (1997). What is political discourse analysis. Belgian journal of linguistics, 11(1), 11-52.

Ward, M. D., Stovel, K., & Sacks, A. (2011). Network analysis and political science. Annual Review of Political Science, 14, 245-264.

Compiling and Exploring a Portuguese Parliamentary Corpus: ParlaMint-PT

José Aires, Aida Cardoso, Rui Pereira, Amália Mendes
University of Lisbon - School of Arts and Humanities / Center of Linguistics
Alameda da Universidade, 1600-214 Lisboa, Portugal
joseaires74@gmail.com, aidacard@gmail.com
ruifilipedebarrospereira@gmail.com, amaliamendes@letras.ulisboa.pt

Abstract

As part of the project ParlaMint II, a new corpus of the sessions of the Portuguese Parliament from 2015 to 2022 has been compiled, encoded and annotated following the ParlaMint guidelines. We report on the contents of the corpus and on the specific nature of the political settings in Portugal during the time period covered. Two subcorpora were designed to enable comparisons of the political speeches between pre- and post-COVID-19 pandemic. We discuss the pipeline applied to download the original texts, ensure their preprocessing and encoding in XML, and the final step of annotation. This new resource covers a period of changes in the political system in Portugal and will be an important source of data for political and social studies. Finally, we have explored the political stance on immigration in the ParlaMint-PT corpus.

Keywords: parliamentary corpus, Portuguese, political views

1. Introduction

Providing access to the sessions of the national parliaments is an important tool for monitoring the democratic system. Session transcripts are frequently available and can be consulted by the population, ensuring greater transparency in a representation system carried out through elections. To ensure access to this data, several initiatives have sought to apply well-established NLP techniques, which use standards in metadata encoding and linguistic annotation, to be carried out on the data. The availability of sessions' transcripts from different national parliaments on a single website adds a new level of transparency and contributes to citizen empowerment. With this objective, the ParlaMint I project (Erjavec et al., 2023) established a first set of corpora with transcriptions of the sessions of 17 European national parliaments, uniformly encoded and with a rich set of metadata and annotated with Universal Dependencies (Erjavec et al., 2021). The first phase of the project was expanded to other European languages (ParlaMint II), including Portuguese. We report on the compilation, preprocessing and annotation of the Portuguese corpus ParlaMint-PT. This new resource allows, on the one hand, to explore the political views of the parties in relation to different topics and, on the other hand, to make contrastive analyses of the policies followed in several European countries, taking advantage of the English translation of the corpora. As a result, it is possible to compare, for example, how parties in southern European countries positioned themselves in relation to vaccination during the COVID-19 pandemic or the position of the different right-wing European parties in relation to immigration.

We present the new ParlaMint-PT corpus and provide information about its content and the levels of annotation added, as well as details about the political situation in Portugal during the period covered by the corpus, that is, from 2015 to 2022. The last legislature of the Portuguese parliament included in the corpus already points to an ongoing change in the Portuguese party system. Until then, and unlike many European countries, the political parties traditionally represented in parliament maintained great influence. The tendency for new parties to be represented in the Parliament starts in 2019 (especially parties on the right of the political spectrum) and increases in the next legislature, in 2022. The corpus thus keeps track of the first moments of an ongoing change in Portuguese politics. The ParlaMint corpora are openly available via the CLARIN.SI repository for download, as well as through the NoSketch Engine and KonText concordancers and the Parlameter interface for online exploration and analysis. The detailed metadata included in the corpora associated with the advanced searches enabled by the query programs allow for intuitive and user-friendly data analysis. We intend to show how the corpus can be useful by presenting preliminary results of work on political stances regarding immigration issues in the parliament sessions.

We discuss in section 2 other initiatives to collect and explore the transcriptions of the sessions of the Portuguese Parliament, and in section 3 we provide information on the political parties represented in the Parliament during the time frame of our corpus.

ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora, pages 12–20
May 20, 2024.

Section 4 addresses the raw data of the ParlaMint-PT corpus: we first provide some quantitative information, then details on the set of metadata in 4.1, and finally information on the structure of the sessions in 4.2. The pipeline for preprocessing, encoding texts in XML and annotation are described in 5. We provide preliminary results of a case study on the parliamentary views about immigration in section 6, and we conclude in 7.

2. Related Work on Portuguese Parliamentary Speeches

A growing number of initiatives have targeted the collection of Parliamentary data for a large set of languages. In this section, we will be concerned specifically with previous resources and projects that have collected and processed data from the Portuguese Parliament.

The first initiative to collect and explore the Portuguese parliamentary speech sessions was undertaken in the framework of a general corpus of Portuguese, the Reference Corpus of Contemporary Portuguese – CRPC (Généreux et al., 2012). The sessions of the Portuguese Parliament, as well as some legislation, are included in a large section called "politics" that contains 163M words. The CRPC corpus can be queried online in the CQPweb platform, and users can restrict their query to the subcorpus "politics"[1]. Although the data cover a large time frame of the XIX century, this corpus lacks detailed metadata that would enable queries regarding time period, political parties and speakers, and the internal sections of the speeches are not structured and properly encoded. A 1M words subset of the section "politics", the PTPARL corpus, is freely available for academic purposes on the PORTULAN CLARIN infrastructure[2], with POS annotation. This subset also lacks detailed metadata and the encoding of the internal structure of the speeches.

The website Demo.cratica was an effort to make accessible the transcripts of the Parliament sessions (A. I Carvalho and R. Lafuente (2010)). Another initiative is the Portuguese Observatory on Parliamentary Dynamics (POPaD)(Giorgi and Dias, 2019). The data has been explored in several analyses of the political system in Portugal and in a contrastive perspective with the patterns found in Europe (De Georgi and Moury, 2015).

More recently, a new compilation of Portuguese Parliamentary speeches, from 2000 to 2015, was used to explore how saliency, government dynamics, and party size affect the use of members of the parliament who specialize in specific areas of expertise in debates (Fernandes et al., 2019). The authors gathered 50,000 speeches and 6,000 bills from the Portuguese Parliament Official Website. A second initiative is the compilation of the corpus PTPARL-D, an annotated corpus of debates in the Portuguese Parliament, covering the years 1976 to 2019 (Almeida et al., 2021).

In spite of these initiatives, there was still no fully accessible set of parliamentary speeches, making use of widely known query tools, providing structured data, following standards established in the community working on parliamentary data, and providing a comparable corpus for Portuguese in line with the growing efforts congregated in the ParlaMint project. We believe that the ParlaMint-PT corpus, by using comparable structure, encoding and annotation to the other corpora of the project, will provide a crucial resource for studies on the Portuguese Parliament and for contrastive studies of the European political system.

3. The Constitution of the Portuguese Parliament – 2015-2022

A single chamber of Members constitutes the Portuguese Parliament (Assembleia da República – AR). The Members of Parliament (MPs) are elected by universal, direct and secret suffrage in legislative elections that take place every four years. The Portuguese Parliament is constituted by the Plenary (corresponding to the elected MPs' seats) and the Bureau. At the start of the legislature, the Assembly elects its President and the remaining members of the Bureau (four vice presidents, four secretaries, and four vice secretaries). The Parliament has a total of 230 seats. Parliamentary proceedings include periods for plenary sittings, parliamentary committee and parliamentary group meetings, and for MPs to spend time on constituency business. The ParlaMint-PT corpus focuses on the plenary sittings and the transcripts of these sessions, which occur, typically, three times a week. Still, it also includes solemn sessions (e.g., commemorative sessions of the *25 de Abril*, or the inauguration session of the President of Portugal).

The ParlaMint-PT corpus was created to cover the temporal period before, during and immediately after the COVID-19 pandemic, which had such an impact on the health and lives of European and global citizens. The data was intended to observe how national parliaments had addressed public health issues and the relationship between political orientation and type of proposals (for example, on vaccination).

In the Portuguese case, in addition to the pandemic period, the years covered by the corpus are also a time of major changes in the configuration of the party system. The corpus covers the last

[1] http://gamma.clul.ul.pt/CQPweb/crpc

[2] https://portulanclarin.net

10 months (January to October 2015) of the XII Legislature, and the full XIII (2015-2019) and XIV (2019-2022) Legislatures. A Legislature (Term of Office) covers the period between legislative elections.

In the 2015 legislative elections, despite the austerity policy imposed by the government of the PSD party (center-right) in the XII Legislature, PSD was surprisingly the most voted party. Nevertheless, it did not succeed in establishing a stable majority in the Parliament, and the PS (socialist party) took office, supported by governance agreements signed with the Left (PCP, PEV, BE). In the 2019 elections, the PS was the party with the most votes, although without an absolute majority. No coalition agreements were signed with parties to the left of the PS, but there were specific agreements in the Parliament for passing bills.

Between 2015 and 2022, the Parliament's configuration underwent major changes. Table 1 presents the number of speakers per party in each Legislature. Parties are identified by their acronym and are listed from Far-Left (FL), Left (L), Center-Left (CL), Center (C), Center-Right (CR), Right (R) and Far-Right (FR). No numerical information is provided in the Table when the party did not exist at the time. The ParlaMint-PT corpus covers the first three Legislatures, from 2015 to 2022. From 2019 to 2022, some of the parties saw their number of speakers decrease. This is the case of the PCP, perhaps reflecting a negative reaction from their electorate to the support given to the PS (De Giorgi and Russo, 2018), its traditional opponent since the revolution of April 25, 1974. Also, the CDS-PP significantly reduces its electorate from 24 speakers in the XII Legislature to 5 speakers in the XIV Legislature. The BE has 19 speakers in this period, surpassing the communist party PCP. The Livre party finally managed to elect a speaker during this period. And several new parties were created and quickly succeeded in electing Parliament members. The PAN party, with environmental concerns, elected 1 speaker in 2015 and increased its representation to 4 speakers in 2019. On the right wing of the political spectrum, two new parties emerged, Iniciativa Liberal (IL) and Chega, which elected 1 speaker each in 2019. With the election of a speaker from the populist party Chega, Portugal ceased to be the only country in Europe that did not have a populist far-right party with parliamentary representation.

When the Left parties refused to approve the budget proposed by the PS in 2022, the President of the Republic dissolved the AR and called elections, resulting in the XV Legislature. This Legislature is not included in the corpus (nor in Table 1). Still, it is worthwhile to provide some information about its composition, as it shows how the 2019 vote was not an isolated moment but rather pointed to trends in the reconfiguration of the political party system. In 2022, the PS has an absolute majority; PCP and BE suffer a drastic reduction to 6 and 5 speakers, respectively; the CDS party no longer has parliamentary representation; on the contrary, the two new parties on the right increase the number of speakers from 1 to 8, in the case of Iniciativa Liberal, and from 1 to 12 in the case of CHEGA. Recently, a corruption investigation in which the Prime Minister's name was mentioned led him to resign, and the President of the Republic dissolved the Parliament. In the elections of March 2024, the parliamentary group of the party CHEGA increased to 50 speakers, a process that is reminiscent of the growth of Marine Le Pen's party in France, and of the political situation in other countries in Europe.

Party	XII Leg.	XIII Leg.	XIV Leg.
PCP (FL)	14	15	10
PEV (FL)	2	2	2
BE (FL)	8	19	19
Livre (L)	0	1	1
PS (CL)	74	86	108
PAN (C)	0	1	4
PSD (CR)	108	79	77
IL (CR)	-	-	1
CDS (R)	24	18	5
Chega (FR)	-	-	1

Table 1: Number of speakers per party in each Legislature
XII=01.01.2015-22.10.2015; XIII=23.10.2015-24.10.2019; XIV=01.11.2019-01.02.2022

It would naturally be interesting to enlarge the corpus in the future to include the XV and XVI Legislatures, to study the evolution of the activities in the Parliament, the topics discussed, and also the type and register of the interventions in the sessions.

4. Parliamentary Raw Data

The Portuguese Parliamentary Corpus' raw data consists of transcripts of Portuguese Parliament sessions. These transcripts were gathered from the official Portuguese Parliament website. On the website, each transcript of the parliamentary sessions is available via the publication of the official journal of the Parliament, the Journal of the *Assembleia da República* (*Diário da Assembleia da República*). The transcripts are available for download in two file formats: TXT and PDF.

The Portuguese Parliamentary Corpus comprehends transcripts of sessions in the time period from 1 January 2015 until 22 March 2022. The cor-

Reference corpus
XII (01.01.2015-22.10.2015)
XIV (01.11.2019-22.03.2022)
XIII (23.10.2015-24.10.2019)
COVID Corpus
XIV (25.10.2019-31.10.2019)

Table 2: Time period of the Reference subcorpus and the COVID subcorpus

	Reference	**COVID**
Session days	499	205
Number of utterances	121,317	49,620
Number of words	11,570,662	5,882,413

Table 3: Contents of the Reference and of the COVID subcorpora

pus was divided into two subcorpora, according to the period each one covers: (i) the reference subcorpus covers sessions from 1st January 2015 until 31st October 2019; (ii) the COVID subcorpus comprehends sessions between 1st November 2019 and 22nd March 2022. The time periods considered, as well as the division into two subcorpora taking into account the start of media coverage about COVID, follow Parla-CLARIN general guidelines and proceedings for parliamentary corpora (Erjavec and Pančur, 2019). The time period of each subcorpus is provided in Table 2. Quantitative information about the number of session days, utterances and words in each subcorpus is given in Table 3.

4.1. Metadata Collection

Regarding metadata, the Portuguese Parliamentary Corpus makes available information concerning the corpus data, the speakers, the political parties, and the session files. More general information is also included, such as the type of parliament (unicameral) and the structure of the proceedings (taxonomy with types of meetings, types of speakers, legislative periods).

The Portuguese corpus provides information regarding the speaker's ID, name and surname(s), birth date, death date, gender, political affiliation (only for MPs, not for occasional speakers), and the status of the speaker (role and role description). The information regarding political parties consists of the abbreviation of the party, the full name of the party (in Portuguese), and the party ID (which is the same as the abbreviation). Finally, the metadata concerning the session files encompasses date-stamped mandates, sessions and speeches. Each session contains the transcripts of the speeches divided into utterances and paragraphs. However, the transcripts also contain the transcribers' commentary, which was retained and encoded. Each speech turn (i.e. utterance) is accompanied by the date, speaker ID, and role of the speaker (chair, regular or guest).

As for the roles attributed to speakers, the *chair* corresponds to the President of the Parliament, designated in Portuguese as *Presidente da Assembleia da República*; *regular* encompasses different situations: the prime minister, ministers and state secretaries (members of the Government), regular MPs from each party elected in legislative elections for the Portuguese Parliament, and MPs that were elected by the Parliament members as vice presidents and secretaries of the Parliament and aid the chair; the term *guest* identifies any visitor, often a member of a foreign country's Government, invited to speak in a Parliament session. The metadata files contain a description of the different roles fulfilled in public office by each member of the Parliament in different time periods and Legislatures.

The information compiled in the metadata was gathered from the official Portuguese Parliament website. This website provides webpages with political and biographic information for each politician who is or was an elected MP, secretary, vice president, or President of the Portuguese Parliament. In a few cases, the information available on the Parliament website was complemented by further research on newspaper articles or on Wikipedia pages of Portuguese politicians.

4.2. Structure of the Portuguese Parliament Plenary Sittings

In building the corpus, we must consider the structure of each plenary session and the particularities of the transcripts published in the Journal of the *Assembleia da República* (*Diário da Assembleia da República*). As it will be made clear, identifying different and regular parts of the political debates and transcriptions was crucial to the production and processing of the XML corpus.

The Portuguese Parliament plenary sittings transcripts are structured in distinctive moments, each providing various types of information that need to be encoded accordingly. The first one is the *Preamble*, which includes the identification of key features and figures in the session: (i) the date, series and number of the Journal of the *Assembleia da República*; (ii) the Legislature and session number; (iii) the date; (iv) the chair, and (v) the secretaries. After the *Preamble*, we find the *Summary*, a brief description of the interventions and votings that took place during the session. Then, we have the *Beginning of the Session*, which overlaps with the chair's first intervention and includes a time stamp. Next is the *Debate*, which corresponds to the core of the session, where we find the different speeches and interventions of the MPs. After the *Debate*, the session usually proceeds to vote on bills, and, thus, we have a section that corresponds to *Voting*. The *Closing* section follows the chair's last intervention, including a time stamp. Finally, some transcriptions end with *Written Voting Declarations*, an appendix to the session. They are not part of the debate itself but correspond to written declarations that the MPs may deliver to the Bureau in order to further justify or explain their voting during the session. We used the linguistic markers that we consistently found associated with each of these moments of the debate sessions to automatically identify the moments in the transcription files, as shown below.

As mentioned, the transcriptions include commentary by the transcribers, which were annotated by type in the XML files. These comments can occur at any moment of the debate. They pertain to pieces of information such as time, date, indication of sections such as summary, or of moments in the debate such as voting, indications of pauses in the debate, and events (e.g., an MP shows a visual aid during their intervention; the chair is replaced by the vice-chair). They may also indicate non-vocalized communicative phenomena (e.g., clapping) or vocalized, but not necessarily lexical, communicative phenomena (e.g., shouting, laughing, protests).

5. Production of the XML Corpus

In this section, we will describe how the Portuguese Parliamentary corpus was prepared for the XML generation, which required information about the actual sessions, as well as all the entities involved in those sessions.

The information about the several entities involved (people, governments, legislatures and so on) required some research so a few TSV files could be compiled and then used as a source for the needed elements. On the other hand, the information about the sessions was only available in text format, which meant they had to be processed in order to produce the corresponding XML files. However, the texts appeared to have been obtained from PDF files, which, in turn, seemed to have been obtained from OCR of the physical paper documents, considering the many issues found in them. Fortunately, after a brief inspection, we realized the texts had a fairly regular structure, with several text sections that could be used as anchors, contributing to simplifying the automation process.

We divided the text processing into several stages, which had the advantage of allowing us to focus on specific issues and keeping them localized. All the stages were carried out iteratively since a failure on a given stage might result from an error on an earlier stage. The several stages are described in the subsections below.

5.1. Preprocessing of Texts

There was a significant number of issues found in the texts, and since we planned on using text markers to identify and extract relevant information, we introduced a first stage in which we focused on fixing those issues.

This way, we could rely more confidently on such markers by ensuring a more uniform structure of the texts and avoiding the introduction of exceptions when looking for such markers. Many of the issues found consisted of cases like the following:

- missing (or extra) spaces, parentheses or dashes;
- mistaking the letter 'o' with the digit '0', and vice-versa;
- Unicode characters which looked similar and required uniformization.

These corrections were accomplished using simple regular expression replacement. Then, we proceeded to discard page headers and footers like numbers, dates, or series, which sometimes ended up between paragraphs spanning more than one page, trying to reestablish text paragraphs. A given number of empty lines, some specific separator symbols, and an initial letter casing were also considered.

Once the paragraphs were identified, we moved on to removing line breaks that did not correspond to new sentences. This was accomplished by checking the end of a line and the start of the next for composed words separated by a dash, letter casing, specific symbols and exception cases.

5.2. Main Sections Identification

At this point, we opted to identify section limits like summaries, interventions, and interruptions, which was done by looking at expressions that would indicate such cases. In our case, we could identify the following main sections:

- head: which in turn had date, session, permanent, title, president and secretary sections;
- summary: which implicated the identification of the time in which the session started;
- main: which implicated the identification of the time in which the session ended;
- final: used for any voting information.

Once these main sections were identified, we were able to improve the paragraphs further by eliminating additional line breaks that did not correspond to new sentences. At this stage, the XML document creation can be carried out in a much simpler way.

5.3. Automatic XML Generation and Checking

The preprocessing of the session files facilitated the implementation of the procedures to produce the XML files.

This time, the information about the entities was also considered to produce the final version. Even though the text files complied with a fairly regular structure, as mentioned above, we had to account for the possibility of errors, which raised the need to check if things were fine. This is why, after creating the XML document, we carried out an additional checking stage that allowed us to identify several situations in which there was missing or unexpected information, which in turn enabled us to look further into the problematic files and fix them.

During this checking stage, we found recurring errors throughout the documents, which affected the identification of utterances, paragraphs, and different types of transcribers' commentaries and events. A close reading of the texts allowed us to identify specific linguistic elements that were consistently used to introduce those sections in the transcriptions (e.g. specific adverbial expressions are used to indicate events or votings, such as *Entretanto* 'In the meanwhile', *Neste momento* 'At this moment' or *De seguida* 'Then') and what specific textual elements were associated with processing errors (e.g. punctuation marks were often associated with errors: every utterance was identified by a colon followed by a dash in the transcription, but these punctuation marks were not always correctly identified as the beginning of an utterance; periods after abbreviations were, in some cases, misidentified as an indicator of the end of paragraph). A set of expressions was compiled from these errors in order to allow an automatic search throughout the XML files. To do so, we automatized the search task by recording a macro using Notepad++, which allowed us to perform searches simultaneously in multiple files. The search results enabled us to focus our attention on a reduced set of possible problematic areas to correct any identified errors manually.

5.4. Syntactic Annotation and Main XML Files

Additional information about the session files needed to be included, namely the POS tagging and Universal Dependency Relations (UDR) identification for the session interventions, which could only be carried out after the basic XML files were produced.

The POS tagging was established using the MBT tagger (Daelemans et al., 1996) trained over the CINTIL corpus (Barreto et al., 2006). We adapted the tagset to be conformant to the UD POS tags used in ParlaMint. The CINTIL corpus includes NER annotation. We lemmatized the corpus with MBLEM (van den Bosch and Daelemans, 1999), which combines a dictionary lookup with a machine learning algorithm to produce lemmas. As a basis for the dictionary, we used a list of wordform – POS-tag combinations mapped to lemmas. This list was produced in-house. The dictionary used in MBLEM contains 102,196 word forms combined with 27,860 lemmas, leading to 120,768 wordform-lemma combinations. The adaptation of the MBT tagger and MBLEM lemmatizer are described in (Généreux et al., 2012).

The UD Relations were established using the LX-UD dependency parser[3], adapted to the set of POS and relation types used in ParlaMint. The UDR tool took a very long time to run, particularly considering the great number of session files, so it became really important to run tasks in parallel. Such parallel processing was implemented within a single file, in which we were able to carry out more than one process per sentence, as well as within a set of files, in which we were able to process several files at once. This approach allowed us to obtain results seven times faster.

6. Using ParlaMint-PT to Explore Political Views on Immigration

The topic of immigration is controversial and is frequently addressed in the programs of the political parties. As such, we expect the discussion of immigration issues and legislation proposals to be identified in the transcriptions of the Parliament sessions and to shed some light on the position of the government and of the opposition regarding the topic. The ParlaMint corpora enable us to test whether some variables are relevant to the political position of the MPs, for instance, political orientation or gender. In Europe, migration routes in the Mediterranean have put pressure on South-East countries, such as Greece and Italy, but they also affect countries in the North. Portugal has not been on the route of this migration, but, according to official numbers in the PORDATA portal[4], the foreign population officially residing in Portugal has been increasing, especially

[3] https://portulanclarin.net/workbench/lx-udparser

[4] https://www.pordata.pt/subtema/portugal/migracoes-34

since 2016, and, in 2022, reached around 800,000 people (with the total population of Portugal being around 10 million). Of the nationalities that immigrate to Portugal, the most notable are immigrants who originate from Portuguese-speaking countries, especially Brazil with 240,000 residents, and more recently, immigrants from the South-East, such as India. The latter work in large agricultural productions and, in some cases, they outnumber the local population, creating some concerns and the need for the local authorities to prepare lines of action for better integration (see, for instance, the town hall program for the integration of immigrants in Odemira (AAVV, 2015-2017)), in the South.

The press and social media have been a frequent source of data related to the perception of migration (Taylor, 2014), but parliamentary speeches are also an interesting source of data, as shown in the project "Who is the enemy now?" based on the UK and Italian ParlaMint corpora (Del Fante and Zorzi, 2023). The project reports similarities between the discourse used in both countries in spite of differences in their political backgrounds, such as the fact that the UK government was of the Conservative Party, while ministers in Italy were mostly from the left wing.

To query the corpus, we establish a list of keywords (and inflected variants) related to the foreign population living in Portugal, such as *imigração* (immigration), *imigrante* (immigrant), *migrante* (migrant), and *refugiado* (refugee). We use the version of the corpus available on Sketch Engine (Kilgarriff et al.) and extract concordances and frequencies. Here, we discuss the word *migrante(s)* that occurs 409 times in the corpus. The list of sessions where the word was most used is presented in Table 4 with the frequency of the word and the relative density (above 100% shows that the word is more frequent in this text type (session) than in the corpus). It shows a significant increase in occurrences from 2016 and 2021. This is in line with the rise of the foreign population in Portugal reported in PORDATA. The results are also aligned with the frequencies found in (Del Fante and Zorzi, 2023) for UK and Italian corpora: the word *migrant* in English and its equivalent in Italian show a strong increase in frequency in both corpora, independently of the political orientation of the government of both countries. This increase is also found in the Portuguese data, where a Center-Left party was in government during the XIII and XIV Legislature, with support from the Left parties.

date	**freq.**	**rel.** (%)
16-03-2016	12	1,675.86%
02-03-2017	11	1,669.85%
22-06-2018	11	2,284.92%
16-12-2020	25	2,841.95%
27-05-2021	37	6,653.24%
09-07-2021	44	8,152.84%

Table 4: Sessions with the higher frequencies of the word *migrante(s)* in ParlaMint-PT

gender	**freq.**	**rel.** (%)
F	243	180.57%
M	166	60.49%

Table 5: Distribution of the word *migrante(s) per gender of the speaker*

Two other variables seem to be related to the use of the word *migrante* 'migrant'. One of them is the gender of the speaker, as reported in Table 5. Speakers of the feminine gender use the word more frequently than speakers of the masculine gender (243 vs. 166). Although feminine members of the Parliament are in the minority, they account for a higher number of occurrences: the relative density shows that the term is not typical of masculine Parliamentary discourse (under 100%), while it is typical of the feminine Parliamentary discourse (above 100%). Another variable is political orientation, as shown in Table 6. As the number of speakers from each party differs considerably (see Table 1), Relative density is a better indicator than raw frequency. The values in Table 6 points to a higher use of the word *migrante* by Left to Far-Left and Center-Left parties. The Right to Far-Right orientation party "Chega" has a single speaker in the Parliament and shows the highest relative density, with 179.43%).

The concordances of the “Chega” party refer to the need to control an unbelievable flux of migrants and connect the reference to migrants to the traffic of human beings, as in example 1.

(1) precisamos de controlar o fluxo inacreditável quer de **migrantes**, quer de *tráfico de seres humanos* ‘we need to control the unbelievable flow of migrants and of the human being

party orientation	**frequency**	**rel** (%)
Left to Far-Left	147	142.08%
Left	10	48.74%
Center-Left	189	131.10%
Center-Right	38	49.03%
Center-Right to Right	18	33.58%
Right to Far-Right	7	179.43%

Table 6: Distribution of the word *migrante(s)* per the political orientation of the speaker

traffic'

The reference to a flow uses a metaphorical representation of migration as a liquid, also present in the UK and IT corpora (Del Fante and Zorzi, 2023). Three other contexts of the party "Chega" refer to the concern of the government and of the Left parties with the life/work conditions of the migrants, in contrast with those that were "born in our land" (*quem nasceu na nossa terra*). It would be interesting to analyse the XV Legislature when the party "Chega" increased its number of speakers from 1 to 12.

While political orientation is certainly important, one also needs to take into consideration the political parties. For instance, the two parties with a Left to Far-Left orientation, the Communist Party and the Bloco de Esquerda, differ in the frequency of use of the word 1. The relative density of 240.27% of the Bloco de Esquerda contrasts with the 37.86% in the case of the Communist party.

7. Final Remarks

The new open-access corpus ParlaMint-PT provides an opportunity to explore the interventions of the speakers of the Portuguese Parliament, by giving information on the topics that are addressed in the Parliament and on the views of individual speakers or political parties, or general patterns of use related to genre, time period and political orientation.

We reported on the contents of the corpus, the metadata, and the syntactic annotation. As a case study using this resource, we provided some data on the political views on immigration in the Portuguese Parliament. The automatic translation to English of the national corpora also enables comparative studies on national views over topics that are of relevance to the social and political situation of Europe today.

8. Acknowledgements

This work was partially supported by CLARIN ERIC, *ParlaMint: Towards Comparable Parliamentary Corpora*, and by Fundação para a Ciência e a Tecnologia, under the grant UIDP/00214/2020. We wish to thank the anonymous reviewers for their helpful comments.

9. Bibliographical References

AAVV. 2015-2017. Odemira integra - plano municipal para a integração dos imigrantes.

P. Almeida, M. Marques-Pita, and J. Gonçalves-Sá. 2021. PTPARL-D: an annotated corpus of forty-four years of Portuguese parliamentary debates. *Corpora*, 16(3):337–348.

F. Barreto, A. Branco, E. Ferreira, A. Mendes, M.F.P. Bacelar do Nascimento, F. Nunes, and J. Silva. 2006. Open resources and tools for the shallow processing of Portuguese. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006). Genoa, Italy*.

W. Daelemans, J. Zavrel, P. Berck, and S. Gillis. 1996. MBT - a memory-based part of speech tagger generator. *Proceedings of the 4th ACL/SIGDAT Workshop on Very Large Corpora. pp. 14–27.*

E. De Georgi and C. Moury. 2015. Government-opposition dynamics in Southern European countries during the economic crisis. great recession, great cooperation? *Journal of Legislative Studies*, 21.

E. De Giorgi and F. Russo. 2018. Portugal: The unexpected path of far left parties, from permanent opposition to government support. In E. De Giorgi and G. Ilonszky, editors, *Opposition parties in European legislatures*. Routledge.

D. Del Fante and V. Zorzi. 2023. ParlaMint - a resource for democracy. Https://www.clarin.eu/impact-stories/parlamint-resource-democracy.

T. Erjavec and A. Pančur. 2019. Parla-clarin: TEI guidelines for corpora of parliamentary proceedings. Technical report.

T. Erjavec et al. 2023. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 57(1):415–448.

J. M. Fernandes, M. Goplerud, and M. Won. 2019. Legislative bellwethers: The role of committee membership in parliamentary debate. *Legislative Studies Quarterly*, 44(2):307–343.

M. Généreux, I. Hendrickx, and A. Mendes. 2012. Introducing the Reference Corpus of Contemporary Portuguese On-Line. In *LREC'2012 – Eighth International Conference on Language Resources and Evaluation*, pages 2237–2244, Istanbul, Turkey. European Language Resources Association (ELRA).

A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlỳ, and journal=Lexicography volume=1 number=1 pages=7–36 year=2014 publisher=Springer Suchomel, V. The Sketch Engine: ten years on.

C. Taylor. 2014. Investigating the representation of migrants in the UK and Italian press: A cross- linguistic corpus-assisted discourse analysis. *International Journal of Corpus Linguistics*, 19(3):368–400.

A. van den Bosch and W. Daelemans. 1999. Memory-based morphological analysis. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, ACL'99. pp. 285–292.*

10. Language Resource References

A. I Carvalho and R. Lafuente. 2010. *Demo.cratica*. PID HTTP://demo.cratica.org/.

T. Erjavec and others. 2021. *Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1*. PID HTTP://hdl.handle.net/11356/1431.

E. D. Giorgi and A. Dias. 2019. *Portuguese Observatory on Parliamentary Dynamics Database (POPAD): information on legislative process, scrutiny activity and speeches in the Portuguese Parliament*. PID HTTPS://popad.org/.

Gender, Speech, and Representation in the Galician Parliament: An Analysis Based on the ParlaMint-ES-GA Dataset

Adina Vladu, Elisa Fernández Rei, Carmen Magariños and Noelia García Díaz
Instituto da Lingua Galega (ILG), Universidade de Santiago de Compostela, Spain
adina.vladu, elisa.fernandez, mariadelcarmen.magarinos, noeliagarcia.diaz@usc.gal

Abstract

This paper employs the ParlaMint-ES-GA dataset to scrutinize the intersection of gender, speech, and representation within the Parliament of Galicia, an autonomous region located in North-western Spain. The research questions center around the dynamics of women's participation in parliamentary proceedings. Contrary to numerical parity, we explore whether increased female presence in the parliament correlates with equitable access to the floor. Analyzing parliamentary proceedings from 2015 to 2022, our quantitative study investigates the relationship between the legislative body's composition, the number of speeches by Members of Parliament (MPs), and references made by MPs in their speeches. The findings reveal nuances in gender representation and participation, challenging assumptions about proportional access to parliamentary discourse.

Keywords: ParlaMint, Parliamentary debates, Gender, Representation

1. Introduction

Parliamentary discourse studies hold a crucial place in understanding the functioning of democratic institutions and the mechanisms that underpin political representation. The debates that take place within legislative bodies can be seen as a reflection of power dynamics, decision-making processes, and the overall health of democratic governance. In this context, an emerging area of focus within the analysis of political discourse centers on the analysis of gendered speech and language, specifically investigating the role of women in parliamentary settings.

Parliamentary debates have the potential to uncover hidden power structures and implicit biases that may affect the equitable participation of diverse voices in the political arena. Beyond the numerical representation of women in parliament, a deeper examination of their linguistic contributions becomes imperative. The frequency of speeches and the references made during parliamentary proceedings offer insights as valuable into the nuanced challenges faced by female representatives as the very content of the words that are spoken in the parliamentary context.

In this context, richly and homogeneously encoded comparable corpora such as ParlaMint (Erjavec et al., 2023), can offer researchers an interface between multiple academic fields and thus open the door to research that combines humanities and computational methods of analysis.

Within the realm of politics and gender studies, an ever growing body of work delves into women's representation in legislative contexts, especially from the perspective of gendered speech and language, emphasizing their pivotal roles in shaping political representation (Raiber and Spierings, 2022).

Though modern democratic systems ensure that both genders are represented in the political system, recent studies show that numeric parity between females and males does not necessarily translate into a more equal parliamentary representation, as women are historically known to give fewer speeches than men (Bäck et al., 2014). That is, being included does not always guarantee being heard (Sanjaume-Calvet et al., 2023). Focusing on the case of the Spanish Parliament, Sanjaume-Calvet et al. (2023) argue that Parliaments perpetuate a gendered political structure, where an increased presence of female representatives is not necessarily indicative of an increased access to the floor. The authors further show that female participation in parliamentary proceedings in Spain is highly constrained by party structures, as access to the floor is always controlled by the organization of parliamentary groups, which represent specific political parties or coalitions.

The power relations in national parliaments are not only reflected in the access to the floor but also in the way speeches influence others and are referenced by others (Skubic et al., 2022). Skubic et al. (2022) argue that gender can affect argumentative power, as comparatively high numbers of female MPs do not generate high numbers of speeches made by female MPs nor high numbers of mentions by fellow speakers.

Another factor that can influence how female MPs interact and make use of their voice in parliamentary debates is their political position. Müller and Pansardi (2023) argue that female leaders usually use more effective communication skills and express either strong support or clear opposition

ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora, pages 21–29
May 20, 2024.

more emphatically than male leaders.

This paper aims to analyze the dynamics of women's participation in parliamentary proceedings by scrutinizing the number of speeches made by female and male participants, comparing it to the number of parliamentary members, and taking a closer look at the relationship between speeches and references or mentions among MPs, in order to shed light on the complex aspects of the power dynamics at play in parliamentary discourse and political deliberations.

The paper is structured as follows: Section 2 presents the research questions addressed. In Sections 3 and 4, we review the data used in our study and the methodology used to analyze it. Section 5 discusses the results of the analysis performed on the dataset. Finally, the conclusions highlight the most important outcome of our work.

2. Research Questions

We were interested in taking a closer look at the concepts of representation and gender (in the traditional terms of male and female) in the Galician Parliament through a quantitative analysis of the relation between the composition of the legislative body and the number of speeches made by MPs, as well as the references made by MPs in their speeches to others, both fellow MPs and persons outside the debate.

We base our interpretation of the concept of representation on the definition by Pitkin (1967), who states that representation is closely connected to the concept of power, which is mainly expressed through descriptive, substantive, and symbolic representation. Furthermore, by representation, following Raiber and Spierings (2022) and Skubic et al. (2022), we understand the following essential issues: the extent to which women are allowed to and do participate in political debate, as mirrored in the quantity of their speeches in the parliamentary context; and the relevance of their participation in such debates, as mirrored in the interaction with other fellow participants.

Thus, we explore the following research questions:

1. Does a relatively balanced male and female presence in the parliament correlate with equitable access to the floor?

2. How can the intersection between the number of speeches, number of mentions, gender, and identity of the speakers who participate in the debates of the Galician Parliament shed light on the representation that is made manifest by this participation?

3. Data

The data analyzed in this paper come from the ParlaMint-ES-GA dataset, a body of parliamentary debates in Galician language spanning over a period of approximately seven years and three legislative terms (2015-2022). The dataset is part of the larger, multilingual ParlaMint 4.0 corpus (Erjavec et al., 2023; Erjavec et al., 2023). In the subsections that follow, we take a closer look at the source of the data, the Galician Parliament, and at the ParlaMint-ES-GA dataset.

3.1. The Galician Parliament

The Galician Parliament is the legislative body governing the autonomous community of Galicia, in North-western Spain. Comprising 75 members, this assembly is elected every four years, known as legislative terms, through a proportional representation system. Its primary functions encompass the enactment of legislation and the oversight of the regional government's activities.

Characterized by a unicameral structure and a multi-party political system, the Galician Parliament is chaired by a President (or Chairperson) elected by the members. The Chairperson assumes the responsibility of ensuring adherence to procedural rules and fostering deliberation and debate among the members. The members of Parliament (MPs), referred to as "deputados" or "deputadas" in Galician, are organized into parliamentary groups, representing their respective parties or electoral coalitions. A "Mixed Parliamentary Group" is available for MPs who do not otherwise meet the requirements to form a parliamentary group. Traditionally, a limited number of parties, usually up to five, secure representation in the Galician Parliament. Members of the regional Government (ministers or MGs), "conselleiros" or "conselleiras" in Galician, can intervene at any time during the debates (Parlamento de Galicia, 2020). The regional Government traditionally comprises up to 12 MGs headed by a President.

The composition of the Parliament and Government tends to be slightly imbalanced in terms of gender. In the three legislative terms that overlap with the ParlaMint-ES-GA dataset, elected female MPs and MGs represent 43.2%, 42.2%, and 45.4%, respectively, whereas male MPs and MGs represent 56.7%, 57.8%, and 54.5%, respectively. The average gender representation throughout the three-term period is 43.6% female and 56.3% male.

3.2. The ParlaMint-ES-GA Dataset

The ParlaMint-ES-GA dataset comprises transcriptions of parliamentary proceedings spanning three legislative terms (2015 - 2022), with a total of 302 files representing individual sittings. The dataset is enriched with metadata about legislative periods, governments, speakers, and political parties. It is encoded in TEI ParlaMint format (Erjavec and Pančur, 2022, 2019) and linguistically annotated following the Universal Dependencies formalism (Nivre et al., 2017) and with named entities recognition (NER).

The dataset contains a total of 83,078 speeches from 227 individuals distributed as follows: 47.1% female speakers and 52.9% male speakers. The total number of speakers in the dataset comprises elected MPs and MGs, substitute MPs and MGs, and guest speakers. Figure 1 illustrates the apparent correspondence between the gender composition of the Galician Parliament and regional Government, and the ParlaMint-ES-GA dataset. In comparing these statistics, a note should be made to the fact that, as mentioned, the ParlaMint-ES-GA dataset also includes interventions from 14 guest speakers, more specifically 8 males and 6 females.

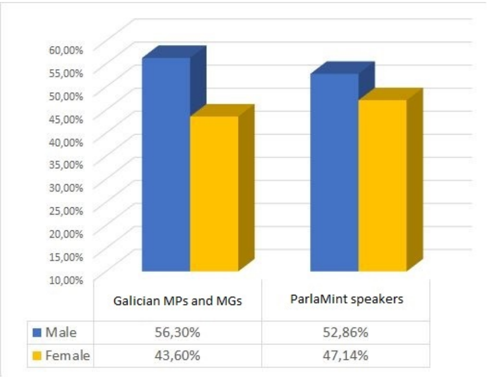


Figure 1: **Gender composition:** MPs elected to the Galician Parliament, together with MGs elected to the Galician regional Government, compared to speakers in ParlaMint-ES-GA.

However, more than 40% of all the speeches in the dataset are uttered by the Chairperson who, nevertheless, pronounces a very small percentage of the total number of words. Table 1 summarizes the relation between gender, role, number of speeches and number of words in ParlaMint-ES-GA.

With regard to political representation, the main parties of the Galician political scene are proportionally represented in the dataset (Partido Popular de Galicia 34.6%, Partido de los Socialistas de Galicia 22.2%, Bloque Nacionalista Galego 17.7%, En Marea 10.4%, and Alternativa Galega de Esquerda 4%).

An interesting particularity of the dataset, which distinguishes it from its Catalan (Pisani et al., 2023) and Basque (Escribano et al., 2022; Alkorta and Quintian, 2022) counterparts, is that ParlaMint-ES-GA can be characterized as almost completely monolingual. With the rare exception of one guest speaker who intervened in Spanish, and leaving aside verbatim quotes from Spanish politicians or media, all speeches are made in Galician language, even though, as described in Vázquez Somoza (2015), the influence of the Spanish language on the speakers' Galician is apparent.

4. Methodology

We were interested in analyzing the number of speeches made by male and female speakers in the ParlaMint-ES-GA dataset, as well as retrieving the mentions made in the dataset to MPs and other individuals.

To this end, we used the TEI annotation of the dataset to retrieve speeches. Each speech is marked with a unique SpeakerID composed of the speaker's first surname and name, as well as with the speaker's role within the Parliament, among other metadata. Once we had carried out a quantitative mapping of the total number of speeches (see Table 1), we filtered out all speeches made by the Chairperson, identified in the dataset with the role "Chair", due to their mainly procedural function in the debates. We also discarded guest speakers, identified as "Guest" in the dataset, given their incidental presence in the debates. We used the metadata provided to divide the data into "male" and "female" speeches, as well as to identify speakers and quantify individual participation in the debates that make up the dataset.

In order to obtain the references or mentions, we used the NER annotation of the dataset to extract the named entities that designated persons from the speeches of MPs/MGs with the role of "Regular" (speaker). Again, as described above, we discarded speeches made by the Chairperson given their strictly procedural role in the debates, and guest speakers, who participation was merely incidental. As we were interested in unequivocally identifying persons mentioned in order to be able to compare data of speakers actively mentioning other persons and persons being mentioned by others, we restricted our analysis to individuals mentioned by name. Thus, using regular expressions, we filtered the results to include only those references that used a proper name. Finally, 45,051 named entities were selected that used the formula "(o/a) S/señor/a [Apelido/Nome e Apelido(s)]", Galician for "Mr./Mrs. [Surname/Name and Surname]". This approach excluded from the results any references that made use of official titles not followed by a proper name (e.g., "señor/a concelleiro/a", "señor presidente") as well as all pronominal references.

In order to match the mentions to the speaker IDs of Galician MPs registered in the dataset meta-

Role	Speeches	Words	% of Total Speeches	% of Total Words
Male Chair	37,885	799k	45.6%	4.5%
Female Chair	7,103	141k	8.5%	0.8%
Male Regular	20,735	9045k	24.9%	51.1%
Female Regular	17,305	7601k	20.8%	43.1%
Male Guest	28	20k	0.03%	0.1%
Female Guest	22	28k	0.03%	0.1%

Table 1: Distribution of speeches and words according to gender and role in the dataset.

data, we used dedicated scripts that measured the Levenshtein distance of the identified mentions and checked gender accuracy with respect to the “señor/señora” gender marker in order to determine a list of probable IDs for each mention. However, the exact identification of each person referenced was not a straightforward task given the particularity of Galician (and generally Iberian) surnames. Individuals are identified by two surnames (one on the paternal side and one on the maternal side). In order to reference a person (e.g., "Paula Prado del Río"), the most common options are to use the paternal surname (e.g., "señora Prado"), the full surname (e.g., "señora Prado del Río"), or a combination of the name and paternal surname (e.g., "señora Paula Prado"). However, the maternal surname can also be used by itself, though less frequently. This complex way of referencing individuals, combined with the repetition of popular surnames (e.g., Rodríguez, Díaz, Sánchez, García) among Galician MPs, and other references such as Spanish politicians and other public figures, made it necessary to add further steps in order to ensure proper identification. If multiple possible matches were detected (i.e., identical surname/s and gender), we checked the names of the MPs that intervened in the Galician Parliament on the corresponding date and, out of the list of possible IDs, selected the ID that appeared closest to the processed mention. Any references that were not identified as matches were considered to be of persons outside the Galician Parliament.

A more complex issue was the surname coincidence between two Galician MPs and the head of the Spanish central Government, frequently referenced in the dataset, all three sharing the surname Sánchez. In the cases where the mention was a direct reference (i.e., the speaker was directly addressing the person referenced), the speaker ID was automatically assigned using the method described above. However, in the cases where there was an indirect mention (e.g., “o señor Sánchez”), the context of the mention was checked for specific terms referencing the Spanish central Government. If none were found in the ten words to the left and the right of the mention, the mention was considered to be of a Galician MP and the most probable ID was determined by the script described in the previous paragraph.

We can argue, in line with Skubic et al. (2022), that a speaker’s relevance, and, thus, representativity, in the parliamentary context can be measured by the number of speeches they pronounce (active relevance or AR) and by the number of times they are referenced by others (passive relevance or PR). In our analysis we go one step forward and take into account the fact that parliamentary debate is by nature dialogic. We expect, then, not only the number of speeches pronounced to be relevant but also the number of times a person mentions another, as this is bound to trigger a response from the person who has been referenced. Thus, in the case of mentions or references, we calculated what we defined as active mention (AM) and passive mention (PM). An active mention is any reference made in a speech or speech fragment to another individual identifiable by name. We expected this metric to differ from AR, as one person can reference multiple other individuals (or none) in one single speech. By passive mention we understand the opposite, that is, quantifying how many times an individual is referenced by others. Again, this metric does not necessarily coincide with the total number of speeches.

5. Results and Discussion

5.1. Gender Representation

Female members represent a total of 43.6% of elected MPs and MGs throughout the three legislative terms analyzed. Similarly, 47.1% of the total number of speakers represented in the ParlaMint-ES-GA dataset are female. If representation were directly proportional with the numbers of male and female speakers who were elected to the Parliament and intervened in the plenary sessions collected in the dataset, we could expect a reasonably similar gender participation in the debates and proportional access to the floor. Indeed, the data in Table 2 seem to support this hypothesis by showing a less than 2% difference between the quantity of speakers and speeches in the dataset, positive in the case of males and negative in the case of females.

However, a more complex analysis is neces-

Gender	Speakers	Speeches
Male	52.9%	54.5%
Female	47.1%	45.5%

Table 2: Gender distribution in speakers and speeches in the dataset

sary in order to determine whether these numbers are actually representative of the distribution of speeches by gender in the dataset. We expected that a relatively small number of speakers would concentrate a large part of the speeches, given that political position greatly determines the number of interventions and amount of speaking time that an individual can benefit from in parliamentary sessions. The data show that fewer than 15% of all MPs and MGs accumulate more than 50% of the total number of speeches pronounced, each with more than 350 speeches amounting to a percentage of between 0.9% and 5.8% of the total number of speeches pronounced by regular speakers (AR). Table 3 below details the relation between number of speeches and gender in this subset of speakers, whereas Figure 2 illustrates the speakers with the highest number of speeches in the dataset (>0.8% of the total number of speeches made by regular speakers). As expected, the most prolific speakers occupy important positions in the Galician political system: head and vice-president of the regional Government (NúñezAlberto, RuedaAlfonso), heads and Parliamentary representatives of different political parties (PontónAnaBelén, SánchezAntón, FernándezXoaquínMaría), ministers of socially relevant ministries such as Education, Finance, or Health (RodríguezRomán, CondeFranciscoJosé, VázquezAlmuíñaJesús), etc. It is important to note that many of these top positions in the fabric of the Galician regional Government and parliamentary group representation are occupied by male politicians. In contrast, some of the more active female speakers are representatives of the main opposition party.

	Speakers (out of total no.)	Speeches (out of total no.)
Male	7.9%	33.6%
Female	5.6%	22.4%

Table 3: Distribution of speakers with >350 speeches in the dataset

5.2. Gender Referentiality

We have already shown that female MPs have a lower active relevance than male MPs, as the former pronounce fewer speeches than the latter (45.5% of the total speeches pronounced by regular speakers in the Galician Parliament, as compared to 54.5% in the case of male MPs), which is also a lower figure than the percentage of female speakers (47.1%) present in the dataset. We have also established that females are less present than males in the top list of speakers (5.6% of the total number of speakers are female MPs who pronounce more than 350 speeches in the dataset, as compared to 7.9% male MPs).

In the case of mentions or references, Table 4 details the numbers and percentages of AMs and PMs in the dataset. While AMs are in line with the general statistics for number of speeches by gender in the corpus, PMs are considerably lower in the case of female MPs, amounting to less than half of the numbers for male MPs. In these results we can safely say that the large number of PMs corresponding to the president of the regional Government (who sums almost as many mentions as the following six highest-ranking individuals combined, and more than triples the number of mentions of the second person in the PM ranking), the vice-president, and the president of the Spanish central Government, all male politicians, play an important role.

Gender	AMs	AM (%)	PMs	PM (%)
Male	25,485	56.6%	31,196	69.2%
Female	19,566	43.4%	13,855	30.8%

Table 4: Active and passive mentions by gender in number of mentions and percentage of the total number of mentions

By calculating AMs and PMs for individual speakers and comparing it to the subset of individual speakers with high AR, it becomes apparent that a small subset of speakers not only has higher access to the floor, and thus a higher number of speeches (AR), but also accumulates a higher number of both active and passive mentions. That is, these individuals are not only referenced, but also actively engage in debate by referencing others. However, out of these 15 speakers, only four are women. Three of these four female MPs also score higher in active mentions than in passive mentions, which means that they actively generate debate by referencing others, but are not similarly referenced. These results are illustrated in Figure 3.

As in the case of AR, high AM and PM results can be explained by various factors. One of the more evident is political position (e.g., the highest-ranking person in all three categories occupied the regional Government Presidency for the best part of the period covered by the dataset, and the second-highest raking is the head of the opposing political party). Another possible factor is political party representation. The 15 higher-raking individuals in all three categories proportionately represent the

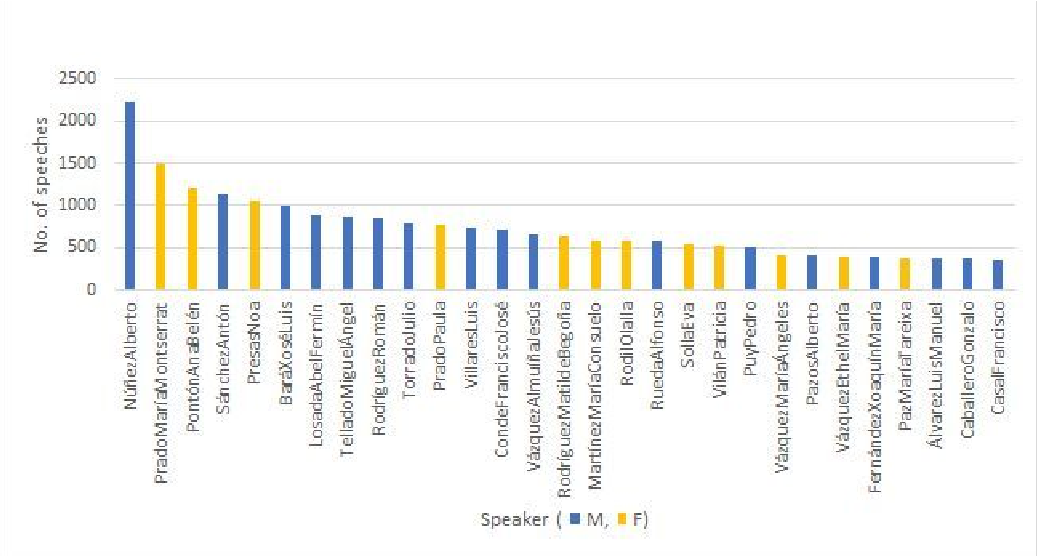


Figure 2: **Speakers with the highest numbers of speeches across the ParlaMint-ES-GA dataset**. The x-axis indicates the speaker ID and gender, while the y-axis indicates the number of speeches.

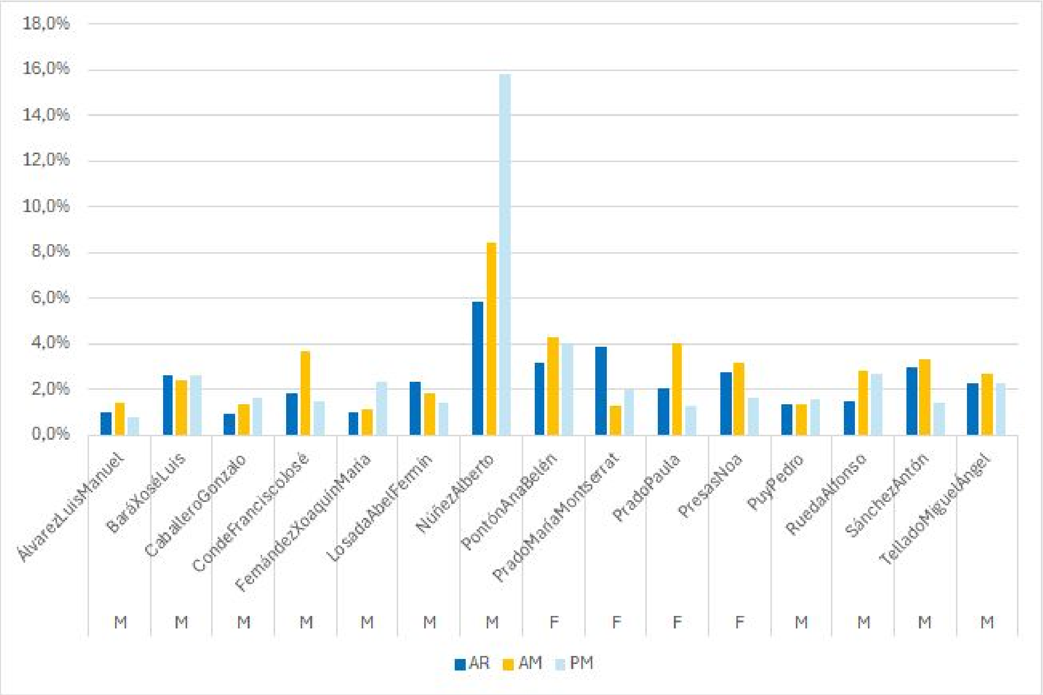


Figure 3: **MPs in the top 25 for all three categories: active relevance (number of speeches), active mentions, and passive mentions**. The x-axis indicates the speaker ID (in alphabetical order) and gender, while the y-axis indicates the percentage of speeches made by the speaker out of the total (AR), as well as the percentage of mentions made by the speaker (AM) or made of the person by others (PM).

main political parties in the Galician political system: six MPs represent the governing party (PPdeG) and nine the opposition (four BNG, four PSdeG-PSOE, and one AGE/En Marea/ANOVA). Finally, gender can play a role in high AM results, as women may feel obligated to generate debate in order to make their voice heard.

6. Conclusions

This paper aimed to analyze gender participation and representation in the parliamentary context. The analysis focused on evaluating the quantity of speeches delivered by both female and male participants in relation to the total number of parliamentary members. Additionally, the study delved into the interplay between speeches and references among MPs by taking a closer look at personal name mentions within the dataset.

Our examination of the Galician Parliament through the ParlaMint-ES-GA dataset supports previous studies stating that gender balance in parliamentary representation does not necessarily ensure equal participation. Despite the proportionate presence of female MPs, our analysis indicates disparities in speech frequency and references. Female MPs are referenced less frequently than their male counterparts, although they can compensate by contributing to the debate through actively referencing other individuals. The disparities relate heavily to the political position of speakers within the Parliament and regional Government. Top political leaders, both present in the debates and referenced, are still predominantly male, which shifts the balance of power.

The study highlights the complexity of gender dynamics in the political context, emphasizing the need to go beyond numerical metrics to assess true parliamentary inclusiveness. The findings also underscore the role of political power structures and party affiliations in shaping participation patterns, which highlights the need for a comprehensive understanding of the multitude of factors influencing parliamentary discourse.

7. Acknowledgements

The work described in this paper was funded by the following grant: Xunta de Galicia: Consolidación 2021 GRC GI-1743 - Filoloxía e Lingüística Galega - FILGA (2021-PG008) Ref.ED431C 2021/20.

The authors would like to thank the anonymous reviewers for their insightful comments and suggestions.

8. Bibliographical References

Jon Alkorta and Mikel Iruskieta Quintian. 2022. Adding the Basque parliament corpus to ParlaMint project. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 107–110, Marseille, France. European Language Resources Association.

Hanna Bäck, Marc Debus, and Jochen Müller. 2014. Who Takes the Parliamentary Floor? The Role of Gender in Speech-making in the Swedish "Riksdag". *Political Research Quarterly*, 67(3):504–518.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darģis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2023. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 57(1):415–448.

Tomaž Erjavec and Andrej Pančur. 2019. Parla-CLARIN: TEI guidelines for corpora of parliamentary proceedings.

Tomaž Erjavec and Andrej Pančur. 2022. The Parla-CLARIN recommendations for encoding corpora of parliamentary proceedings. *Journal of the Text Encoding Initiative (Selected Papers from the 2019 TEI Conference)*, (14).

Nayla Escribano, Jon Ander Gonzalez, Julen Orbegozo-Terradillos, Ainara Larrondo-Ureta, Simón Peña-Fernández, Olatz Perez-de Viñaspre, and Rodrigo Agerri. 2022. BasqueParl: A bilingual corpus of Basque parliamentary transcriptions. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3382–3390, Marseille, France. European Language Resources Association.

Henriette Müller and Pamela Pansardi. 2023. Women Leading the Opposition: Gender and Rhetoric in the European Parliament. *Politics and Governance*, 11(1):164–176.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Cristina Bosco, Gosse Bouma, Sam Bowman, Marie Candito, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Fabricio Chalub, Jinho Choi, Çağrı Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Linh Hà M~y, Dag Haug, Barbora Hladká, Petter Hohle, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Natalia Kotsyba, Simon Krek, Veronika Laippala, Phng Lê Hồng, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Lng Nguy~ên Thị, Huyền Nguy~ên Thị Minh, Vitaly Nikolaev, Hanna Nurmi, Stina Ojala, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Livy Real, Laura Rituma, Rudolf Rosa, Shadi Saleh, Manuela Sanguinetti, Baiba Saulīte, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Jonathan North Washington, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. 2017. Universal Dependencies 2.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Parlamento de Galicia. 2020. Regulamento do Parlamento de Galicia.

Marilina Pisani, Rodolfo Zevallos, and Núria Bel. 2023. Catalan Parliamentary Plenary Session Transcriptions from 2015 to 2022. The ParlaMintCAT Corpus. *Procesamiento del Lenguaje Natural*, 71(0):125–136.

Hanna F. Pitkin. 1967. *The Concept of Representation*. University of California Press, Berkeley/Los Angeles.

Klara Raiber and Niels Spierings. 2022. An agnostic approach to gender patterns in parliamentary speech: a question of representation by topic and style. *European Journal of Politics and Gender*, 5(3):361–381.

Marc Sanjaume-Calvet, Joan-Josep Vallbé, and Marina Muñoz-Puig. 2023. Can women take the floor in parliament? Evidence from the Spanish lower chamber. *Women's Studies International Forum*, 97:102694.

Jure Skubic, Jan Angermeier, Alexandra Bruncrona, Bojan Evkoski, and Larissa Leiminger. 2022. Networks of power: Gender analysis in selected european parliaments. In *Proceedings of the 2nd Workshop on Computational Linguistics for Political Text Analysis (CPSS-2022)*, Potsdam, Germany. CPSS.

Xosé Luís Vázquez Somoza. 2015. *A lingua no Parlamento de Galicia*. Universidade de Santiago de Compostela. Facultade de Filoloxía. Departamento de Filoloxía Galega.

9. Language Resource References

Tomaž Erjavec, Matyáš Kopp, Maciej Ogrodniczuk, Petya Osenova, Rodrigo Agerri, Manex Agirrezabal, Tommaso Agnoloni, José Aires, Monica Albini, Jon Alkorta, Iván Antiba-Cartazo, Ekain Arrieta, Mario Barcala, Daniel Bardanca, Starkaður Barkarson, Roberto Bartolini, Roberto Battistoni, Nuria Bel, Maria del Mar Bonet Ramos, María Calzada Pérez, Aida Cardoso, Çağrı Çöltekin, Matthew Coole, Roberts Darģis, Jesse de Does, Ruben de Libano, Griet Depoorter, Katrien Depuydt, Sascha Diwersy, Réka Dodé, Kike Fernandez, Elisa Fernández Rei, Francesca Frontini, Marcos Garcia, Noelia García Díaz, Pedro García Louzao, Maria Gavriilidou, Dimitris Gkoumas, Ilko Grigorov, Vladislava Grigorova, Dorte Haltrup Hansen, Mikel Iruskieta, Johan Jarlbrink, Kinga Jelencsik-Mátyus, Bart Jongejan, Neeme Kahusk, Martin Kirnbauer, Anna Kryvenko, Noémi Ligeti-Nagy, Nikola Ljubešić, Giancarlo Luxardo, Carmen Magariños, Maans Magnusson, Carlo Marchetti, Maarten Marx, Katja Meden, Amália Mendes, Michal Mochtak, Martin Mölder, Simonetta Montemagni, Costanza Navarretta, Bartlomiej Nitoń, Fredrik Mohammadi Norén, Amanda Nwadukwe, Mihael Ojsteršek, Andrej Pančur, Vassilis Papavassiliou, Rui Pereira, María Pérez Lago, Stelios Piperidis, Hannes Pirker, Marilina Pisani, Henk van der Pol, Prokopis Prokopidis, Valeria Quochi, Paul Rayson, Xosé Luís Regueira, Michal Rudolf, Manuela Ruisi, Peter Rupnik, Daniel Schopper, Kiril Simov, Laura Sinikallio, Jure Skubic, Minna Tamper, Lars Magne Tungland, Jouni Tuominen, Ruben van Heusden, Zsófia Varga, Marta Vázquez Abuín, Giulia Venturi, Adrián Vidal Miguéns, Kadri Vider, Ainhoa Vivel Couso, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, Rodolfo Zevallos, and Fišer, Darja (2023). Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 4.0. Slovenian language resource repository CLARIN.SI. Retrieved from http://hdl.handle.net/11356/1860.

Tomaž Erjavec, Matyáš Kopp, Maciej Ogrodniczuk, Petya Osenova, Rodrigo Agerri, Manex Agirrezabal ... and Fišer, Darja (2023). Multilingual comparable corpora of parliamentary debates ParlaMint 4.0. Slovenian language resource repository CLARIN.SI. Retrieved from http://hdl.handle.net/11356/1859.

Bulgarian ParlaMint 4.0 Corpus as a Testset for Part-of-Speech Tagging and Named Entity Recognition

Kiril Simov, Petya Osenova
Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
{kivs, petya}@bultreebank.org

Abstract

The paper discusses some fine-tuned models for the tasks of part-of-speech tagging and named entity recognition. The fine-tuning was performed on the basis of an existing BERT pre-trained model and two newly pre-trained BERT models for Bulgarian that are cross-tested on the domain of the Bulgarian part of the ParlaMint corpora as a new domain. In addition, a comparison has been made between the performance of the new fine-tuned BERT models and the available results from the Stanza-based model which the Bulgarian part of the ParlaMint corpora has been annotated with. The observations show the weaknesses in each model as well as the common challenges.

Keywords: BERT model, Bulgarian, parliamentary sessions, domain cross-validation, POS tagging, NER

1. Introduction

The Bulgarian Parliamentary Corpus is part of the ParlaMint 4.0 multilingual corpus of 29 national and regional parliaments in Europe (Erjavec et al., 2023). The data is publicly available for usage through the CLARIN.SI repository[1]. Our plans for the further development of the Bulgarian ParlaMint Corpus (BParlC) go in two main directions: an extension of the phenomena covered by the annotations of the corpus as well as an extension of BParlC with additional data. Concerning the latter direction, the additional data will include: (1) diachronically available editions of parliament activity to cover the period after the liberation of Bulgaria from Ottoman Empire in 1878; (2) additional documents such as records of debates in the Parliamentary Committees; (3) similar corpora related to Municipality Councils for some economically influential cities in Bulgaria; (4) documents of political parties; (5) linking to the Bulgaria-centric Knowledge Graph (BGKG); and others.

Concerning the former direction, our first goal is to perform linking of the named entities within the texts of BParlC with the Bulgaria-centric Knowledge Graph. However, in order to achieve this task in the best possible way, we decided to check the quality of the annotation already present in BParlC, and to implement some processing tools for Bulgarian which to improve the current annotations — especially the part-of-speech tagging (POS) (UPOS for tagging with Universal Universal POS tags[2] and the XPOS for tagging with BulTreeBank POS tags[3]) as well as the Named Entity Recognition (NER).

Thus, our immediate aim is to test the BERT pre-trained models for POS tagging and NER tasks in a cross-domain setting. We had at our disposal a BERT model, already pre-trained over newsmedia texts – BERT-WEB-BG[4] — see Marinova et al. (2023) – and two new BERT models BERT-NEWS-LIT-BG-1 and BERT-NEWS-LIT-BG-2 which were pre-trained especially for the purposes of these experiments on more and other newsmedia texts as well as on additions of fictional texts (original Bulgarian and Translated into Bulgarian Foreign Literature).

The fine-tuning was performed on the two datasets for POS tagging (the BulTreeBank Dataset and the CLaDA-BG-POS Dataset[5]) as well as on the two datasets for NER (the BulTreeBank dataset and the Bulgarian Balto-Slavic Dataset). We first fine-tuned the models for each task mentioned above, and evaluate them with respect to the test sets within the corresponding dataset. Additionally, we evaluate the fine-tuned models on a new genre of texts — parliament debates.

Our work somewhat relates to the task of Domain Adaptation (DA) in sequential labeling tasks. However, at this stage we did not use the strategy of adding some in-domain data either in the pre-training phase or in the fine-tuned model to check whether it would improve for the target domain. This

[1] https://www.clarin.eu/parlamint#parlamint-corpora

[2] https://universaldependencies.org/u/pos/all.html

[3] http://bultreebank.org/wp-content/uploads/2017/04/BTB-TR03.pdf

[4] https://huggingface.co/usmiva/bert-web-bg

[5] This dataset is under development within the Bulgarian Infrastructure project CLaDA-BG: https://clada-bg.eu/en/

setting remains for future work. In addition, a semi-automatic comparison through human checks has been made between the performance of our BERT model and the Stanza-based one.

The motivation behind such a task includes the following aspects: i) improving the quality and the coverage of the BERT models for the two above-mentioned tasks, and (ii) evaluating the applicability of the models with respect to a different domain.

At first sight it might seem that part-of-speech tagging and named entity recognition are already solved tasks to a great extent. And this is true already for many languages and domains since the SOTA results are beyond 90 % F-measure (even beyond 95 %). However, it would be useful to track the systemic and occasional errors in the remaining percentages of unrecognized or wrongly annotated tokens in the data.

The paper is structured as follows: in the next section a brief overview is given of related work. Section 3 outlines the experimental setting. Section 4 discusses the results and provides quality comparison between the two models. Section 5 presents the conclusions.

2. Related Work

For the POS tagging there are a number of works that evaluate taggers' F-measure out-of-domain. For example, Schnabel and Schütze (2013) show that there is no single representation and method that works equally well for all target domains. In the target domains that were considered in the paper there is politics or parliamentary data.

Then, Hansen and van der Goot (2023) evaluate the performance of two taggers for English in a domain different from the Wall Street Journal section of the Penn Treebank, namely – on video games related dataset. Authors conclude that the accuracy on unknown tokens decreases and that the main problems are with the proper nouns and inconsistent capitalization.

In (Kübler and Baucom, 2011) a fast method for adding in-domain training data has been proposed that uses three taggers trained on the source data and run on the target unannotated data. The source domain is the Wall Street Journal part of the Penn Treebank, and the target one consists of dialogues in a collaborative task. The authors add sentences to the training data only when the majority of the taggers agree on the POS tags.

For NER also there is a lot of work devoted to its handling in a cross- and/or out-of-domain setting. For example, Liu et al. (2020) introduce a cross-NER dataset that comprises five domains among which politics. The authors provide specific NE for each domain. For the politics they are: politician, person, organization, political party, event, election, country, location, miscellaneous. The authors find that this domain overlaps mostly with the Reuters domain, i.e. newsmedia (35.7%). With BERT on English they report an integrated F1 on token level of 68.83.

Later on Zheng et al. (2022) use a graph matching method that learns graph structure via matching label graphs from source to target domain, and improve these results on all domains among which the domain of politics. In contrast to Liu et al. (2020) we do not use parliament data in the training phase but only in the pre-training one, thus making the task slightly harder. On the other hand, we facilitate our work by applying the standard set of categories for NER used in the corresponding datasets: Person, Organization, Location, and Other for Bultreebank dataset and Person, Organization, Location, Event, and Product for Balto-Slavic dataset.

3. The Experimental Setting

In this section we present the characteristics of the models as well as the specifics of the datasets that were used in the experiments.

3.1. BERT Pre-trained Models

In the experiments we exploited one of the existing models released for free public usage — BERT-WEB-BG. This model has been pre-trained on 30 GB of text which we estimated to comprise 3 536 668 132 tokens. The domain was newsmedia data. However, in order to check the impact of the text types used in the pre-training, we decided to pre-train a new model with a size of 2 192 734 242 tokens, from which about 800 000 000 tokens are fiction and the rest are newsmedia data. The other parameters have not been changed, including the number of epochs. In Table 1 the characteristics of the pre-training datasets are given.

3.2. BERT Fine-tuning Datasets

BulTreeBank Datasets In our experiments the following datasets were used:

1. *BulTreeBank-UD (BTB-UD)*. The BulTreeBank in its Universal Dependency format comprises data of 156K in tokens. It contains annotation for POS tagging divided into two: UPOS - annotation with Universal parts-of-speech and XPOS - annotation with the original BulTreeBank tags. The dataset follows the division into training and test sets in the Universal Dependencies package[6].

2. *BulTreeBank-NER (BTB-NER)*. The BTB-NER used the original BulTreeBank resource that

[6] https://universaldependencies.org/

Dataset	Size in GB	Size in tokens	Loss	Accuracy
BERT-WEB-BG	30.0	3 536 668 132	1.451	0.6906
BERT-NEWS-LIT-BG-1	18.6	2 192 734 242	2.153	0.5593
BERT-NEWS-LIT-BG-2	18.6	2 192 734 242	1.414	0.6913

Table 1: Characteristics of the pre-trained models. The dataset for the training of BERT-WEB-BG is proprietary and for that reason the exact size in tokens is not available to us. Thus, we estimated it on the basis of the other datasets. The main differences between BERT-NEWS-LIT-BG-1 and BERT-NEWS-LIT-BG-2 are the following hyper parameters: the hidden size of the first model is the default 768, but it is 1024 in the second model; the number of the attention heads is 12 for the first model, and 16 for the second one; the intermediate size of the first model is 3072, and 4096 for the second one respectively. The sizes of the parameters in the models are as follows: 109 113 649 parameters for the first one, and 183 485 745 parameters for the second.

is constituency-based and dependency-aware. Thus, it used the data of 256K in tokens. It includes four kinds of Named Entities: PER (persons), LOC (locations), ORG (organization), and OTH (other names). The treebank consists of 40 sets of sentences. Some of these sets are just small segments of texts extracted from different sources like Bulgarian grammar books, random paragraphs from corpora. Other sets are whole articles or other genres like newspaper articles, chapters of books, Bulgarian constitution, etc. The division in training, development, and test sets was performed on the basis of the whole sets of the treebank in the proportion of 80 % training set, 10 % development set and 10 % test set.

The additional two datasets used for fine-tuning are the following:

3. *Bulgarian Balto-Slavic Dataset (BS-NER)*. The datasets are from years 2019 (Piskorski et al., 2019) and 2021 (Piskorski et al., 2021). This integrated dataset contains annotations of Named Entities in the following categories: PER (person), ORG (organization), LOC (location), EVN (event) and PRO (product). The dataset follows the division of training and test sets as described in (Hardalov et al., 2023).

4. *CLaDA-POS Dataset*. This is a newly annotated dataset created within CLaDA-BG research infrastructure. The texts are collected from different sources. They include all the definitions from BTB-WN: the BTB Bulgarian Wordnet — see (Simov and Osenova, 2023), all the examples related to meanings from BTB-WN, newsmedia documents, first paragraphs of about 1000 articles from the Bulgarian Wikipedia. The dataset is divided into training, development and test sets by us for the purposes of these experiments.

It can be seen that no parliamentary data was used in the fine-tuning step due to the lack of sufficient gold data. At the same time, only some of the characteristics that can be found in the parliament corpora, are already present albeit in small portions in the above-mentioned datasets. This means that there are politically oriented topics and named entities that refer to politicians, especially in the newsmedia data.

4. Results

In order to evaluate different models for POS tagging and NER over ParlaMint data we performed fine-tuning of the two pre-trained models on the above described datasets for these tasks. The results from the first experiments are given in Table 2 where some evaluation was performed within the same datasets.

It can be seen that the best F1 measure metrics for all the tasks was achieved by BERT-NEWS-LIT-BG-2 model. These results reflect the increase in both metrics - Precision and Recall. This means that the more parameters and the more context included, the better the results. Table 2 also shows that concerning Precision and Recall, the two new models outperform the previous one in all tasks with the exception of the results on the BS-NER dataset by BERT-NEWS-LIT-BG-1 model.

For the task of XPOS tagging we could compare our results with the state-of-the-art performance reported in (Georgiev et al., 2012). There the authors report a method based on Guided Learning with results 95.72 % Accuracy; Guided Learning + Lexicon 97.83 % Accuracy; and Guided Learning + Lexicon + Rules 97.98 %. The results are achieved by training on the constituent version of BulTreeBank, because at that time BulTreebank-UD has not existed yet. Thus, we consider our current models comparable to the state-of-the-art. Since the best results with Guided Learning were achieved by the inclusion of an inflectional lexicon, as a further step we plan to encode this lexicon in the pre-trained models.

With respect to the NER Task, our results are comparable with the results given in (Marinova

Pre-trained Models	Task	Classes	Dataset	Precision	Recall	F1
BERT-WEB-BG	NER	11	BS-NER	0.986718	0.991105	0.988907
BERT-WEB-BG	NER	11	BTB-NER	0.810180	0.813631	0.811902
BERT-WEB-BG	UPOS	16	BTB-UD	0.987725	0.987725	0.987725
BERT-WEB-BG	XPOS	546	BTB-UD	0.943907	0.943907	0.943907
BERT-WEB-BG	XPOS	674	CLaDA-POS	0.948318	0.948318	0.948318
BERT-NEWS-LIT-BG-1	NER	11	BS-NER	0.983014	0.988522	0.985760
BERT-NEWS-LIT-BG-1	NER	11	BTB-NER	0.837433	0.833865	0.835645
BERT-NEWS-LIT-BG-1	UPOS	16	BTB-UD	0.991668	0.991668	0.991668
BERT-NEWS-LIT-BG-1	XPOS	546	BTB-UD	0.953256	0.953256	0.953256
BERT-NEWS-LIT-BG-1	XPOS	674	CLaDA-POS	0.952327	0.952327	0.952327
BERT-NEWS-LIT-BG-2	NER	11	BS-NER	0.993962	0.996836	**0.995397**
BERT-NEWS-LIT-BG-2	NER	11	BTB-NER	0.869374	0.843450	**0.856216**
BERT-NEWS-LIT-BG-2	UPOS	16	BTB-UD	0.992877	0.992877	**0.992877**
BERT-NEWS-LIT-BG-2	XPOS	546	BTB-UD	0.977995	0.977995	**0.977995**
BERT-NEWS-LIT-BG-2	XPOS	674	CLaDA-POS	0.954940	0.954940	**0.954940**

Table 2: Fine-tuning tasks performance for the two pre-trained models. The number of epochs for the NER tasks is 7 and for the POS tasks – 10.

Task	BothTrue	CLTrueBTBFalse	CLFalseBTBTrue	BothFalse	CLASSLA	BERT-NEWS-LIT-BG-1
	#	#	#	#	Accuracy	Accuracy
NER	1079	55	38	60	**92,04 %**	90.66 %
UPOS	1162	21	28	21	96.02 %	**96.59 %**
XPOS	959	8	16	57	92.98 %	**94.51 %**

Table 3: Evaluation over the ParlaMint data of the BERT-NEWS-LIT-BG-1-based models.

et al., 2020) and (Marinova et al., 2023). This similarity is obvious since we also rely on their BERT pre-trained model. The main difference between the two models is that the division of the BS-NER data into training and test subsets is not the same. We were surprised to see that the model performance on the BTB-NER dataset was quite poor. The analysis shows that this is due to selection mainly literature tests for the test set. The category OTHER is problematic, because it practically covers a very diverse set of named entities like names of books, movies, and similar names that sometimes are long phrases or even full sentences. Later on, it was discovered that this type of names were also the largest problem with respect to the NER performance within the Bulgarian part of the ParlaMint corpora.

Evaluation over the ParlaMint data. At the moment we do not have a gold standard dataset for POS tagging and NER over the Bulgarian ParlaMint corpus that is significant in size. Thus, direct measurements of the performance of the train models are not possible. However, in order to perform some initial evaluation, the debates from three days (27/28/29.07.2022) were selected and then annotated automatically with the best one from the above fine-tuned models, based on BERT-WEB-BG and BERT-NEWS-LIT-BG-1 pre-trained models[7]. Then the annotations were manually checked for about 1000 occurrences of NEs and POS tags. More precisely — 1232 for named entities and for UPOS task, and 1040 for the XPOS task. Since the ParlaMint corpora of South Slavic languages were already annotated by Nikola Ljubešić with the CLASSLA models, we were able to compare the results from the models.

The evaluation was performed by our best annotator and the process was executed by the usage of the following categories:

- *BothTrue*: this label means that both CLASSLA and our model took the same decision.
- *CLTrueBTBFalse*: this label means that the decision of CLASSLA was correct and the decision of our model was wrong.
- *CLFalseBTBTrue*: this label means that the decision of CLASSLA was wrong and the decision of our model was correct.
- *BothFalse*: this label means that the decision of both – CLASSLA and our model – was wrong.

The following annotations were considered: UPOS Tagging, XPOS Tagging and NER. For the UPOS Tagging and XPOS Tagging tasks we used the fine-tuned model on BTB-UD dataset. The NER task used the fine-tuned model on BS-NER dataset. The results are given in Table 3 for the models that were fine-tuned on the BERT-NEWS-LIT-BG-1 pre-trained model, and in Table 4 for the models that

[7] We did not have the same evaluation based on the BERT-NEWS-LIT-BG-2 model.

Task	BothTrue	CLTrueBTBFalse	CLFalseBTBTrue	BothFalse	CLASSLA	BERT-WEB-BG
	#	#	#	#	Accuracy	Accuracy
NER	1055	79	33	65	**92,04 %**	89.12 %
UPOS	1168	15	30	19	96.02 %	**97,24 %**
XPOS	959	8	16	57	92.98 %	**93.75** %

Table 4: Evaluation over the ParlaMint data of the BERT-WEB-BG-based models.

were fine-tuned on the BERT-NEWS-LIT-BG-1 pretrained model.

It can be seen that the biggest drop of the performance is on the NER tasks – with about 8-9 %. The manual check shows that the most problematic cases are the names of documents/regulations/laws that have been discussed during the debates in the Parliament. Here is an example of such a name: “Zakon za ratifitsirane na Memoranduma za razbiratelstvo otnosno podkrepa za proekti na Evropeiskiya sayuz mezhdu pravitelstvoto na Republika Bulgaria i Evropejskata investitsionna banka.” (Law on the ratification of the Memorandum of Understanding regarding support for the European Union projects between the Government of the Republic of Bulgaria and the European Investment Bank.) Another type of problematic names are the names of some parties. For example, compare the name “Ima takav narod” (There is Such a People) which constitutes a complete sentence. The same holds for the party “Prodalzhavame promyanata” (We continue with the changes).

These above examples illustrate two issues: i) some names are very long and thus, the usual encoding in a BIO format is not appropriate for them, and ii) there is a big density of novel complex names where the recursive chain is very deep since one name can contain a number of other names. In our view, such newly generated and long names also require new approaches and strategies for the domain adaptation of the existing NER models.

As for the POS annotation, the BERT-NEWS-LIT-BG-1 model performs better on the UPOS tags than on the in-house XPOS ones. However, it must be noted that the UD tags are 16, while the number of XPOS classes are much higher.

To conclude the section, not surprisingly, the out-of-domain data are predominantly sensitive to the named entities and not so much to the POS tags. Despite this it can be seen that POS tagging also drops. This means that there are morphosyntactic specifics in the parliamentary domain that have to be addressed.

5. Conclusions

At this stage of our work no in-domain data was used either in pre-training or fine-tuning phases. The reason for not including data in pre-train stages was that we considered the available one not large enough.

The reason for not including data during the fine-tuning process is the fact that the semi-automatic morphosyntactic disambiguation and the NER checking on the parliamentary sessions has not been finished yet. Only some manual inspection was made on POS tags and NER labels from our fine-tuned BERT models and the CLASSLA Stanza-based model. However, these cross-checks were sufficient to give some main orientation to us about the sources of the drops in the respective results.

For future work we plan to specialize the NER labels towards the parliamentary data. Our idea is to explore the following places for getting domain information on named entities: i) within the specific structure of the sessions such as the interaction formulas; ii) through the referring to various legislative acts and iii) through the discussion over various topics where topic modeling experiments might be applied in advance.

6. Acknowledgements

The reported work has been partially supported by CLaDA-BG, *the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH*, funded by the Ministry of Education and Science of Bulgaria (support for the Bulgarian National Roadmap for Research Infrastructure). We would like to thank Nikolay Paev, Veneta Kireva and Silvia Petrova for their help with training and testing the models. We would like to thank also the reviewers for their valuable and insightful comments.

7. Bibliographical References

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešic, Kiril Simov, Andrej Pancur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çagrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul

Rayson, Vaidas Morkevicius, Tomas Krilavicius, Roberts Dargis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fiser. 2023. The parlamint corpora of parliamentary proceedings. *Lang Resources and Evaluation*, 57:415–448.

Georgi Georgiev, Valentin Zhikov, Kiril Simov, Petya Osenova, and Preslav Nakov. 2012. Feature-rich part-of-speech tagging for morphologically complex languages: Application to Bulgarian. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 492–502, Avignon, France. Association for Computational Linguistics.

Kia Kirstein Hansen and Rob van der Goot. 2023. Cross-domain evaluation of pos taggers: From wall street journal to fandom wiki.

Momchil Hardalov, Pepa Atanasova, Todor Mihaylov, Galia Angelova, Kiril Simov, Petya Osenova, Veselin Stoyanov, Ivan Koychev, Preslav Nakov, and Dragomir Radev. 2023. bgGLUE: A Bulgarian general language understanding evaluation benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8733–8759, Toronto, Canada. Association for Computational Linguistics.

Sandra Kübler and Eric Baucom. 2011. Fast domain adaptation for part of speech tagging for dialogues. In *Proceedings of the International Conference Recent Advances in Natural Language 2011*, pages 41–48, Hissar, Bulgaria.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2020. Crossner: Evaluating cross-domain named entity recognition.

Nikola Ljubešić and Kaja Dobrovoljc. 2019. What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 29–34, Florence, Italy. Association for Computational Linguistics.

Iva Marinova, Laska Laskova, Petya Osenova, Kiril Simov, and Alexander Popov. 2020. Reconstructing NER corpora: a case study on Bulgarian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4647–4652. European Language Resources Association.

Iva Marinova, Kiril Simov, and Petya Osenova. 2023. Transformer-based language models for Bulgarian. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 712–720, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Petya Osenova, Kiril Simov, Iva Marinova, and Melania Berbatova. 2022. The Bulgarian event corpus: Overview and initial NER experiments. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3491–3499, Marseille, France. European Language Resources Association.

Jakub Piskorski, Bogdan Babych, Zara Kancheva, Olga Kanishcheva, Maria Lebedeva, Michał Marcińczuk, Preslav Nakov, Petya Osenova, Lidia Pivovarova, Senja Pollak, Pavel Přibáň, Ivaylo Radev, Marko Robnik-Sikonja, Vasyl Starko, Josef Steinberger, and Roman Yangarber. 2021. Slav-NER: the 3rd cross-lingual challenge on recognition, normalization, classification, and linking of named entities across Slavic languages. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 122–133, Kiyv, Ukraine. Association for Computational Linguistics.

Jakub Piskorski, Laska Laskova, Michał Marcińczuk, Lidia Pivovarova, Pavel Přibáň, Josef Steinberger, and Roman Yangarber. 2019. The second cross-lingual challenge on recognition, normalization, classification, and linking of named entities across Slavic languages. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 63–74, Florence, Italy. Association for Computational Linguistics.

Marko Prelevikj and Slavko Zitnik. 2021. Multilingual named entity recognition and matching using BERT and dedupe for Slavic languages. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 80–85, Kiyv, Ukraine. Association for Computational Linguistics.

Tobias Schnabel and Hinrich Schütze. 2013. Towards robust cross-domain domain adaptation for part-of-speech tagging. In *International Joint Conference on Natural Language Processing*.

Kiril Simov and Petya Osenova. 2023. Recent developments in BTB-WordNet. In *Proceedings of the 12th Global Wordnet Conference*, pages 220–227, University of the Basque Country, Donostia - San Sebastian, Basque Country. Global Wordnet Association.

Junhao Zheng, Haibin Chen, and Qianli Ma. 2022. Cross-domain named entity recognition via graph matching. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2670–2680, Dublin, Ireland. Association for Computational Linguistics.

Resources and Methods for Analysing Political Rhetoric and Framing in Parliamentary Debates

Ines Rehbein
University of Mannheim
Data and Web Science Group
ines.rehbein@uni-mannheim.de

Recent work in political science has made extensive use of NLP methods to produce evidential support for a variety of analyses, for example, inferring an actor's ideological positions from textual data or identifying the polarisation of the political discourse over the last decades. Most work has employed variations of lexical features extracted from text or has learned latent representations in a mostly unsupervised manner. While such approaches have the potential to enable political analyses at scale, they are often limited by their lack of interpretability. In the talk, I will instead look at semantic and pragmatic representations of political rhethoric and ideological framing and present several case studies that showcase how linguistic annotation and the use of NLP methods can help to investigate different framing strategies in parliamentary debates.

The first part of the talk investigates populist framing strategies, specifically, the use of pronouns to create in- and out-groups and the identification of people-centric messages. The second part of the presentation focusses on framing strategies on the pragmatic level.

Modelling populist rhetoric in text. A rhetoric strategy often used in political debates is *Othering*, a technique that aims at describing a person or minority group as distant and different from what is considered as "the norm", i.e., the speaker's own in-group. To better understand how political actors use Othering, we developed a compositional annotation scheme to capture the clusivity properties of personal pronouns in context, that is their ability to construct and manage in-groups and out-groups (Rehbein and Ruppenhofer, 2022). Our exploratory analysis of pronoun use in the parliamentary setting provides some face validity for our schema, that I will discuss in the talk.

Another prominent feature of populist discourse is the use of people-centric messages, also referred to as *thin* populism (Jagers and Walgrave, 2007). To automatically identify *thin* populism in text, we combine insights from political science (Mudde, 2017; Wirth et al., 2019) with quantitative text analysis and NLP methodologies (Klamm et al., 2023). In a first step, we identify the core protagonistis of populist rhetoric, i.e., mentions of *The People* (such as: Germans, tax payers, Muslims, etc.) and of *The Elite* (e.g., the government, media, politicians, etc.). Aggregating the extracted information, we are able to measure the use of *thin* populism for different parties in parliament and show that our measure correlates with experts' ratings from the Populism and Political Parties Expert Survey 2018 (POPPA) (Meijers and Zaslove, 2021).

Pragmatic framing in political debates. On the pragmatic level, the analysis of speech acts can provide rich information on how political actors frame their messages. Kondratenko et al. (2020) present a linguo-pragmatic taxonomy for speech acts in political discourse. On the highest level, their taxonomy distinguishes cooperation from conflict communication which, on the next level, are further divided into six subclasses. Extending their work, we develop a fine-grained speech act annotation scheme for German parliamentary debates and automatically predict speech acts in a corpus of Bundestag debates, ranging from 2003 to 2023. Our initial analysis confirms our expectations regarding the different rhetorical strategies used by political actors in government and in opposition (Reinig et al., 2024).

Another rhetorical strategy related to epistemological bias (Recasens et al., 2013) is to frame a proposition as a fact or part of the common ground rather than presenting it as personal opinion. Our case study shows how we can identify epistemological bias, based on the identification of events of speech, thought and writing in debates, together with their corresponding roles (e.g., speaker, addressee, message), and combining this information whith clustering techniques (Rehbein et al., 2024).

Finally, I will discuss ongoing work on the annotation of moral frames in political communication and highlight the challenges and potentials of this type of analysis.

1. Bibliographical References

Jan Jagers and Stefaan Walgrave. 2007. Populism as Political Communication Style: An Empirical Study of Political Parties' Discourse in Belgium. *European Journal of Political Research*, 46(3):319–345.

Christopher Klamm, Ines Rehbein, and Simone Paolo Ponzetto. 2023. Our kind of people? Detecting populist references in political debates. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1227–1243, Dubrovnik, Croatia. Association for Computational Linguistics.

Natalia V. Kondratenko, Anastasiia A. Kiselova, and Liubov V. Zavalska. 2020. Strategies and tactics of communication in parliamentary discourse. *studies about languages*, 36:17–29.

Maurits J. Meijers and Andrej Zaslove. 2021. Measuring populism in political parties: Appraisal of a new approach. *Comparative Political Studies*, 54(2):372–407.

Cas Mudde. 2017. Populism: An Ideational Approach. In C. Rovira Kaltwasser, P. Taggart, and et al. Ochoa Espejo, P., editors, *The Oxford Handbook of Populism*, pages 27–47. Oxford: Oxford University Press.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, Sofia, Bulgaria. Association for Computational Linguistics.

Ines Rehbein and Josef Ruppenhofer. 2022. Who's in, Who's out? Predicting the Inclusiveness or Exclusiveness of Personal Pronouns in Parliamentary Debates. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5849–5858, Marseille, France. European Language Resources Association.

Ines Rehbein, Josef Ruppenhofer, Annelen Brunner, and Simone Paolo Ponzetto. 2024. Out of the mouths of MPs: Speaker Attribution in Parliamentary Debates. In *Proceedings of the Fourteenth Language Resources and Evaluation Conference*, Torino, Itala. European Language Resources Association.

Ines Reinig, Ines Rehbein, and Simone Paolo Ponzetto. 2024. How to do politics with words: Investigating speech acts in parliamentary debates. In *Proceedings of the Fourteenth Language Resources and Evaluation Conference*, Torino, Itala. European Language Resources Association.

Werner Wirth, Martin Wettstein, Dominique Wirz, Nicole Ernst, Florin Büchel, Anne Schulz, Frank Esser, and et al. 2019. *Codebook: NCCR democracy Module II: The Appeal of populist Ideas and Messages*. Unpublished paper.

PTPARL-V: Portuguese Parliamentary Debates for Voting Behaviour Study

Afonso Sousa, Henrique Lopes Cardoso
Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC)
Faculdade de Engenharia da Universidade do Porto, Portugal
{ammlss,hlc}@fe.up.pt

Abstract

We present a new dataset, PTPARL-V, that is a valuable resource for advancing discourse analysis of parliamentary debates in Portuguese and their alignment with voting behaviour. This is achieved by processing the open-access information available at the official Portuguese Parliament website and scraping the debate minutes concerning legislative initiatives, together with meta-data related to voting positions. Our dataset includes interventions from 547 different deputies of all major Portuguese parties, from 736 legislative initiatives spanning five legislatures from 2005 to 2021. We present a statistical analysis of the dataset compared to other publicly available Portuguese parliamentary debate corpora. Finally, we provide baseline performance analysis for voting behaviour classification.

Keywords: Portuguese debates, Discourse analysis, Parliamentary data, Voting behaviour

1. Introduction

Parliamentary corpora are essential language resources that can be approached from various research perspectives, including political science, sociology, history, and psychology. Parliamentary and legislative debate transcripts provide access to information concerning elected politicians' opinions, positions, and policy preferences. This kind of information can be used for a variety of computational tasks and natural language applications, such as critical discourse analysis (Van Dijk, 1993), sentiment analysis (Abercrombie and Batista-Navarro, 2020), argument detection (Cabrio and Villata, 2018) or stance detection (Schiller et al., 2021).

The digitization of parliamentary documents and the advancement of computer tools have created interesting opportunities for political data analysis. In recent years, efforts have been made to compile well-structured corpora of parliamentary debates. At the time of writing, the CLARIN infrastructure offers access to 35 parliamentary corpora[1], covering most languages spoken in European countries.

However, almost all corpora are solely comprised of text passages and tags extracted from postprocessing said text (e.g., POS tagging or NER). Moreover, large-scale research on voting discipline and behaviour does not compare discourses, instead solely focusing on the scattering of votes through the various parties (Kam, 2009).

There are many compilations of parliamentary debates. To our knowledge, specifically for Portuguese there is PTPARL (Généreux et al., 2012), a compendium of Portuguese parliamentary debates from 1970 to 2008; PTPARL-D (Almeida et al., 2021), a compilation of all debates of the Third Portuguese Republic, spanning 44 years; and another speech compilation (Fernandes et al., 2021) comprised of speeches from 1999 to 2017. None of these includes annotations for any NLP task.

We introduce PTPARL-V, a new Portuguese dataset that addresses the voting behaviour of members of the Portuguese Parliament. We compiled interventions across five legislatures and extracted associated metadata. We gathered information on the initiatives voted in favour, against, or abstained by all major parties in the Portuguese Parliament. We expect this work to help produce more thorough studies regarding voting behaviour from the different parties and their members.

In summary, the contributions of this paper are: (i) a new Portuguese dataset for voting behaviour analysis of political debates; (ii) a statistical analysis of the newly created dataset; and (iii) a preliminary baseline performance benchmark for forecasting voting behaviour.[2]

2. About Legislative Initiatives

The Portuguese Parliament (*Assembleia da República*) provides open-access data on parliamentary activities on its official website[3]. We next describe the source of information and how the interventions are selected.

Many different activities are conducted in parliament. We focus on *legislative initiatives*: proposals for new laws. These initiatives can be proposed

[1] https://www.clarin.eu/resource-families/parliamentary-corpora

[2] The dataset and code were made available at https://github.com/afonso-sousa/pt_parliamentary_minutes

[3] https://www.parlamento.pt/

by members of the parliament (MPs), parliamentary groups or groups of voting citizens – draft law (*projeto de lei*) – or by the Government or the Regional Legislative Assemblies (RLA) – proposed law (*proposta de lei*). After being admitted by the President of the Assembly, the initiative is subjected to an assessment by the specialised Commission to which it has been assigned, followed by its general debate in a plenary meeting, which ends with a voting process. Further steps may be taken for an initiative to be considered law. For voting behaviour and discourse analysis, we built our dataset by collecting the plenary debate and the general voting information. We discarded joint initiatives (multiple initiatives discussed in the same plenary meeting) because the respective transcripts are cluttered with different subjects and themes, making their automatic parsing and clear distinction of initiatives unfeasible. These initiatives are published in *Diário da República* – the official Portuguese journal where laws, decisions by the Constitutional Court and other relevant texts are published.

The represented parties in *Assembleia da República* that intervened in the plenary meetings to discuss the initiatives mentioned above are briefly summarised in Table 1.

3. Dataset Compilation

We next describe what attributes were selected and how the PTPARL-V dataset was built.

While the open-access data is available in common formats, like XML or JSON, processing the free-text concerning MP speeches is not trivial, as these are contained within PDF files embedded in the website:

- We first downloaded all the published transcripts matching our time span: legislatures X to XIV, spanning from 2005 to 2021.
- Then, from the open-access data, we collected initiatives that matched our previously settled requirements: legislative initiatives (avoiding joint ones) with plenary debate and a general vote. From these, we collected relevant attributes to characterize an entry in the dataset.
- We extracted the text from the transcripts related to the collected initiatives. Each initiative has annotations of the pages within *Diário da República* with the discussion of the initiative. We used text extraction tools to retrieve the text from the designated PDF pages.
- From the retrieved pages, we matched the deputy's name in the metadata with the speaker's name at the beginning of the paragraph (see bold in Figure 1) and concatenated the collected paragraphs. This step produces a multi-sentence text passage comprised of all paragraphs of a deputy's speeches in the discussion of the initiative.

The above-mentioned steps produce text about each MP's stance towards a given initiative – an *intervention*. This information, along with the corresponding metadata, makes up an entry in the dataset. The metadata serves to characterise the intervention and covers three main concepts: the intervention, the initiative for which the intervention was made, and the legislature in which the initiative was proposed. The *intervention* is made by an MP, who has a name and a party they belong to. The intervention also has information on the MP's vote on the initiative being discussed: in favour, against or abstention. The *initiative* has information about the proponents, the type of initiative, and a summary description of the topics being discussed. Lastly, the initiative is proposed in a given legislative session within a *legislature*, identified by a Roman numeral and temporally framed.

4. Data Analysis

We analyse some properties of our dataset.

4.1. Basic Statistics

In Table 2, we compare basic statistics between PTPARL (Généreux et al., 2012) and PTPARL-V. After cleaning, PTPARL-V has a total of 736 initiatives and 5833 interventions (see Table 3 for a distribution over legislatures). To the best of our knowledge, PTPARL is the only previously publicly available compilation of interventions in the Portuguese parliament. PTPARL-V is much larger than PTPARL, with the added benefit of having the accompanying metadata (including voting behaviour).

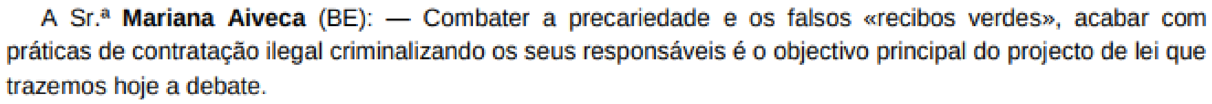
As for general metadata statistics, Table 4 shows some overall information on per-party initiatives and interventions. There are approximately 10 interventions per party per initiative.

4.2. Exploratory Data Analysis

From the metadata alone, we can judge the political scene in Portugal for the dataset time frame. By aggregating similar votes for each initiative, Figure 2 shows the eight sets of parties with the highest similar vote frequency. This means that if an initiative was voted in favour by, say, both PSD and CDS-PP, it would count +1 towards the 'in favour' bar of the "PDS,CDS-PP' set. From the plot, we see that parties often vote in favour of the proposed initiatives, as given by the overall higher frequencies in the respective bars. Additionally, we can see that parties closer in the political spectrum (namely PSD and CDS-PP, or BE, PCP and PEV, see Table 1)

Party Initials	Full Name	Main Ideology	Position
PCP	Portuguese Communist Party, *Partido Comunista Português*	Marxism-Leninism	Left-wing to far-left
BE	Left Bloc, *Bloco de Esquerda*	Democratic socialism	Left-wing to far-left
PEV	Ecologist Party "The Greens", *Partido Ecologista "Os Verdes"*	Eco-socialism	Left-wing
PS	Socialist Party, *Partido Socialista*	Social Democracy	Centre-left
PAN	People Animals Nature, *Pessoas-Animais-Natureza*	Environmentalism	Centre-left
PSD	Social Democratic Party, *Partido Social Democrata*	Liberal conservatism	Centre-right
CDS-PP	Democratic and Social Centre - People's Party, *Centro Democrático e Social – Partido Popular*	Conservatism	Centre-right to right-wing
IL	Liberal Initiative, *Iniciativa Liberal*	Classical liberalism	Centre-right to right-wing
CH	ENOUGH, *CHEGA*	Right-wing populism	Right-wing to far-right

Table 1: General information on the Portuguese parties that have/had representation in *Assembleia da República* (retrieved from Wikipedia).



A Sr.ª **Mariana Aiveca** (BE): — Combater a precariedade e os falsos «recibos verdes», acabar com práticas de contratação ilegal criminalizando os seus responsáveis é o objectivo principal do projecto de lei que trazemos hoje a debate.

Figure 1: Sample paragraph from an intervention in *Diário da República*.

Dataset	# tokens	# sentences
PTPARL	975 806	48 911
PTPARL-V	3 790 086	111 614

Table 2: Basic dataset statistics for PTPARL-V and PTPARL (retrieved from PORTULAN Clarin).

legislature	# initiatives	# interventions
X	211	1609
XI	46	416
XII	267	2061
XIII	152	1239
XIV	60	508
Total	736	5833

Table 3: Distribution of initiatives and interventions per legislature in the PTPARL-V dataset.

often vote together. While a centre-left party, PS is often seen voting alone, explained by the fact that PS is often the governing party, sometimes with an absolute majority.

In Figure 3, we see a distribution of initiatives and interventions per year. Legislatures X and XII were the ones where more initiatives were proposed, spanning from 2006 to 2009 and from 2011 to 2015, respectively.

Party	# initiatives	# votes (favour/against/abst)
Government	443	–
RLA Madeira	29	–
RLA Açores	13	–
PCP	38	512/357/162
BE	59	529/297/146
PEV	21	135/90/36
PS	38	726/367/101
PAN	4	45/7/13
PSD	32	642/336/247
CDS-PP	31	473/293/237
IL	0	12/9/4
CH	0	17/12/8
Mixed	27	–
Citizens	1	–

Table 4: Initiatives and intervention votes in the PTPARL-V dataset. 'Mixed' refers to initiatives authored by deputies of different parties.

5. Predicting Voting Behaviour

In this section, we model the task of predicting voting behaviour as a supervised multiclass classification problem. We try to predict if a given speaker will be voting 'in favour', 'against', or 'abstaining' based on the contents of their interventions.

We randomly split the dataset into train and test

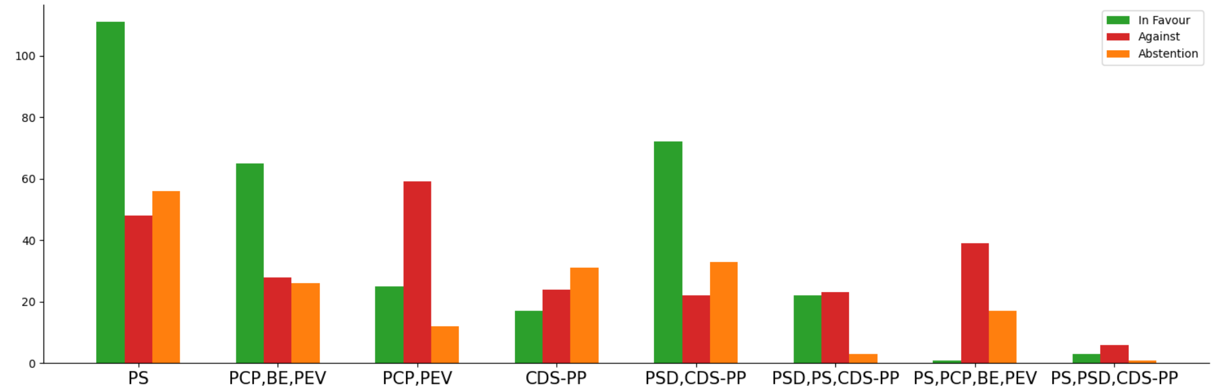


Figure 2: *Who votes with whom?* These bar charts show the parliamentary sets of parties that most frequently voted together. This data was compiled using the frequency of initiative votes for every combination of parties in the dataset.

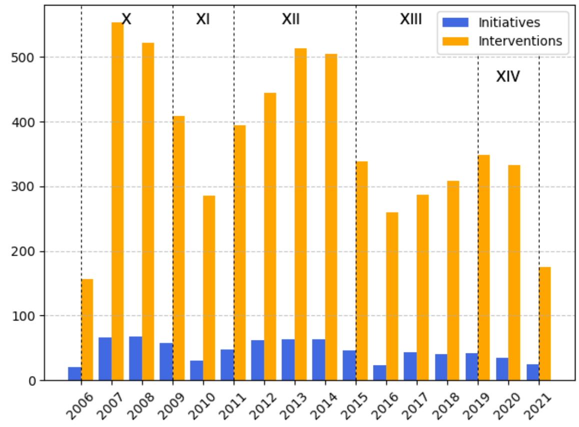


Figure 3: Initiatives and interventions per year.

sets in a stratified fashion (i.e., we keep the original distribution of labels in each set). The splits contain roughly 80% and 20% of the total entries, respectively. We trained a Naive Bayes classifier with TF-IDF features and a Logistic Regression classifier with word embedding features. We also fine-tuned a pretrained multilingual DistilBERT (Sanh et al., 2019)[4] model, the Portuguese encoder BERTimbau (Souza et al., 2020)[5] base model, and two versions of ALBERTINA (900M and 1.5B parameters versions[6], the latter being fine-tuned with LoRA (Hu et al., 2022)).

For feature-based models (Naive Bayes and Logistic Regression), we preprocessed the data:

- We removed all special characters.
- We converted all the text to lowercase.
- We removed generic stopwords (e.g., determiners, conjunctions and prepositions) and domain-specific stopwords (e.g., 'Sr.' (Mr.), 'secretário' (secretary), etc.). These domain-specific words are very prevalent in this compilation of parliamentary debates since there exists a standardized introductory etiquette for addressing the assembly.

The word embeddings used for the Logistic Regression were FastText CBOW with 50 dimensions[7] and were averaged to produce document-level embeddings. We relied on scikit-learn's[8] implementations of the feature-based models.

Model	Accuracy	Precision	Recall	F1
Naive Bayes	0.5904	0.4004	0.4267	0.3934
Logistic Regression	0.5687	0.5012	0.4189	0.3908
DistilBERT	0.5915	0.3830	0.4480	0.4123
BERTimbau	**0.6075**	**0.5240**	**0.4739**	**0.4711**
ALBERTINA-900M	0.5409	0.4416	0.3962	0.3790
ALBERTINA-1.5B-LoRA	0.4989	0.2363	0.3333	0.2639

Table 5: Multiclass classification performance on PTPARL-V.

Looking at Table 5, we find that all models can produce results better than a majority baseline for our dataset, that is, given the three-class split of our dataset is around 53-30-17, respectively for in-favour, against and abstention labels (see Figure 4 for details), the performance of our models is superior to just predicting the majority class, which would give an accuracy of around 53%. As such, we can assume some knowledge is contained within the text passages that can convey the voting behaviour of the speakers.

Interestingly, larger models did not perform better than the 110M parameter BERTimbau. We address this to the somewhat limited amount of training samples. Other issues with the dataset and models

[4] https://huggingface.co/distilbert/distilbert-base-multilingual-cased

[5] https://huggingface.co/neuralmind/bert-base-portuguese-cased

[6] https://huggingface.co/PORTULAN

[7] http://nilc.icmc.usp.br/embeddings

[8] https://scikit-learn.org/stable/supervised_learning.html

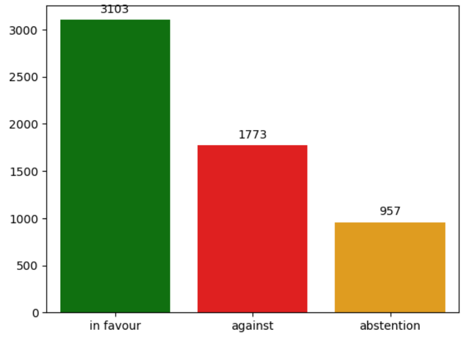


Figure 4: Number of instances per category.

used are as follows. The data contains repetitive passages of formally addressing the President and utterances of disagreement that do not contribute to the argument being made. Additionally, the texts are too long for the 512-token context cap of all the transformer-based models we tested. For reference, this dataset has an average word count of 712 words, meaning nearly half of the text is truncated. We believe a more careful consideration or text preprocessing will alleviate this issue.

6. Discussion and Conclusion

We introduced a new dataset, PTPARL-V, built from interventions of MPs in the Portuguese Parliament for six legislatures. We also briefly show the potential of such a dataset for political debate analysis – with some examples from exploratory data analysis showing the behavioural patterns of voting in the Portuguese Parliament – and vote behaviour forecasting – with a baseline classifier for vote prediction. Future improvements may still be made to the dataset. As for political debate analysis, we just scratched the surface of the insights that can be uncovered from a dataset like this, so we encourage anyone using this dataset to further the research on the Portuguese political scene. Finally, for vote prediction, thoroughly cleaning the text passages can significantly improve the performance of the classifiers. Additionally, using argument mining may be an interesting direction to uncover the most relevant discourse units that best indicate the voting preferences. The dataset and code to reproduce results were made available.

Acknowledgments

Afonso Sousa is supported by a PhD studentship (reference 2022.13409.BD), funded by Fundação para a Ciência e a Tecnologia (FCT). This research was supported by Base Funding (UIDB/00027/2020) and Programmatic Funding (UIDP/00027/2020) of the Artificial Intelligence and Computer Science Laboratory (LIACC) funded by national funds through FCT/MCTES (PIDDAC).

Bibliographical References

Gavin Abercrombie and Riza Batista-Navarro. 2020. Sentiment and position-taking analysis of parliamentary debates: a systematic literature review. *Journal of Computational Social Science*, 3(1):245–270.

Paulo Almeida, Manuel Marques-Pita, and Joana Gonçalves-Sá. 2021. Ptparl-d: an annotated corpus of forty-four years of portuguese parliamentary debates. *Corpora*, 16(3):337–348.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *IJCAI*, volume 18, pages 5427–5433.

Jorge M Fernandes, Mariana Lopes da Fonseca, and Miguel Won. 2021. Closing the gender gap in legislative debates: The role of gender quotas. *Political Behavior*, pages 1–25.

Michel Généreux, Iris Hendrickx, and Amália Mendes. 2012. A large portuguese corpus online: Cleaning and preprocessing. In *Computational Processing of the Portuguese Language*, pages 113–120, Berlin, Heidelberg. Springer Berlin Heidelberg.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Christopher J Kam. 2009. *Party discipline and parliamentary politics*. Cambridge University Press.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance detection benchmark: How robust is your stance detection? *KI-Künstliche Intelligenz*, 35(3):329–341.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.

Teun A Van Dijk. 1993. Principles of critical discourse analysis. *Discourse & society*, 4(2):249–283.

Polish Round Table Corpus

Maciej Ogrodniczuk[1], Ryszard Tuora[1] and Beata Wójtowicz[1,2]

[1]Institute of Computer Science, Polish Academy of Sciences;
[2]University of Warsaw;

maciej.ogrodniczuk@ipipan.waw.pl; ryszardtuora@gmail.com;
b.wojtowicz@uw.edu.pl

Abstract

The paper describes the process of preparation of the Polish Round Table Corpus (Pol. *Korpus Okrągłego Stołu*), a new resource documenting negotiations taking place in 1989 between the representatives of the communist government of the People's Republic of Poland and the Solidarity opposition. The process consisted of OCR of graphical transcripts of the talks stored in the form of parliament-like stenographic transcripts, carrying out their manual correction and making them available for search in a concordancer currently used for standard parliamentary transcripts.

Keywords: parliamentary data, Polish Round Table negotiations, contemporary history

1. Introduction

In 1988, against the backdrop of a growing wave of strikes and social protests, the authorities in communist People's Republic of Poland entered into negotiations with a section of the opposition (Solidarity movement, led by Lech Wałęsa) to resolve a simmering political conflict. Their final phase was the so-called 'Round Table', held in 1989 between 6 February and 5 April, with representatives of the Catholic Church acting as mediators. These talks marked the beginning of major political changes in Poland and accelerated the collapse of the entire communist bloc in Europe which makes them an important event in the recent history.

Round tables were about building a community of all people being equal. During the meeting, three main negotiating committees (the so-called tables) were established. The first was devoted to discussing the issue of trade union pluralism, the second one dealt with problems of economy and social policy, while the third team focused on the issue of political reforms. In addition to the committees, sub-committees (the so-called sub-tables) were also created. They were engaged in agriculture, mining, law and court reforms, associations and local governments, youth, mass media, housing, science, education and technical progress, health, ecology, wage and income indexation. A total of eleven sub-teams worked simultaneously headed by the country's main political leaders of that time. Several hundred people (participants, experts and observers) took part in the deliberations of all the teams, sub-teams and working groups (Polak and Galij-Skarbińska, 2021).

Although the Round Table negotiations were not part of the official parliamentary debate, they were documented in a form identical to the Polish parliamentary transcripts and are officially available on the Sejm website[1] as graphic PDF documents, without the text layer. This motivated us to make them available for searching in the concordance similarly (though separately from) the Polish Parliamentary Corpus (Ogrodniczuk, 2012, 2018)[2].

2. Data Preparation

2.1. Original Data Format

The original dataset consists of 96 documents contained in nearly 14,500 (A4) pages. Each (sub)table produced 1 to 13 meeting transcripts written on a typewriter (see Fig. 1). The documents vary in size from couple of dozen to 270 pages. They also vary in quality, due to unequal print visibility, writing errors, and handwritten notes that make the document less readable.

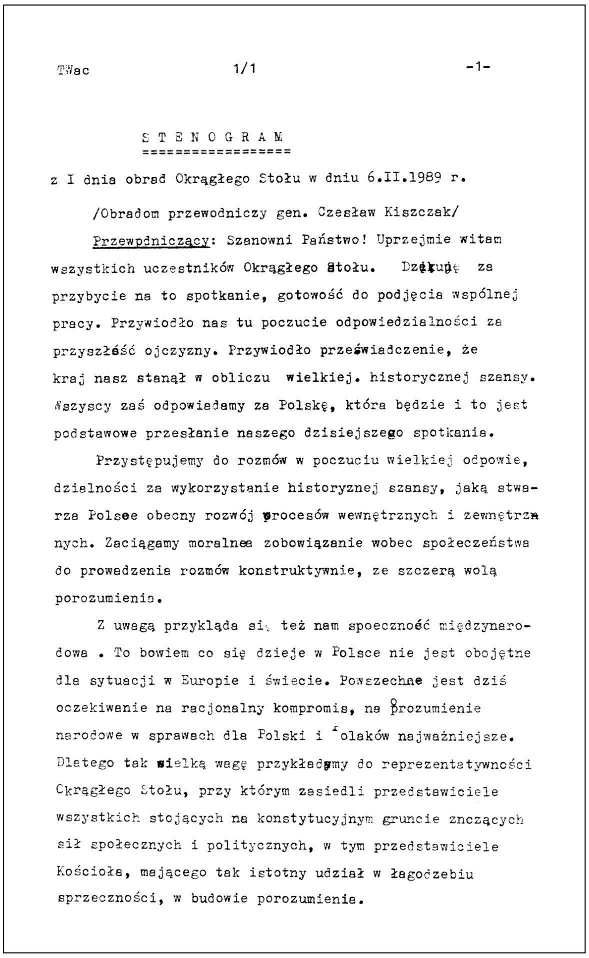
The documents follow a fairly consistent format for specifying the metadata, speakers' name, or interruptions, compatible with the one used while recording parliamentary sessions.

Fig. 1 illustrates well the quality of the transcript; already on its first page the number of problems of various kind is very high:

- 9 words with overwritten wrong characters
- one case of missing hyphenation (`odpowie, dzialności`)

[1] https://www.sejm.gov.pl/sejm7.nsf/stenOkrStol.xsp

[2] https://kdp.nlp.ipipan.waw.pl/



TWac 1/1 -1-

STENOGRAM
=================

z I dnia obrad Okrągłego Stołu w dniu 6.II.1989 r.

/Obradom przewodniczy gen. Czesław Kiszczak/

Przewpdniczący: Szanowni Państwo! Uprzejmie witam
wszystkich uczestników Okrągłego Stołu. Dziękuję za
przybycie na to spotkanie, gotowość do podjęcia wspólnej
pracy. Przywiodło nas tu poczucie odpowiedzialności za
przyszłość ojczyzny. Przywiodło przeświadczenie, że
kraj nasz stanął w obliczu wielkiej. historycznej szansy.
Wszyscy zaś odpowiadamy za Polskę, która będzie i to jest
podstawowe przesłanie naszego dzisiejszego spotkania.

Przystępujemy do rozmów w poczuciu wielkiej odpowie,
dzialności za wykorzystanie historyznej szansy, jaką stwa-
rza Polsee obecny rozwój procesów wewnętrznych i zewnętrzn
nych. Zaciągamy moralnee zobowiązanie wobec społeczeństwa
do prowadzenia rozmów konstruktywnie, ze szczerą wolą
porozumienia.

Z uwagą przykląda si, też nam spoeczność międzynaro-
dowa . To bowiem co się dzieje w Polsce nie jest obojętne
dla sytuacji w Europie i świecie. Powszechne jest dziś
oczekiwanie na racjonalny kompromis, na $\overset{o}{p}$rozumienie
narodowe w sprawach dla Polski i ^{P}olaków najważniejsze.
Dlatego tak wielką wagę przykładamy do reprezentatywności
Okrągłego Stołu, przy którym zasiedli przedstawiciele
wszystkich stojących na konstytucyjnym gruncie znczących
sił społecznych i politycznych, w tym przedstawiciele
Kościoła, mającego tak istotny udział w łagodzebiu
sprzeczności, w budowie porozumienia.

Figure 1: The first page of the transcript of the inaugural meeting of the Round Table on February 6, 1989.

- almost unreadable end character in a word (się)

- 6 words with various uncorrected typos (historyznej → historycznej, moralnee → moralne, przykląda → przygląda, spoeczność → społeczność, znczących → znaczących, łagodzebiu → łagodzeniu)

- one correction made by adding the missing character over the word (prozumienie → porozumienie)

- one character typed over the line (^{P}olaków)

- one case of wrong punctuation (dot in place of a comma: wielkiej. historycznej szansy)

All writing flaws and text imperfections had a negative impact on the quality of the OCR process. Therefore an additional phase of manual correction had to be introduced.

2.2. Data Conversion and Annotation

The transcripts were OCR-ed with ABBYY FineReader 12 under manual supervision, and initially reviewed by an annotator (all errors noted down in the previous section were successfully corrected in this process). After this phase, the documents were converted to HTML format, to preserve their structural features. Subsequently, the data was cleaned and homogenized using semi-manual techniques (e.g. regular expressions). Additionally normalization of speakers was applied to around 200 most frequent vocal participants (originally multiple aliases[3] per person were used).

The data was then processed using the pl_nask model[4] for spaCy (Honnibal et al.), to provide POS tagging, lemmatization, dependency parsing and named entity recognition.

2.3. Data Statistics

The corpus consists of 96 documents, which amount to 3 272 149 tokens, 162 595 sentences, 67 185 paragraphs and 23 437 speeches.

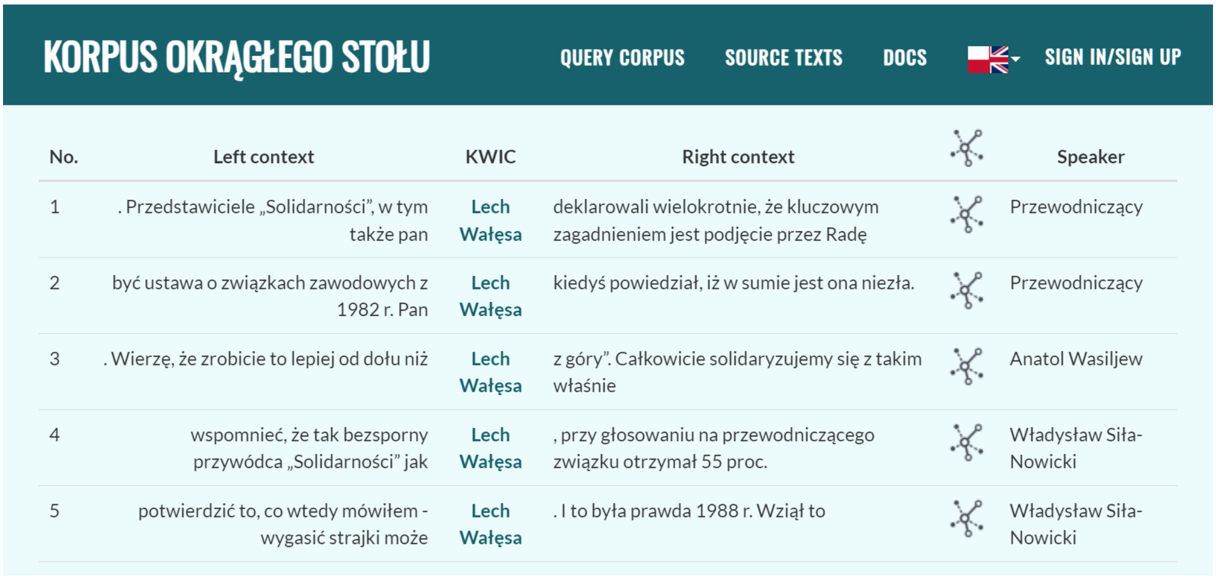
The most frequent speaker designation is 'Chairman', decoded in the initial section of the transcript. It is also very common that speeches are not attributed to anyone (see Table 1).

Speaker	Speeches
Chairman	7507
Missing	2997
Jerzy Kołodziejski	744
Władysław Baka	610
Łukasz Balcer	407
Chairwoman	375
Stefan Kozłowski	325
Jan Brol	322
Adam Strzembosz	310
Alojzy Pietrzyk	258
Voice from the audience	244
Witold Trzeciakowski	242
Rajmund Moric	211
Tadeusz Mazowiecki	202
Bronisław Geremek	189

Table 1: Most frequent speakers.

[3] For instance: Aleksander Kwaśniewski, the then minister-without-portfolio in the People's republic of Poland figures in text as 'minister Kwaśniewski', 'colleague Kwaśniewski', 'deputee Kwaśniewski' etc. It was not possible to provide full normalization of speakers, as in some cases (e.g. when two participants share a surname, or a private person is speaking, with no full name given) attributions are ambiguous.

[4] https://huggingface.co/ipipan/pl_nask



No.	Left context	KWIC	Right context	Speaker
1	. Przedstawiciele „Solidarności”, w tym także pan	Lech Wałęsa	deklarowali wielokrotnie, że kluczowym zagadnieniem jest podjęcie przez Radę	Przewodniczący
2	być ustawa o związkach zawodowych z 1982 r. Pan	Lech Wałęsa	kiedyś powiedział, iż w sumie jest ona niezła.	Przewodniczący
3	. Wierzę, że zrobicie to lepiej od dołu niż	Lech Wałęsa	z góry”. Całkowicie solidaryzujemy się z takim właśnie	Anatol Wasiljew
4	wspomnieć, że tak bezsporny przywódca „Solidarności” jak	Lech Wałęsa	, przy głosowaniu na przewodniczącego związku otrzymał 55 proc.	Władysław Siła-Nowicki
5	potwierdzić to, co wtedy mówiłem - wygasić strajki może	Lech Wałęsa	. I to była prawda 1988 r. Wziął to	Władysław Siła-Nowicki

Figure 2: A KWIC index offered by Korpusomat.

The statistics of speeches within specific committees and sub-committees (see Table 2) illustrates the importance of the topics discussed.

Committee	Speeches
Economy and Social Policy	4874
Law and Court Reform	2756
Ecology	2650
Health	1899
Union Pluralism	1987
Political Reform	1758
Mining	1602
Associations and Local Gov.	1504
Agriculture	1071
Housing Policy	1005
Science, Education and Technical Progress	850
Youth Affairs	584
Mass Media	449
Wage and Income Indexation	391
Plenary Sessions	57

Table 2: Statistics of speeches within specific committees.

3. Searching the Corpus

Finally the documents were indexed in Korpusomat (Kieraś et al., 2018; Saputa et al., 2023) — an established Web application for accessing and working with corpus data (see Figure 2). The transcripts are searchable, using both the annotation layers, and metatextual information (i.e. speaker names, or metadata such as committee name).

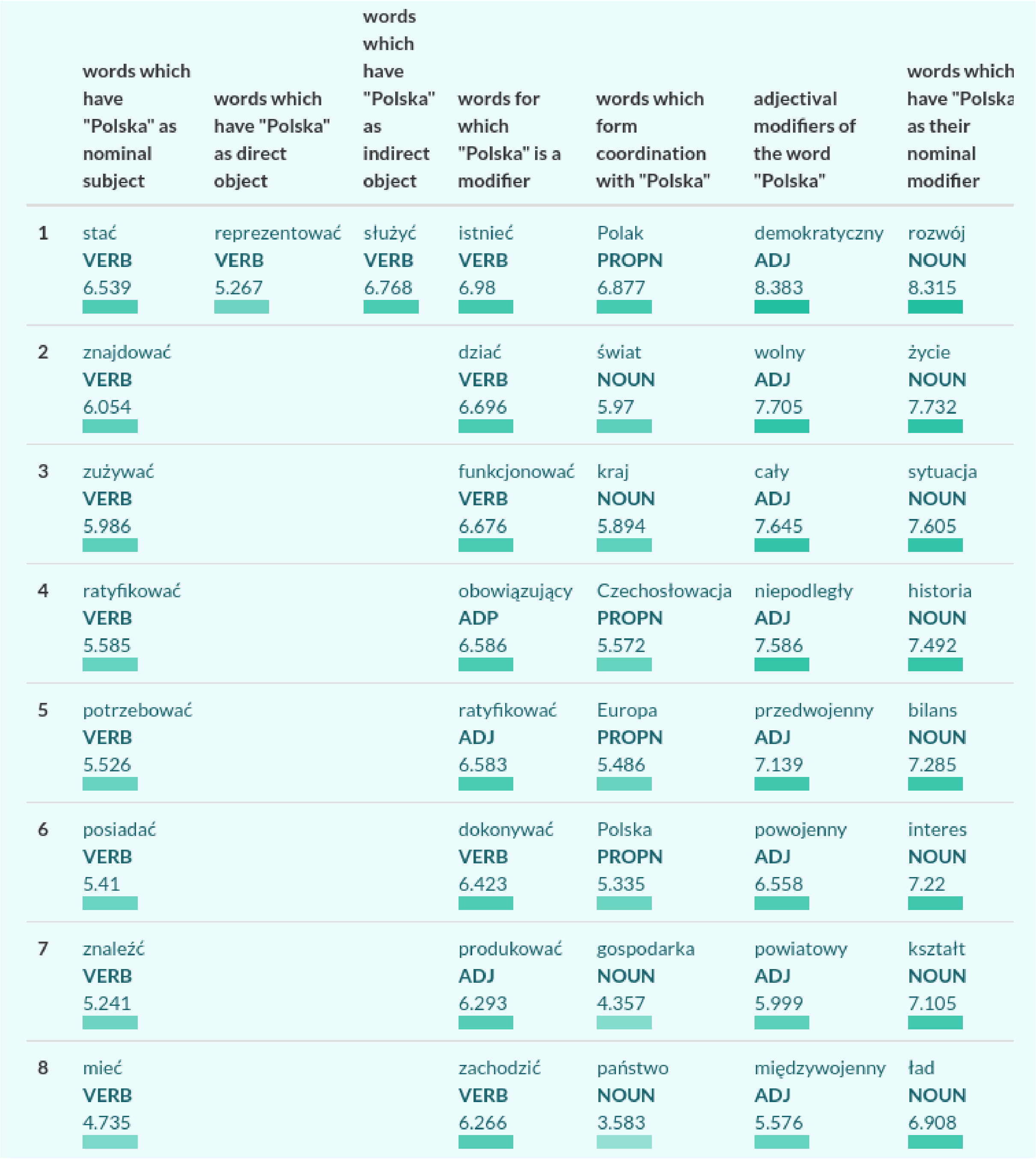
Additionally, the ‘word profile’ functionality was employed, which allows to visualize how a particular word is used in the corpus (see Figure 3) by surveying regularities in grammatical connections it enters into with other words.

4. Future Work

Even though the data conversion process involved manual interventions at various stages, the data still needs many manual updates. Known types of errors include:

- typos introduced by the stenographer and corresponding to in-vocabulary Polish words (such as `patynie` instead of `pytanie` on page 4 in the first session), undetectable without careful revision of the text
- wrong recognition of mostly Polish characters during the OCR process (such as `sie` instead of `się`, `l` instead of `i`), which are difficult to spot
- obvious slips which are always corrected in the official transcript (e.g. `w sprawach najważniejsze` → `najważniejszych`)
- typographical errors, including editing errors, e.g. introducing unnecessary characters, like extra spaces, in the text
- names of speakers’ functions (e.g. “chairman”) used in place of their names after the function assignment to the speaker is recorded in the commentary on the earlier part of the transcript (see Fig. 1, line in typewriter brackets directly over the underlined designation).

Despite such errors, the transcripts make a valuable documentation of the Polish bloodless road



	words which have "Polska" as nominal subject	words which have "Polska" as direct object	words which have "Polska" as indirect object	words for which "Polska" is a modifier	words which form coordination with "Polska"	adjectival modifiers of the word "Polska"	words which have "Polska as their nominal modifier
1	stać VERB 6.539	reprezentować VERB 5.267	służyć VERB 6.768	istnieć VERB 6.98	Polak PROPN 6.877	demokratyczny ADJ 8.383	rozwój NOUN 8.315
2	znajdować VERB 6.054			dziać VERB 6.696	świat NOUN 5.97	wolny ADJ 7.705	życie NOUN 7.732
3	zużywać VERB 5.986			funkcjonować VERB 6.676	kraj NOUN 5.894	cały ADJ 7.645	sytuacja NOUN 7.605
4	ratyfikować VERB 5.585			obowiązujący ADP 6.586	Czechosłowacja PROPN 5.572	niepodległy ADJ 7.586	historia NOUN 7.492
5	potrzebować VERB 5.526			ratyfikować ADJ 6.583	Europa PROPN 5.486	przedwojenny ADJ 7.139	bilans NOUN 7.285
6	posiadać VERB 5.41			dokonywać VERB 6.423	Polska PROPN 5.335	powojenny ADJ 6.558	interes NOUN 7.22
7	znaleźć VERB 5.241			produkować ADJ 6.293	gospodarka NOUN 4.357	powiatowy ADJ 5.999	kształt NOUN 7.105
8	mieć VERB 4.735			zachodzić VERB 6.266	państwo NOUN 3.583	międzywojenny ADJ 5.576	ład NOUN 6.908

Figure 3: Word profile generated for the word "Polska" (Poland), which occurs a total of 2030 times. The figures correspond to logDICE values for each collocation.

to democracy and its searchable variant will definitely help the digital humanities researchers in their work.

Acknowledgements

The work was financed by the European Regional Development Fund as a part of the 2014–2020 Smart Growth Operational Programme, CLARIN — Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00–00C002/19[5], the Polish Ministry of Education and Science grant 2022/WK/09 and as part of the investment CLARIN ERIC — European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure (period: 2024–2026) funded by the Polish Ministry of Science and Higher Education (Programme: "Support for the participation of Polish scientific teams in international research infrastructure projects"), agreement number 2024/WK/01.

[5] https://clarin.biz/

5. Bibliographical References

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python.

Witold Kieraś, Łukasz Kobyliński, and Maciej Ogrodniczuk. 2018. Korpusomat — a tool for creating searchable morphosyntactically tagged corpora. *Computational Methods in Science and Technology*, 24(1):21–27.

Maciej Ogrodniczuk. 2012. The Polish Sejm Corpus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2219–2223, Istanbul, Turkey. European Language Resource Association.

Maciej Ogrodniczuk. 2018. Polish Parliamentary Corpus. In *Proceedings of the LREC 2018 Workshop* ParlaCLARIN: Creating and Using Parliamentary Corpora, pages 15–19, Paris, France. European Language Resources Association.

Wojciech Polak and Sylwia Galij-Skarbińska. 2021. The Round Table in 1989 — Consequences and Evaluation. *Polish Political Science Yearbook*, 50:149–156.

Karol Saputa, Aleksandra Tomaszewska, Natalia Zawadzka-Paluektau, Witold Kieraś, and Łukasz Kobyliński. 2023. Korpusomat.eu: A multilingual platform for building and analysing linguistic corpora. In *Computational Science – ICCS 2023. 23rd International Conference, Prague, Czech Republic, July 3–5, 2023, Proceedings, Part II*, number 14074 in Lecture Notes in Computer Science, pages 230–237, Cham. Springer Nature Switzerland.

Investigating Multilinguality in the Plenary Sessions of the Parliament of Finland with Automatic Language Identification

Tommi Jauhiainen, Jussi Piitulainen, Erik Axelson, Ute Dieckmann, Mietta Lennes, Jyrki Niemi, Jack Rueter, Krister Lindén
Department of Digital Humanities, University of Helsinki

Abstract

In this paper, we use automatic language identification to investigate the usage of different languages in the plenary sessions of the Parliament of Finland. Finland has two national languages, Finnish and Swedish. The plenary sessions are published as transcriptions of speeches in Parliament, reflecting the language the speaker used. In addition to charting out language use, we demonstrate how language identification can be used to audit the quality of the dataset. On the one hand, we made slight improvements to our language identifier; on the other hand, we made a list of improvement suggestions for the next version of the dataset.

Keywords: language identification, multilinguality, plenary sessions

1. Introduction

In this paper, we use automatic language identification to investigate the usage of different languages in the plenary sessions of the Parliament of Finland. The plenary sessions are published as transcriptions of speeches given in Parliament, reflecting the language the speaker actually used. Finland has two national languages, Finnish and Swedish, as well as several minority languages, such as the Sami languages and the Finnish Romani.

Language identification can be used to bring forth many kinds of problems in the corpus processing pipeline for the dataset at hand. Instead of trying to circumvent all the problems by tweaking the language identifier, we record the issues that we can correct earlier in the pipeline.

In Section 2, we introduce some work on multilingual parliamentary proceedings. The details of the corpus we are focusing on in this paper are given in Section 3. Section 4 is a detailed description of our language identification process and how it can be used to improve the quality of both the corpus and the language identifier. In Section 5, we present the results of the language identification experiments, e.g., details on the languages used in Parliament. Section 6 is dedicated to investigating the sentences tagged as written in an undetermined language. In Section 7, we discuss the process and list our improvement suggestions.

2. Previous Work

There are several state bodies similar to the Parliament of Finland where the use of several languages is permitted. One of the most prominent bodies is the Canadian Parliament, where both English and French enjoy equal status and use (Hudon, 2022). Another source for multilingual parliamentary data is the Catalan Parliament, where discussions can include Spanish and Aranese Occitan interventions in addition to Catalan (Kulebi et al., 2022). The Belgian federal Parliament uses both Dutch and French, which were automatically identified on the paragraph level for the ParlaMint corpora (Erjavec et al., 2023).

The language use in the European Parliament is on a totally different level of multilingualism, currently with 24 official languages.[1]

As far as we are aware, this is the first study where fine-grained language identification is performed and the results analyzed on any of these corpora.

We have previously done similar experiments with the Newspaper and Periodical Corpus of the National Library of Finland (NLF) [2] and the Suomi24 Sentences Corpus 2001-2017 (suomi24-2001-2017)[3] (Jauhiainen et al., 2022b).

3. Corpus

The focal dataset of this paper is the Plenary Sessions of the Parliament of Finland, Downloadable Version 1.5 (The Parliament of Finland, 2017-01-01).

The dataset is available at the Language Bank of Finland (LBF).[4] The Language Bank is a comprehensive service suite for researchers utilizing linguistic resources. It hosts an extensive collection

[1] https://european-union.europa.eu/principles-countries-history/languages_en

[2] http://urn.fi/urn:nbn:fi:lb-2021092404

[3] suomi24-2001-2017-korp-v1-1, http://urn.fi/urn:nbn:fi:lb-2020021803

[4] https://www.kielipankki.fi/language-bank/

of text and speech datasets, enabling diverse use. Users can explore and process these datasets using the Language Bank's online tools or download them to their personal computers.

The services of the Language Bank are overseen by the national FIN-CLARIN consortium, which consists of Finnish universities and research organizations.[5] FIN-CLARIN is part of the international CLARIN ERIC research infrastructure.[6] Researchers and research groups can arrange with FIN-CLARIN for the storage and distribution of their own research datasets.

3.1. Plenary Sessions of the Parliament of Finland

The proceedings of the plenary sessions of the Parliament of Finland are documented in minutes, which include information on the content of discussions, details of decisions made, and all speeches given. These minutes are prepared in both Finnish and Swedish. However, the speeches are recorded and published in the language in which they were originally delivered. The preparation of the minutes occurs in real-time during the session, and they are made available on the Parliament's website as soon as they are ready.[7]

3.2. Speech and Text Alignment

The Parliament of Finland's original written records have been synchronized with the audio from the video footage of the plenary meetings. The synchronization process involved aligning the spoken words of each individual speaker separately. This task was accomplished using automated tools developed by Aalto University.[8]

It's important to be aware that the synchronized transcripts might include inaccuracies, and unnecessary tags could have been added to the text as a result of the automated synchronization and voice recognition procedures. In instances where there was no corresponding text for the original audio in the transcripts, the speech was automatically transcribed, which could lead to unusual or incorrect entries.

3.3. eduskunta-v1.5-dl

The verticalized text (VRT) version of the Eduskunta corpus consists of one 1.9-gigabyte CWB-VRT (The IMS Open Corpus Workbench-VRT, (Evert and Team, 2022)) file comprising 28 million lines, organized into 1,009 text elements that mirror video files. These elements are further broken down into paragraphs (111,097), utterances (1,499,627, linked to specific video timestamps), and sentences, which are sequences of tokens. Each token is on its own line, together with the linguistic analysis of the sentence as token annotations. The sentence length distribution in Table 1 was computed with one of the vrt-tools developed in the Language Bank.

Total	Nobs	Mean
22,458,581	1,499,627	14.98

Min	D1	D2	LoQ	D3	D4	
1	4	6	7	9	11	
Med	**D6**	**D7**	**HiQ**	**D8**	**D9**	**Max**
13	15	18	20	22	28	406

Table 1: The distribution of sentence lengths measured in tokens, as segmented in eduskunta.vrt (v1.5). There are over 22 million tokens in 1.5 million sentences, with a mean sentence length of just below 15 tokens and a median of 13. The quantile points (deciles and the low and high quartile) are represented by the observed value at or above the point.

Table 2 shows the distribution of the videos over the time covered by the corpus.

videos	year
46	2008
120	2009
132	2010
124	2011
130	2012
134	2013
130	2014
116	2015
77	2016

Table 2: The 1,009 videos (by the attributes in the text element tags in the VRT file) counted by year.

The LBF has invested in the ability to annotate a single file format (CWB-VRT) with different tools, which is facilitated by adding *field names* to the otherwise purely positional token records. The names are declared in a comment at the beginning of the file, leaving token lines in the form of tab-separated values. The various VRT tools[9] can then refer to the input and output fields by name regardless of

[5] https://www.kielipankki.fi/organization/

[6] https://www.clarin.eu

[7] https://www.eduskunta.fi/FI/taysistunto/Sivut/Taysistuntojen-poytakirjat.aspx

[8] http://urn.fi/urn:nbn:fi:lb-2014091904

[9] https://github.com/CSCfi/Kielipankki-utilities/tree/master/vrt-tools

6,917,510	N
4,697,950	V
2,523,037	Adv
2,485,364	Pron
1,729,493	C
1,589,639	A
1,499,627	Punct
348,303	Num
315,561	Foreign
281,789	Adp
52,092	Symb
18,216	Interj

Table 3: The counts of the "parts of speech" of the tokens in the corpus file, as identified by the annotation pipeline.

their actual position on the line.

The sentences were annotated in the LBF with the old TurkuNLP Finnish dependency parser pipeline, adapted for the VRT file format.[10] The pipeline consists of two uses of a lexical transducer, OmorFI (Pirinen, 2015), first to look up all possible lemmatizations and some corresponding morpho-syntactic features for each word form in a sentence, disambiguated with a MarMot model (Mueller et al., 2013) trained by the Turku group as part of their pipeline. This is followed by another OmorFi lookup to fill in the features of the contextually selected reading of each token, and finally syntactic dependency analysis corresponding to the Turku Dependency Treebank (TDT) (Haverinen et al., 2014) with a trained model that uses MaTe tools (Björkelund et al., 2010).[11] The annotation model predates the Universal Dependencies effort (De Marneffe et al., 2021).

Further variants of the base forms were added afterward to enable certain features in the Korp platform, where the corpus is made available for the search of examples.[12]

The corpus was further annotated with FiNER (Ruokolainen et al., 2020) to annotate the tokens that were recognized to be parts of names (or some other expressions) by their classes (like person, organization, location).

Table 3 shows how many times the annotation pipeline classified a token as noun, verb, and so on. The number of "Foreign" words may or may not be an indication of the proportion of non-Finnish language in the corpus.

The sentence-per-line view of the VRT file used in the following experiments was extracted with a relatively straightforward VRT tool that, by default, lists the token forms of each sentence on the same line, separated by space characters.

4. Language Identification

Our language identification experiments were conducted on a sentence level using the HeLI-OTS language identifier (Jauhiainen et al., 2022a).[13] We are currently using this language identifier on our standard corpus creation pipeline (Jauhiainen et al., 2022c; Dieckmann et al., 2023). However, the level on which the language identification is sensible differs from one dataset to another. For example, the optical character recognition (OCR) quality of the Newspaper and Periodical Corpus of the National Library of Finland (NLF) [14] is in places so terrible that out of the box identification results for the sentences can be very exotic (Jauhiainen et al., 2022b).

For development purposes, we have an internal test set for HeLI-OTS. The test set contains more than 1.2 million lines of text written in one of the 200 languages HeLI-OTS has in its repertoire. Whenever we modify the software or its language models, we investigate the effects of these changes by considering the recall, precision, and F-scores before and after the change. We look at these scores on the overall average level for all languages as well as on the level of individual languages if needed. The test set is not an independent entity, and it has not been manually verified, so whenever we make changes that cause the error rates to increase for some languages, we may take a look at the misidentified sentences in order to check their validity and remove them from the test set. We may also add new text lines to the test set when developing the identifier system as part of a specific investigation similar to what is described in this paper.

As of the writing of this paper, the current published version of the identifier is HeLI-OTS 1.5.[15] On the internal test set, it attains a macro F1 score of 99.21% over the 200 languages and a micro F1 score of 99.62% over the c. 1.2 million lines.

4.1. Experiments with HeLI-OTS 1.5

At first glance, the quality of the sentences in the corpus at hand seems to be far superior to the one in the NLF corpus. However, sentence-level monolingual language identification still comes up with sentences in 129 different languages. Table 4

[10] https://github.com/TurkuNLP/Finnish-dep-parser

[11] https://www.ims.uni-stuttgart.de/en/research/resources/tools/matetools/

[12] https://www.kielipankki.fi/korp

[13] http://urn.fi/urn:nbn:fi:lb-2022011801

[14] http://urn.fi/urn:nbn:fi:lb-2021092404

[15] https://doi.org/10.5281/zenodo.10071264

lists the ten languages with the most identifications on the initial language identification run.

# sentences	ISO 639-3	Language
1,460,998	fin	Finnish
17,408	swe	Swedish
9,454	ido	Ido
1,840	hat	Haitian Creole
1,306	izh	Ingrian
1,044	ewe	Ewe
1,041	vot	Votic
756	und	Undetermined
512	kal	Greenlandic
468	olo	Livvi

Table 4: Top 10 identified languages with the number of sentences for each. HeLI-OTS 1.5 language identifier.

6,449 of the sentences identified as Ido were simply “Ed .”. This is an abbreviation for “Edustaja” meaning “representative”. Most of the rest of the sentences identified as Ido ended with “ ed .”. The problem here seems to be on the sentence tokenization level, as the sentences have been cut using the period after the abbreviation in a way it should not have been done.

The 576 sentences identified as Haitian Creole were “Värderade talman .”, meaning "Honored Speaker" in Swedish. Most of the other sentences identified as Haitian Creole included the word “talman” as well. “talman” seems to be a common word ending in the HeLI-OTS training corpus for Haitian Creole, whereas the word “talman” is so rare in the Swedish training corpus that the word has not made it to the word level language model for Swedish. Both training corpora are based on web crawls and originate from the Leipzig corpora collection (Goldhahn et al., 2012).[16] As this is a clear language identification error on a correctly tokenized sentence, we decided to switch to our development version of the HeLI-OTS language identifier featuring individual confidence thresholds for each language. We expected that sentences like “Värderade talman .” and other short sentences in Swedish identified as Haitian Creole would not have high confidence scores.

The unpublished version of the HeLI-OTS used in these experiments had been modified from the 1.5 version in the context of performing language identification on an excerpt of 10,000 Tweets from the Sydney area. In addition to the new confidence thresholds, the modifications included cleaning English material from the training corpora of other languages. On the internal test set, this version attained a macro F1 score of 99.59% and a micro F1 score of 99.66%.

4.2. Experiments with Confidence Thresholds

The development version came up with a slightly lower number of languages for the dataset: 112. The renewed top 10 language list can be seen in Table 5.

# sentences	ISO 639-3	Language
1,462,709	fin	Finnish
17,523	swe	Swedish
13,802	und	Undetermined
1,514	hat	Haitian Creole
711	izh	Ingrian
666	vot	Votic
331	ido	Ido
275	lud	Ludic
194	est	Estonian
99	pol	Polish

Table 5: Top 10 identified languages with the number of sentences for each. Development version of the HeLI-OTS language identifier.

The number of sentences in the undetermined category rose drastically due to the introduction of the confidence thresholds. Some of the language models in the development version have also been improved, so the number of sentences identified as Finnish and Swedish also rose slightly. Surprisingly, the number of sentences identified as Haitian Creole did not decrease as much as expected. “Värderade talman .”, “Ärade talman .”, and “Herr talman .” were still identified as Haitian Creole.

4.3. Increasing Swedish Vocabulary

Developing a general-purpose language identification system is always a compromise between the compactness of the system and the number of features retained for each language. At this point, the 10,000 most common features were retained in each feature category for Swedish.[17] The number of features retained is an individual setting for each language, currently spanning from 5,000 to 50,000 features. Based on these perfectly Swedish, and surely not Haitian Creole, sentences being misidentified, we increased the number of retained features to 30,000 for Swedish. The updated list of the top 10 languages and the number of sentences identified as each is shown in Table 6.

The number of sentences identified as Haitian Creole decreased so much that the language

[16] The corpora are “hat-ht_web_2015_30K” for Haitian Creole and the “swe_web_2002_1M” for Swedish.

[17] The feature categories in the off-the-shelf HeLI-OTS are words and character n-grams from one to six.

# sentences	ISO 639-3	Language
1,462,451	fin	Finnish
19,303	swe	Swedish
13,787	und	Undetermined
711	izh	Ingrian
666	vot	Votic
331	ido	Ido
275	lud	Ludic
200	est	Estonian
98	pol	Polish
79	eng	English

Table 6: Top 10 identified languages with the number of sentences for each on the third LI round.

dropped out of the top 10, with the number of sentences identified as Swedish increasing accordingly. On the internal test set, this version attains a macro F1 score of 99.21% and a micro F1 score of 99.64%. Increasing the Swedish vocabulary resulted in more of the sentences marked as Norwegian or Danish in the internal test set to be identified as Swedish. However, we considered it less of an error to confuse between these close Scandinavian languages than between Scandinavian languages and Haitian Creole. We should also be able to rectify this problem later by increasing the size of the Danish and Norwegian language models similarly.

The next language on the list is Ingrian, an underresourced Finnic language that is rather similar to Finnish. The most common sentences that had been identified as Ingrian were: "Otan esimerkin .", "Minä kysyn .", and "Pulliaiselle .", in English "I take an example.", "I ask.", and "To Pulliainen." These are perfectly all-right sentences in spoken Finnish, but the problem is that they could also be so in Ingrian.

4.4. Confusion between Ingrian Dialects and Ingrian

After closer examination of the sentences identified as Ingrian in the dataset as well as the Ingrian training corpus for HeLI-OTS, we came to the conclusion that, unfortunately, a long transcribed interview of a Finnish Ingrian dialect speaker had ended up in the Ingrian corpus. The Ingrian dialects[18] are considered Finnish, whereas Ingrian[19] itself is a separate language by the ISO 639-3 standard. After cleaning the Ingrian training corpus and recalculating its language models, we arrived at a list shown in Table 7.

At this point, we also audited the results on the internal test set and produced an updated set (already version 30 for the 200 languages). This was our last modification of the HeLI-OTS in the experiments described in this paper. On the new internal test set, the macro F1 over the 200 languages was 99.61%, and the micro F1 over the c. 1.2 million sentences was 99.68%.

# sentences	ISO 639-3	Language
1,462,016	fin	Finnish
19,304	swe	Swedish
14,883	und	Undetermined
668	vot	Votic
331	ido	Ido
274	lud	Ludic
203	est	Estonian
99	pol	Polish
79	eng	English
76	gsw	Swiss German

Table 7: Top 10 identified languages with the number of sentences for each on the fourth LI round.

Closer inspection of the sentences identified as Votic and Ludic revealed that most of them had the same " ed ." abbreviation problem as the sentences identified as Ido.

4.5. Common Abbreviation Handling

As the " ed ." abbreviation seemed to be responsible for the majority of remaining incorrect language identification, we decided to simulate the situation where the problem would have been corrected earlier in the pipeline. We rejoined the sentences where they had been cut off after the abbreviation. We also corrected an encoding issue, which was observed on c. 400-500 lines. The total number of sentences dropped from 1,499,627 to 1,474,286, which meant that 25,341 additional sentences had been created due to the abbreviation.

With this change, the number of different languages dropped from 112 to 111, and the top ten languages with the number of sentences can be seen in Table 8.

The number of sentences with undetermined language was more than halved, and the number of sentences identified as Votic, Ido, and Ludic was drastically reduced.

The corpus description[20] of the Korp version declares, "For portions where the original audio track did not have matching text in the transcript, the speech signal was recognized automatically using a Finnish language model, and such portions may contain strange or erroneous content." This declaration is missing from the metadata of the

[18] https://en.wikipedia.org/wiki/Ingrian_dialects

[19] https://en.wikipedia.org/wiki/Ingrian_language

[20] http://urn.fi/urn:nbn:fi:lb-2019101621

# sentences	ISO 639-3	Language
1,446,338	fin	Finnish
19,286	swe	Swedish
6,626	und	Undetermined
187	est	Estonian
111	vot	Votic
78	eng	English
75	gsw	Swiss German
57	kal	Greenlandic
56	pol	Polish
54	roh	Romansh

Table 8: Top 10 identified languages with the number of sentences for each on the fifth LI round.

downloadable version.[21] It is an important piece of information as this behavior seems to be the cause of most of the "sentences" erroneously identified as Estonian. Unfortunately, this information is not currently provided on the utterance or word level.

4.6. Automatic Speech Recognition

Even though the metadata did not indicate whether the utterances were transcribed using Finnish automatic speech recognition (ASR), we observed that in all those cases we encountered, the sentences started with a lowercase letter. Identifying the origin language of the ASR-generated sentences is a very different task from general language identification and would require the use of other kinds of tools. As most of the observed identification errors in the top 10 languages seemed to originate from exotic utterances created by ASR, we decided to filter out all those sentences starting with a lowercase letter. This operation reduced the number of "sentences" by 6.4% and the number of tokens by 5.3%, indicating that the ASR-generated texts were shorter than average.

The total number of identified languages went down from 111 to 77. The updated list of sentences per language is shown in Table 9.

The final list of languages is missing the lone Northern Sami utterance we discovered during the described process. It came from Oras Tynkkynen[22] in 2013. In the middle of his speech, he re-saluted the speaker of the house in four languages. Finnish, Swedish, Northern Sami, and Russian: "Arvoisa puhemies. Ärade talman. arvvus adnon sagadoalli. Uvazhajemyi predsedatel." The Russian version was transcribed using Latin characters and was thus not identified as Russian but as Slovakian. The Sami version was lost when we discarded all sentences beginning with a lowercase letter. We inspected the transcript on the Parliament site and found that for that sentence, it reads: "Árvvus adnon ságadoalli!". It seems our corpus preparation process has dismissed the accents in this case. On this occasion, we noticed that all punctuations other than periods had also been either removed or transformed into periods.

# sentences	ISO 639-3	Language
1,358,878	fin	Finnish
18,030	swe	Swedish
3,053	und	Undetermined
84	vot	Votic
43	est	Estonian
36	kal	Greenlandic
36	eng	English
35	roh	Romansh
24	lat	Latin
23	ido	Ido

Table 9: Top 10 identified languages with the number of sentences for each on the sixth LI round.

5. Results

The actual languages attested in the dataset were very few: Finnish, Swedish, English, Latin, French, German, Spanish, Italian, and Northern Sami. Table 10 gives the number of "sentences" containing languages other than Finnish or Swedish observed in the dataset. Some sentences were well formed, but others were only single-word or partial sentences, as well as multilingual sentences containing Finnish and the indicated language.

# sentences	ISO 639-3	Language
34	eng	English
21	lat	Latin
2	fra	French
2	deu	German
2	spa	Spanish
1	ita	Italian

Table 10: The number of sentences in languages other than Finnish or Swedish that were actually observed and correctly identified in the dataset.

The longest English sentence we have found was uttered by Jacob Söderman[23] in 2011: "Whistleblowing is the popular term used when someone who works in or for an organisation raises a concern about a possible fraud crime danger or other serious risk that could threaten customers colleagues shareholders the public or the organisations own reputation.". He used this definition when explain-

[21] http://urn.fi/urn:nbn:fi:lb-2019101721

[22] https://www.eduskunta.fi/EN/kansanedustajat/Pages/846.aspx

[23] https://www.eduskunta.fi/FI/kansanedustajat/Sivut/311.aspx

ing the concept of whistleblowing to the Parliament in an otherwise Finnish speech.

The only real sentence in Latin we found was "Navigare necesse est.", which was said by Astrid Thors[24] in 2012. Her speech about the state of Finnish seafaring was mostly in Swedish but contained two longer passages in Finnish as well.

Our only French sentence comes from Timo Soini[25] in 2014: "Un pere une mere cest elementaire.". He was talking about participating in protests in France in the context of a discussion about same-sex marriage in Finland. The only lone (real) sentence identified as German comes from the same political party, the Finns: "Kein Geld fur Merkel nicht mehr.". It was uttered by Juha Väätäinen[26] in December 2011 in the context of European monetary policy. His previous sentence in the same speech is one of the two Spanish sentences we found in the corpus: "No mas dinero para Espana no mas euros para Italia.". A week later, he said the only more than one-word sentence identified as Italian: "Bravo bravissimo.".

6. Undetermined Languages

During the language identification experiments, we were especially focused on minimizing the number of false positives in languages that were not actually attested in the dataset. In the final list, shown in Table 9, we additionally had a little over three thousand sentences tagged as written in an undetermined language.

745 of these did not contain any alphabetical characters but consisted only of number characters or a single dot. Furthermore, c. 200 "sentences" consisted only of a personal name. These we consider to be correctly identified when tagged with an "und" label.

We collected the top 10 Finnish sentences left undetermined in Table 11. Most of these gain similar scores for other close Finnic languages as they do for Finnish. In cases like "Ei." e.g., "No", or "On.", e.g., "is", they could correctly be tagged with several languages such as Finnish and Estonian. A more advanced way of handling multi-lingual words would be to tag them with several language labels or with a label of the language group the languages belong to. A notable difference in the list is the last example, which contains the abbreviation "Ed." again favoring the identification as the Ido language. The "Ed." abbreviation followed by an inflected personal name is found in a further 250 sentences.

#	Sentence
273	Hyvät kollegat.
133	Hyvät edustajat.
88	Hyvät edustajakollegat.
62	Ei ole.
50	Ei.
28	Näin ei voi jatkua.
20	Kysynkin ministeri Risikolta.
17	On.
14	Hyvät ystävät.
12	Ed. Ukkolalle.

Table 11: The counts of the top 10 Finnish sentences tagged with undetermined language.

HeLI-OTS has the option to perform language set identification, which means that in the case of multilingual sentences, it can give several tags to the sentence. We have not yet experimented with this feature on this corpus, but there is a clear need for it as the next most common sentences left undetermined were multilingual Finnish-Swedish sentences. We give the top eight multilingual sentences with their counts in Table 12. The rest of the multilingual sentences did not occur more than once. Seven out of the eight repeating sentences are multilingual only due to the decision made by the transcriber. They could as well have been transcribed as two separate sentences, e.g., the first part of the most common multilingual sentence "Värderade herr talman" occurs 80 times as a lone sentence and the latter part "Arvoisa herra puhemies" 18,552 times. The word "Eli", e.g., "So", followed by the Latin phrase "summa summarum", could perhaps be considered a real code-switched sentence.

#	Sentence
36	**Värderade herr talman** arvoisa herra puhemies.
34	Arvoisa puhemies **herr talman**.
32	**Herr talman** arvoisa puhemies.
11	Arvoisa herra puhemies **värderade herr talman**.
8	Eli **summa summarum**.
4	**Fru talman** rouva puhemies.
2	**Värderade herr talman** ar voisa herra puhemies.
2	Hyvät edustajat **bästa riksdagsledamöter**.

Table 12: The counts of the top eight multilingual sentences tagged with undetermined language. The non-Finnish parts are indicated by boldface type.

The next notable group of sentences with undetermined languages consists of two to three-

[24] https://www.eduskunta.fi/FI/kansanedustajat/Sivut/770.aspx

[25] https://www.eduskunta.fi/FI/kansanedustajat/Sivut/767.aspx

[26] https://www.eduskunta.fi/FI/kansanedustajat/Sivut/1139.aspx

letter sentences containing the word "ministeri", e.g., "minister". For some reason, "ministeri" is a very common word in the Greenlandic training corpus, which resulted in a high number of short sentences being identified as Greenlandic, as can be seen in Table 4. Now, 231 of these sentences containing "Ministeri" or "ministeri" are tagged with an undetermined language.

Additionally, there were still a few long sentences containing Finnish and Swedish words that were clearly produced by the ASR that we had not managed to filter out. While perusing the three thousand undetermined sentences, we did not notice any written in languages not already mentioned.

7. Discussion and Conclusions

After the modifications during the described experiments, the general results on the internal test set of the development version of the HeLI-OTS remained at the same level. However, the identification accuracy on the dataset at hand was clearly improved.[27]

The following is a list of improvement ideas specific to the corpus at hand, which we noticed while inspecting the results of the language identification process. In addition to guiding us in preparing the next edition of the corpus, it functions as a general example of what kind of issues can be brought to light when fine-grained language identification is performed on this kind of corpora.

- Add "ed." to the list of known abbreviations after which the sentence should not be cut. More generally, any domain-specific text corpus can contain a disproportionate number of abbreviations not attested in a more general text corpus for the same language.

- Some of the parliamentary sessions are very long. At least one observed session (on the 20th of December 2011) lasted for more than 12 hours. However, the metadata for that session in Korp says it is 10 hours longer. This might be a systematic error when the metadata is created.

- In some cases, the metadata indicated that the utterance happened later than the end of the session, even though the metadata reflected the correct duration for the session.

- The encoding for common Scandinavian characters "ä" and "ö" was messed up in some of the sentences (less than 500). For example, "käy myöskin" had changed to: "kÃ€y myÃ¶skin".

- Add metadata indicating whether the utterances, sentences, and tokens were automatically generated by ASR during the text alignment process.

- Consider retaining manually transcribed accents and punctuation.

- Use language identifier with confidence thresholds.

- Add a Latinized version of Russian as one of the languages in order to detect further use of Russian.

The problems we encountered pertaining to the ASR-generated texts were similar in nature to the OCR problems we encountered with the NLF corpora, albeit less severe (Jauhiainen et al., 2022b). In both cases, language identification brought to light parts where OCR and ASR had been especially underperforming. With the Suomi24 corpus, we suggested leaving close Finnish-related languages out of the language repertoire when performing the language identification (Jauhiainen et al., 2022b). In this work, we were able to improve the quality of the Ingrian training corpus and use confidence thresholds to bring down the number of sentences that needed to be manually verified.

In this paper, we have demonstrated how a fine-grained language identification system can be used to find rare usage of foreign languages amongst a large number of sentences. We have also demonstrated how inspecting the language identification results with unexpected languages can bring forth problems in the corpus.

8. Acknowledgements

This project has received funding from the European Union – NextGenerationEU instrument and is funded by the Research Council of Finland under grant number 358720 (FIN-CLARIAH – Developing a Common RI for CLARIAH Finland).

We thank the anonymous reviewers, especially for pointing out the need to examine the sentences tagged as written with an undetermined language.

9. Bibliographical References

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstrations*, pages 33–36, Beijing, China. Coling 2010 Organizing Committee.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021.

[27] The newest version of HeLI-OTS including changes described in this article is available at `https://doi.org/10.5281/zenodo.10907468`.

Universal dependencies. *Computational linguistics*, 47(2):255–308.

Ute Dieckmann, Mietta Lennes, Jussi Piitulainen, Jyrki Niemi, Erik Axelson, Tommi Jauhiainen, and Krister Lindén. 2023. The pipeline for publishing resources in the Language Bank of Finland. In *CLARIN Annual Conference*, pages 33–43.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, et al. 2023. The ParlaMint corpora of parliamentary proceedings. *Language resources and evaluation*, 57(1):415–448.

Stephanie Evert and The CWB Development Team. 2022. *The IMS Open Corpus Workbench (CWB) Corpus Encoding and Management Manual*.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. 2014. Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, 48:493–531.

Marie-Ève Hudon. 2022. Official languages and parliament. *Ottawa, Canada: Library of Parliament.*

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2022a. HeLI-OTS, off-the-shelf language identifier for text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3912–3922, Marseille, France. European Language Resources Association.

Tommi Jauhiainen, Jussi Piitulainen, Erik Axelson, and Krister Linden. 2022b. Language diversity in the newspaper and periodical corpus of the National Library of Finland. Digital Research Data and Human Sciences (DRDHum) ; Conference date: 01-12-2022 Through 03-12-2022.

Tommi Jauhiainen, Jussi Piitulainen, Erik Axelson, and Krister Lindén. 2022c. Language identification as part of the text corpus creation pipeline at the Language Bank of Finland. In *The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022)*, pages 251–259, Uppsala, Sweden.

Baybars Kulebi, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. 2022. ParlamentParla: A speech corpus of Catalan parliamentary sessions. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 125–130, Marseille, France. European Language Resources Association.

Thomas Mueller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA. Association for Computational Linguistics.

Tommi A Pirinen. 2015. Omorfi — free and open source morphological lexical database for Finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 313–315, Vilnius, Lithuania. Linköping University Electronic Press, Sweden.

Teemu Ruokolainen, Pekka Kauppinen, Miikka Silfverberg, and Krister Lindén. 2020. A Finnish news corpus for named entity recognition. *Language Resources and Evaluation*, 54:247–272.

10. Language Resource References

The Parliament of Finland. 2017-01-01. *Plenary Sessions of the Parliament of Finland, Downloadable Version 1.5*. Kielipankki.

Exploring Word Formation Trends in Written, Spoken, Translated and Interpreted European Parliament Data – A Case Study on Initialisms in English and German

Katrin Menzel
Saarland University
Campus A2.2, 66123 Saarbrücken, Germany
k.menzel@mx.uni-saarland.de

Abstract

This paper demonstrates the research potential of a unique European Parliament dataset for register studies, contrastive linguistics, translation and interpreting studies. The dataset consists of parallel data for several European languages, including written source texts and their translations as well as spoken source texts and the transcripts of their simultaneously interpreted versions. The paper presents a cross-linguistic, corpus-based case study on a word formation phenomenon in these European Parliament data that are enriched with various linguistic annotations and metadata as well as with information-theoretic surprisal scores. The paper specifically addresses the questions of how initialisms are used across languages and production modes in the English and German corpus sections of these European Parliament data and whether there is a correlation between the use of initialisms and the use of their corresponding multiword full forms in the analysed corpus sections. The correlation analysis particularly addresses the question of whether initialisms in the analysed discourse types function as synonymous alternatives used in alternation with their full forms or primarily as replacements increasing compactness and lexical economy, but not necessarily transparency. Additionally, the paper explores what insights might be gained from an analysis of information-theoretic surprisal values with regard to the informativity and possible processing difficulties of initialisms. The results show that English written originals and German translations are the corpus sections with the highest frequencies of initialisms. The majority of cross-language transfer situations lead to fewer initialisms in the target texts than in the source texts, which means that they are either entirely omitted or other means are used to replace them in mediated discourse, e.g. hypernyms as less specific terms or multiword terms as semantically more explicit variants. In the English data, there is a positive correlation between the frequency of initialisms and the frequency of the respective full forms. There is a similar correlation in the German data, apart from the interpreted data. Additionally, the results show that initialisms represent peaks of information with regard to their surprisal values within their segments. Particularly the German data show higher surprisal values of initialisms in mediated language than in non-mediated discourse types, which indicates that in German mediated discourse, initialisms tend to be used in less conventionalised textual contexts than in English.

Keywords: European Parliament data, translation and interpreting data, corpus analysis

1. Introduction

1.1 Background and Motivation

This paper presents an example of the research potential of a unique European Parliament dataset consisting of parallel data for several European languages, including written source texts and their translations as well as spoken source texts and the transcripts of their simultaneously interpreted versions. The paper presents a cross-linguistic, corpus-based case study on a word formation phenomenon in these data that are enriched with various linguistic annotations and metadata as well as with information-theoretic surprisal scores calculated from the probabilities output of a 4-gram model trained for each language on an external domain-comparable resource. It thus applies language modelling to the recently published multilingual resource for the cross-lingual retrieval and analysis of selected word formation types within their contexts in the respective discourse types.

The case study presented in this paper compares English and German in EU parliamentary debate speeches. Furthermore, written and spoken mode are compared as the dataset includes edited, published records from the parliamentary debates and verbatim transcripts of the debates reflecting spoken language features such as repetitions, unfinished sentences, reformulations etc. Additionally, the non-mediated language of source texts is compared to mediated, i.e., translated and interpreted language. Here the focus is on initialisms as a particular type of word formation choices in European Parliament texts that are characterised by informative and persuasive messages and that need to be transferred to other languages.

Only a few case studies have discussed morphology and word formation in the context of contrastive research and translation studies (e.g. Cartoni and Lefer, 2011; Lefer, 2012; Defrancq and Rawoens, 2016; Berg, 2017). Ström Herold et al. (2021) specifically looked at initialisms in parallel data. Nevertheless, word formation remains an understudied area in corpus work on specialised registers and on translated and interpreted discourse. Moreover, there is still a research gap on initialisms in corpus linguistics, contrastive linguistics and translation and interpreting studies that this paper aims to address. The theoretical morphological literature has often treated initialisms as peripheral, marginal or extra-grammatical word formation patterns (cf. Menzel, forthcoming, for a literature summary). However, initialisms are a very interesting and unique strategy for shortening multiword terms to word-like units in a one-token format which gives them higher syntactic flexibility than their underlying full expressions. Initialisms have more complex functions and features than mere abbreviations, and they therefore deserve a much more prominent role in theoretical morphology and in corpus-based work.

The purpose of the analysis is to show how initialisms as a specific type of word formation and shortening strategy for multiword expressions (MWE) are used across languages and production modes in the English and German corpus sections of the selected datasets and whether there is a correlation between the use of initialisms and the use of their corresponding full forms in the analysed corpus sections. The correlation analysis particularly addresses the question of whether initialisms in the analysed discourse types function as synonymous alternatives used in alternation with their full forms or primarily as replacements increasing compactness and lexical economy, but not necessarily transparency. Additionally, the analysis presents insights on the informativity and on possible processing difficulties of initialisms gained from information-theoretic surprisal values. The data used for the surprisal calculations and for the corpus-linguistic analysis are the EuroParl_UdS[1] (Karakanta et al., 2018) and the EPIC-UdS corpora (Przybyl et al., 2022a/b, Menzel et al., forthcoming).

1.2 Initialisms

Initialisms can be defined as combinations of initial letters of multiword sequences of words functioning as shortened, more word-like forms of their spelt-out forms. Examples are the letter-by-letter initialism *EPA* in which each letter corresponds to a part of the full multiword expression ***E**conomic **P**artnership **A**greement* or the acronymic initialism *CITES* with a word-like pronunciation as a short form of ***C**onvention on **I**nternational **T**rade in **E**ndangered **S**pecies*. In a broader sense, initialisms also include shortenings of multimorphemic individual orthographic words that contain more than one meaningful part. By using this broader definition, we may include initialisms of multiword expressions that contain closed compounds (e.g. *EFSF* for ***E**uropäische **F**inanz**s**tabilisierungs**f**azilität*) that are often found in German where English typically prefers open compounds although shortening processes may lead to similar reduced forms in both languages (e.g. *EFSF* for ***E**uropean **F**inancial **S**tability **F**acility*). On the basis of this broader definition, we also include shortenings of expressions that contain individual words with combining forms whose initial letters are used in abbreviated forms as is often the case in technical and scientific concepts (e.g. *PCB* for ***p**oly**c**hlorinated **b**iphenyl* or *AIDS* for ***a**cquired **i**mmuno**d**eficiency **s**yndrome*).

Initialisms are productive in specialised registers such as political, administrative, military and business language. They function as insiders' code words giving shorter labels and an intended flavour of familiarity to concepts that already have multiword designations (Mattiello, 2013: 66). The shortened form is the result of the compression of a semantically equivalent multi-word denomination that refers to the same referent. Both the full and the short form continue to coexist as absolute synonyms, but their formal and stylistic features may make them suitable for different contexts.

Many initialisms in the register of EU parliamentary debates replace multiword proper nouns referring to institutions, groups, projects and policies that are important for the internal structure and the networks of the EU as the organisation in which the discourse takes place. The texts also contain initialisms for geographical entities and for technical and scientific concepts that play a role in the parliamentary debates.

2. Data

The written dataset EuroParl_UdS consists of parallel, sentence-aligned corpora for English, German and Spanish, and the source side contains texts only by native speakers of the respective languages. The corpus has been enriched with various metadata that were not available in previous European Parliament corpora. The EuroParl_UdS data are based on speeches adapted to the requirements of written language. They contain edited and published records of debates that took place in the European Parliament and they also contain their officially published translations. Like data from other parliamentary records such as the British Hansard (SAMUELS Consortium, 2015), they also include some written statements to the Parliament from parliamentary sessions.

The spoken dataset EPIC-UdS is also a multilingual parallel corpus of political debates from the European parliament for English, German and Spanish. Here, the release version V3 (Przybyl et al., 2022b) is used. Like in EuroParl_UdS, various metadata have been added to the EPIC-UdS texts (for instance, the speed of the speeches in words per minute and the topics of the texts). The EPIC-UdS data are unedited verbatim transcripts of what was said in parliamentary debates, and they also include simultaneous interpreting transcripts. For various written corpus texts, there are also the corresponding spoken ones in EPIC-UdS, but of course not for all of them as the spoken sections are smaller than the written ones. This paper focusses on the data from the German-English language pair and the respective corpus sections in the analysis (cf. Table 1).

[1] UdS stands for 'Universität des Saarlandes' (Saarland University)

	Corpus section	Tokens
English	EPIC-UdS EN orig. (spoken)	68.548
	EPIC-UdS EN interpr.	59.100
	EuroParl_UdS EN orig. (written)	8.693.135
	EuroParl_UdS EN transl.	6.260.869
German	EPIC-UdS DE orig. (spoken)	57.049
	EPIC-UdS DE interpr.	58.218
	EuroParl_UdS DE orig. (written)	7.869.289
	EuroParl_UdS DE transl.	3.100.647

Table 1: Corpus size of EuroParl_UdS and EPIC-UdS V3[2]

EuroParl_UdS and EPIC-UdS complement each other. Additionally, they complement other European Parliament datasets that contain translated or interpreted texts such as the EuroParl Simultaneous Interpreting Corpus (ESIC, Macháček et al., 2021), the Hungarian European Parliamentary Intermodal Corpus (HEPIC, Götz, 2020), the Polish Interpreting Corpus (PINC, Chmiel et al., 2022) and the EP-Poland Interpreting Corpus (Bartłomiejczyk et al., 2022). EPIC-UdS in particular builds on the experience of existing EPIC[3] parallel corpora developed at the University of Bologna (cf. Bendazzoli and Sandrelli, 2005; Russo et al., 2012, Bernardini et al., 2018) and EPICG at Ghent University (Defrancq et al., 2015) by using similar standards and transcription guidelines, and it extends them with the German-English language pair. There is also EPTIC (the European Parliament Translation and Interpreting Corpus), a bidirectional English-Italian corpus of interpreted and translated EU Parliament proceedings aligned to each other and to their corresponding source texts, i.e. the transcripts of the speeches and their edited and published written versions (Bernardini et al., 2016). The range of these corpora can be used to test hypotheses from translation studies in translated and / or simultaneous interpreted language. Some of these datasets have been used, for instance, to look at lexical and syntactic simplification processes, but the role of word formation patterns in parliamentary discourse and in translated or interpreted speech has not yet been a major research focus despite its potential significance in this context.

Table 2 contains example extracts from the spoken and written versions of a speech that illustrate the use of initialisms in the different corpus section types used for the analysis in this paper. In this table, we see that there are not many differences from the transcript of the live speech to the written and published version in the German example, only a grammatically correct form of the definitive article "der" replaces "des" before *EFSF* and "Einsatz" is used instead of "Nutzen" in this nominal group to use a more conventionalised context in front of the initialism. The examples in Table 2 illustrate what we might generally expect: nominal groups with initialisms sometimes become longer in translations via explicitation. Here, the full term for *EFSF* is added in the English translation before the initialism is introduced. In interpreted texts, nominal groups with initialisms remain short. Explicitation of initialisms is rare in interpreting, and these forms are used in contexts of more general vocabulary than in the other corpus sections (e.g. *"help from the EU"* in the English interpreted version vs. *"remedial measures from the EU"* in the translated version).

EPIC-UdS DE orig. (spoken)	EPIC-UdS EN interpr.
[…] erscheinen konzertierte Hilfsmaßnahmen von EU und IWF das Nutzen des EFSF unausweichlich zu werden	*[…] agreed help from the EU and the IMF the use of the EFSF seem to be unavoidable*
EuroParl_UdS DE orig. (written)	**EuroParl_UdS EN transl.**
[…] erscheinen konzertierte Hilfsmaßnahmen von EU und IWF und der Einsatz der EFSF unausweichlich zu werden.	*[…] concentrated remedial measures from the EU and the IMF and the use of the European Financial Stability Facility (EFSF) appear inescapable.*

Table 2: Example extracts from EPIC-UdS and EuroParl_UdS with initialisms

Tables 3 and 4 with longer extracts from the different versions of a parliamentary speech illustrate other examples of initialisms in the dataset that show that these forms are part of lexical chains and contribute to the network of cohesive ties in the texts.

EPIC-UdS DE orig. (spoken)	EPIC-UdS EN interpr.
*[…] und uns Gedanken machen wie dieses in dem Zusammenspiel mit dem **Europäischen Sozialfonds** möglicherweise noch effizienter gestaltet werden kann*	*we want to see how we can make this even more efficient together with the **ESF** as well*
*was die Finanzierungsquellen angeht haben Sie natürlich Recht was die Zahlungsermächtigung aus dem **ESF** angeht*	*you're quite right when it comes to payment appropriations from the **ESF***
*aber am Ende möchte ich schon dass das Gesamtspiel der Verpflichtung und der Zahlung sowohl für die **Strukturfonds** als auch für den **ESF** dann so ausgeht wie wir es in den Gesamtzahlen vereinbart haben*	*however what I would like to see is that the commitment appropriations and the payment appropriations should actually happen with the **European Structural Funds** as we've set out in the interinstitutional agreement*

Table 3: Example extracts from EPIC_UdS with an initialism (*ESF*) in lexical chains establishing cohesive links between textual elements

[2] EN = English, DE = German, orig. = original (source) texts, transl. = translations, interpr. = interpreted texts

[3] European Parliament Interpreting Corpus

EuroParl_UdS DE orig. (written)	EuroParl_UdS EN transl.
Wir müssen uns Gedanken machen, wie dies im Zusammenspiel mit dem ***Europäischen Sozialfonds (ESF)*** *möglicherweise noch effizienter gestaltet werden kann.*	*We need to contemplate how this interaction with the* ***European Social Fund (ESF)*** *could possibly be better shaped.*
Was die Finanzierungsquellen angeht, haben Sie, was die Zahlungsermächtigungen aus dem ***ESF*** *angeht, natürlich Recht.*	*As far as the sources of funding are concerned you were, of course, absolutely correct in what you said about the payment appropriations from the* ***ESF****.*
Aber am Ende möchte ich schon, dass das Gesamtspiel der Verpflichtungen und der Zahlungen sowohl für die ***Strukturfonds*** *als auch für den* ***ESF*** *dann so ausgeht, wie wir es in den Gesamtzahlen vereinbart haben.*	*Ultimately, however, I would like the overall picture for the obligations and the payments, both for the* ***structural funds*** *and for the* ***ESF****, to be as we agreed in the overall figures.*

Table 4: Example extracts from EuroParl_UdS with an initialism (*ESF*) in lexical chains establishing cohesive links between textual elements

The extracts in Tables 3 and 4 illustrate general differences between the written and spoken versions of the parliamentary debate speeches. In these extracts, the initialism *ESF* is used several times in lexical chains to create lexical cohesion between different segments via repetition, the use of the synonymous full form and other semantic relations such as hyponym-hypernym relations. At the beginning of the written and the translated versions of the speech in both languages in the EuroParl_UdS data, we have the first use of the initialism after the use of its full MWE, similarly to what we would find in many other formal written registers in both languages. In the transcribed spoken text in EPIC-UdS, only the full form is used by the speaker at the beginning as it is less common to give a pair of a short and long form of the same concept in spoken language. The listeners have to make the implicit connection between the full and the short form in this spoken text on their own. In the interpreted version in EPIC-UdS, only the initialism is used in the first segment as it is faster to pronounce than *European Social Funds,* and we may assume that the interpreter is familiar with the term to make the connection between the full term and its short form during the interpreting process. The interpreter also seems to expect the audience to understand what *ESF* stands for. However, the word formation choices of the interpreter lead to a different cohesive structure of the target text – the full term is not mentioned before the interpreter starts using the short form. Later, the hyponym *European Structural Funds* is used in the interpreted version (in fact, the *ESF* is one of the *European Structural and Investment Funds*). Understanding the network and chains of lexical relations in the interpreted version is more demanding for the audience than in the other text versions.

3. Analysis and Results

3.1 Query Design

The retrieval process for initialisms can be compared to developing annotation guidelines for a pattern that might seem rather fuzzy in the existing literature. It involved linguistic work of decisions with regard to relevant categories and subcategories. There are various phenomena that look similar on the surface, but decisions need to be taken to determine which ones have to be excluded due to their irrelevance. Some decisions need to be made in order to optimise precision and recall (e.g. to exclude forms which are theoretically possible and exist in various text types but are marginal for our dataset). As a first step, rather broad CQP queries (Corpus Query Processor, cf. Evert 2005) were used for words containing capital letters. Irrelevant forms were excluded via refined queries, e.g. abbreviations such as *EUR*, actual words spelt with capital letters such as *CARS* (an EU Action Plan on the car industry), forms that contain splinters from source words that fall under blends (e.g. *ALTENER* for *Alternative Energy Programme*) and mixed forms with only some letters as in initialisms (e.g. *REACH* that contains more than one letter from a source word [CH = Chemicals]). As they are marginal for this dataset due to spelling rules in EU style guides,[4] initialisms with small letters or periods (e.g. *aids* or *G.M.T.*) were excluded from the queries. They would occur more frequently in other text types or in older data. Hyphenated and open compounds that start with an initialism (e.g. *HACCP-based, EIB operations*) represent an interesting case from a cross-linguistic perspective due to different compounding strategies in English and German, but they will not be a particular focus of the analysis in this paper.

3.2 The Usage of Initialisms across Languages and Production Modes

One expectation for the analysis of initialisms across languages and production modes is to find that interpreters use many initialisms instead of multiword terms to save time. However, the spoken original data usually have more initialisms than the interpreted speeches. The German interpreted data have the lowest number of initialisms of all corpus sections (cf. Fig. 1). If we look at translated and interpreted data, we also have to take the influence of the respective source texts and the frequencies of initialisms in them into account. For instance, the differences between the English original spoken data and interpreted German are more pronounced than between the German original spoken data and the

[4] cf. for instance, *English style guide – A handbook for authors and translators in the European Commission*, [Latest PDF version: https://commission.europa.eu/document/download/c45f5b70-2d0e-4da7-b181-b5fe3a16c4bb_en]

interpreted English data. Thus, initialisms as word formation patterns are not just copied one to one as anglicisms in spoken language transfer from English to German.

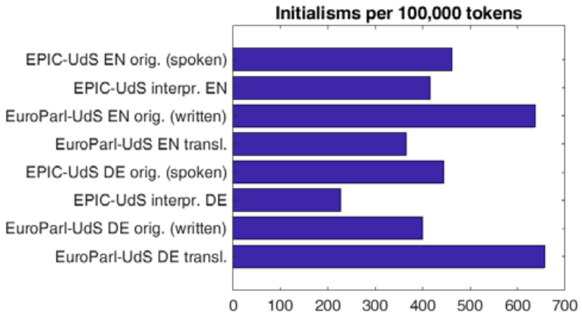


Figure 1: Usage of initialisms across languages and production modes

English written originals and German translations are the corpus sections with the highest frequencies among all. The German translations even have slightly more initialisms than the original English written texts, and German originals have a considerably lower frequency. In summary, we see the following trends in language transfer: English to German leads to fewer initialisms in spoken language transfer, but to more initialisms in written language transfer. In both spoken and written language transfer from German to English, fewer initialisms are found in the target texts than in their originals. Generally, that means that in most language transfer situations from the English-German language pair, and particularly in spoken language transfer, some initialisms from the source texts are either entirely omitted or other means are used to replace them in the target texts.

3.3 Frequency of Usage of Initialisms and their Corresponding Full Forms

This section addresses the question of whether there is a correlation between the use of initialisms and the use of their corresponding full forms in the analysed corpus sections. Measuring this correlation will reveal whether initialisms mainly function as synonyms to their MWE in this register, namely if the MWE have similar or higher frequencies than the initialisms themselves. From time to time, especially in debates on a variety of specialised topics, speakers may want to remind the audience of what an initialism as a potentially ambiguous form consisting of letters as submorphemic elements stands for. Additionally, short and long forms referring to the same concepts may be used in alternation in the texts to function like synonyms, as other types of synonyms for specialised multi-word terms or multi-word named entities are not necessarily available. However, if many initialisms are very conventionalised in this register, their full expressions may occur rarely or not at all in the texts. If this is mainly the case in the data, the most important function of the initialisms would be to give more efficient labels and an intended flavour of familiarity to specialised concepts that have lengthy multiword designations. Therefore, the more often conventionalised initialisms are used in the EU parliament discourse community, the less often the community might need to express their underlying full forms. A slightly positive correlation between frequently used initialisms and their MWE in both languages and all production modes can be expected.

Table 5 shows the correlation coefficients and the p-values in order to investigate the relationship between the normalised frequencies of initialisms and their corresponding MWE in the data. The 30 most frequent types of initialisms and their corresponding full forms in each corpus section were taken into account for this analysis. In contrast to all other forms found in the data, the initialism *"EU"* represents an extreme outlier. It is always used much more frequently than the second most frequent initialism in the respective datasets. It would have such a strong influence on the calculations and subsequent interpretation that it is excluded here in order to obtain more fine-grained insights on the other initialisms that are not characterised by such extreme values.[5]

	Corpus	**Correlation coefficient *r***	***p*-value**
English	EPIC-UdS EN orig. (spoken)	0.40	0.03
	EPIC-UdS EN interpr.	0.17	0.39
	EuroParl_UdS EN orig. (written)	0.88	4.83e-10
	EuroParl_UdS EN transl.	0.51	0.005
German	EPIC-UdS DE orig. (spoken)	0.05	0.76
	EPIC-UdS DE interpr.	0.40	0.03
	EuroParl_UdS DE orig. (written)	-0.0007	0.99
	EuroParl_UdS DE transl.	0.09	0.063

Table 5: Pearson correlation coefficients between normalised frequencies of most frequent initialisms and their corresponding MWE and significance level

[5] In all corpus sections, both the form "*EU"* and "*European Union"* were used with similar frequencies like synonyms (between 150 and 200 times per 100.000 tokens). Including these exceptions here would give us a correlation coefficient of almost 1 in all sections due to their high frequencies.

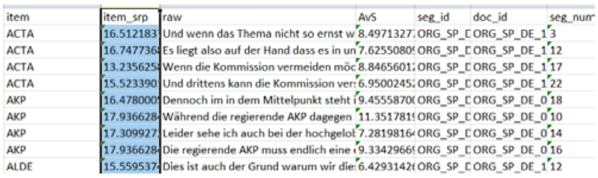
Table 5 shows that the English spoken and interpreted data have a slightly positive correlation for the frequency of initialisms and the frequency of the respective MWE, and the English written and translated data have a stronger positive correlation. There is not really any correlation to see in the German data, apart from a slightly positive one in the interpreted data. German uses some frequent multiword expressions in the original written and spoken data whose shortened forms are also among the top abbreviated forms in these data, but the usage of the full form is considerably more important than the usage of the initialism in some cases in German compared to English, while in other cases, the full form of a frequent initialism is not used at all or rarely in the German data. An obvious difference to English is that more initialisms in the German data originally represent foreign multiword expressions, but native equivalents for the full forms may exist as well. For instance, *"UN"* is used in German, but it is unusual to use the full English term in the German text. Additionally, the initialism has no visible link to the semantically equivalent German multiword term *"Vereinte Nationen"*. This may explain why in some cases neither the original full form nor an equivalent MWE is used frequently when a borrowed initialism has become conventionalised in the target language. Overall, the full expressions for frequently used initialisms seem to have become more unusual alternatives in German than in English.

3.4 Analysis of Surprisal Values

The data have been annotated with surprisal scores. Surprisal (S) has been calculated as the negative log (base 2) probability of each token (t) given its preceding context of three tokens measured in bits of information as in the following equation: $S(t_i) = -\log_2 p(t_i|(t_{i-1}\ t_{i-2}\ t_{i-3}))$. The values were calculated from the probabilities output by a KenLM 4-gram model, i.e. the model considers the three preceding words of each word to predict its surprisal. It was trained for each language on a domain-comparable resource. The data was balanced with regard to the size of the different corpus sections by discarding a number of random document pairs from the larger, written ones.

From an information-theoretic perspective, processing effort is related to surprisal that can be measured in bits (Hale, 2001; Degaetano-Ortlieb and Teich, 2022). For instance, the initialism *"CAP"* (Common Agricultural Policy) after the 3-token-sequence *"context of the"* is rather predicable with lower surprisal values in our data than *"CAP"* after a sequence such as "*be driven by*". The assumption here was that initialisms, apart from the extremely frequent example *"EU",* would represent peaks of information with regard to their surprisal values within their segments.

Surprisal scores for all initialisms regardless of their frequencies were identified in the data and the average surprisal scores of the respective text segments were extracted together with the text of the segments (Fig. 2).



item	item_srp	raw	AvS	seg_id	doc_id	seg_num
ACTA	16.512183	Und wenn das Thema nicht so ernst w	8.4971327	ORG_SP_D	ORG_SP_DE_1	3
ACTA	16.747736	Es liegt also auf der Hand dass es in un	7.6255080	ORG_SP_D	ORG_SP_DE_1	12
ACTA	13.235625	Wenn die Kommission vermeiden möc	8.8465601	ORG_SP_D	ORG_SP_DE_1	17
ACTA	15.523390	Und drittens kann die Kommission ver	6.9500245	ORG_SP_D	ORG_SP_DE_1	22
AKP	16.478000	Dennoch im in dem Mittelpunkt steht i	9.4555870	ORG_SP_D	ORG_SP_DE_0	18
AKP	17.936628	Während die regierende AKP dagegen	11.351781	ORG_SP_D	ORG_SP_DE_0	10
AKP	17.309927	Leider sehe ich auch bei der hochgelol	7.2819816	ORG_SP_D	ORG_SP_DE_0	14
AKP	17.936628	Die regierende AKP muss endlich eine	9.3342966	ORG_SP_D	ORG_SP_DE_0	16
ALDE	15.559537	Dies ist auch der Grund warum wir die	6.4293142	ORG_SP_D	ORG_SP_DE_1	12

Figure 2: Extract from table with extracted surprisal scores for initialisms (item_srp), the text segments (raw) and their average surprisal (AvS)

Figure 3 shows the range of the surprisal values of initialisms in the English data.[6] In the English translated data, surprisal is significantly higher than in the English written originals. We do not see the same difference between original spoken and interpreted discourse. Surprisal here is also generally higher in the spoken than in the written data.

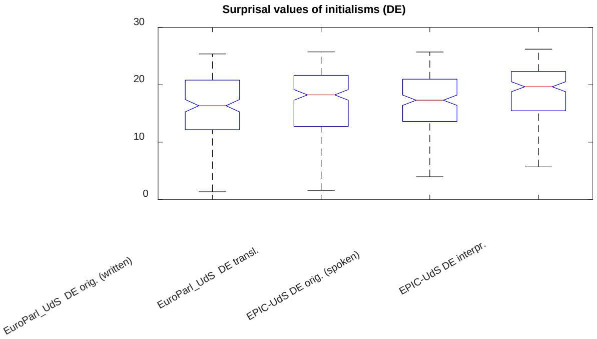


Figure 3: Surprisal values of initialisms in English

The German data in Figure 4 look slightly similar, but both types of mediated language production have significantly higher mean values than the respective non-mediated forms, which we can conclude from the plotted notches that represent the confidence interval around the median. This indicates that in German mediated discourse, be it written or spoken, initialisms are generally used in less conventionalised contexts than in original texts. Overall, the written and translated sections here in the German data turn out to be closer to the spoken and interpreted ones from the same language than in English.

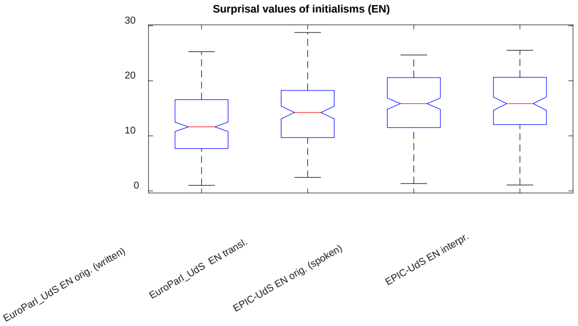


Figure 4: Surprisal values of initialisms in German

[6] Due to its exceptional frequency in all corpus sections, *"EU"* has again been excluded in this step.

The average surprisal of the entire segments in which initialisms occur is typically between 6 and 9 in all corpus sections (not plotted here). In most cases in both languages and all production modes, initialisms as condensed word-like forms of multiword terms indeed represent elements with high or very high surprisal compared to the average of their segments.[7] Interestingly, many examples do not have fixed sequences of part-of-speech patterns such as preposition + determiner before the initialism. There is generally a great variety of part-of-speech patterns in the preceding contexts, including content words such as verbs, nouns and adjectives. Generally, surprisal values for fixed elements in the full forms of the corresponding multiword expressions tend to be lower.

Initialisms achieve higher syntactic flexibility than MWE due to their one-word format. A qualitative analysis of initialisms in their contexts shows that untypical local context occur, for instance, if an initialism represents a MWE from a different language. Table 6 shows two examples of initialisms with very high surprisal values.

EPIC-UdS EN orig. (spoken): *And are we really happy that somebody who will be in charge of our overseas security policy was an activist a few years ago in an outfit like **CND***
EPIC-UdS DE interpr.: *Und sind wir wirklich glücklich darüber, dass jemand, der für unsere außenpolitische Sicherheit zuständig ist, vor ein paar Jahren aktiv war in **CND**.*

Table 6: Examples of initialisms with very high surprisal values (>20).

In the examples in Table 6, "*Campaign for Nuclear Disarmament"* is shortened, and already in the English original text, the value for "*CND"* was very high (20.45) due to an untypical context of the three preceding words, but in the German interpreted data, its value was one of the highest (25.64) as the form is not used in a great variety of contexts. From a cognitive perspective, reproducing a similar sequence of letters to produce a fluent target text might be less capacity-demanding in mediated discourse for the interpreters than replacing it with another structure. Nevertheless, an initialism like this might not be so common in the target language, and a different expression might normally be preferred by the target audience.

4. Conclusion and Outlook

To sum up, the case study presented in this paper has demonstrated the utility and a research context of the EuroParl_UdS and EPIC-UdS data that consist of written, spoken, translated and interpreted European Parliament texts for different languages. The case study on initialisms in English and German as a particular type of word formation and shortening strategy for MWE has shown differences and similarities between the languages and production modes in the data and provides valuable insights for the fields of register studies, contrastive linguistics, translation and interpreting studies. Some differences between the spoken and interpreted versions and the written and translated versions of parliamentary debate speeches may be due to the fact that the two former production modes directly address experts taking part in debates on specialised topics, while the two latter ones function as written documentation like reports. They address a larger, more heterogeneous audience of people including all those who did not take place in the actual debate. This explains some of the choices in the written texts, e.g. to restructure the elements and types of semantic relations in lexical chains in a different way than in the spoken texts or not to start right away with an initialism without mentioning the full form. Other strategies with regard to fixed multiword expressions and less explicit initialisms consisting of submorphemic elements reflect general mediated language effects and some are specific to interpreting due to high time pressure and cognitive effort in this language transfer task. In the annotated data, all segments have been extracted that contain no initialism, but the aligned source or target segment does contain one. Therefore, in a future analysis, it would also be useful to focus on specific contexts where initialisms were omitted or added in the translated or interpreted speeches and to analyse the types of translation/interpreting procedures in more detail. Generally, we can expect to see an overall trend towards explicitation in written translations (e.g. EuroParl_UdS DE orig. [written]: *das SIS* -> EuroParl_UdS EN transl.: *the Schengen Information System*) and the usage of less specific vocabulary, i.e. fewer initialisms and fewer multiword terms, in interpreting (e.g. EPIC-UdS DE orig. [spoken]: *das SWIFT-Abkommen* -> EPIC-UdS EN interpr.: *the agreement*).

One could further look into the subtypes of the initialisms, considering, for instance, their length, whether they have to be pronounced as one-word acronyms or letter by letter, what type of MWE they stand for (e.g. technical term or named-entity, foreign or native origin) and whether they are used as the head of a nominal group or as a premodifier of another noun as in that case they often cannot easily be replaced by the full form. A larger size of the spoken original and interpreted data would be useful for this type of analysis. Additionally, one might control for specific metadata when comparing word formation choices in the different production modes. What makes this challenging is that some types of metadata are not available although they would be relevant for particular questions (e.g. specific background information on the translators and interpreters). Other metadata types are not available for all types of production modes or difficult to use for specific studies in their current form. EPIC-UdS, for instance, contains information on the general topics and the titles of the debate as indicated by the European Parliament. However, the

[7] The outlier *"EU"* has low to medium surprisal values.

debates represent a huge variety of topics that are rather difficult to assign to overarching clear-cut categories (e.g. a debate on the beekeeping sector has been assigned the topic of "Economy", the situation in the Middle East/Gaza Strip falls under "International affairs", the democratic process in Turkey under "Politics" and "Food distribution to the most deprived persons in the Community (amendment of the Single CMO Regulation)" has been labelled with "Health". Enlarging the spoken part and further enriching and enhancing the metadata would therefore be an opportunity to facilitate follow-up studies.

5. Acknowledgments

This work is based on research funded by the Deutsche Forschungsgemeinschaft (German Research Foundation) – SFB 1102 / Project-ID 232722074, project B7 (Translation as Rational Communication). Additionally, the author is greatly indebted to Maria Kunilovskaya who calculated the surprisal scores and extracted them with the respective corpus segments.

6. Bibliographical References

Berg, T. (2017). Compounding in German and English – A quantitative translation study. In *Languages in contrast*, 17(1): 43–68.

Cartoni, B. and M. Lefer (2011). Negation and lexical morphology across languages: Insights from a trilingual translation corpus. In *Poznan studies in contemporary linguistics*, 47(4): 795–843.

Defrancq, B., K. Plevoets and C. Magnifico (2015). Connective items in interpreting and translation: Where do they come from? In J. Romero-Trillo (ed.) : *Yearbook of corpus linguistics and pragmatics 2015: Current approaches to discourse and translation studies.* Cham: Springer International Publishing, 195–222.

Defrancq, B. and G. Rawoens (2016). Assessing morphologically motivated transfer in parallel corpora. In *Target*, 28(3): 372–398.

Degaetano-Ortlieb, S. and Teich, E. (2022). Toward an optimal code for communication: The case of scientific English. In *Corpus Linguistics and Linguistic Theory*, 18(1): 175–207.

Evert, S. (2005). *The CQP query language tutorial.* IMS: Stuttgart University.

Götz, A. (2020). Discourse markers and connectives in interpreted Hungarian discourse: A corpus-based investigation of discourse properties and their interdependence. In *Speech Science* 2020(1): 259–284.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, 1–8, Pittsburgh, Pennsylvania, June 2001, ACL.

Lefer, M. (2012). Word formation in translated language – The impact of language-pair specific features and genre variation. In *Across languages and cultures,* 13(2): 145–172.

Mattiello, E. (2013). *Extra-grammatical morphology in English. Abbreviations, blends, reduplicatives and related phenomena.* Berlin: Mouton de Gruyter.

Menzel, K. (forthcoming). Initialisms in scientific writing in the 19th and early 20th centuries. *Zeitschrift für Wortbildung,* 2/24.

Przybyl, H., A. Karakanta, K. Menzel and E. Teich, (2022a). Exploring linguistic variation in mediated discourse: Translation vs. interpreting. In M. Kajzer-Wietrzny, A. Ferraresi, I. Ivaska and S. Bernardini (eds.): *Mediated discourse at the European Parliament: Empirical investigations.* Berlin: Language Science Press, 191–218.

Shannon, C. E. (1948): A mathematical theory of communication. *Bell Systems Technical Journal* 27: 379–423.

Ström Herold, J., M. Levin and J. Tyrkkö (2021). RAF, DNA and CAPTCHA: English acronyms in German and Swedish translation. *Bergen Language and Linguistics Studies,* 11(1): 163–184.

7. Language Resource References

Bartłomiejczyk, M., E. Gumul and D. Koržinek (2022). EP-Poland: Building a bilingual parallel corpus for interpreting research*.* In *GEMA, Online Journal of Language Studies,* 22(1): 110–126.

Bendazzoli, C. and A. Sandrelli (2005). An approach to corpus-based interpreting studies: Developing EPIC (European Parliament Interpreting Corpus). In H. Gerzymisch-Arbogast and S. Nauert (eds.): *Mutra2005 – Challenges of Multidimensional Translation. Proceedings of the Marie Curie Euroconferences.* Saarbrücken. May 2005. https://www.euroconferences.info/proceedings/2005_Proceedings/2005_proceedings.html

Bernardini, S., A. Ferraresi and M. Miličević (2016). From EPIC to EPTIC – Exploring simplification in interpreting and translation from an Intermodal perspective. In *Target,* 28: 61–86.

Bernardini, Silvia, A. Ferraresi, M. Russo, C. Collard and B. Defrancq (2018). Building interpreting and intermodal corpora: A how-to for a formidable task. In M. Russo, C. Bendazzoli and B. Defrancq (eds.): *Making way in corpus-based interpreting studies.* Singapore: Springer Nature, 21–42.

Chmiel, Agnieszka, D. Koržinek, M. Kajzer-Wietrzny, P. Janikowski, D. Jakubowski and D. Polakowska (2022): Fluency parameters in the Polish Interpreting Corpus (PINC). In M. Kajzer-Wietrzny, A. Ferraresi, I. Ivaska and S. Bernardini (eds.): *Mediated discourse at the European Parliament: Empirical Investigations.* Berlin: Language Science Press, 63–91.

Heafiel, K. (n.d.). KenLM toolkit https://kheafield.com/code/kenlm/.

Karakanta, A., M. Vela and E. Teich (2018). Europarl-UdS: Preserving metadata from parliamentary debates. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018).* Miyazaki, May 2018.

Resource available at: https://fedora.clarin-d.uni-saarland.de/europarl-uds/

Macháček, D., M. Žilinec and O. Bojar (2021). *ESIC 1.0 - Europarl Simultaneous Interpreting Corpus,* LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-3719.

Menzel, K., H. Przybyl and E. Lapshinova-Koltunski (forthcoming). EPIC-UdS – ein mehrsprachiges Korpus als Grundlage für die korpusbasierte Dolmetsch- und Übersetzungswissenschaft. In *Proceedings of the 4th TRANSLATA Conference, 2021*, Innsbruck.

Przybyl, H., E. Lapshinova-Koltunski, K. Menzel, S. Fischer and E. Teich (2022b). EPIC UdS – Creation and applications of a simultaneous interpreting corpus. In *Proceedings of 13th Conference on Language Resources and Evaluation (LREC 2022)*, pp. 1193–1200, Marseille, June 2022. Resource available at: https://fedora.clarin-d.uni-saarland.de/epic-uds/index.html

Russo, M., C. Bendazzoli, A. Sandrelli and N. Spinolo (2012). The European Parliament Interpreting Corpus (EPIC): Implementation and developments. In F. Straniero Sergio and C. Falbo (eds.): *Breaking ground in corpus-based interpreting studies.* Bern: Peter Lang, 53–90.

SAMUELS Consortium (2015). Hansard corpus. SAMUELS Project, available via https://www.english-corpora.org/hansard/

Quantitative Analysis of Editing in Transcription Process in Japanese and European Parliaments and its Diachronic Changes

Tatsuya Kawahara
School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501 Japan
kawahara@i.kyoto-u.ac.jp

Abstract

In making official transcripts for meeting records in Parliament, some edits are made from faithful transcripts of utterances for linguistic correction and formality. Classification of these edits is provided in this paper, and quantitative analysis is conducted for Japanese and European Parliamentary meetings by comparing the faithful transcripts of audio recordings against the official meeting records. Different trends are observed between the two Parliaments due to the nature of the language used and the meeting style. Moreover, its diachronic changes in the Japanese transcripts are presented, showing a significant decrease in the edits over the past decades. It was found that a majority of edits in the Japanese Parliament (Diet) simply remove fillers and redundant words, keeping the transcripts as verbatim as possible. This property is useful for the evaluation of the automatic speech transcription system, which was developed by us and has been used in the Japanese Parliament.

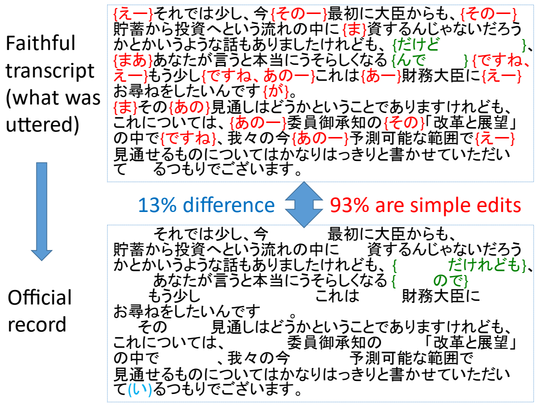
Keywords: Parliamentary record, Japanese Diet, European Parliament

1. Background

Transcription is a process of converting speech into text, and there are two goals: accuracy, or faithfulness to speech, and readability, or easiness of reading. They are often in trade-off relationships. Thus, standards or guidelines on transcription, including editing, have been strictly designed and enforced in Parliamentary reports compared with private sectors. One of the well-known is the Hansard of the British Parliament (Mollin 2007). They are, however, different across languages and countries and also change over time. They may be affected by other factors such as TV broadcasting, SNS, and the use of automatic speech recognition (ASR) technology. In this study, a quantitative analysis of the editing process is conducted for Japanese and European Parliamentary meetings by comparing the faithful transcripts of audio recordings and the official transcript records.

2. Edits in Transcription Process

An example of a faithful transcript and an official record is shown in Figure 1. There are many factors requiring edits in the transcription process. First of all, disfluency must be removed. Other kinds of redundancy need to be removed. Then, grammatical errors must be corrected. Some colloquial expressions should be rephrased into formal expressions. Last but not least, speech does not have explicit punctuations, unlike text, so it is necessary to insert periods and commas in appropriate places. In addition to these edits, some structural modifications are sometimes made to improve readability. Moreover, some semantic corrections are made for apparent mistakes. These are explained one by one.



{えー}それでは少し、今{そのー}最初に大臣からも、{そのー}
貯蓄から投資へという流れの中に{ま}資するんじゃないだろう
かとかいうような話もありましたけれども、{だけど　　　　　}、
{まあ}あなたが言うと本当にうそらしくなる{んで　　　}{ですね、
えー}もう少し{ですね、あのー}これは{あー}財務大臣に{えー}
お尋ねをしたいんです{が}。
{ま}その{あの}見通しはどうかということでありますけれども、
これについては、{あのー}委員御承知の{その}「改革と展望」
の中で{ですね}、我々の今{あのー}予測可能な範囲で{えー}
見通せるものについてはかなりはっきりと書かせていただい
て　　るつもりでございます。

　　　それでは少し、今　　　　　最初に大臣からも、
貯蓄から投資へという流れの中に　　資するんじゃないだろう
かとかいうような話もありましたけれども、{　　　だけれども}、
　　　あなたが言うと本当にうそらしくなる{　　　ので}
　　　もう少し　　　　　　　　　これは　　　財務大臣に
お尋ねをしたいんです　　。
　　その　　　見通しはどうかということでありますけれども、
これについては、　　　　委員御承知の　　　「改革と展望」
の中で　　　　、我々の今　　　　予測可能な範囲で
見通せるものについてはかなりはっきりと書かせていただい
て(い)るつもりでございます。

Figure 1: Example of a faithful transcript and an official record in the Japanese Diet

2.1 Removal of Redundancy

Fillers, such as "um" and "ahh" in English, must be definitely removed. They are not transcribed by human stenographers in the first place. They can also be automatically eliminated by ASR systems. Repeats and repairs must also be removed, but their automatic removal is difficult.

Discourse markers, such as "OK" and "yes" in English, can be kept, but too many tokens reduce readability. Other extraneous expressions, such as "Thank you," can also be kept, but the removal of them would improve readability.

2.2 Correction of Errors and Colloquial Expressions

There are some kinds of grammatical errors whose correction is mandatory, for example, missing or incorrect articles such as "a" and "an," and improper use of prepositions such as "in" and "on." Some kinds of colloquial expressions should also be corrected; for example, "was like" changed to "said" and "but" changed to "however." But we note language use changes over time. Handling of dialect is also an issue. While some dialects cannot be understood by many readers, dialect is often used to express the identity of the speaker.

2.3 Structural and Semantic Corrections

Some structural reordering is conducted; for example, “Finnish incoming presidency” is changed to “incoming Finnish presidency.” It is often necessary to split a long sentence into a sequence of plain sentences.

On the other hand, semantic correction needs attention. While apparent errors such as mistakes of “billion” and “million” should be corrected via a proper process, it is a question whether errors of proper names or fact errors should be corrected because MPs should be responsible for their statements. Especially when the errors affected the following interaction in the meeting, they should not be corrected.

3. Corpus Analysis in European Parliament and Japanese Diet

3.1 Used Corpora

A corpus-based analysis was conducted using transcripts from the European Parliament (Koehn 2005) and the Japanese Diet (the House of Representatives) (Akita 2006). From the European Parliament proceeding, English-speaking parts in some plenary sessions in 2007 were selected. With regard to the Japanese Diet, a number of sessions in committee meetings held during 2005-2007 were selected. They were selected to cover all major meetings in a good balance. In addition to the official proceeding text, faithful transcripts of spoken words, including fillers and disfluencies, were manually prepared for the analysis. In fact, these faithful transcripts were prepared for the development of the ASR system. The general statistics of the two corpora are shown in Table 1.

The overall edit distance in words between the faithful transcript and the official record is approximately 13% in the Japanese Diet, while it is over 20% in the European Parliament. The larger difference in the European Parliament is attributed to grammatical strictness in the English language compared to Japanese. For example, a subject and prepositions are often omitted in Japanese, while they cannot be omitted in English. There are also many non-native English speakers in the European Parliament.

3.2 Analysis of Edit Categories

Table 2 lists the statistics of edit categories described in the previous section. A large majority (93%) of edits in the Japanese Diet are simple and can be classified as deletion, insertion, or substitution (correction) of words. Almost 90% of them are deletions, and almost half of them involve filler removal. On the other hand, there are much more complex corrections in the European Parliament because English needs many grammatical corrections and syntax reordering. Thus, there is a different tendency according to the language.

Here are typical edit patterns observed for English in the European Parliament. The most frequently removed words other than fillers are “thank you,” “I think,” and “also,” while the most frequently inserted ancillary and functional words are “the,” “that,” “a,” “also,” and “and.” The most frequently corrected patterns are “but → however,” “thank you → Mr.,” “would → should,” “our → the,” and “this → that.”

Table 1 General statistics of corpora

	European Parliament	Japanese Diet
#words (faithful transcript)	30.9K	418K
#words (official record)	27.1K	379K
% of edited words	20.5%	12.9%

Table 2 Statistics of edit categories

		European Parliament	Japanese Diet
Remove	Fillers	11.6%	46.7%
	Repeats/repairs	11.0%	9.4%
	Discourse markers	1.8%	18.4%
	Extraneous expressions	16.8%	3.0%
Correct	Grammatical errors	20.1%	7.5%
	Colloquial expressions	18.0%	8.4%
Reorder		19.6%	5.9%

3.3 Analysis per Meeting Category and Diachronic Changes

The occurrence ratio of edits per committee in the Japanese Diet is shown in Figure 2. There was a tendency in 2007 that fewer edits were made in the Commission on the Constitution, the Committee on Budget, and the Question Time. While one-on-one interaction is a norm in other committees, the Commission on the Constitution adopts the style of free discussions by all members. This style affects the transcription process. The Committee on Budget and the Question Time are usually broadcasted on the national TV channel, and this may affect the editing process.

In Figure 2, we can also observe a significant change from 2007 to 2016. The ratio of edits has been reduced by 40% over the ten years.

4. Discussions

There are several causes for the reduction of edits. Most significantly, phrase reordering is not done anymore. Some discourse markers are now kept, and some repeats are allowed, such as those expressing emphasis. Moreover, many colloquial expressions are getting accepted. These suggest that the transcripts become more verbatim than before.

There are some possible reasons for this trend. First is the deployment of the ASR system since 2011. Reporters now edit a faithful transcript, which is

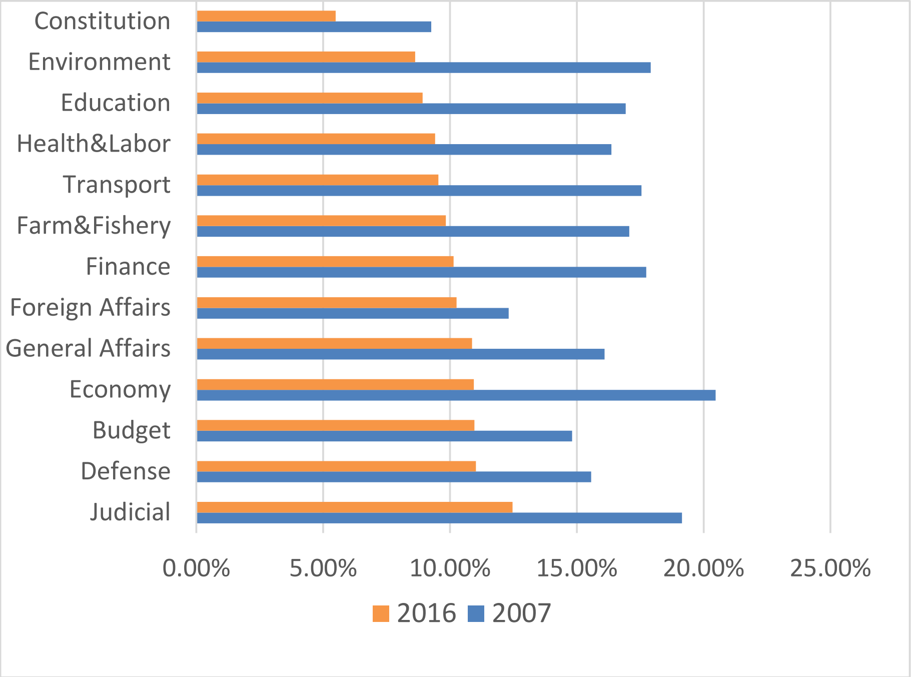


Figure 2: Statistics of edits per meeting category in the Japanese Diet in 2007 and 2016

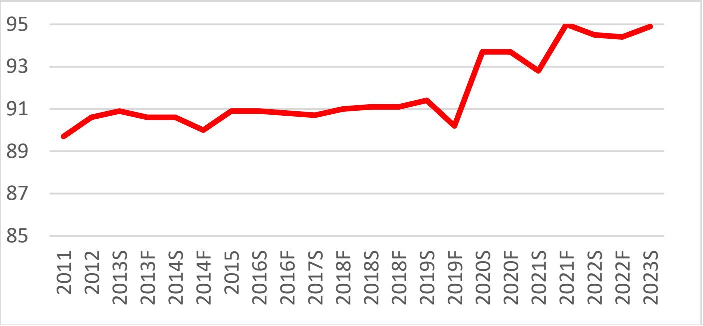


Figure 3: Character correct rate (%) of automatic speech recognition for Japanese Diet

generated by the ASR system. In the old system based on stenography, they typed in text with editing in mind. The second factor is Internet broadcasting. All meetings are broadcast via the Internet. They are also archived and can be accessed at any time; thus, they can be referred to on social media. With these factors, the guidelines by editors may have been changed, although there is no written guideline in the Japanese Diet.

Besides these factors, it is pointed out that there has been a global trend toward writing more verbally or in a closer way to speech in recent decades, even in parliaments (Korhonen 2023).

5. Evaluation of Automatic Transcription System

Since 2011, the House of Representatives of Japan has adopted the ASR system, which was developed by the author's lab (Kawahara 2012). The acoustic model was trained with 1000 hours of parliamentary speech, and the language model was trained with 25 years of meeting records.

It is found that a large majority of edits for the transcripts of the Japanese Diet are the removal of fillers and discourse markers. This property makes it easy to automatically evaluate the performance of the ASR system without preparing the faithful transcripts. The word/character correct rate is defined by the edit distance minus insertion errors, which counts only substitution and deletion errors. Notice again that a majority of insertions in automatic transcription are due to fillers and redundant words, which must be omitted in the final transcript. The effect of other kinds of edits is smaller than 1%.

The character correct rate for each session/year is plotted in Figure 3. It had been steady at around 91% before the ASR system was improved by adopting a deep learning model in 2020, which significantly improved the accuracy to 95%.

6. Conclusions and Future Perspectives

As the ASR system shows very useful performance, the next step will be to automate the post-editing process. This study has been conducted before with many approaches (Charniak 2001, Honal 2003, Hori 2003, Maskey 2006, Neubig 2012, Shitaoka 2004, Yeh 2006), but it has never met the satisfactory level required by the Parliament.

However, recent large language models such as GPT-4 show the functionality of cleaning transcripts either by human or ASR systems. It is a time to revisit the problem.

7. Acknowledgments

This study is supported by JST A-STEP JPMJTR23RD.

8. Bibliographical References

Mollin, S. (2007). The Hansard hazard: Gauging the accuracy of British parliamentary transcripts. Corpora 2(2): 187--210.

Akita, Y., and Kawahara, T., (2010). Statistical transformation of language and pronunciation models for spontaneous speech recognition. IEEE Transactions on Audio, Speech, and Language Processing 18 (6), 1539–1549.

Charniak, E. and Johnson, M. (2001). Edit detection and parsing for transcribed speech. Proceedings of NAACL.

Honal, M. and Schultz, T., (2003). Correction of disfluencies in spontaneous speech using a noisy-channel approach. Prof. EuroSpeech, pp. 2781–2784.

Hori, T., Willett, D., and Minami, Y., (2003). Paraphrasing spontaneous speech using weighted finite-state transducers. In: ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition.

Kawahara. T. (2012). Transcription system using automatic speech recognition for the Japanese Parliament (Diet). In Proc. AAAI/IAAI, pp.2224—2228.

Koehn, P. (2005): Europarl: A Parallel Corpus for Statistical Machine Translation, Proc. MT Summit, pp. 79–86.

Maskey, S., Zhou, B., and Gao, Y. (2006). A phrase-level machine translation approach for disfluency detection using weighted finite state transducers. Proceedings of InterSpeech, pp. 749–752.

Neubig, G., Akita, Y., Mori, S., and Kawahara T. (2012). A monotonic statistical machine translation approach to speaking style transformation. Computer Speech and Language, Vol.26, No.5, pp.349--370, 2012.

Shitaoka, K., Nanjo, H., and Kawahara, T. (2004). Automatic transformation of lecture transcription into document style using statistical framework. InterSpeech, pp. 2169–2172.

Yeh, J.-F. and Wu, C.-H. (2006). Edit disfluency detection and correction using a cleanup language model and an alignment model. IEEE Transactions on Audio, Speech, and Language Processing 14 (5), 1574–1583.

Korhonen, T., Kotze H., and Tyrkkö J. eds. (2023). Exploring language and society with big data: Parliamentary discourse across time and space. John Benjamins.

9. Language Resource References

Koehn, P. (2005): European Parliament Proceedings Parallel Corpus 1996-2011 (https://www.statmt.org/europarl/)

Proceedings of Japanese Diet. 国会会議録検索システム (https://kokkai.ndl.go.jp/#/)

Automated Emotion Annotation of Finnish Parliamentary Speeches Using GPT-4

**Otto Tarkka, Jaakko Koljonen, Markus Korhonen,
Juuso Laine, Kristian Martiskainen, Kimmo Elo, Veronika Laippala**
University of Turku
20014 University of Turku, Finland
ohitar@utu.fi

Abstract

Annotating datasets can often be prohibitively expensive and laborious. Emotion annotation specifically has been shown to be a difficult task in which even trained annotators rarely reach high agreement. With the introduction of ChatGPT, GPT-4 and other Large Language Models (LLMs), however, a new line of research has emerged that explores the possibilities of automated data annotation. In this paper, we apply GPT-4 to the task of annotating a dataset, which is subsequently used to train a BERT model for emotion analysis of Finnish parliamentary speeches. In our experiment, GPT-4 performs on par with trained annotators and the annotations it produces can be used to train a classifier that reaches micro F1 of 0.690. We compare this model to two other models that are trained on machine translated datasets and find that the model trained on GPT-4 annotated data outperforms them. Our paper offers new insight into the possibilities that LLMs have to offer for the analysis of parliamentary corpora.

Keywords: emotion analysis, parliamentary speeches, annotation, chatgpt

1. Introduction

Recent years have shown growing interest in the studies of sentiments and emotions in politics (see e.g., Fraccaroli et al. 2022; Orellana and Bisgin 2023). In the Finnish context, however, this is still an underdeveloped field of study. Koljonen et al. (2022) analyze emotion in post-WWII party manifestos, but analysis on modern plenary speeches in Finland is a largely unexplored territory. While sentiment analysis typically aims to categorise texts into two or three categories (positive, negative + neutral), emotion analysis aims at a more fine-grained classification, where texts are divided into emotion categories based on the emotion(s) they reflect. Sentiment and emotion classification have traditionally been done with dictionary based methods but they have given way to deep-learning approaches, which have shown greater classification accuracy (Widmann and Wich, 2023; Borst et al., 2023).

The downside of deep-learning is that it requires annotated training data. Data annotation is notoriously laborious, time consuming and often expensive. Crowd-sourcing platforms, such as Amazon's Mechanical Turk (MTurk) can be used to cut costs but there has been growing concern over both the quality of annotations and the ethical questions that using MTurk raises (e.g. Chmielewski and Kucker, 2020; Shmueli et al., 2021). When a ready-made dataset in the desired language is not available, and there are not sufficient resources to build a dataset from the ground up, there are few options available for researchers. One option is to machine translate an existing dataset to the desired language (Eskelinen et al., 2023). Very recently, a new option has emerged, which is leveraging ChatGPT or other similar *large language models* (LLMs) to do the previously laborious and costly annotation quickly and relatively cheaply.

This paper explores the possibilities of using GPT-4 to annotate a dataset that is used to train a BERT-based classifier for analysing emotion in Finnish parliamentary speeches. We create and evaluate an emotion annotated dataset and show that a BERT model trained on this data outperforms models trained on machine translated datasets. Our results show that GPT-4 is a promising tool for creating datasets for emotion analysis in parliamentary speeches. All training scripts and annotated data are available on GitHub.[1]

2. Background

GPT is a family of LLMs trained on massive natural language datasets that continue a given prompt with words that have the statistically best fit (Floridi and Chiriatti, 2020). ChatGPT has been further trained with conversational data to produce coherent responses to questions and to follow instructions. The version of ChatGPT that is most commonly used is also known as GPT-3.5. GPT-4 is an even larger and more capable multimodal model that performs well even in many academic and professional exams (OpenAI et al., 2023).

BERT is a language representation model which was first introduced in 2018 and outperformed state-of-the-art models in several *Natural Language Pro-*

[1] https://github.com/TurkuNLP/FinParl-emotion

cessing (NLP) tasks (Devlin et al., 2018). It is still today the standard in many NLP tasks. There are often two stages in the BERT algorithm workflow: first, pre-training which uses masked language modelling and next sentence prediction, and second, fine-tuning (Rogers et al., 2020). We use FinBERT (Virtanen et al., 2019) as our base model, which we fine-tune with data annotated by GPT-4.

Using ChatGPT for automating the annotation process is not a wholly original idea. In earlier research, ChatGPT has been used successfully in annotation tasks. For example, Gilardi et al. (2023) compare annotations between trained annotators and ChatGPT. They show that ChatGPT outperforms both crowd workers and trained annotators in a number of tasks with regard to inter annotator agreement. Malik et al. (2024), also, use ChatGPT to create annotated data to train a multi-label emotion classification model. Their model trained with data annotated by ChatGPT achieves satisfactory performance when using 8 emotion categories to classify emotions in tweets.

There is no one set of emotions that is universally used in emotion analysis, and, instead, papers in the field use a wide set of emotions. Bostan and Klinger (2018) compile and compare 14 datasets built for emotion classification using 12 different annotation schemes. Many papers use either Ekman's six basic emotions (Ekman, 1992) or Plutchik's wheel of emotions (Plutchik, 1982) as the basis of their set of emotions but often modify the taxonomy somewhat to suit the needs of the study. Others use a whole different set of emotions, such as the GoEmotions dataset, which employs a 28 category taxonomy (Demszky et al., 2020). The only pre-existing Finnish resources for emotion annotation that we are aware of are the XED corpus, which contains sentence-level multi-label emotion annotations for movie subtitles (Öhman et al., 2020) and the emotion lexicon SELF (Öhman, 2022).

3. Data

ParlamenttiSampo (Semantic Computing Research Group, 2021) contains the transcribed records of all plenary sessions of the Finnish Parliament (*Eduskunta*). To create our own dataset of emotion annotated plenary debates, we handpicked a number of plenary sessions discussing the reports of the Parliamentary Committees of the Finnish Parliament between the years 2017 and 2020. The 17 permanent Committees play an influential role in the decision-making in the Parliament. The Committees prepare e.g., legislative initiatives, government bills and reports for handling in plenary sessions. MPs are divided to the Committees proportionally in a way that reflects the strength of each party in the Parliament.

Each Committee works within their own field of expertise within the scope of a corresponding ministry. Thus, by choosing speeches from different committee reports, we assure that the speeches in our training data cover a variety of topics, terms and perspectives, which might evoke different emotional responses from MPs. This leads to a more representative dataset as parties tend to be more active in policy areas that are important to the party's key voter clientele (Bäck and Debus, 2016). We choose plenary debates from two different parliamentary terms to combat any bias caused by the changing dynamics between parties within parliamentary terms. Opposition politicians are inclined to have a greater incentive to persuade voters and reclaim their position as a credible alternative to become the governing party (Russell et al., 2017). In a competitive parliamentary system, opposition politicians tend to criticise government policies and, thus, their status of as an opposition MP is likely to affect their behaviour and rhetoric (Tuttnauer, 2018).

Our final data comes from 15 Committee reports consisting of 529 speeches, which were split into 6025 sentences using the Python NLP toolkit Trankit (Nguyen et al., 2021). We use the sentence as the unit of observation.

4. Methods

The steps we took in the creation of our dataset and model are as follows: First, we manually annotated a small set of sentences from parliament that act as the gold standard against which all evaluation is done. Then, we used GPT-4 to annotate the same set and evaluate its performance. Over multiple iterations and prompt engineering we reached results that are comparable to human performance. We then used GPT-4 to annotate a larger set of sentences using the same prompt. This data was then used as training data for a BERT model.

ID	Emotion	*N*	%
0	neutral	153	53
1	happiness/success	17	6
2	hopefulness/optimism/trust	33	11
3	love/praise	37	13
4	surprise (positive)	3	1
5	sadness/disappointment	3	1
6	fear/concern/mistrust	17	6
7	hate/disgust/derision	21	7
8	astonishment (negative)	6	2

Table 1: Emotion categories and gold standard labels.

4.1. Data Annotation

251 sentences from three plenary debates were manually annotated. Initially, we planned to use Ekman's six basic emotions to categorise the sentences but testing showed that the annotators struggled to assign sentences to these categories consistently. Hence, we chose to create our own set of emotions based on the emotions that we observed in the data. After further test rounds and discussion, we decided on a final set of 8 emotions + neutral (see Table 1). The final evaluation data, which we refer to as the gold standard, was annotated by four expert annotators (ann1-ann4). The annotators were native speakers of Finnish and all were familiar with the practices and typical rhetoric of the Finnish Parliament. The emotion label of a sentence was chosen by majority vote. If a sentence did not have a single winning label, all winning labels were accepted as a possible labels, which is why the numbers in Table 1 add up to more than 251. 31 sentences in the gold standard have more than one label, 27 of which have two and four have four labels.

The emotion categories in Table 1 were chosen because annotation tests showed that they reflected the data well and to create an annotation scheme where the emotions are balanced in terms of sentiment: four emotions express positive and another four negative sentiment. This is to prevent the formation of catch-all categories, which might oversimplify and distort the analysis. To ease annotation and make the emotions more clearly defined, we decided to refine the emotion categories by specifying their different manifestations: for example, love in the context of parliamentary speech can also be understood as praise or admiration. A challenge that emerged was to distinguish between true emotion and rhetorical strategy. In other words, what should be classified as an emotion, instead of a mere performance? We followed a definition commonly employed by psychologists in viewing emotion as a subcategory of *affect*, wherein affect is embodied and unconscious, while emotions are more structured and patterned expressions of affects, anchored in language.

4.2. Prompting and Model Training

Interacting with LLMs requires prompt engineering, which refers to formulating and manipulating the model input in such a way that desired results are achieved. We tried multiple different prompts and compared the results to our gold standard before settling on the final version. We found that writing the prompt in English, even though our data is in Finnish, improved the results. This effect is likely explained by the fact that GPT-4 performs worse on low resource languages, such as Finnish, compared to its performance on English (OpenAI et al., 2023). We also tried including the preceding and following sentences for each example as context but found that this only confused the model and led to worse results. We noted that keeping the instructions short and concise led to higher inter-annotator agreement than including detailed explanations for each class. Finally, using GPT-4 gave better results than the standard GPT-3.5, which is why this is the model we decided to use despite its higher cost. To save some cost, the re-occurring formulaic greeting *Arvoisa puhemies!* ('Honoured chairman!') and its variations were automatically given the "neutral" label. In total, the cost of annotating our data using the OpenAI API was around $60.

annotators	κ	F1 micro	F1 macro
ann1-ann2	0.406	0.602	0.379
ann1-ann3	0.476	0.685	0.429
ann1-ann4	0.145	0.590	0.355
ann2-ann3	0.553	0.713	0.529
ann2-ann4	0.624	0.729	0.590
ann3-ann4	0.518	0.673	0.481
average	0.499	0.655	0.416
gold-GPT-4	0.554	0.725	0.495

Table 2: Inter-annotator agreement between different human annotators and between gold standard and GPT-4.

GPT-4 was then used to annotate a dataset of 6025 sentences. The specific version of the model used is *gpt-4-0125-preview*, which is the most recent version of the model at the time of writing. The 251 sentences used for annotation evaluation were kept separate and the remaining 5774 sentences were split into train and validation sets with a 90-10 split. These data were used to train a BERT model. The model was evaluated against the gold standard annotations. We used a grid search to optimize the hyperparameters of the training stage. We used a learning rate of 3.16e-05, batch size of 32 and a label smoothing factor of 0.1.

As a baseline for our model, we trained two other BERT models using two machine translated datasets. Machine translating datasets has been shown to be a resource efficient way to create datasets that can produce better results than using multilingual models (Eskelinen et al., 2023). We produced the Finnish translations using DeepL[2]. The first baseline model is trained on the Many Emotions (ME) dataset. ME combines emotion annotated sentences from three separate datasets: Daily Dialog (Li et al., 2017), GoEmotions (Demszky et al., 2020) and Emotion (Saravia et al., 2018). These datasets source from transcriptions of casual conversations, Reddit posts and Twitter mes-

[2] https://www.deepl.com/translator

sages, respectively. The second baseline model is trained on the HunEmPoli dataset, which contains emotion annotated sentences from the Hungarian parliament (Üveges and Ring, 2023). We test the performance of these baseline models against the gold standard. Since the labels in the datasets differ slightly from our labels, we harmonise the labels before comparison by removing sentences and combining labels where necessary.

5. Results

We use Cohen's Kappa (κ) and F1 metrics to evaluate *inter-annotator agreement* (IAA). The numbers in Table 2 attest to the difficulty of the annotation task: despite many test rounds and discussion, IAA remained modest. When discussing the annotation results, we noticed that in many cases there is no single correct label for a given sentence and, instead, different interpretations are equally valid. The subjectivity of emotion annotation and subsequent low IAA has been noted before in the literature (Öhman, 2020).

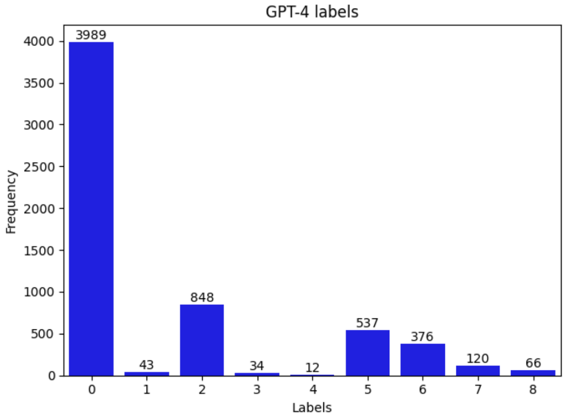


Figure 1: Distribution of labels in final GPT-4 annotated dataset.

IAA between human annotators and GPT-4 was calculated by comparing GPT-4 annotations to the gold standard. For sentences with multiple labels in the gold standard, any of the possible labels are counted as correct since GPT-4 agrees with at least one human annotator. As the numbers in Table 2 show, GPT-4 reaches human level accuracy in the task.

The model trained on GPT-4 annotated data, which we here call the GPT-4 model, reaches a macro F1 of 0.411 and a micro F1 of 0.690 meaning that the model performs well overall but struggles with some classes. This is understandable considering the distribution of labels in the datasets shown in Figure 1. The plot in Figure 2 shows that the model tends to over-predict class 0 (neutral) and seems to combine most sentences with positive sentiment in class 2 as is the case with the GPT-4 annotations, too. This suggests that the results could improve via further prompt engineering, although positive classes were also difficult for human annotators to distinguish. The comparison between models in Table 3 shows that the baseline models perform much worse. Surprisingly, even the in-domain HunEmPoli dataset does not seem to fit our data well. This might be because of differing annotation schemes and instructions, or due to cultural differences between the two parliaments. The ME model only predicts the emotions *neutral* and *joy* in our evaluation set, suggesting that casual conversation and internet discourse are too distinct from parliamentary discourse to be used as our training data. These results support the use of GPT-4 as a resource efficient method of creating training data.

GPT-4 model	**ME model**	**HunEmPoli model**
GPT-4 annotated parliamentary speeches	machine translated Many Emotions	machine translated parliamentary speeches
c. 6,000 sentences	c. 550,000 sentences	c. 19,000 sentences
9 labels	7 labels	6 labels
micro **0.690** macro **0.411**	micro 0.574 macro 0.138	micro 0.261 macro 0.182

Table 3: Model comparison

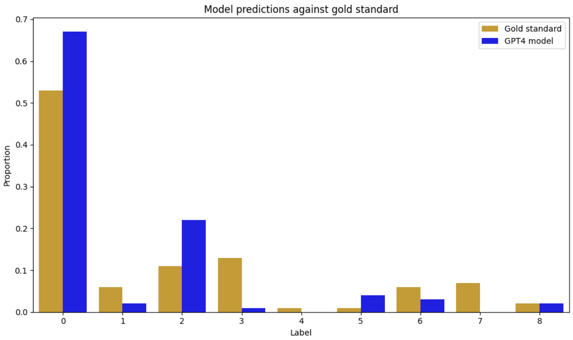


Figure 2: Model predictions vs gold standard.

6. Conclusion

In this paper, we have shown that GPT-4 can be used to create an emotion analysis dataset that can then be used to train an emotion classifier. The work presented in this paper is still ongoing as we continue refining the annotation, prompting and training procedures in the near future. This emerging methodology shows much promise as it makes the previously expensive and time consuming process of manual annotation much faster and cheaper. Machine translating existing datasets can still be a useful method for obtaining training data but, depending on the task, domain and availability of datasets, using an AI assistant such as GPT-4 might be a viable option. In the future, as the technology matures and costs are reduced, their use in data annotation could become commonplace, although they do raise their own set of challenges that must be overcome (see Ziems et al. 2024).

Many open questions still remain and there is much research being done is this emerging field. One open question is the viability of using AI assistants for other annotation tasks, as there is no guarantee that quality annotation is possible for all tasks and datasets. In fact, Heseltine and von Hohenberg (2024) show that GPT-4 annotations vary between tasks and languages. Additionally, much more is still to be learned about optimal prompting strategies. For example, Hu and Collier (2024) measure the effect of introducing persona variables, such as gender, political orientation and level of education in the prompt. We encourage other researchers in the field to continue experimenting with similar methods to advance resource efficient data annotation.

7. Acknowledgements

This research was funded by the Research Council of Finland [grant number 353569]. The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources and professor Filip Ginter for his sage advice and support in the undertaking of this research.

8. Bibliographical References

Hanna Bäck and Marc Debus. 2016. *Political Parties, Parliaments and Legislative Speechmaking*. Palgrave Macmillan UK.

Janos Borst, Jannis Klähn, and Manuel Burghardt. 2023. Death of the dictionary? – The rise of zero-shot sentiment classification. In *Computational Humanities Research Conference (CHR 2023)*.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Oswald Campesato. 2021. *Natural Language Processing Fundamentals for Developers*. Mercury Learning & Information.

Michael Chmielewski and Sarah C. Kucker. 2020. An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, 11(4):464–473.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. ArXiv preprint.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Anni Eskelinen, Laura Silvala, Filip Ginter, Sampo Pyysalo, and Veronika Laippala. 2023. Toxicity detection in Finnish using machine translation. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 685–697, Tórshavn, Faroe Islands. University of Tartu Library.

Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4).

Nicolò Fraccaroli, Alessandro Giovannini, Jean-François Jamet, and Eric Persson. 2022. Ideology and monetary policy. the role of political parties' stances in the European Central Bank's parliamentary hearings. *European Journal of Political Economy*, 74.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Michael Heseltine and Bernhard Clemm von Hohenberg. 2024. Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1).

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. ArXiv preprint.

Juha Koljonen, Emily Öhman, Pertti Ahonen, and Mikko Mattila. 2022. Strategic sentiments and emotions in post-second world war party manifestos in Finland. *Journal of Computational Social Science*, 5.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of The 8th International Joint Conference on Natural Language Processing (IJCNLP 2017)*.

Usman Malik, Simon Bernard, Alexandre Pauchet, Clément Chatelain, Romain Picot-Clémente, and Jérôme Cortinovis. 2024. Pseudo-labeling with large language models for multi-label emotion classification of French tweets. *IEEE Access*, 12:15902–15916.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*.

Emily Öhman. 2020. Emotion annotation: Rethinking emotion categorization. In *Post-Proceedings of the 5th Conference Digital Humanities in the Nordic Countries (DHN 2020)*.

Emily Öhman. 2022. SELF & FEIL: Emotion lexicons for Finnish. In *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries conference*.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. XED: A multilingual dataset for sentiment analysis and emotion detection. In *The 28th International Conference on Computational Linguistics (COLING 2020)*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston

Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. GPT-4 technical report.

Salomon Orellana and Halil Bisgin. 2023. Using natural language processing to analyze political party manifestos from New Zealand. *Information*, 14(3).

Robert Plutchik. 1982. A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5):529–553.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Meg Russell, Daniel Gover, Kristina Wollter, and Meghan Benton. 2017. Actors, motivations and outcomes in the legislative process: Policy influence at Westminster. *Government and Opposition*, 52(1):1–27.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of NLP crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Or Tuttnauer. 2018. If you can beat them, confront them: Party-level analysis of opposition behavior in European national parliaments. *European Union Politics*, 19(2):278–298.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. ArXiv preprint.

Tobias Widmann and Maximilian Wich. 2023. Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in German political text. *Political Analysis*, 31(4):626–641.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? ArXiv preprint.

István Üveges and Orsolya Ring. 2023. HunEmBERT: A fine-tuned BERT-model for classifying sentiment and emotion in political communication. *IEEE Access*, 11:60267–60278.

9. Language Resource References

Semantic Computing Research Group. 2021. *Parlamenttisampo*. PID https://parlamenttisampo.fi/fi/.

Making Parliamentary Debates More Accessible: Aligning Video Recordings with Text Proceedings in Open Parliament TV

Olivier Aubert, Joscha Jäger
LS2N - Nantes Université, Open Parliament TV
contact@olivieraubert.net, joscha.jaeger@openparliament.tv

Abstract
We are going to describe the Open Parliament TV project and more specifically the work we have done on alignment of video recordings with text proceedings of the German Bundestag. This has allowed us to create a comprehensive and accessible platform for citizens and journalists to engage with parliamentary proceedings. Through our diligent work, we have ensured that the video recordings accurately correspond to the corresponding text, providing a seamless and synchronised experience for users. In this article, we describe the issues we were faced with and the method we used to solve it, along with the visualisations we developed to investigate and assess the content.

Keywords: video, text proceedings, alignment, data

1. Introduction

While parliamentary discourse analysis has traditionally been text-based, over the last 5 years the research community has seen a slow shift towards incorporating audiovisual information into parliamentary datasets.

Enriching parliamentary datasets with multimodal information allows new methods of analysis, like non-verbal cues, gestures/mimical information eg. to gain insights into their influence on perceived trust and/or confidence in politicians.

Additionally the audio information can help identify important events that were not transcribed or can be used as supplementary cues, e.g. for sentiment analysis.

Beyond the academic realm, video recordings of parliamentary debates hold great untapped potential for digital democracy. They serve as a tangible and contemporary interface to the daily work of parliaments. The recordings and live streams are not just video collections for journalists or corpora for scientific research but a direct application of the guiding principle “the parliament negotiates in public“.

Open Parliament TV uses this potential by developing a search engine and interactive video platform, in which speeches are searchable, linkable, citable and shareable beyond the boundaries of single parliaments.

1.1 Background

Almost every parliament publishes video recordings and text proceedings of sessions. But despite comparable structures and similar workflows, parliamentary proceedings are published in various, incompatible formats and parliament tv contents are only accessible via proprietary platforms. With Open Parliament TV we are developing a parliament independent open source solution which makes the video recordings searchable, shareable and citable via an automatic synchronisation of video recordings and text proceedings.

Our work is thereby focused on live data, which is made accessible via an easy to use platform interface[1], show in figure 1, as well as a standardised and well documented open data api[2]. By implementing parliament independent data processing workflows we aim to interconnect political discourse between parliaments on national, regional as well as supranational (eg. EU Parliament) levels.

We have created a reference implementation with data from the German Bundestag, through which more than 60k speeches spanning over 10 years of parliamentary history are accessible (from 2013 until today).

In contrast to efforts like Open Discourse (Richter et al, 2023) and GermaParl (Blätte & Blessing, 2018; Blätte et al, 2022) who also work with a German Bundestag corpus, we focus on the audiovisual representation of speeches and work with archived and live data. The proceedings are hereby a means of making the videos more accessible, not vice versa.

1.2 Web Platform

Via the automated synchronisation of video recordings and official text proceedings we enable a full text search of the videos on the Open Parliament TV platform. By force aligning text fragments in the proceedings with specific points of time in the video recordings, we can additionally provide

- Interactive Transcripts
  (click on a sentence > jump to point of time in the video)
- Additional Information
  (show relevant documents and links at specific points of time in the speech)
- Improved means of participation
  (cite, embed and share video segments in the context of the full speech)

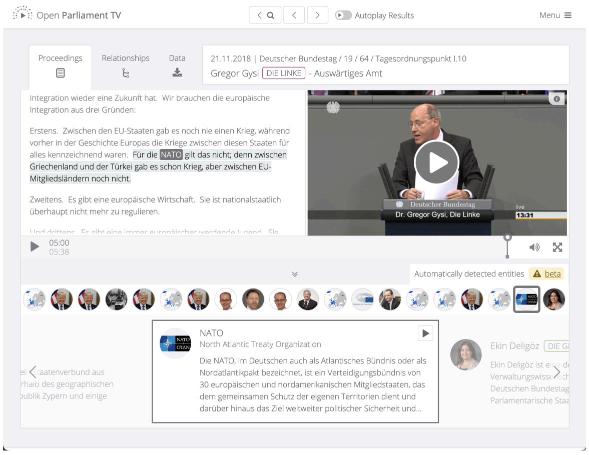
The platform significantly simplifies finding, sharing, embedding and citing specific video segments of political speeches and thus makes

[1] https://de.openparliament.tv
[2] https://de.openparliament.tv/api

ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora, pages 77–83
May 20, 2024.

parliamentary processes more transparent and accessible.

By providing an easy-to-use platform interface we make parliamentary work more accessible for researchers but also for journalists, political activists, educational institutions and the general public.



21.11.2018 | Deutscher Bundestag / 19 / 64 / Tagesordnungspunkt I.10
Gregor Gysi DIE LINKE - Auswärtiges Amt

Integration wieder eine Zukunft hat. Wir brauchen die europäische Integration aus drei Gründen:

Erstens. Zwischen den EU-Staaten gab es noch nie einen Krieg, während vorher in der Geschichte Europas die Kriege zwischen diesen Staaten für alles kennzeichnend waren. Für die NATO gilt das nicht; denn zwischen Griechenland und der Türkei gab es schon Krieg, aber zwischen EU-Mitgliedsländern noch nicht.

Zweitens. Es gibt eine europäische Wirtschaft. Sie ist nationalstaatlich überhaupt nicht mehr zu regulieren.

Und drittens. Es gibt eine immer europäischer werdende Jugend. Sie

NATO
North Atlantic Treaty Organization
Die NATO, im Deutschen auch als Atlantisches Bündnis oder als Nordatlantikpakt bezeichnet, ist ein Verteidigungsbündnis von 30 europäischen und nordamerikanischen Mitgliedstaaten, das dem gemeinsamen Schutz der eigenen Territorien dient und darüber hinaus das Ziel weltweiter politischer Sicherheit und...

Figure 1: the Open Parliament TV platform

The Open Parliament TV platform additionally serves as a tangible open data use case, which is regularly used to advocate for better open data policies, standards and implementations on parliament level.

Linking platform contents directly with the respective parliamentary sources provides an additional layer of transparency and makes it easier for users to track and cite the original source as well as the context of quotes by politicians.

Especially in the context of citing quotes from political speeches, the official proceedings are a valuable source. The fact that the stenographic protocols don’t exactly match the actual spoken word is in our case a feature, as the proceedings function as an immediately citable trusted source and will always be more reliable than transcripts generated by automatic speech-to-text (ASR) systems.

We do however use automated processes in order to annotate and enrich platform content with additional information. Based on the proceedings text, we extract Wikidata entities for people, organisations, laws and specific terms via Named Entity Recognition and provide information like Wikipedia abstracts or links to additional sources right inside the video player.

2. Related Work

In the ParlaMint community, several efforts have been made to use a combination of audio plus proceedings or transcripts to train ASR models (Ogrodniczuk et al, 2022).

A common issue is the alignment of incoherent source data for text proceedings and video recordings. This issue can be broken down into 2 main challenges:

- Finding a common identifier for both sources
- Determining common speech boundaries

Parliaments usually publish proceedings and video recordings via separate platforms, managed by different departments. In some cases the publication of video recordings is even outsourced to third party companies or media partners. This leads to differences in naming conventions of speakers and agenda items, making it difficult to identify the correct video resource for a specific speech in the proceedings (Ljubešić et al, 2022; Kulebi et al 2022).

One approach to deal with these inconsistencies is applying fuzzy match algorithms for the names of speakers (Kulebi et al, 2022). Beyond naming conventions both modalities sometimes have a different segmentation of speeches (specifically regarding speech items by the president), making it difficult to apply the otherwise feasible solution of comparing the two sources by the order / indices of items (Kulebi et al, 2022). In the ParlamentParla project, the two speaker sources have additionally been aligned using the Smith-Waterman sequence matching algorithm (Smith & Waterman, 1981).

Subsequently there is no common understanding of the beginning and end of speeches and agenda items. This is specifically relevant when using a combination of audio streams and proceedings in order to train ASR models (Ljubešić et al, 2022) as well as with automatic video subtitling systems (Alkorta, J., & Quintian, 2022). To determine common boundaries, some use automatic transcripts derived from speech-to-text systems and compare those with the official proceedings text via a forced alignment process (Hladká et al, 2020).

One solution to the problem of incoherent sources are machine readable proceedings, which contain references to the respective audiovisual resources in the metadata, as can be found (in non-standardised formats) in some parliament's data, like the Czech parliament (Hladká et al, 2020), the French Assemblée Nationale[3] or more recently the Austrian Parliament[4].

In recent years the extension of proceedings data with video recordings and the subsequent publication of multi-modal aligned (research) corpora has increasingly been mentioned as future work (Ogrodniczuk et al, 2022; Agnoloni et al,

[3] https://www.assemblee-nationale.fr/dyn/16/comptes-rendus/seance

[4] https://www.parlament.gv.at/recherchieren/protokolle/index.html

2022). This would allow annotating corpora with physical communicative features like gestures and facial expressions (Ogrodniczuk et al, 2022; Ménard & Aleksandrova, 2022).

3. Automating AV Alignment

Access to video material depends on some kind of discretization to facilitate indexing and navigation. In the case of parliaments, the proceedings are an official source of textual data that should match the video feeds. There is not yet any standard shared by all parliaments, therefore the Open Parliament TV has to conceive an ingest infrastructure dedicated to handle the specificities of each parliament and convert its data into its own common model.

3.1 Context: the Bundestag Plenary Sessions

In the Bundestag case, the video stream is broadcasted live on the https://www.bundestag.de/ website. Some textual metadata is associated with it before the recording, based on the agenda of the session. The interface displays the title of each intervention - current and forecoming - as well as the speaker name (with the planned time of speaking), and features references to additional material.

In addition to the frontend web interface, the video feed is provided as a video podcast, i.e. a RSS stream of mp4 files. Each item features the session identification with its date, the intervention title and the speaker name.

Official minutes are provided through the website as well, with a delay of 2 to 3 days. The main web interface features links to related documents, as well as a link to a summary of each part. The official plenary minutes[5] are provided as PDF files. A session is divided into multiple agenda items. Each item provides a link to the corresponding PDF file and to the video of the session. API-wise, the Bundestag proposes the Documentation and Information system for Parliamentary materials (DIP) which provides structured access to a query interface into the document material (as text fragments or PDF documents), but does not give access to proceedings themselves in a structured format.

An additional opendata API is also available[6]. It provides a stream of plenary proceedings in XML format, structured using a dedicated dbtplenarprotokoll DTD.

The goal of the data ingestion phase is to provide for the Open Parliament TV platform a unified data source combining both video streams and text proceedings, enriched with Wikidata IDs, so that they can be presented in meaningful ways through the platform interface. This process has to be able to process both old data, but also to run unattended to provide an as-live-as-possible experience to the users: the video data is presented as soon as it is available, and later enriched with the text proceedings when the data becomes available.

The data is organised in electoral terms. The current one, the 20th, started on 26/09/2021 and is still ongoing. In order to give an idea of the corpus dimensions, we will focus on the preceding term, for which we have the complete data. The 19th electoral term ran from 24 October 2017 until 26 October 2021. The 736 representatives attended 239 plenary sessions, which produced 35.86 hours of video.

3.2 Architecture of the Code

The data processing code is published on github[7]. It is free software, licensed under the General Public License.

It is divided into fetcher modules that download updated data (media and proceedings) in raw XML or json format, and parser modules that massage the data into the unified model of the Open Parliament TV platform. Then a merger module, which we will more precisely describe in this article, matches data from both sources in order to produce a unified format mixing both video and textual information.

Once the media and proceedings items are aligned, additional processing takes place. Speaker names are linked with their corresponding Wikidata id (in nel module) and forced alignment is applied on the video fragments and transcript in order to provide a more fine-grained association of the text proceedings.

The different modules (fetcher, parser, nel, forced alignment…) can be used independently, and their orchestration is implemented in a workflow script.

3.3 Identified Alignment Issues

The main key for aligning items between the video feed and the OpenData XML proceeding feed is the title of the item and the speaker name. However, similarly to [Kulebi et al, 2022], a number of mismatches plague the data. They may come from human errors or the application of transcription conventions that remove some text for the sake of clarity. In the German Bundestag , speakers are also given the opportunity to amend the transcribed version, as described in rules 116-119 of (Deutscher Bundestag, 2022).

First, there are small transcription errors in speaker names and speech titles, e.g. putting the title (Dr., Prof.) in front of the name in media data but not in

[5] Official plenary minutes https://www.bundestag.de/dokumente/protokolle/plenarprotokolle

[6] Opendata API https://www.bundestag.de/services/opendata

[7] Processing tools repository https://github.com/OpenParliamentTV/OpenParliamentTV-Tools

proceeding data. Then, there are larger and more systematic errors, often occurring in batches, where a whole agenda item title is wrongly assigned. Similar issues can be found in the time segmentation: speech boundaries do not always match, the session president introduction being sometimes included in the preceding speech in the video capture. More importantly, completely different segmentations may occur, resulting in different amounts of media and proceeding items for the same session, and increasing the difficulty for the matching process.

3.4 First Approach based on Speaker/Title Similarity

A first naive matching approach was first used, based on using speaker and title - after a small normalizing process - for generating a key identifying each media and proceedings item. Collisions were handled by adding an incremental index.

To try to alleviate small transcription errors, common similarity measures like the Levenshtein distance were experimented to compare the generated keys, but the nature of the underlying data, where agenda items can often share the same base title and differ only with an index or a reference number, made this approach inappropriate.

Moreover, the discrepancy between the number of media items and the number of proceeding items made this approach inherently fragile. Figure 1 presents a scatter plot of each session in the 19th term with its number of proceeding and media items on the horizontal and vertical axis. We can see that the majority of sessions have the same number of media and proceeding items, being concentrated on the diagonal, but the number of non-matching sessions is important.

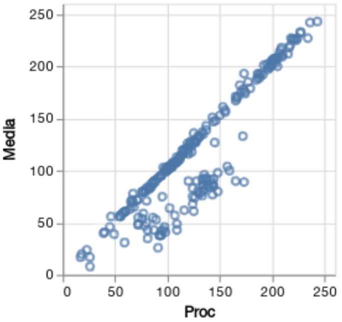


Figure 2: representation of proceeding items count vs media items count

To investigate further, we built a new visualisation, using small multiple scatter plots, as presented in figure 2 : x-axis represents the index of the media item in the "media" sequence. The y-axis represents the index of the corresponding proceeding item in the proceedings sequence. In the ideal case, both axes have the same length (same number of media items wrt. proceeding items) and the representation should be a diagonal, meaning that media item number N matches proceeding item number N for all items. This allows to visually quickly discriminate against misaligned sessions. Additionally, this gives information about the alignment symptoms. For instance, for the first session in the second line, we see a horizontal line with a diagonal starting approximately in the middle x axis, with a low media index. The interpretation, corroborated by examination of the actual data, is that the proceeding segmentation has been more fine-grained than the media segmentation. Hence, the same media item (first one) has been aligned with multiple proceeding items, which gives a horizontal line. Upon examination of the data, this shape often occurs in sessions of questions to the government, which generates a single, long media item with the president of parliament as indicated speaker, while the proceedings have split the questions by speaker, producing multiple shorter items in proceedings.

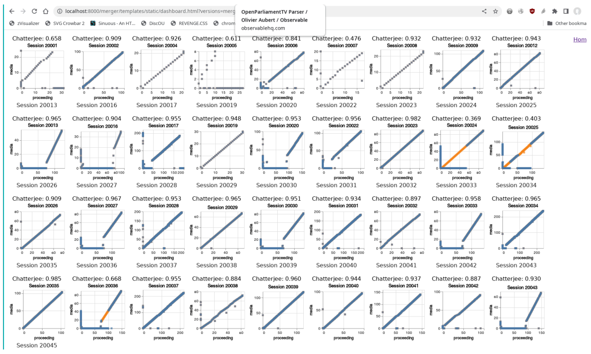

localhost:8000/merger/templates/static/dashboard.html?versions=merg
zVisualizer
SVG Crowbar 2
Sinuous - An HT...
DiscOU
REVENGE.CSS
chrom
OpenParliamentTV Parser /
Olivier Aubert / Observable
observablehq.com
Other bookma
Hom
Chatterjee: 0.658 Session 20001
Chatterjee: 0.909 Session 20002
Chatterjee: 0.926 Session 20004
Chatterjee: 0.611 Session 20005
Chatterjee: 0.841 Session 20006
Chatterjee: 0.476 Session 20007
Chatterjee: 0.932 Session 20008
Chatterjee: 0.932 Session 20009
Chatterjee: 0.943 Session 20012
Session 20013
Session 20016
Session 20017
Session 20019
Session 20020
Session 20022
Session 20023
Session 20024
Session 20025
Chatterjee: 0.965 Session 20013
Chatterjee: 0.904 Session 20016
Chatterjee: 0.955 Session 20017
Chatterjee: 0.948 Session 20019
Chatterjee: 0.953 Session 20020
Chatterjee: 0.956 Session 20022
Chatterjee: 0.982 Session 20023
Chatterjee: 0.369 Session 20024
Chatterjee: 0.403 Session 20025
Session 20026
Session 20027
Session 20028
Session 20029
Session 20030
Session 20031
Session 20032
Session 20033
Session 20034
Chatterjee: 0.909 Session 20026
Chatterjee: 0.967 Session 20027
Chatterjee: 0.953 Session 20028
Chatterjee: 0.965 Session 20029
Chatterjee: 0.951 Session 20030
Chatterjee: 0.934 Session 20031
Chatterjee: 0.897 Session 20032
Chatterjee: 0.958 Session 20033
Chatterjee: 0.965 Session 20034
Session 20035
Session 20036
Session 20037
Session 20038
Session 20039
Session 20040
Session 20041
Session 20042
Session 20043
Chatterjee: 0.985 Session 20035
Chatterjee: 0.668 Session 20036
Chatterjee: 0.955 Session 20037
Chatterjee: 0.884 Session 20038
Chatterjee: 0.960 Session 20039
Chatterjee: 0.944 Session 20040
Chatterjee: 0.937 Session 20041
Chatterjee: 0.887 Session 20042
Chatterjee: 0.930 Session 20043
Session 20045
media
proceeding

Figure 3: small-multiple scatter plot visualisation of sessions

This visualisation moreover allows comparison between results of different alignment algorithm/parameters: respective outputs are plotted in different colors (here, yellow and blue), which allows to quickly identify the sessions where changes occurred. For instance in the image, the alignment for the last 2 sessions of the second line improved greatly, giving an adequate diagonal.

This lead us to the conclusion that the problem was not simply tackable through simple index matching, especially because of the segmentation difference. It occurred regularly in specific groups of sessions, like the questions to parliament. The index matching approach also had the inconvenience of ignoring the sequencing of items, while the data, being the recording of an event, implies that item sequences must be preserved.

In essence, we have 2 sequences of "alphabet" that should somehow match. There can be cases of insertion of sequences in one side (shorter segments for instance) or of deletion (longer segments). There can also be complete changes (like in human transcription errors), which we can call mutation.

This similarity to DNA-alignment led us to investigate this direction as a new approach.

3.5 Needleman-Wunsch Algorithm

We investigated with colleagues from a bioinformatics team, in order to explain the issue and find similarities and solutions that could be transferred from this domain. Indeed,the Needleman-Wunsch algorithm, a classical algorithm from the 1970s, can be used to align DNA sequences, trying to preserve global order, with parameterized costs for insertion or deletion. Another algorithm, the Smith-Waterman Algorithm (Smith & Waterman, 1981) has been used in (Kulebi et al, 2022) for similar purposes, but focused on local sequences.

In our implementation the algorithm has 4 parameters. Two, *speaker_weight* and *title_weight*, are related to the similarity measure between 2 items. As with our previous experiments, we noticed that the data specificities made common string approximations like Levenshtein inappropriate, and we chose to do basic string comparison, weighted by parameters. The other two parameters, *merge_penalty* and *split_penalty*, are used by the algorithm itself to parameterize.

The Needleman-Wunsch algorithm is a dynamic programming algorithm used in bioinformatics to perform global sequence alignment between two sequences. It starts by creating a matrix, typically called the scoring matrix, where the columns represent items from proceedings and the rows represent items from the media source. A similarity function is defined, as the ponderated sum of the string similarities of speaker and titles. The matrix is initialized along the first row and column with the similarity between corresponding items.

To fill the matrix, we recursively calculate scores, starting from the first cell, and computing the score of the neighboring cells, comparing the hypotheses of moving to one of the horizontal, vertical or diagonal neighbors, and keeping the hypothesis with the highest valued. The horizontal neighbor hypothesis adds an increment of *merge_penalty*, since it represents the cell that would be reached by merging two proceeding items. The vertical neighbor hypothesis adds an increment of *split_penalty*, since it represents the fact that a proceeding should be split between two media items. The diagonal hypothesis provides an increment of the similarity score between its items, representing the "normal" hypothesis. We iterate through the matrix, calculating scores for each cell based on the recurrence relation until the entire matrix is filled.

Once the matrix is filled, as presented in figure 3, traceback is performed to find the optimal alignment between the two sequences. We start at the highest score in the top-right corner of the matrix, and trace back to the bottom-left corner, following the path of highest scores. This traceback process identifies the alignments that maximizes similarity between the sequences.

As a result, the algorithm outputs the optimal item alignments along with their corresponding scores. The graphical and interactive representation, linked with the transcript and the video, allowed us to validate the efficiency of the approach.

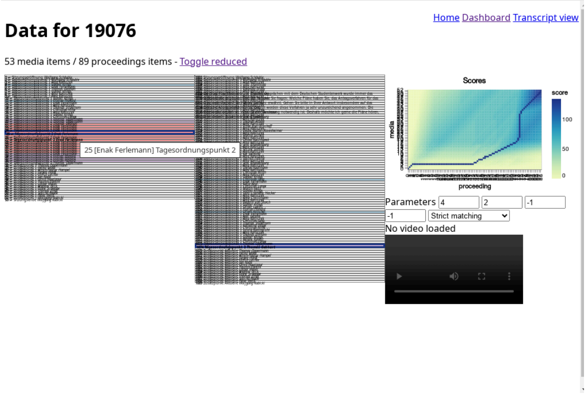
Overall, the Needleman-Wunsch algorithm efficiently finds the optimal alignment(s) between two sequences by considering all possible alignments and scoring them based on a defined scoring scheme, making it a fundamental tool in bioinformatics for sequence analysis and comparison.

3.6 Dashboard and Visualisations

A dashboard presenting visualisations of the processed corpus is available at https://openparliamenttv.github.io/OpenParliamentTV-Tools/optv/parliaments/DE/dashboard/dashboard.html

The data hosted on github is subjected to download rate limits though, and also does not have the intermediary parsed media and proceeding files, limiting the use of some visualisations. Hence, we are also hosting the same dashboard on a dedicated server at https://optv.olivieraubert.net/ with the whole data.

The dashboard proposes to select subsets of the whole corpus, incrementally loading their data to present it. Once a group is selected, the different scatter plot visualisations are presented. Clicking on the title of each graph leads to more precise visualisations, in order to provide better context for exploring data and issues. The Session word is linked to the "block view" described below, while the session number is linked to the "transcript view".



Data for 19076

53 media items / 89 proceedings items - Toggle reduced

Parameters 4 2 -1

-1 Strict matching

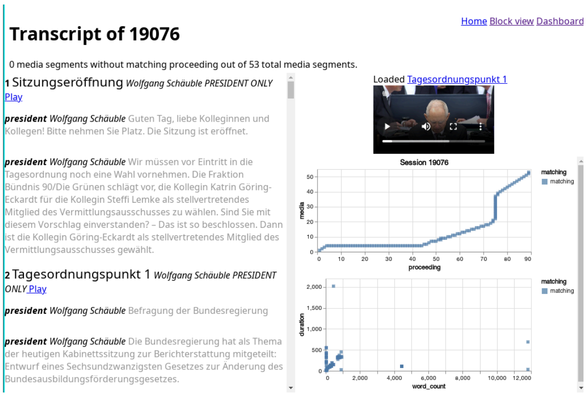
No video loaded

Home Dashboard Transcript view
25 [Enak Ferlemann] Tagesordnungspunkt 2
Scores
media
proceeding
score
100
50
0

Figure 4: "block visualisation" with dynamic parameterization of the matching algorithm

Figure 4 presents an interactive visualisation, called "block view", that was built to validate and fine-tune parameters of the algorithm. It takes as input a session identifier and gets its data from the media and processing parsed files. It presents in the first two columns a visualisation of the media and item blocks, with their information (index, speaker and title) readable on mouse over. As a synchronized view, it highlights in the other column the items having a matching speaker name, speech title or having both. This view implements a javascript version of the Needleman-Wunsch algorithm, and presents a live-generated matrix of its output, with the ability to dynamically tune the 4 algorithm parameters and the string similarity method, in order to assess their influence.

Figure 5 presents a second interactive visualisation that was developed to explore and evaluate the actual output of the processing and merging workflow. It uses the merged output file data as input. On the left of the page, the transcript - generated from the proceeding data - is presented, along with an affordance to play the aligned video. It also offers a visualisation of the path produced by the algorithm in the right-hand side, along with a scatter-plot visualisation of the word count (from proceedings) vs duration (from media) of the aligned items, in order to explore other indicators.



Home Block view Dashboard

Transcript of 19076

0 media segments without matching proceeding out of 53 total media segments.

1 Sitzungseröffnung *Wolfgang Schäuble PRESIDENT ONLY* Play

president *Wolfgang Schäuble* Guten Tag, liebe Kolleginnen und Kollegen! Bitte nehmen Sie Platz. Die Sitzung ist eröffnet.

president *Wolfgang Schäuble* Wir müssen vor Eintritt in die Tagesordnung noch eine Wahl vornehmen. Die Fraktion Bündnis 90/Die Grünen schlägt vor, die Kollegin Katrin Göring-Eckardt für die Kollegin Steffi Lemke als stellvertretendes Mitglied des Vermittlungsausschusses zu wählen. Sind Sie mit diesem Vorschlag einverstanden? – Das ist so beschlossen. Dann ist die Kollegin Göring-Eckardt als stellvertretendes Mitglied des Vermittlungsausschusses gewählt.

2 Tagesordnungspunkt 1 *Wolfgang Schäuble PRESIDENT ONLY* Play

president *Wolfgang Schäuble* Befragung der Bundesregierung

president *Wolfgang Schäuble* Die Bundesregierung hat als Thema der heutigen Kabinettssitzung zur Berichterstattung mitgeteilt: Entwurf eines Sechsundzwanzigsten Gesetzes zur Änderung des Bundesausbildungsförderungsgesetzes.

Loaded Tagesordnungspunkt 1

Figure 5: transcription view, presenting the result of the alignment process

3.7 Contributions

The code and live-updated data are publicly available as repositories hosted on Github[8]. They can be interacted with through the Open Parliament TV platform, and assessed through a dashboard. The fetching and parsing code are specific to the concerned parliament, but the whole suite has been designed to be also used as much as possible on other parliament's data.

As a methodological contribution, we identified a number of issues like transcription errors and segmentation issues that will be common to other similar projects, as can be seen in (Kulebi et al, 2022). This lead us to implement and evaluate the adequacy of the Needleman-Wunsch sequence alignment algorithm.

Moreover, we produced a number of interactive visualisations for the data, either global (the dashboard view) or more specific (the block and transcript view), which could also be used as an inspiration in other projects.

4. Conclusion / Future Work

The Open Parliament TV project proposes a user-oriented interface for making parliamentary debates more accessible to the public and the media. By unifying video recordings with text proceedings, we have created a valuable resource for understanding the intricacies of legislative discussions. This work paves the way for expansion to other parliaments, thereby improving the parliament independent workflow and interconnecting discourse beyond national borders. While we have been using our own unified data model, we would like to move towards more standardised models in order to foster interoperability. Collaborating with organisations such as Open Discourse or GermaParl would offer an opportunity to integrate historical debates into the platform, extending its reach and value further.

While alternative approaches, such as complete transcription through speech-to-text algorithms and automatic translation, could have been considered, the availability of performant and robust models at the time of our research necessitated a different approach, and we wanted to be able to process data on standard computers. Today, however, advancements in this technology make it an exciting prospect for future exploration and further refining the alignment results.

In addition to checking discrepancies between official proceedings and actual discourse, increased accessibility through automatic speech translation into multiple languages would open new possibilities for users following debates in other parliaments. Furthermore, the potential to interconnect parliamentary discourse beyond language and parliamentary boundaries enables more comprehensive search and analysis capabilities. Overall, the Open Parliament TV project signifies a crucial advancement towards making parliamentary proceedings more accessible, transparent, and globally interconnected.

5. Acknowledgements

The authors wish to thank Guillaume Fertin and Geraldine Jean from LS2N, Nantes Université, for their advice in bioinformatics inspiration.

6. Bibliographical References

Agnoloni, T., Bartolini, R., Frontini, F., Montemagni, S., Marchetti, C., Quochi, V., Ruisi, M.. & Venturi, G. (2022, June). Making Italian Parliamentary

[8] Data repository (see above footnote for tools repository): https://github.com/OpenParliamentTV/OpenParliamentTV-Data-DE/

Records Machine-Actionable: The Construction of the ParlaMint-IT Corpus. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference* (pp. 117-124).

Alkorta, J., & Quintian, M. I. (2022, June). Adding the Basque Parliament Corpus to ParlaMint Project. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference* (pp. 107-110).

Blätte, A., & Blessing, A. (2018, May). The germaparl corpus of parliamentary protocols. In proceedings of the eleventh international conference on language resources and evaluation (LREC 2018).

Blätte, A., Rakers, J., & Leonhardt, C. (2022, June). How germaparl evolves: Improving data quality by reproducible corpus preparation and user involvement. In Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference (pp. 7-15).

German Bundestag (2022). Rules of Procedure of the German Bundestag and Rules of Procedure of the Mediation Committee. https://www.btg-bestellservice.de/pdf/80060000.pdf (visited on 2024.03.29)

Hladká, B., Kopp, M., & Straňák, P. (2020, May). Compiling Czech parliamentary stenographic protocols into a corpus. In *Proceedings of the Second ParlaCLARIN Workshop* (pp. 18-22).

Kulebi, B., Armentano-Oller, C., Rodríguez-Penagos, C., & Villegas, M. (2022, June). ParlamentParla: A speech corpus of catalan parliamentary sessions. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference* (pp. 125-130).

Ljubešić, N., Koržinek, D., Rupnik, P., & Jazbec, I. P. (2022, June). ParlaSpeech-HR-a freely available ASR dataset for croatian bootstrapped from the parlaMint corpus. In Proceedings of the workshop ParlaCLARIN III within the 13th language resources and evaluation Conference (pp. 111-116).

Ménard, P. A., & Aleksandrova, D. (2022, June). A French Corpus of Québec's Parliamentary Debates. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference* (pp. 25-32).

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, *48*(3), 443-453.

Ogrodniczuk, M., Osenova, P., Erjavec, T., Fišer, D., Ljubešić, N., Çöltekin, Ç., Kopp, M. & Katja, M. (2022, June). ParlaMint II: The Show Must Go On. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference* (pp. 1-6).

Richter, F., Koch, P., Franke, O., Kraus, J., Warode, L., Kuruc, F., Heine, S., Schöps, K. (2023, January 21). Open Discourse: Towards the first fully Comprehensive and Annotated Corpus of the Parliamentary Protocols of the German Bundestag. https://doi.org/10.31235/osf.io/dx87u

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, *147*(1), 195-197.

Russia and Ukraine through the Eyes of ParlaMint 4.0: A Collocational CADS Profile of Spanish and British Parliamentary Discourses

María Calzada Pérez
Departamento de Traducción y Comunicación, Universitat Jaume I (Castellón, Spain)
calzada@uji.es

Abstract

This article resorts to mixed methods to examine British and Spanish parliamentary discourse. The quantitative corpus-assisted (lexical priming) theory and data are complemented by the qualitative discourse historical approach. Two CLARIN ParlaMint corpora – ParlamMint-GB and ParlaMint-ES – are queried in the analysis, which focuses on English ("Rusia" and "Ukraine") and Spanish ("Rusia" and "Ucrania") nodes and collocations. In sum, the analysis sketches a brief profile of each corpus. The British House of Commons is more homogenous, strongly associating "Russia" and "Ukraine" with their participation in the war. Furthermore, this chamber shows a greater interest in "Russia. The Spanish Congreso de los Diputados indicates greater quantitative differences (heterogeneity). Here, "Russia" clearly transcends its role as a military contender and is also portrayed as an economic competitor for the West. Unlike in Britain, the Spanish lower house shows more mentions of "Ucrania", which is assigned just one role – as an invasion victim. In conclusion, the productivity of corpus-assisted mixed methods is confirmed along with the precious value of the ParlaMint constellation.

Keywords: Parliamentary Discourse, ParlaMint, Russia/Ukraine.

1. Introduction

Parliaments are institutions of the utmost importance. Democratic systems count on them to safeguard political representation and accountability. They are not just a mirror on which societies look but also spaces where politicians propose, discuss, and justify their actions. Most importantly, they are responsible for drafting and passing the laws citizens abide by. Their function is, therefore, essential to uphold equality, transparency, and fairness.

It is no wonder they have already attracted attention from a wide variety of areas, notably political science (see, for instance, Box-Steffensmeier, Brady, and Collier, 2008; Hix, Noury, and Roland, 2006; Bütikofer and Hug, 2015) and sociology. Skubic and Fišer (2022) are particularly illuminating for a literature review on the latter discipline since they identify the most prominent topics discussed in sociology on parliamentary discourse. Furthermore, they list the prolific methods to do so, among which the gamut of (critical) discourse studies excel, informing over 60% of the sociological analyses reviewed. The authors (Skubic and Fišer, 2022: 82) advocate that "the goal of sociological research of parliamentary discourse is to analyze political discourse and language". Hence, it is hardly surprising that they highlight the role of linguistics when approaching parliaments, and they recommend synergies with language-oriented studies.

Linguistics has also taken an interest in parliamentary/political discourse, as Calzada Pérez (2018) serves to testify. This work points to a growing pool of analyses approaching lower and upper houses from various prisms and targeting the micro- and macro-levels of parliamentary texts and contexts. Moreover, it confirms that, on this topic, linguistics also favors (critical) discourse studies.

When sociology and linguistics examine parliamentary interventions – as attested by both Skubic and Fišer (2022) and Calzada Pérez (2018) – they tend to draw on qualitative methodologies, with the highest potential for exposing descriptive results. However, they risk falling into subjectivism due to the small number of textual samples that are often analyzed.

A potential way to avoid subjectivism in parliament-related research is by advocating mixed methods, which boost qualitative results with quantitative data. Corpus-assisted studies (or CADS) do precisely this with "impressive results" (Garzone and Santulli, 2004: 353). With its name first coined by Partington (2004), CADS has been defined as "that set of studies into the form and/or function of language as communicative discourse which incorporates the use of computerized corpora in their analyses" (Partington, Duguid, and Taylor, 2013: 10). In other words, CADS uses corpus linguistics as a means to produce and dissect textual data for discourse studies.

Nevertheless, only a handful of analyses resort to CADS to examine parliamentary communication (e.g., Baker, 2006; 2010; Bayley, Bevitori, and Zoni, 2004; Bayley and San Vicente, 2004; Bevitori, 2004; Calzada Pérez, 2017; Calzada Pérez, 2017; Calzada Pérez, 2020; Dibattista, 2004; Garzone and Santulli, 2004; Vasta, 2004). This is partly because CADS depends on corpora, and researchers may find compilation and annotation somewhat cumbersome.

To aid experts in examining parliamentary discourse, in 2020, CLARIN vouched for the scholarly initiative led by Tomaž Erjavec, Maciej Ogrodniczuk and Petya Osenova, resulting in the ParlaMint project[1]. As stated on their website, at the time, ParlaMint-I managed to muster the efforts of at least 17 groups of scholars, which compiled parliamentary corpora with debates from 2015 to 2021 from 17 different

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

countries, such as the British House of Commons and the Spanish Congreso de los Diputados (the two corpora analyzed in this article). Erjavec et al. (2023) describe the project's rationale, the compilation and annotation stages, and the resulting corpora, which are "uniformly encoded, contain rich meta-data about 11 thousand speakers, and are linguistically annotated following the Universal Dependencies formalism and with named entities." (Erjavec et al. 2023, 415). At this stage, the totality of the 17 corpora amounted to almost half a billion words, and each of them was split into two specific subcorpora: a reference compilation (with texts from 2015 to 30th January 2020) and a Covid-19 corpus (with texts from 31st January 2020). Covid-19 is, as seems clear, a focal point for project researchers.

A Parlamint-II phase followed with data from 2022 and 2023. Subsequent phases are foreseen because scholars such as "sociologists are predominantly interested in current events, which means that it is of crucial importance for ParlaMint corpora to be updated on a regular basis" (Skubic and Fišer 2022, 89). ParlaMint-II has enlarged the time span of existing corpora, added new parliaments (there are now 29 parliaments from different countries and regions), upgraded the mark-up and annotation guidelines, tagged new metadata, and improved a common (Github-based) workflow. In practice, versions 3.0 and 4.0 were released in 2023, with yet another subcorpus under the label of "war". Thus, ParlaMint II adds another focal point of analysis: parliamentary texts around the Russia – Ukraine war.

As a result, the ParlaMint constellation is a robust tool to look into parliamentary discourse from a (quantitative) corpus-driven standpoint or to back up (qualitative) discourse studies. It may add to the complexity of the field since experts can now dissect texts and contexts according to a range of parameters: speakers, affiliations, positions, and gender, among others. Most importantly, it is a powerful artefact to aid researchers in their comparative and chronological studies. Thanks to compilation and annotation uniformity, comparability and interoperability, it is now possible to go beyond the national level and contrast results between and among different parliaments. It is also possible to carry out Modern-Diachronic Corpus Discourse Studies (following Partington, Duguid, and Taylor, 2013).

Against this background, the present article looks into parliamentary discourse from a CADS perspective. After ParlaMint II, attention is devoted to the way Russia and Ukraine are represented in two of their (2015-2022) full corpora: ParlaMint-GB (with interventions from the British House of Commons) and ParlaMint-ES (with interventions from the Spanish Congreso de los Diputados). In other words, in this article, the quantity afforded by corpus linguistics is nuanced by the quality of discourse studies. The former provides data and the notion of *lexical priming*. The latter contributes with the discourse historical approach (Wodak and Meyer, 2009). All this is explained further right below.

2. Priming Theory and Discourse Historical Approach

2.1 Priming Theory with Collocations

Corpus Linguistics is not just the source of quantitative data and corpus-based or corpus-driven methods (see McEnery and Hardie, 2012). It is also the realm that has seen the emergence of linguistic theories, among which *Priming*, it may be argued, is its most decisive one.

Priming theory is the work of Michael Hoey (2005: 8) (2005: 8), for whom "[a]s a word is acquired through encounters with it in speech and writing, it becomes cumulatively loaded with the contexts and co-texts in which it is encountered."

In priming theory, concordances and collocations play an essential role. Concordances (also known as keywords in context, KWIC) are lines of text around a certain node, like the example below:

nes to Russia and its invasion of **Ukraine** . That is why we have brought it

Figure 1: a concordance line

Hoey (2013, 155) implies that, on the one hand, "the brain must be storing language in a manner analogous to (though obviously not identical to) the way a concordance represents language" and, on the other hand, that:

> "when we encounter language we store it much as we receive it, at least some of the time, and that repeated encounters with a word (or syllable or group of words) in a particular textual and social context, and in association with a particular genre or domain, prime us to associate that word (or syllable or group of words) with that context and that genre or domain." (Hoey 2013: 155)

McEnery and Hardie (2012: 123) define a collocation as "a co-occurrence pattern that exists between two items that frequently occur in proximity to one another – but not necessarily adjacently or, indeed, in any fixed order". It is these surrounding patterns (i.e. the co-text) that end up transferring a great deal of meaning to the central "node" in context.

Collocations are built upon concordances, which means that scholars must generate concordances first and then identify collocations, either manually (by counting and listing the words around the node) or automatically, using statistics measures. Some important measures for collocation generation are logDice, MI, MI3, T-score, Z-score, etc. For a particularly clear, in-depth explanation of corpus statistics, see Brezina (2018).

There are at least two ways scholars may examine cumulative exposure (hence lexical priming potential) to repetitive contextual and cotextual patterns (such as collocation): (a) by focusing on the primed items ("for example […] all the lexical primings associated with the word consequence," Hoey 2005: 14); and (b) by identifying relationship among lexical primings ("all the primings that contribute to the production of a sentence," Hoey 2005: 14). Calzada Pérez (2017) mentions a third path that Hoey seems to have overlooked: that of the prime *per se* (such as the word "consequence" in our previous example). Nevertheless, regardless of how collocation is approached, it is a prominent gateway into lexical priming. The present paper opts for the first alternative and proposes a collocational analysis as the bulk of its quantitative examination.

2.2 Discourse Historical Approach

The present study opts for the discourse historical approach (DHA), which, in principle, advocates a top-down analysis that starts with an exhaustive ethnographic examination of the historical and generic contexts in which the texts under discussion are produced.

Then, researchers turn to the actual texts and move from means and forms of realization through strategies to content, which they see as closely associated with the context already studied (Wodak et al., 1999: 36–42). Contents, strategies and means are three analytical dimensions that are 'closely interwoven' (Wodak et al., 1999: 30) and are particularly relevant to my work here. The content dimension is straightforward, pointing to the thematic areas of the objects of study. Means and forms of realization are also easy to comprehend since they refer to the different linguistic features (or textural traits) that make up texts. In fact, in the present article, means and forms are the collocational patterns surrounding the central nodes under study.

DHA's strategies, however, require further explanation and may be classified under several labels. For the purposes of this study, I highlight the operationality of the following two for this research:

1. Nomination: 'discursive construction of social actors, objects/phenomena/events and processes/actions' (Wodak and Meyer, 2009: 94). This strategy seems to take place within the area covered by Halliday's (1985) ideational meaning and, more specifically, concerning participants and processes. It is prominent in the present study.

2. Predication: 'discursive qualification of social actors, objects, phenomena, events/processes and actions (more or less positively or negatively)' (Wodak & Meyer, 2009: 94). Adjectives and other modifiers (such as appositions, relative clauses, and prepositional phrases) are the means to convey this strategy. Predication is also in the chambers under study, though not as frequently as the previous strategy.

In sum, this article proposes a DHA-inspired examination as part of the qualitative analysis. Nevertheless, here, the order of analysis is reversed and proceeds from means and forms (in our case, collocations) to content through strategies. At the same time, and due to space constraints, the content–context connection is kept to the minimum and is left for further research.

3. Methodology

In agreement with the great interest ParlaMint assigns to the Russia-Ukraine war, this article aims to identify collocations associated with the main central nodes of "Russia"/"Ukraine" (in English) and "Rusia"/"Ucrania" (in Spanish) within the British and Spanish Chambers. This war is not solely of academic interest for ParlaMint but is one of the hottest issues in today's global world, attracting attention from an ample range of media and (economic, cultural, and societal) circles. In a way, it might be argued that it is one of those historical events that determine the standpoint of societies as a whole and individuals in particular.

To fulfil our goal, we queried ParlaMint-GB v.4.0 (with 2015 to 2022 interventions from Britain's lower chamber – the House of Commons) and ParlaMint-ES v.4.0 (with 2015 to 2022 interventions from Spain's lower chamber – the Congreso de los Diputados)(Erjavec, Kopp and Ogrodniczuk, et al.)[2]. As per lexical priming, we did this to discuss the cumulative meaning that is transferred from the co-text to the nodes in these parliamentary settings. Notice that we examined both ParlaMint-GB and ParlaMint-ES corpora in full in search of collocations rather than focus on the "war" subcorpus (containing material from 24th February 2022). This decision is explained by the fact that there has been a non-stop military conflict between the two countries from 12th April 2014 (with the war in Donbas) to now, hence almost perfectly overlapping ParlaMint's time span (2015-2022).

Collocations are generated with NoSketch Engine[3], a free concordancer prepared to query all ParlaMint corpora in a comparable fashion. Three measures were used for collocation generation: LogDice, MI, and T-score. Collocations will be sorted in descending LogDice order. Using these three measures is not only a NoSketch Engine default but also a technique to find a suitable combination of frequent and strong collocations. Following Brezina (2018: 74), statistical details about collocation generation may be found in Table 1:

Statistics name	L and R span	Minimum Collocate Frequency (NC)	Filter
Log Dice MI T-score	-5 +5	5	lemma

Table 1: Collocation statistics.

In brief, the stages of analysis were as follows:

[2] http://hdl.handle.net/11356/1860

[3] https://www.clarin.si/ske/#open

- Computerized identification of all concordances of “Russia”/“Rusia”, “Ukraine”/“Ucrania” lemmas in ParlaMint-GB and ParlaMint-ES.
- Computerized generation of collocations of “Russia”/“Rusia”, “Ukraine”/“Ucrania” lemmas in ParlaMint-GB and ParlaMint-ES.
- Selection of the top 50 collocations of both “Russia”/“Rusia” and “Ukraine”/”Ucrania” lemmas in ParlaMint-GB and ParlaMint-ES.
- Quantitative discussion of top 50 collocations of both “Russia”/“Rusia” and “Ukraine”/“Ucrania” lemmas in ParlaMint-GB and ParlaMint-ES with special reference to implications drawn from lexical priming theory.
- Qualitative discussion of the top 50 collocations of both “Russia”/“Rusia” and “Ukraine”/“Ucrania” lemmas in ParlaMint-GB and ParlaMint-ES with special reference to implications drawn from DHA.

Notice that space constraints limit the extension and depth of the analysis described here. This is why only 50 collocations are considered. Further studies will go beyond the conclusions drawn here.

4. Analysis. Russia, Ukraine; Ucrania, Rusia: Same Difference in Britain and Spain?

4.1 Preliminary Data

First, below are some of the most basic quantitative data regarding the full size of both ParlaMint-GB and ParlaMint-ES.

ParlaMint-GB: includes speeches from the House of Commons (and House of Lords) from 2015-2022 (see Table 2).

Tokens	139,686,402
Words	124,744,599
Sentences	5,323,032
Paragraphs	1,406,962
Documents	670,912

Table 2: ParlaMint-GB in figures.

We have just used speeches from the House of Commons – Britain’s lower chamber – for collocation generation to make material comparable.

ParlaMint-ES, in full, contains speeches from Spain’s lower chamber – the Spanish Congreso de los Diputados – from January 2015 to 23rd February 2023 (see Table 3).

Tokens	22,118,291
Words	19,423,835
Sentences	770,424
Paragraphs	243,994
Documents	76,351

Table 3: ParlaMint-ES in figures.

We have only queried interventions from 2015 to 2022 for collocation generation to make material comparable.

Table 4 contains data about collocation generation in ParlaMint-GB:

	Lemma Russia	Lemma: Ukraine
Number of collocations	1893	1514
Number of hits for lemma	7328	6091
Number of node/lemma hits per million	52.46	43.6
Percent of the corpus	0.005246 %	0.004360 %
Corpus size	139,686,402	139,686,402

Table 4: Collocations in ParlaMint-GB.

Table 5 contains data about collocation generation in ParlaMint-ES.

	Lemma: Rusia	Lemma: Ucrania
Number of collocates	192	399
Concordance size (number of lemma hits)	444	1181
Number of node/lemma hits per million	20.07	53.39
Percent of the whole corpus	0.002007%	0.005339%
Corpus size	22,118,291	22,118,291

Table 5: Collocations in ParlaMint-ES.

4.2 Quantitative Analysis and Priming Theory

Tables 2-5 show that the ParlaMint-GB corpus (124,744,599 words) is much larger than the ParlaMint-ES corpus (19,423,835). This size divergence is due to the fact that sessions convened in the House of Commons are much more frequent and longer than those in the Congreso de los Diputados. In effect, this means that members of the British parliament are exposed to a greater amount of linguistic data than their Spanish counterparts in general. Lexical priming inputs are bound to be greater in the former than in the latter.

When analyzing collocations, and precisely because of the difference in the size of corpora, we must now refer to comparable figures – those pointing at the number of times that nodes appear per million words or pmw. Otherwise, corpora cannot be compared on equal terms. In this case, British MPs are exposed to a greater amount of the "Russia" node (52.46 pmw) than Spanish MPs (20.07 pmw). This cumulative exposure to references implies that British MPs are bound to have a stronger (more vivid, more linguistically informed, more ingrained by frequency) image of "Russia" than the Spanish MPs. If we turn to Ukraine, we realize the situation is very different. British MPs are comparably less exposed to the "Ukraine" node (46.3 pmw) than Spanish deputies (53.39 pmw). On this occasion, the latter receive more cumulative exposure and are bound to have more ingrained perceptions in their minds.

Notice also that British parliamentarians are more exposed to mentions of "Russia" (52.46 pmw) than "Ukraine" (43.6 pmw). The difference is 8.86 points. Apart from the fact that it is quite the opposite in the Spanish Parliament, the cumulative exposure to the "Ucrania" node (53.39 pmw) more than doubles the exposure to the "Rusia" node (20.07 pmw). The gap in exposure between the two nodes in the Spanish chamber (33.32 pmw) is, thus, especially wide (in statistics, this is measured via effect size measures such as LogR: 1.91) and statistically significant (LL: 15.11; p<0.001) vis-à-vis what happens in the House of Commons.[4]

When focusing on the number of collocates that accompany and prime the nodes, higher figures are observed in ParlaMint-GB than in ParlaMint-ES. The British chamber has 1893 collocates for "Russia" and 1514 for "Ukraine". As is clear, the raw variety of potential lexical priming transfer is larger for the first node than for the second. This difference is statistically significant (LL:174.14; p<0.0001). However, the size of this raw difference (known in statistics as effect size) is virtually non-existent (LogR:0.34). On the contrary, in the Congreso de los Diputados, potential lexical priming is more intense for "Ucrania" (with 399 different collocates) than for "Rusia" (199). In this case, the collocates of "Ucrania" are more than double those of "Rusia". Resorting to statistics again, this difference is significant (LL:74.06; p<0.0001) and with a large effect size (LogR: 1.06).

In sum, when it comes to "Russia/Rusia" and "Ukraine/Ucrania", linguistic behavior quantitatively differs in both the British and Spanish chambers not only in the amount of exposure to the nodes pmw but also in the range size of collocates that are bound to impregnate these nodes. If we go beyond raw data and examine the statistics, the collocate span (or range size) difference is especially heterogeneous in the Spanish Chamber. This difference is statistically significant (LL:174.14; p<0.0001) and of a great effect size (LogR: 1.06).

At this point, only hypotheses are possible. The wide gap detected between the nodes in the Spanish house and its greater heterogeneity in collocates show less convergence in this chamber than in the British house. This recalls prior research (Calzada Pérez, 2023), which discusses other cases where the Spanish Congreso de los Diputados is shown to be more heterogeneous (and prone to contextual events) than the British House of Commons. Something like this may be happening here. Also worth noting is that the main node interest shifts from "Russia" (in ParlaMint-GB) to "Ucrania" (in ParlaMint-ES). The different mention of "Ukraine" (in ParlaMint-GB) and "Ucrania" (in ParlaMint-ES) is larger (heading towards twice the amount of difference with a LogR of 0.73) and undoubtedly significant (LL; 73.91).

4.3 Qualitative Analysis and DHA

For a qualitative analysis of collocations, we have to go beyond figures and examine them in a rather more manual fashion. Indeed, this has advantages as it allows researchers to go deeper into lexical priming (or potential meaning transfer from the collocates to the node). However, the main disadvantage of any manual work is that we need to downsize linguistic samples. For instance, it would be difficult for scholars to focus on 1893 different ParlaMint-GB collocates of "Russia". It would be even less feasible to report on this extensive work in an article with the space limitations of the present one. This is why this section reports on the top 50 collocations of "Russia" and "Ukraine" (from ParlaMint-GB) and "Rusia" and "Ucrania" (from ParlaMint-ES). These collocations are grouped in Tables 9 and 10.

These tables arrange collocates in three categories for each "Russia"/"Ukraine" node: (a) common collocates for both nodes; (b) common collocates which appear in the top 50 rank in the case of one of the nodes but not the other; and (c) specific collocates for each node.

For example, "invasion" is a top 50 collocate of (and primes) both "Russia" and "Ukraine" in ParlaMint-GB, as seen in Table 6.

Node	Freq	Coll. freq.	T-score	MI	logDice
Russia	126	1742	11.21683	10.42916	8.83039
Ukraine	360	1742	18.96966	12.21048	10.5565

Tabla 6: "Invasion" as a collocate of "Russia" and "Ukraine."

The term "China" is a top 50 collocate of "Russia" but appears in position 506 as a collocate of "Ukraine". See statistics in Table 7.

Node	Freq	Coll. freq.	T-score	MI	logDice
Russia	346	10922	18.57027	9.23809	9.27902
Ukraine	8	10922	2.66005	4.0702	3.94565

Table 7: "China", as collocate of "Russia" and "Ukraine."

[4] Statistics data are calculated using the https://ucrel.lancs.ac.uk/llwizard.html.

Finally, Table 8 shows specific collocates of “Russia” and “Ukraine.”

Coll.	Node	Fq	Coll. fq.	T-score	MI	logDice
Assad	Russia	346	10922	18.57027	9.23809	9.27902
Zelensky	Ukraine	8	10922	2.66005	4.0702	3.94565

Table 8: Specific collocates of “Russia” and “Ukraine.”

Table 9 registers the top 50 collocations of ParlaMint_GB. Notice that collocates are sorted according to Log-Dice (the higher the Log-Dice, the higher the word appears in the table).

Collocation Type	Russia Collocations	Ukraine Collocations
Common	Ukraine Putin invasion Crimea aggression invade Russia NATO illegal Russian Belarus war eastern ally attack condemn President Military weapon incursion	invasion Russia Putin russian eastern aggression war invade Crimea incursion NATO military Ukraine illegal weapon President ally attack condemn Belarus
Common but in different ranks	China sanction Today Iran threat pose Syria annexation regime Security States against influence Turkey US gas action behaviour intelligence pressure annex India disinformation Germany	territorial Sovereignty integrity defend Georgia troops unprovoked Poland sovereign border humanitarian conflict brutal situation crisis scheme Russians stand brave
Totally Specific	Assad veto resurgent Sanctions Korea Brazil	Homes flee solidarity Ukrainian grain lethal refugee visa Zelensky Ukrainians aid

Table 9: Collocations in ParlaMint-GB.

Table 9 may be analyzed in line with DHA methodology: moving from means through strategies to content. It shows that, in ParlaMint-GB, “Russia” and “Ukraine” share 20 collocates within the top 50 rank. Most contribute to nomination strategies, which characterize participants, processes, and objects. See in alphabetical order:

- States or institutions: “Belarus”, “Crimea”, “Russia”, “Ukraine”, and “NATO”.
- Human participants: “ally” and “Putin.”
- Phenomena and processes: “aggression”, “attack”, “condemn”, “incursion”, “invade”, “invasion”, “military”, “war”, “weapon.”

By way of illustration (and for reasons of space), here are only two examples of common collocates of “Russia” and “Ukraine”.

- Even in Russia , Putin’s invasion is now having disastrous consequences. (HC20220616)
- As we have heard today, the destabilization resulting from Putin’s invasion of Ukraine continues, bringing with it humanitarian crises that go way beyond the region in which we see military action. (HC20220721)

There are also 3 predication-related collocates through which participants, processes and objects are characterized: “eastern”, “illegal”, and “Russian”. Below are some examples of “eastern”:

- Although it is important that we take Russian security concerns seriously, we must resist at all costs any attempts by Russia to re-imperialize eastern Europe. (HC20220117)
- The war in eastern Ukraine drags on; the Nord Stream pipeline has been shut down; flights are being cancelled left, right and centre; and Britain is facing an unprecedented heat wave as our climate changes in front of our very eyes. (HC20220718)

In short, common collocates tend to be directly associated with the war, as is particularly clear when the focus is set on phenomena and processes, which almost all are (near) synonyms or may be placed in the same semantic realm: aggression, attack, etc. Thus, through common collocates, “Russia” and “Ukraine” are primed to be understood as contenders in the military conflict.

Many conclusions emerge when the eyes are turned to those common collocates spaced out in the ranking list. Two are especially relevant for the present paper. While “Russia” is primed by its connections to other countries, some of which are not necessarily allies of Great Britain (“China”, “Iran”, “Syria”, “Turkey”), “Ukraine” is particularly predicated with evaluative adjectives such as “brutal”, “brave”, “humanitarian”, “unprovoked”, resorting to a more affective discourse that places the node in a more friendly position.

- There is no doubt that revanchist Russia and Iran have grown closer under Putin’s leadership. (HC20220630)
- Putin’s war on Ukraine is brutal, illegal and a calculated attack on peace and stability in Europe. (HC20220224)

Though a handful, specific collocates portray a different image of both nodes. “Russia” is linked to what seems to be a lexical priming trend, through which it is connected to allies such as “[Bashar Al-] Assad”, “Korea”, or “Brazil”, in an “othering” technique, which ends up separating Russia from the West, in general, and Britain, in particular. In the meantime, “Ukraine” is primed in the opposite direction, and a different trend (among others) is spotted. This trend (see Table 9 above) shows the node as associated with Ukranian refugees that flee from a lethal war and receive Britain’s aid and solidarity through the concession of visa(s) and the application of the Homes for Ukranian scheme. As in the following example:

- This is a whole Government effort, as well as a UK-wide effort to support families and the Homes for Ukraine scheme. (HC20220620)

For its part, Table 10 registers the top 50 collocations of ParlaMint_ES. Again, notice that collocates are sorted according to Log-Dice (the higher the Log-Dice, the higher the word appears in the table).

Collocation Type	Rusia Collocations	Ucrania Collocations
Common	Ucrania invasión invadir Rusia agresión Putin guerra OTAN provocado ucraniano conflicto frontera ataque Europa	invasión guerra Rusia Putin invadir agresión provocado conflicto ataque Ucrania ucraniano frontera Europa OTAN
Common but in diffent ranks	Gas amenaza tensión Estados depender parte importar rechazar relación Unión afectar solamente Europea impacto	Consecuencia
Totally Specific	China sanción proveedor exportación Crimea India Turquía exportador dependencia pétroleo suministro agresor dependiente carbón procedente comprar natural Unidos energético 2020 demanda convertir	ruso tropa derivado enviar arma Moldavia bélico envío Georgia agravado RUSA militar pueblo refugiado armamento material desestabilización Palestina primo integridad defensivo paz crisis resistencia Embajada Bielorrusia liberado agravar desplazado ayudar criminal terrible Taiwán brutal Minsk

Table 10: Collocations in ParlaMint-GB.

As Table 10 shows, ParlaMint-ES projects a very different image of the nodes. On this occasion, what is particularly striking is that there are many more specific collocates for each node (22 for “Russia” and 35 for “Ukraine). Hence, while in the House of Commons (overlapping or spaced out) similarities are “the norm” when referring to node collocates, in the Congreso de los Diputados specificities dominate. Now, the number of collocations for each node differs strikingly, and the nature of such collocations is also idiosyncratic. This reaffirms the intuition/ hypothesis/ previous results that suggest that the House of Commons is more homogenous and stable than the Congreso de los Diputados.

The nodes “Rusia” and “Ucrania” share 14 collocates within the top 50 rank. Most contribute to nomination strategies, through which participants, processes and objects are characterized. Among them (in alphabetical order):

- States or institutions: “Rusia” [Rusia], “Ucrania” [Ukraine], “Europa” [Europe], “OTAN” [NATO],
- Human participants: “Putin”
- Phenomena and processes: “agredir” [to carry out aggression], “ataque” [attack], “agresión” [aggression], “conflict” [conflict], “frontera” [border], invadir [“invade”], invasion [“invasion]”, “guerra” [“war”], “weapon”.

By way of illustration (and for reasons of space), here are only two examples of common collocates of “Rusia” and “Ucrania”.

- El empobrecimiento de Ucrania, Europa y Rusia será la consecuencia de estas sanciones, como la propia Unión Europea ya está advirtiendo. (CD20220302)
- Hoy, en España y en Europa sufrimos economía de guerra porque a España y a Europa la guerra de Ucrania no nos es ajena.(CD20220309)

There is only 1 common predication-related collocate (the lemma “provocado” [“provoked”]), pointing at the reasons for the conflict. The examples below represent this predication: responsibility is assigned to Russia, while Ukraine is portrayed as the invasion victim. Alternatively, Spain is also seen as suffering the consequences of the war.

- El trasfondo de la subida de precios de la energía hay que buscarlo en la situación provocada de manera intencionada por Rusia para tensionar los mercados del gas y de la electricidad en la Unión Europea, con el único objetivo, señorías, de minar la recuperación económica europea. (20220316)
- Entonces reparé en el añadido del enunciado en el orden del día: para informar sobre las medidas económicas y sociales adoptadas por el Gobierno para dar respuesta a la crisis provocada por la guerra en Ucrania. (CD20221013)

The number of common collocates that rank far apart in the collocational list is now less frequent than in the case of the House of Commons. Space constraints lead us to mention this category in passing, pointing out that in the case of “Ucrania”, a particularly strong collocate is “consecuencias” [consequences].

Node	Freq	Coll. freq.	T-score	MI	logDice
Rusia	6	9308	2.37321	5.00503	4.33348
Ucrania	37	9308	6.00106	6.21814	6.85286

Table 11: “Consecuencia” as lemma collocate in ParlaMint-ES.

Like with the lemma “provocado”, the way “consecuencia” is used with “Ucrania” suggests that MPs are concerned about the impact of the war (not only on Ukraine itself) but also (especially?) on Spain.

- En definitiva, financiar políticas públicas para hacer frente a las consecuencias de la guerra de Ucrania y lograr un pacto de rentas. (CD20220913)

Specific collocates now abound and portray very different images of both nodes. With its foes and friends, “Russia” is primed as a major world economy, with a great potential impact upon the West. See the clearest collocational trend below:

- Geopolitical spaces: “Crimea”, “India”, “Turquía” [Turkey]
- Economic terms: “carbón” [coal], “comprar” [buy], “demanda” [demand], “dependencia” [dependency], “dependiente” [dependent], “energético” [energy], “exportación” [exports], “exportador” [exporter], “(gas) natural” [natural (gas)], “suministro” [supply], “petróleo” [oil], “proveedor” [supplier].

The following example provides an illustration:

- Usted ha convertido a Rusia en el tercer proveedor de gas en España. (CD20221221)

Through specific collocates, Ukraine, in turn, is reduced to its military role and linked to other, very concrete, geopolitical world regions with which the country is identified (in Spain). See the main trend below:

- Nomination collocates highlighting Ukraine’s role as war participant: “arma” [arm], “armamento” [weaponry], “desestabilización” [destabilization], “desplazado” [displaced], “liberado” [liberated], “militar” [military], “paz” [peace], “refugiado” [refugee], “resistencia” [resistance], tropa” [troop].
- Adjectival collocates with an affective value: “brutal” [brutal], “defensivo” [defensive], “liberado” [liberated], and “terrible” [terrible].
- Nomination strategies placing Ukraine in relation to friends and enemies: “Bielorrusia” [Belarus], “Palestina” [Palestine], “Taiwán” [Taiwan].

In sum, through especially nomination and predication strategies, in ParlaMint-GB, “Russia” and “Ukraine” are associated with the war through common (semi-)common, and specific collocates. However, ParlaMint-ES has a very different portrayal of “Rusia” and “Ucrania.” The former is depicted as an important economic competitor, transcending its participation in the conflict. Othering strategies are spotted in the analysis (by association with allies that are enemies or adversaries of the West). The latter is reduced to its role as the invasion victim, and the Congreso de los Diputados takes sides with it not just through affective predication but also by sharing the consequences of such an invasion with Ucrania.

5. Conclusions

This paper examines British and Spanish parliamentary discourse around the nodes “Russia” and “Ukraine” (in English) and “Rusia” and “Ucrania” (in Spanish). To do so, quantitative CADS is complemented by qualitative DHA. The results of the combination are certainly illuminating. Furthermore, CLARIN ParlaMint-GB and ParlaMint-ES are queried with free NoSketch Engine. After this study, it is advocated here that the ParlaMint constellation is a powerful tool for research into parliamentary discourse.

Concerning quantity, (some of) the lexical priming potential of both parliamentary chambers is revealed in the analysis. Quantitative raw data suggests that British MPs are more exposed to nodes and collocations. However, when looking into statistics, British deputies are seen to be particularly primed to the node "Russia". By contrast, their Spanish counterparts show greater interest in "Ucrania". The gap between exposure to both nodes is particularly wide in the Spanish Congreso de los Diputados, where "Ucrania" has double the number of hits than "Rusia". In fact, this gap difference (or effect size) between the two chambers is large enough to be mentioned here and statistically significant. Also, the range of collocates is particularly heterogeneous (with greater effect sizes) and significant in the Spanish Congreso de los Diputados. According to prior studies (Calzada Pérez, 2023), heterogeneity is a "common" feature in the Spanish Parliament and often suggests that this chamber is more exposed to context than its British equivalent. This result adds to the conclusion drawn in past studies. Yet further research is required.

Regarding qualitative results, the nature of MPs lexical priming to nodes and (common, quasi-common and specific) collocates of "Russia"/"Rusia" and "Ukraine"/"Ucrania" differs in ParlaMint-GB and ParlaMint-ES drafting two different profiles for the nodes. British MPs are primed to see "Russia" and "Ukraine" in again a more homogenous manner, as participants in a war. Spanish MPs boost "Ukraine"'s victim role and sympathize with it. In the Congreso de los Diputados, Russia is seen as a (economic and fighting) contender whose activity may have "terrible" "consequences" (to use some of the collocates discussed above) not just for "Ucrania" but also for Spain and its allies.

6. Acknowledgments

This research was carried out within the framework of the project *Original, translated and interpreted representations of the refugee cris(e)s: methodological triangulation within corpus-based discourse studies* (PID2019-108866RB-I00 / AEI / 10.13039/501100011033), funded by the Spanish Ministry of Science and Innovation.

ParlaMint-ES v. 4.0 has been compiled and marked up by María Calzada Pérez, Rubén de Líbano and Monica Albini with funding from CLARIN. Annotation and indispensable guidance and aid are due to Tomaž Erjavec and Matyáš Kopp.

Special thanks are due to Tomaž Erjavec, Maciej Ogrodniczuk and Petya Osenova for their inspiration and most efficient guidance within the ParlaMint project. Carmen Pérez Basanta read the first draft of this article and provided most valuable insights.

7. Bibliographical References

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London; New York: Continuum.

———. (2010). *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Bayley, P., Bevitori, C. and Zoni, E. (2004). Threat and Fear in Parliamentary Debates in Britain, Germany and Italy. In P. Bayley (Ed.), *Cross-Cultural Perspectives on Parliamentary Discourse*. Amsterdam; Philadelphia: John Benjamins, pp. 185–236

Bayley, P. and San Vicente, F. (2004). Ways of Talking about Work in Parliamentary Discourse in Britain and Spain'. In P. Bayley (Ed.), *Cross-Cultural Perspectives on Parliamentary Discourse*. Amsterdam; Philadelphia: John Benjamins, pp. 237–69.

Bevitori, C. (2004). Negotiating Conflict: Interruptions in British and Italian Parliamentary Debates. In P. Bayley (Ed.), *Cross-Cultural Perspectives on Parliamentary Discourse*. Amsterdam; Philadelphia: John Benjamins, pp. 87–109.

Box-Steffensmeier, J.M., Brady, H.E. Collier, D. (Eds). (2008). *The Oxford Handbook of Political Methodology*. Oxford; New York: Oxford University Press.

Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: University Printing House [Kindle version].

Bütikofer, S, and Hug. S. (2015). Strategic Behaviour in Parliament. *The Journal of Legislative Studies*, 21(3): 295–322.

Calzada Pérez, M. (2017). Corpus-Based Methods for Comparative Translation and Interpreting Studies. *Translation and Interpreting Studies*, 12(2): 231–252.

———. (2017). Five Turns of the Screw: A CADS Analysis of the European Parliament. *Journal of Language and Politics*, 16(3): 412–33.

———. (2018). Researching the European Parliament with Corpus-Assisted Discourse Studies: From the Micro- and Macro-Levels of Text to the Macro-Context. *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics,* 30(2): 465–90.

———. (2020). A Corpus-Assisted SFL Approach to Individuation in the European Parliament: The Case of Sánchez Presedo's Original and Translated Repertoires. *Meta: Journal Des Traducteurs/Meta: Translators' Journal*, 65(1): 142–67.

———. (2023). The Representation of Migration in Parliamentary Settings: Critical Cross-Linguistics Corpus-Assisted Discourse Analyses. *Humanities and Social Sciences Communications*, 10(1): online. https://doi.org/10.1057/s41599-023-02496-y.

Dibattista, D. (2004). Legitimizing and Informative Discourse in the Kosovo Debates in the British House of Commons and the Italian Chamber of Deputies. In P. Bayley (Ed.), *Cross-Cultural Perspectives on Parliamentary Discourse*. Amsterdam; Philadelphia: John Benjamins, pp. 151–84.

Erjavec, T., Ogrodniczuk, M, Osenova, P. et al. (2023). The ParlaMint Corpora of Parliamentary Proceedings. *Language Resources and Evaluation*, 57(1): 415–48.

Garzone, G. and Santulli, F. (2004). What Can Corpus Do for Critical Discourse Analysis? In A. Partington, J. Morley and L. Haarman (Eds.), *Corpora and Discourse*, Bern: Peter Lang, pp. 351–68.

Hix, S., Noury, A. and Roland, G. (2006). Dimensions of Politics in the European Parliament. *American Journal of Political Science*, 50(2): 494–520.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London; New York: Routledge.

———. 2013. Lexical Priming and Translation. In A. Kruger, K. Wallmach and J. Munday *Corpus-Based Translation Studies: Research and Applications*, London: Continuum, pp. 153–68.

McEnery, T. and Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: CUP.

Partington, A. (2004). Corpora and Discourse: A Most Congruous Beast. In A. Partington, J. Morley and L. Haarman (Eds.), *Corpora and Discourse* Bern: Peter Lang, pp.11–20.

Partington, A., Duguid, A. and Taylor, C. (Eds.), (2013). *Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS)*. Amsterdam: John Benjamins.

Skubic, J. and Fišer, D. (2022). Parliamentary Discourse Research in Sociology: Literature Review. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, edited by D. Fišer, M. Eskevich, J. Lenardič and F. de Jong, pages 81–91. Marseille, France: European Language Resources Association.

Van Eemeren, F.H. and Houtlosser, P. (2006). The Case of Pragma-Dialectics. In S. Parsons, N. Maudet, P. Moraitis, and I. Rahwan (Eds.), *Argumentation in Multi-Agent Systems*, Berlin; Heidelberg: Springer, pp. 1–28.

Vasta, N. (2004). Consent and Dissent in British and Italian Parliamentary Debates on the 1998 Gulf Crisis. In P. Bayley (Ed.), *Cross-Cultural Perspectives on Parliamentary Discourse*. Amsterdam; Philadelphia: John Benjamins, pp. 111–49.

Wodak, R., De Cillia, R. Reisigl, M. and Liebhart, K. (1999). *The Discursive Construction of National Identity*. Edinburgh: Edinburgh University Press.

Wodak, R. and Meyer, M. (2009). *Methods for Critical Discourse Analysis*. London; Thousand Oaks California: SAGE Publications Ltd.

8. Language Resource References

Erjavec, T., Kopp. M. and Ogrodniczuk, M., et al. (2023) Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 4.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1860.

Multilingual Power and Ideology Identification in the Parliament: a Reference Dataset and Simple Baselines

Çağrı Çöltekin[1], Matyáš Kopp[2], Katja Meden[3,5], Vaidas Morkevicius[4], Nikola Ljubešić[3], Tomaž Erjavec[3]

[1]University of Tübingen, Tübingen, Germany, [2]Charles University, Prague, Czech Republic,
[3]Jožef Stefan Institute, Ljubljana, Slovenia, [4]Kaunas University of Technology, Kaunas, Lithuania,
[5]Jožef Stefan International Postgraduate School, Slovenia,
ccoltekin@sfs.uni-tuebingen.de, kopp@ufal.mff.cuni.cz, katja.meden@ijs.si,
vaidas.morkevicius@ktu.lt, nikola.ljubesic@ijs.si, tomaz.erjavec@ijs.si

Abstract

We introduce a dataset on political orientation and power position identification. The dataset is derived from ParlaMint, a set of comparable corpora of transcribed parliamentary speeches from 29 national and regional parliaments. We introduce the dataset, provide the reasoning behind some of the choices during its creation, present statistics on the dataset, and, using a simple classifier, some baseline results on predicting political orientation on the left-to-right axis, and on power position identification, i.e., distinguishing between the speeches delivered by governing coalition party members from those of opposition party members.

Keywords: ideology, power, parliamentary corpus, ParlaMint

1. Introduction

Parliaments are one of the most important institutions in modern democratic states where issues with high societal impact are discussed. The decisions made in a national parliament affect the citizens of its country on fundamental aspects of their life. The societal importance of parliamentary discourse requires a better understanding and analysis of parliamentary debates. As a result, there has been a recent increase in the number of resources (Fišer and Lenardič, 2018; Lenardič and Fišer, 2023) and (computational) linguistic analyses of parliamentary debates (see Glavaš et al., 2019; Abercrombie and Batista-Navarro, 2020, for recent reviews). The impact of the decisions made in a parliament often goes beyond their borders, and may even have global effects. Hence, comparative studies of parliamentary debates across countries and in multiple languages is also important.

The dataset described here is derived from the ParlaMint corpora, a collection of comparable corpora of transcribed parliamentary speeches from 29 national and regional parliaments, covering at least the period from 2015 to 2022 (Erjavec et al., 2022). The dataset is prepared for a shared task on two important aspects of a political discourse, *political orientation* and *power* (Kiesel et al., 2024).[1] Although a simplification, political orientation on the left-to-right spectrum has been one of the defining properties of political ideology (Arian and Shamir, 1983; Vegetti and Širinić, 2019). Power is another factor that shapes the political discourse (van Dijk, 2008; Fairclough, 2013a,b). Despite its central role in critical discourse analysis, to the best of our knowledge, power was not studied computationally earlier.[2] We provide a reference dataset of parliamentary speeches for both tasks, which we expect to be instrumental for quantitative and computational studies on ideology and power in parliamentary debates beyond the present shared task as well.

Both tasks are formulated as binary classification tasks. For the power position identification task, this choice is mostly straightforward, as the distinction we want to make is between the speeches delivered by governing party members and those given by opposition party members.

Classifying political orientation is more complex, as it can be expressed in many ways. In fact, ParlaMint provides annotations from two sources (Erjavec et al., 2023b): Wikipedia and the Chapel Hill Expert Survey Europe (CHES, Jolly et al., 2022). Wikipedia classifies the political orientations of parties into 13 categories on the left-to-right spectrum, as well as five other values that do not fit into this axis (e.g., ‘Big Tent’, or ‘Single Issue Politics’ values). Conversely, CHES gives political orienta-

[1] Further practical information about the shared task can be found on the shared task web page at `https://touche.webis.de/clef24/touche24-web/ideology-and-power-identification-in-parliamentary-debates.html`.

[2] Our definition of power for the present data set is also simplified. As suggested by an anonymous reviewer, other power roles, such as being a (shadow) cabinet member, or the role in the party may manifest differently in the speech. We leave such aspect of power in speech for future research.

tion along a large number of dimensions (85 in total, e.g., stance towards European integration, but also the general left-to-right position of a party), with the numeric values based on averaged scores of expert surveys. For the left-to-right position experts assigned a numeric score between 0 to 10 (far left to far right) based on a party's general ideological stance. Not all parties have political orientation annotations in ParlaMint, but the coverage of the Wikipedia annotations is more comprehensive than that of the CHES annotations. As a result, we use orientation values from Wikipedia.

To facilitate graded predictions on the left-to-right scale, we use labels 0 for left, and 1 for right-wing parties. We mark Wikipedia categories from 'far-left' (FL) to 'centre to centre-left' (CCL) as *left*, and those from 'far-right' (FR) to 'centre to centre-right' (CCR) as *right*. We exclude the speeches from the members of the parties marked as centre and parties whose orientation does not fit into the left-to-right continuum.

For both tasks, the main challenge in the creation of a dataset is to minimize the effects of covariates. Even though the instances to classify are speeches, the annotations are based on the party membership of the speaker. As a result, underlying variables like party membership, or speaker identity perfectly covary with ideology and power in most cases. The sampling procedure described in Section 2 below aims to reduce these correlations, and encourage systems trained on the data to generalize to the particular task, rather than predictions based on easier-to-guess covariates.

ParlaMint is a multilingual dataset of transcribed speeches delivered in different regional and national parliaments. As a result, it also offers opportunities to investigate similarities and differences of ideology and power in varying cultures and parliamentary traditions, as well as their reflection in different languages. Even though the shared task does not offer a cross-lingual evaluation track, the uniformly encoded data allows participants to exploit 'universal' aspects of ideology and power through, for example, transfer learning. To encourage participation in multiple languages, and help participants build (simple) multilingual classifiers easily, we also include automatic English translations of the speeches.

Our aim in this paper is to describe the process and rationale behind the dataset construction, as well as providing an overview of the resulting data. We also describe a trivial baseline and the results of experiments with this baseline.

2. Data

The data is a subset of ParlaMint version 4.0 (Erjavec et al., 2023a). For the shared task, we split the data into training and test sets (without a fixed validation set), and share them via https://zenodo.org/records/10450640. We also provide English translations provided in the ParlaMint distribution (Kuzman et al., 2023). The main motivation for the subsampling is to reduce the effects of covariates explained above. Furthermore, since ParlaMint contains over 1.2 billion words, and more than 7.7 million speeches (more correctly 'utterances' in ParlaMint TEI annotations), sampling also results in a more manageable dataset for machine-learning experiments, promoting inclusion of participants without access to high-performance computing facilities.

Before sampling the speeches, we join the utterances by the same speaker when they were interrupted by a single utterance of another speaker, and we filter out speeches that are shorter than 500 characters, and longer than 20 000 characters. The former is intended for the inclusion of the interrupted speeches as a whole.[3] The latter, filtering by size, removes short interruptions and very long speeches. On average, the lengths of the selected speeches are between 200 and 1 000 words, approximately corresponding to speeches of 2 to 10 minutes. The utterances of the session chairs, which are typically about procedural matters, are always filtered out.

The only preprocessing steps we apply are replacing the party names or abbreviations as listed in ParlaMint with a placeholder <PARTY>, and using a <p> tag to indicate paragraph boundaries in the original transcripts. Masking the party references eliminates some trivial cues, as in 'I am speaking on behalf of <PARTY>'. We only replace the party names and abbreviations as given in ParlaMint metadata, which do not cover some of the alternative names or abbreviations of the parties, as well as (consistent) mistranslations in the automatically translated texts. We leave the rest of the named entities intact. Even though (stance towards) some of the named entities may also provide strong cues for power and ideology, many of these cues will be legitimate, and we expect the models to discover and make use of them (e.g., the stance towards a particular event, like Brexit, may genuinely stem from a speakers' relation with the government or their political orientation). Future releases of the data may improve on eliminating the obvious cues for power or ideology.

We also include the sex of the speaker, an anonymised speaker ID, and automatic translation to English in the training data. The gender information in ParlaMint was collected from var-

[3] It is common for the speeches to be interrupted by the chair, often asking the speaker to finish in the allotted time. Unauthorized interruptions from the audience are also common.

	Orientation						Power					
	Training			Test			Training			Test		
	n	L%	tokens	n	L%	tokens	n	O%	tokens	n	O%	tokens
Austria (AT)	7 879	32.6	535.4	2 002	44.7	566.6	15 971	58.8	568.1	2 181	49.0	598.5
Bosnia and Herzegovina (BA)	1 301	20.9	375.4	2 014	28.9	348.2	2 531	16.8	351.5	1 992	16.9	355.0
Belgium (BE)	2 276	32.1	403.9	2 018	38.2	378.4	4 765	47.4	397.1	1 973	47.4	398.2
Bulgaria (BG)	3 907	32.3	447.9	2 006	36.0	444.8	6 699	52.8	444.6	1 981	46.1	456.9
Czechia (CZ)	4 137	39.0	356.9	2 002	18.8	386.9	6 744	47.8	376.2	1 965	42.9	406.5
Denmark (DK)	3 069	57.1	457.2	2 015	56.6	465.7	5 493	37.2	498.8	1 971	47.4	529.7
Estonia (EE)	2 595	36.4	243.6	2 012	38.9	247.5	-	-	-	-	-	-
Spain (ES)	4 770	44.9	938.2	2 003	53.8	956.3	7 198	29.3	935.7	1 930	40.9	960.5
Catalonia (ES-CT)	2 077	46.6	915.2	2 007	47.5	921.0	1 525	34.8	896.0	1 999	35.3	904.1
Galicia (ES-GA)	943	54.1	1 072.1	2 010	58.2	1 144.2	953	42.5	1 138.0	2 000	43.5	1 164.0
Basque Country (ES-PV)	-	-	-	-	-	-	1 031	43.7	962.6	1 989	46.3	981.9
Finland (FI)	1 179	42.7	233.2	2 001	45.5	219.8	6 111	55.4	227.3	1 986	49.6	219.3
France (FR)	3 618	30.2	275.3	2 002	28.2	292.8	9 813	63.0	272.3	1 996	66.5	275.3
Great Britain (GB)	24 239	48.8	438.5	2 017	44.7	465.9	33 257	43.6	455.0	1 996	31.9	485.7
Greece (GR)	5 639	46.9	959.8	2 013	56.7	959.7	6 389	37.3	971.0	1 972	42.8	966.4
Croatia (HR)	8 322	22.8	489.7	2 016	26.9	504.2	10 741	60.3	503.9	1 989	58.8	525.8
Hungary (HU)	2 935	24.2	581.3	2 020	24.0	633.0	2 597	59.1	598.8	2 000	57.7	585.7
Iceland (IS)	536	48.0	470.0	2 015	38.3	552.5	-	-	-	-	-	-
Italy (IT)	3 367	38.3	696.5	2 014	45.8	707.4	7 848	62.5	671.7	1 971	56.8	704.5
Latvia (LV)	798	21.3	357.9	2 008	19.5	303.9	1 410	67.0	317.5	1 990	70.5	303.3
The Netherlands (NL)	5 657	38.4	502.5	2 001	37.8	473.0	7 906	58.5	484.5	1 986	59.4	500.7
Norway (NO)	10 998	50.4	457.1	2 009	40.8	475.7	-	-	-	-	-	-
Poland (PL)	5 489	11.1	356.4	2 014	16.9	359.6	9 705	45.2	329.8	2 000	46.3	340.1
Portugal (PT)	3 464	57.7	459.3	2 001	56.1	464.9	7 692	58.7	458.6	1 958	43.2	451.9
Serbia (RS)	9 914	16.1	652.9	2 015	14.1	594.5	15 114	72.9	650.4	1 990	65.7	659.2
Sweden (SE)	8 425	46.3	675.2	2 011	47.4	702.1	-	-	-	-	-	-
Slovenia (SI)	2 726	73.4	516.4	2 002	63.5	519.5	9 040	62.5	533.6	2 014	49.7	526.7
Turkey (TR)	16 138	41.8	410.3	2 008	45.7	413.7	17 384	48.6	418.5	1 990	44.5	430.3
Ukraine (UA)	2 545	16.2	232.3	2 001	14.8	242.4	11 324	68.8	224.5	2 182	35.6	233.3

Table 1: Statistics of the dataset. For each dataset, the number of speeches (n), the class imbalance (L% – the percentage of *left* for orientation, O% – the percentage of *opposition* for power), and the average number of tokens are reported.

ious sources, typically from the information provided on the web pages of the parliaments, or from Wikipedia, while in a small number of cases, the gender is unknown. Similarly, the machine translations are also not available in a small number of instances, mostly due to technical problems. The motivation for including speaker ID is to provide informed ways of dividing the available data as training and validation sets. The speaker ID is not included in the test set.

Sampling For ideal datasets for both tasks, we would need a large variation with respect to political party affiliations and speaker identities. For example, we would want multiple disjoint left-wing and right-wing political parties to be present in the training set and the test set so that the models could be evaluated for their ability to predict political orientation without relying on party affiliation. However, the nature of the ParlaMint data (in fact, any realistic corpus of parliamentary debates) prevents having such a dataset. For many parliaments, the number of political parties of a particular orientation is limited to a small number. For the power identification tasks, this is even more severe since a single party or only a few parties are in power in some countries throughout the time period covered in ParlaMint.

As a trade-off between data size, and for reducing the effect of covariates, we opt for a speaker-based sampling. First, to discourage, to some extent, the classifiers from relying on author identification, we sample maximally 20 speeches of a single speaker. This is also important for introducing variation into the dataset, as the number of speeches from each speaker follows a power-law distribution. While a small number of speakers tend to deliver most of the speeches, e.g., party or party group leaders, most speakers have relatively few speeches. The distribution of speeches or speakers to include in training and test sets is also important for proper evaluation. For the ideology task, the set of speakers in the training and test sets are disjoint. For a reasonably accurate evaluation, we set the test set size to 2 000 instances (about 100 to 200 speakers depending on the individual corpus and the task). Despite multiple speeches from each speaker, due to missing annotations and the lack of diversity of orientation in some parliaments, the disjoint training/test constraint above results in a small number of training instances, leaving a small number of instances in the training set for some of the parliaments.

Ideally, power identification requires a different constraint. That is, the same speaker should be present in both training and test sets such that speeches from one set should be when the speaker was in power, and the other set should contain the speeches while the same speaker is part of the opposition. This constraint is too difficult, or impossible, to satisfy for many parliaments in the ParlaMint data. For example, in Poland, only a single party is in power throughout the period covered by the corpus. Similarly, even when there is some variation, only a small number of speakers often serve both in governing coalitions and opposition. As a result, we use a best-effort train–test split, where if possible, we make sure that the speakers in the test set are also available in the training set with the opposite power role.[4] Otherwise, we randomly sample more speakers to obtain approximately 2 000 instances in the test set. Political systems in some countries do not have a formal coalition–opposition distinction. As a result, we leave these parliaments out of the dataset.

Statistics The procedure described above results in training sets from 28 parliaments for the ideology identification task, and 25 parliaments for the power identification task. Table 1 provides some statistics on the training and test datasets. In general, there is a varying class imbalance in both datasets, but class distribution and speech lengths between training and test sets are similar. For some parliaments, the sampling procedure results in rather small training sets. Better classification of these datasets may be achieved by techniques like cross-lingual transfer and data augmentation.

3. Baselines

The main purpose of this paper is to introduce the dataset. However, we also report results from a simple baseline which is provided for the shared task. The baseline uses TF-IDF weighted character n-gram features with a simple logistic regression classifier. The motivation for such a simple baseline is twofold. First, since it will be used as the baseline for the shared task, a competitive baseline may intimidate some of the potential participants, particularly students and early researchers. Second, since the baseline only uses ‘surface’ features, with no claim of ‘language understanding’, it also provides initial data about how much of ‘the politics is about the words’.

Table 2 presents the F1-scores of the baseline for both tasks and for all parliaments. Most scores are better than a random baseline (which would result in a 50 % F1-score). Most of the lower scores are the result of relatively high precision and low recall,[5] clearly showing the lack of hyperparameter tuning. The mild correlation between the F1-scores and the training set size (0.53 and 0.36 on orientation and power detection tasks respectively) and weak but significant correlation of the class imbalance and the scores (−0.21 and −0.16 on orientation and power detection tasks respectively) also indicate that the data size and class imbalance are important factors for the success of the present classifier. However, these are not the only sources of difficulty. Despite relatively

	Orientation		Power	
	dev	test	dev	test
AT	59.1	51.9	68.5	65.0
BA	42.4	41.6	46.0	45.9
BE	55.6	56.7	58.3	63.4
BG	53.7	53.7	61.8	64.7
CZ	54.0	51.1	59.0	62.0
DK	50.9	54.0	51.7	53.4
EE	47.5	47.4	-	-
ES	72.1	71.7	61.2	65.0
ES-CT	72.8	66.4	68.6	76.7
ES-GA	62.4	70.5	74.3	70.7
ES-PV	-	-	66.3	68.9
FI	59.4	52.6	55.9	52.1
FR	43.9	45.0	64.1	66.1
GB	75.9	74.9	74.4	70.9
GR	72.5	75.2	66.9	64.0
HR	43.8	43.2	60.2	59.4
HU	56.2	55.8	81.8	84.9
IS	41.6	46.2	-	-
IT	57.3	50.9	47.0	43.9
LV	42.8	44.6	42.0	52.3
NL	51.4	54.4	60.9	64.5
NO	60.9	63.0	-	-
PL	46.4	45.4	74.6	75.6
PT	61.7	63.7	67.5	63.4
RS	47.9	51.6	69.7	62.7
SE	75.5	75.5	-	-
SI	44.5	40.7	53.1	53.7
TR	85.8	83.6	84.4	81.9
UA	56.7	58.9	59.4	45.4

Table 2: Macro-averaged F1-scores of the baseline on (dev)elopment and test sets on all development and test sets. All scores are averages of five random splits of the provided training data as 80 % for training and 20 % for validation. The scores above were obtained without any hyperparameter tuning.

[4] The data from only three parliaments (AT, SI, UA) satisfy this constraint, while there are no speakers that changed their roles in ES-GA, HU and PL.

[5] Since F1-score favours similar precision and recall values.

large datasets, for example, AT and NO are classified rather poorly for political orientation (and also the F1-score drops substantially in the test set compared to the development set), which may be because of better separation of speakers across training and test sets. On the other hand, the success of the baseline on both tasks on TR is unlikely to be explainable by the size and the class imbalance. One can perhaps relate these to political polarization, rather than the technical reasons we list above.[6]

4. Conclusions

The paper presents a dataset derived from the ParlaMint corpora, meant for studying automatic methods for detecting political orientation and power position in parliamentary debates. We believe it could be a valuable resource for studying these phenomena and other aspects of political discourse in multiple political and parliamentary cultures/traditions, and in multiple languages. Since measuring power and ideology on an individual basis is difficult, we use the well-known sources of party orientation and power position information to label individual speeches. This introduces some strong covariates of the ideology and power in any dataset that is derived from existing resources. Instead of a more restrictive setting where covariates are more strictly eliminated, we opted for a more inclusive dataset of including many parliaments and languages. We intend to improve the existing dataset by increasing its coverage and quality and by adding more metadata.

5. Limitations

The orientation and power based on party affiliation may not always reflect the individuals' positions at the time of their speeches. However, this is unlikely to be resolved easily without restricting the number of speakers drastically. A possible solution, as suggested by an anonymous reviewer, is to do manual annotations of the individual politicians by the experts, which would definitely be costly, and may also have its own limitations, such as changing positions in time.

We did not include the centre even though it clearly falls within the left–right spectrum of political orientation. This decision was motivated by simplicity. The inclusion of a centre in a binary classification scheme is not trivial, and not all parliamentary corpora include parties annotated as centre. For the future, multi-class classification, or a form of ordinal regression/classification may be interesting alternatives against this limitation.

In the current version of the data, some procedural aspects of speech may also provide trivial, unwanted, cues for power and orientation. More rigorous identification and elimination of these cues in a big multilingual corpus is a difficult undertaking, that we leave for a potential new version of the corpus.

6. Acknowledgements

This work has been supported by CLARIN ERIC, ParlaMint: Towards Comparable Parliamentary Corpora.

7. Bibliographical References

Gavin Abercrombie and Riza Batista-Navarro. 2020. Sentiment and position-taking analysis of parliamentary debates: a systematic literature review. *Journal of Computational Social Science*, 3(1):245–270.

Asher Arian and Michal Shamir. 1983. The primarily political functions of the left-right continuum. *Comparative politics*, 15(2):139–158.

Tomaž Erjavec, Matyáš Kopp, Maciej Ogrodniczuk, Petya Osenova, Manex Agirrezabal, Tommaso Agnoloni, José Aires, Monica Albini, Jon Alkorta, Iván Antiba-Cartazo, Ekain Arrieta, Mario Barcala, Daniel Bardanca, Starkaður Barkarson, Roberto Bartolini, Roberto Battistoni, Nuria Bel, Maria del Mar Bonet Ramos, María Calzada Pérez, Aida Cardoso, Çağrı Çöltekin, Matthew Coole, Roberts Darģis, Ruben de Libano, Griet Depoorter, Sascha Diwersy, Réka Dodé, Kike Fernandez, Elisa Fernández Rei, Francesca Frontini, Marcos Garcia, Noelia García Díaz, Pedro García Louzao, Maria Gavriilidou, Dimitris Gkoumas, Ilko Grigorov, Vladislava Grigorova, Dorte Haltrup Hansen, Mikel Iruskieta, Johan Jarlbrink, Kinga Jelencsik-Mátyus, Bart Jongejan, Neeme Kahusk, Martin Kirnbauer, Anna Kryvenko, Noémi Ligeti-Nagy, Nikola Ljubešić, Giancarlo Luxardo, Carmen Magariños, Måns Magnusson, Carlo Marchetti, Maarten Marx, Katja Meden, Amália Mendes, Michal Mochtak, Martin Mölder, Simonetta Montemagni, Costanza Navarretta, Bartłomiej Nitoń, Fredrik Mohammadi Norén, Amanda Nwadukwe, Mihael Ojsteršek, Andrej Pančur, Vassilis Papavassiliou, Rui Pereira, María Pérez Lago, Stelios Piperidis, Hannes Pirker, Marilina Pisani, Henk van der Pol, Prokopis Prokopidis, Valeria

[6] A proper investigation of this is beyond the scope of the current paper. Hence this statement should only be taken as a potential future direction for research.

Quochi, Paul Rayson, Xosé Luís Regueira, Michał Rudolf, Manuela Ruisi, Peter Rupnik, Daniel Schopper, Kiril Simov, Laura Sinikallio, Jure Skubic, Lars Magne Tungland, Jouni Tuominen, Ruben van Heusden, Zsófia Varga, Marta Vázquez Abuín, Giulia Venturi, Adrián Vidal Miguéns, Kadri Vider, Ainhoa Vivel Couso, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, Rodolfo Zevallos, and Darja Fišer. 2023a. Multilingual comparable corpora of parliamentary debates ParlaMint 4.0. Slovenian language resource repository CLARIN.SI.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Vladislava Grigorova, Michał Rudolf, Andrej Pančur, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Henk van der Pol, Griet Depoorter, Jesse de Does, Bart Jongejan, Dorte Haltrup Hansen, Costanza Navarretta, María Calzada Pérez, Luciana D. de Macedo, Ruben van Heusden, Maarten Marx, Çağrı Çöltekin, Matthew Coole, Tommaso Agnoloni, Francesca Frontini, Simonetta Montemagni, Valeria Quochi, Giulia Venturi, Manuela Ruisi, Carlo Marchetti, Roberto Battistoni, Miklós Sebők, Orsolya Ring, Roberts Darģis, Andrius Utka, Mindaugas Petkevičius, Monika Briedienė, Tomas Krilavičius, Vaidas Morkevičius, Sascha Diwersy, Giancarlo Luxardo, and Paul Rayson. 2021. *Multilingual comparable corpora of parliamentary debates ParlaMint 2.1*. Slovenian language resource repository CLARIN.SI.

Tomaž Erjavec, Katja Meden, and Jure Skubic. 2023b. Adding political orientation metadata to ParlaMint corpora. In *CLARIN annual conference 2023, book of abstracts*. https://office.clarin.eu/v/CE-2023-2328_CLARIN2023_ConferenceProceedings.pdf.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darģis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2022. The ParlaMint corpora of parliamentary proceedings. *Language resources and evaluation*, 57:415–448.

Norman Fairclough. 2013a. *Critical Discourse Analysis: The Critical Study of Language*. Longman applied linguistics. Taylor & Francis.

Norman Fairclough. 2013b. *Language and Power*. Language In Social Life. Taylor & Francis.

Darja Fišer and Jakob Lenardič. 2018. CLARIN Corpora for Parliamentary Discourse Research. In *Proceedings of the LREC2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora*. European Language Resources Association. http://lrec-conf.org/workshops/lrec2018/W2/summaries/14_W2.html.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2019. Computational analysis of political texts: Bridging research efforts across communities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 18–23, Florence, Italy. Association for Computational Linguistics.

Seth Jolly, Ryan Bakker, Liesbet Hooghe, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen, and Milada Anna Vachudova. 2022. Chapel Hill Expert Survey trend file, 1999–2019. *Electoral Studies*, 75:102420.

Johannes Kiesel, Çağrı Çöltekin, Maximilian Heinrich, Maik Fröbe, Milad Alshomary, Bertrand De Longueville, Tomaž Erjavec, Nicolas Handke, Matyáš Kopp, Nikola Ljubešić, Katja Meden, Nailia Mirzhakhmedova, Vaidas Morkevičius, Theresa Reitis-Münstermann, Mario Scharfbillig, Nicolas Stefanovitch, Henning Wachsmuth, Martin Potthast, and Benno Stein. 2024. Overview of touché 2024: Argumentation systems. In *European Conference on Information Retrieval*, pages 466–473. Springer.

Taja Kuzman, Nikola Ljubešić, Tomaž Erjavec, Matyáš Kopp, Maciej Ogrodniczuk, Petya Osenova, Paul Rayson, John Vidler, Rodrigo Agerri, Manex Agirrezabal, Tommaso Agnoloni, José Aires, Monica Albini, Jon Alkorta, Iván Antiba-Cartazo, Ekain Arrieta, Mario Barcala, Daniel Bardanca, Starkaður Barkarson, Roberto Bartolini, Roberto Battistoni, Nuria Bel, Maria del Mar Bonet Ramos, María Calzada Pérez, Aida Cardoso, Çağrı Çöltekin, Matthew Coole, Roberts Darģis, Jesse de Does, Ruben de Libano, Griet Depoorter, Katrien Depuydt, Sascha Diwersy, Réka Dodé, Kike Fernandez, Elisa Fernández Rei, Francesca Frontini, Marcos Garcia, Noelia García Díaz, Pedro García Louzao, Maria Gavriilidou, Dimitris Gkoumas, Ilko Grigorov, Vladislava Grigorova, Dorte Haltrup Hansen, Mikel Iruskieta, Johan Jarlbrink, Kinga Jelencsik-Mátyus, Bart Jongejan, Neeme Kahusk, Martin Kirnbauer, Anna Kryvenko, Noémi Ligeti-Nagy, Giancarlo Luxardo, Carmen Magariños, Måns Magnusson, Carlo Marchetti, Maarten Marx, Katja Meden, Amália Mendes, Michal Mochtak, Martin Mölder, Simonetta Montemagni, Costanza Navarretta, Bartłomiej Nitoń, Fredrik Mohammadi Norén, Amanda Nwadukwe, Mihael Ojsteršek, Andrej Pančur, Vassilis Papavassiliou, Rui Pereira, María Pérez Lago, Stelios Piperidis, Hannes Pirker, Marilina Pisani, Henk van der Pol, Prokopis Prokopidis, Valeria Quochi, Xosé Luís Regueira, Michał Rudolf, Manuela Ruisi, Peter Rupnik, Daniel Schopper, Kiril Simov, Laura Sinikallio, Jure Skubic, Minna Tamper, Lars Magne Tungland, Jouni Tuominen, Ruben van Heusden, Zsófia Varga, Marta Vázquez Abuín, Giulia Venturi, Adrián Vidal Miguéns, Kadri Vider, Ainhoa Vivel Couso, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, Rodolfo Zevallos, and Darja Fišer. 2023. Linguistically annotated multilingual comparable corpora of parliamentary debates in English ParlaMint-en.ana 4.0. Slovenian language resource repository CLARIN.SI.

Jakob Lenardič and Darja Fišer. 2023. CLARIN Resource Families: Parliamentary Corpora. https://www.clarin.eu/resource-families/parliamentary-corpora, accessed on 2024-01-20.

T.A. van Dijk. 2008. *Discourse and Power*. Bloomsbury Publishing.

Federico Vegetti and Daniela Širinić. 2019. Left–right categorization and perceptions of party ideologies. *Political Behavior*, 41(1):257–280.

IMPAQTS: A Multimodal Corpus of Parliamentary and Other Political Speeches in Italy (1946-2023), Annotated with Implicit Strategies

Federica Cominetti, Lorenzo Gregori, Edoardo Lombardi Vallauri, Alessandro Panunzi
University of L'Aquila, University of Florence, University of Roma Tre, University of Florence
federica.cominetti@univaq.it, lorenzo.gregori@unifi.it, edoardo.lombardivallauri@uniroma3.it, alessandro.panunzi@unifi.it

Abstract

The paper introduces the IMPAQTS corpus of Italian political discourse, a multimodal corpus of around 2.65 million tokens including 1,500 speeches uttered by 150 prominent politicians spanning from 1946 to 2023. Covering the entire history of the Italian Republic, the collection exhibits a non-homogeneous consistency that progressively increases in quantity towards the present. The corpus is balanced according to textual and socio-linguistic criteria and includes different types of speeches. The sociolinguistic features of the speakers are carefully considered to ensure representation of Republican Italian politicians. For each speaker, the corpus contains 4 parliamentary speeches, 2 rallies, 1 party assembly, and 3 statements (in person or broadcasted). Parliamentary speeches therefore constitute the largest section of the corpus (40% of the total), enabling direct comparison with other types of political speeches. The collection procedure, including details relevant to the transcription protocols, and the processing pipeline are described. The corpus has been pragmatically annotated to include information about the implicitly conveyed questionable contents, paired with their explicit paraphrasis, providing the largest Italian collection of ecologic examples of linguistic implicit strategies. The adopted ontology of linguistic implicitness and the fine-grained annotation scheme are presented in detail.

Keywords: political discourse, multimodal corpus, pragmatic annotation, implicit content

1. The IMPAQTS corpus

1.1. Introduction

Linguistic implicit communication is a powerful means of persuasion, extensively characterizing manipulative discourse: indeed, it is used to convey deceptive content by reducing the receiver's attention to it, leading to its passive acceptance (Lombardi Vallauri, 2016a; Morency et al., 2008). This property makes linguistic implicit communication a potentially dangerous tool when it is used to influence people's choices and behaviors. The IMPAQTS project (Implicit Manipulation in Politics – Quantitatively Assessing the Tendentiousness of Speeches) is focused on this manipulative use of implicit content in political speeches: it aims to build a large multimodal corpus of Italian political discourse and annotate it per implicitly conveyed questionable content.

At the moment of writing, the corpus collection and annotation have been completed but the data processing and the building of a fully searchable web resource are still in progress.

The IMPAQTS corpus includes 1,500 speeches uttered by 150 Italian politicians throughout the history of the Italian Republic (1946-2023), totaling around 2.65 million tokens. Accordingly, the "political discourse" portrayed by the corpus is to be intended in the strict sense of "discourse by politicians", and not in the loose sense of "discourse on political issues" (Van Dijk et al., 1997). Even in its strict sense, political discourse is a wide text genre, including very different textual and communicative types ranging from interventions in the Houses of Parliament to live recordings on social media. In the IMPAQTS corpus, political speeches have been classified according to channel (in presence vs. broadcast) and addressees (institutions, supporters, general public). Only monologues have been collected, thus focusing on the most typical structure of political discourse, excluding political dialogues and conversations.

1.2. Types of Speeches

Six types of political monologue were pinpointed and included in the corpus:

- Parliamentary speech (IMPAQTS_PARL): speech given in presence, addressing institutions, typically in the Chambers or local councils. It is normally characterized by a formal or solemn register and a very high degree of planning.
- Rally: speech given in presence, addressing an audience mainly of supporters, typically during an election campaign or a public event. Degrees of formality and planning can vary; this variability is linked to the personal style of the speaker and the specific communicative situation, but it also appears highly sensitive

to diachronic variation. The rallies of the so-called "First Republic" (cf. §1.3) tend to be much more formal than the more recent ones. In any case, rallies are usually less formal and less planned than parliamentary interventions.

- Party assembly: speech given in presence, addressing an audience of party colleagues, typically during a party congress. As in the case of the rally, notable interpersonal, intertextual, and intratextual variations can be observed in the register and are further influenced by the diachronic component. Not unlike what was observed for rallies, party assemblies in the First Republic tend to stick to a formal register, while the tone in more recent assemblies can be much more informal.
- Statement in presence: speech given in presence, before an institutional and/or general audience, typically including journalists, as in the case of statements released at a press conference. They may be well-planned speeches or spontaneous declarations; the register can also be more or less formal depending on the situation.
- Broadcast statement: speech delivered for video/audio transmission, intended for the general public, as in the case of messages to the nation from the President of the Republic or the Prime Minister and self-promotional messages broadcast by politicians on television or radio; the register is often medium-formal and the degree of planning tends to be high.
- New media statement: speech recorded and/or broadcast via new media, intended for an audience mainly of followers, such as in Facebook live broadcasts; the register is normally medium, and informal traits are possible; the degree of planning is usually low.

As the descriptions show, the Italian political language represented by the IMPAQTS corpus is not a monolithic entity but portrays instances of monologic speech of medium and even informal register.

To take into account the role of personal style in the linguistic phenomena witnessed by the corpus, 10 speeches for each speaker are included, balanced according to the text-type scheme: for each speaker, the corpus includes 4 parliamentary speeches, 2 rallies, 1 party assembly, and 3 declarations. Considering this, 150 politicians were selected, totaling 1,500 speeches.

Table 1 reports the number of speakers, speeches, tokens, and words per speech type.[1]

1.3. Diachrony

The IMPAQTS corpus covers the entire history of the Italian Republic from its foundation to the year of resource release. To ease research taking into account the diachronic variable, the corpus has been divided into three sub-sections:

- the speeches delivered between June 25th 1946, the day of establishment of the republican institutions, and May 24th 1972, the closing day of the fifth legislature;
- the period from May 25th 1972 to April 14th 1994, corresponding to legislatures VI to XI;
- from April 15th 1994 to the spring of 2023, representing the legislatures from XII to XIX, still in progress at the time of project closure.

The first breaking point was set to account for the change in the themes and tones of the political debate observed in Italy in the early 1970s, in particular with the first bill for the regulation of abortion. The second breaking point corresponds to the transition from the proportional to the majoritarian electoral system, which marks in Italy a crucial turning point defined as the beginning of the so-called Second Republic. The consistency of the 3 sub-sections is not homogeneous but progressive towards contemporaneity, as shown in Table 2.

This responds to different needs. Firstly, the availability of audio-video resources (and even mere transcripts of speeches) falls dramatically the further we move away from the present. The limited availability of texts becomes even more significant if we consider the balancing between different types of political discourse described in Table 1. Secondly, the greater emphasis given to contemporaneity responds to one of the aims of the IMPAQTS project, namely the dissemination of the themes of linguistic implicitness and the education towards it. Such endeavor was reckoned to be more effective if applied to recent political texts, produced in cultural contexts better known by citizens and more impacting on their lives.

1.4. Political Orientation

For each period, well-known figures were favored, and the selection was also respectful of the composition of the parliamentary assemblies in the period considered in terms of gender and political affiliation. As a consequence of Italian political history, women are not represented in the corpus until the 60s. The whole corpus includes 340 speeches by women (23%, corresponding to 34 speakers) and 1,160 speeches by men (77%, 116 speakers). The average age of the speakers is 56. The youngest age is 27, while the oldest is 89.

[1] These numbers refer to the part of IMPAQTS corpus processed so far, i.e. 1403 speeches out of 1,500 (93.5%).

Speech Type	Speakers	Speeches	Token	Words
Parliamentary speech	150	561 (39.99%)	1,015,495	889,769 (43,11%)
Rally	150	283 (20.17%)	557,902	480,983 (23,30%)
Part assembly	137	137 (9.76%)	264,920	229,379 (11,11%)
Statement in person	133	231 (16.46%)	345,558	299,404 (14,51%)
Broadcast statement	108	164 (11.69%)	145,286	126,427 (6,13%)
New media statement	24	27 (1.92%)	44,429	37,971 (1,84%)
Total	**150**	**1403**	**2,373,590**	**2,063,933**

Table 1: IMPAQTS numbers per speech type (data derived from 93.5% of total corpus)

Period	Speakers	Speeches
1946-1972	25	88 (6.27%)
1972-1994	57	327 (23.31%)
1994-2023	124	988 (70.42%)

Table 2: Consistency of the diachronic sub-section (data derived from 93.5% of total corpus)

Orientation	Speakers	Speeches
Independent	23	153 (10.91%)
Left	28	199 (14.18%)
Center-Left	50	385 (27.44%)
Center	34	236 (16.82%)
Center-Right	38	294 (20.96%)
Right	20	136 (9.69%)

Table 3: Speeches by political orientation (data derived from 93.5% of total corpus)

The political affiliation was expressed with reference to the party to which the speaker belonged at the time of utterance. Due to the remarkable fragmentation of the Italian political history, no less than 65 different parties were included in the metadata. To ease research, the additional metadatum "political orientation" was added, including six possible values: left, center-left, center, center-right, right, independent. The distribution of the speeches according to this variable is reported in Table 3.

1.5. Multimodality

The IMPAQTS corpus was conceived as, and mainly is, a multimodal corpus. However, the ambition to cover the entire history of the Italian Republic in diachrony made it necessary to include in the corpus some speeches for which no video nor audio recording is available. Specifically, this is the case with 63 speeches, whose transcripts were found only in parliamentary stenographs or in printed publications. Numerous speeches – around 600 – are available only in audio format, which means that over 800 speeches are available in video format. Recordings were sourced from different archives, including the Chambers' web TVs and parties and politicians' YouTube channels. An invaluable source for old speeches was Radio Radicale's archive, a very large collection including not only parliamentary recordings starting from 1976 (while the Chambers' web TVs are available only from the XIV legislature, i.e., from 2001) but also a very large collection of rallies, party assemblies, broadcast messages, and press conferences, some of which dating back to the 60s.

2. Criteria for the Annotation of Implicitness

2.1. General Aims and Motivation

The IMPAQTS corpus is entirely annotated with information about the implicitly conveyed questionable contents. The collection of a large catalog of spontaneous, ecologic examples of linguistic implicit strategies in Italian is indeed one of the main aims of the project. Political discourse is a text genre particularly suitable for the collection of linguistic implicit strategies. Theoretical and experimental accounts have shown that implicit strategies have strong persuasive power, being able to reduce the critical vigilance that addressees use, as compared to when they are aware of being the target of persuasion attempts (typically, explicit). Accordingly, linguistic implicit strategies are extensively used in text genres characterized by persuasive aims, of which political discourse is a typical representative (Van Dijk, 1992, 1997; Van Dijk et al., 2000; Van Dijk, 2011; Sbisà et al., 1999; Chilton, 2005; Danler, 2005; Rocci, 2002; Charaudeau, 2005; Reisigl, 2008; Lombardi Vallauri et al., 2020; Cominetti et al., 2022, 2023).

2.2. Implicit Strategies

The model adopted for the annotation of the implicit strategies includes four main categories:

- presupposition;
- implicature;
- vagueness;
- topicalization.

In the following, the annotated categories are presented with examples extracted from the parliamentary section of the IMPAQTS corpus.

Presupposition The presupposition is an implicit strategy included in practically all analyses and taxonomies on implicitness (Bertuccelli Papi, 2009).

> To presuppose something is to take it for granted, or at least to act as if one takes it for granted, as background information - as common ground among the participants in the conversation (Stalnaker, 2002)

(1) Il rapporto tra individuo e Stato con un rafforzamento degli elementi di dialogo e di consulenza preventiva per i cittadini, con una sottolineatura del principio di irretroattività delle norme di sfavore, quindi davvero elementi di fisco amico e di *uno Stato che deve smettere non solo di essere ma anche di apparire sleale e nemico rispetto al cittadino contribuente*.
The relationship between the individual and the State with a strengthening of the elements of dialogue and preventive consultancy for citizens, with an underlining of the principle of non-retroactivity of unfavorable regulations, therefore true elements of friendly taxation and of a State that must stop not only being but also appearing disloyal and hostile towards the tax-paying citizen. [DCAP13-A1]

In (1), the change of state predicate *smettere* ("to stop") presupposes that the State is currently being disloyal and hostile towards the tax-paying citizen.

Change of state predicates (Sellars, 1954; Karttunen, 1973) is only one of many presupposition triggers pinpointed in the literature, including factive predicates (Kiparsky and Kiparsky, 1971; Karttunen, 1971), verbs of judgment (Fillmore, 1969), iteratives (Levinson, 1983), some adverbial clauses (Frege, 1892; Lombardi Vallauri, 2000, 2009), definite descriptions (Frege, 1892), etc.

Implicature Implicatures are the second cornerstone of linguistic implicitness, famously defined by Grice (1975) as propositions that can be communicated through an utterance without being explicitly said, as in (2).

(2) Dobbiamo uscire da questa crisi e dobbiamo uscirne più forti come italiani. Tutti sapete che la corsa non di un governo, ma di una lunga fase politica, durata quindici anni, è finita. *Lo dicono quei sondaggi che un tempo venivano tanto citati e oggi tanto nascosti*.
We must emerge from this crisis and we must emerge stronger as Italians. You all know that the race not of a government, but of a long political phase, which lasted fifteen years, is over. Those polls that were once so often cited and today are so hidden say so. [WVEL11-A1]

In (2), the speaker – a member of the opposition – is implying that the majority is aware of its loss of consensus and is deliberately hiding polls to conceal it. This is an example of conversational implicature, a type of implicit content arising as a consequence of the obedience in discourse to the four maxims of conversation (the Gricean Maxims, Grice 1975), which jointly express a general cooperative principle. In the specific case, the utterance in itself would violate the maxim of quantity, by apparently giving insufficient information about why polls today are hidden. The maxim is only respected if the mentioned implicature is added to the explicit content of the message.

The literature pinpoints two other types of implicature: conventional and generalized implicatures. The former arise from the use of certain expressions (often connectives and adverbs) to which they are conventionally associated. Generalized implicatures are conversational implicatures that tend to apply frequently in the same way, also in different contexts.

Vagueness Vagueness is an implicit strategy contiguous to implicatures, in that it also leaves the completion of the explicitly expressed content to addressees. More specifically, persuasive vagueness is based on the deliberate omission of a relevant detail to assure an advantage to the source (Lombardi Vallauri, 2016a,b, 2019). Typically, speakers resort to vagueness when they want to charge rivals with (often exaggerated) accusations, or when they are making (often exaggerated) promises. Vagueness can be obtained through semantic means, as in (3), or syntactic means.

(3) Qua c'è *gente* che chiacchiera di mafia ma poi se la dà a gambe quando si deve intervenire con durezza contro la mafia, eh. Qua l'antimafia dei chiacchieroni.
Here there are people *who chat about mafia but then run for the hills when it is time to intervene with rigidity against mafia, huh. Here, the big mouths' antimafia.* [MSAL20-A1]

In (3), the collective noun "gente" (*people*) is used to avoid explicitly mentioning the actual people who are supposedly responsible for the mentioned behavior.

Topicalization Finally, topicalization is a category of implicitness based on the prosodic and/or syntactic framing of some content as a topic information unit. The topic is defined in opposition to the comment, the part of the utterance that realizes the informative purpose of the utterance and conveys the utterance's illocutionary force. Not differently from presuppositions, topics tend to receive shallower processing, because they tend to encode information already active in the short-term memory of the addresses (Lombardi Vallauri, 2009; Lombardi Vallauri and Masia, 2014). Accordingly, they can be considered an implicit strategy. Specifically (and not differently from presupposition), what topicalization leaves implicit is the epistemic responsibility of the source for introducing its content. An example of a tendentious topic is presented in (4).

(4) Abbiamo compiuto un gesto vero, immaginando sensatamente di confrontarci con interlocutori veri. *Poiché siamo condotti a constatare che le cose non stanno così e che non ci si vuole più paragonare su una misura di verità*, non possiamo avere più dubbi sulla inesorabile esigenza di un gesto reciso.

We made a real gesture, sensibly imagining that we were dealing with real interlocutors. Since we are led to realize that things are not like this and that we no longer want to compare ourselves on a measure of truth, *we can no longer have doubts about the inexorable need for a decisive gesture.* [MMAR87-A1][2]

In the IMPAQTS corpus, only potentially manipulative contents are annotated. In fact, linguistic implicitness is not per se a dishonest linguistic device. On the contrary, it may be a legitimate strategy allowing for conciseness and politeness. For example, it is licit on the part of a source to presuppose that the Italian Republic exists: on the contrary, it would be uneconomical to state it explicitly. The criterion adopted to distinguish potentially manipulative from legitimate implicitness relies on the concept of *bona fide* true information, which applies to contents that any speaker can legitimately think to be shared by any other. Accordingly, the mentioned implicit strategies are annotated only when they convey non-*bona fide* true contents.

The table 4 presents the full set of pragmatic annotation classes.

2.3. Communicative Functions

Following the model proposed by Brocca et al. (2016), and Garassino et al. (2022), any implicitly conveyed questionable content is reckoned to perform some communicative function. In particular, five possible functions are identified:

- Stance-taking: conveying one's position or stand on a particular issue (Evans, 2016);
- Attack: a blast of unfavorable characteristics or flaws of a political opponent or group (Lee and Xu, 2018);
- Self-praise: a positive content about oneself or one's own (or one's allies') policy (Dayter, 2014);
- Praise to others: a positive content about other people's ideas, intentions, or deeds (Garassino et al., 2019);
- Defence: conveying one's righteousness and non-guilt (Cominetti et al., 2022).

Accordingly, implicit strategies in the corpus are tagged for the communicative function(s) they perform. For example, the conversational implicature described in (2) functions as an attack towards the majority.

3. Building and Annotating the IMPAQTS Corpus

3.1. Processing Pipeline

Even if the core part of the corpus collection and annotation are made manually by experts, a set of computational linguistics tools is used during the corpus creation process. Each video or audio source passes through the following steps:

1. Transcription of the speech source
2. Time-alignment of the transcription to the source
3. Cooperative pragmatic annotation and curation
4. Export of the XML file with annotation.

3.2. Transcription and Alignment

The spoken datum is the obvious starting point of a spoken corpus. Nonetheless, for the large part of the IMPAQTS corpus consisting of parliamentary speeches (IMPAQTS_PARL), obtaining transcripts was eased by the availability of the stenographic reports of all parliamentary sessions. For the other

[2] The typical intonation of topic in Italian can be appreciated in the corresponding audio fragment: https://www.radioradicale.it/scheda/22225?p=2&s=1528&t=1550&f=2.

Implicatures (IMPL)
Conventional implicature
Generalized implicature
Conversational (particularized) implicature
Conversational implicature by metaphor
Conversational implicature by list
Presuppositions (PPP)
Pragmatic presupposition
Semantic presupposition by definite description
Sem. pres. by restrictive relative clause
Sem. pres. by anaphoric indefinite description
Sem. pres. by adverbial subordinate clause
Sem. pres. by second term of comparison
Sem. pres. by change of state predicate
Sem. pres. by factive predicate
Sem. pres. by adverb
Sem. pres. by adjective
Sem. pres. by wh- question
Sem. pres. by alternative question
Sem. pres. by counterfactual construct
Vagueness (VAG)
Syntactically triggered vagueness
Semantically triggered vagueness
Vagueness triggered by metaphor
Topicalization (TOP)
Syntactically triggered topicalization
Prosodically triggered topicalization

Table 4: Types of implicit annotation in IMPAQTS corpus.

types of text, the speeches were automatically transcribed through the Google Speech-to-Text tool.[3] Both types of transcripts – stenographic reports and automatic transcripts – were then reviewed by at least two members of the IMPAQTS team to eliminate errors and deliberate interventions by stenographers.

Two versions of the written section of the corpus will be released: in the first one, orthographic punctuation is inserted to ease readability(see below); in the other one, prosodic breaks are inserted according to the Lablita/C-ORAL-ROM conventions (Cresti and Moneglia, 2005).

Transcribed texts are automatically aligned to their audio through Aeneas, an open-source tool[4] that performs forced alignment.

3.3. Protocol for Implicit Annotation

Pragmatic annotation is a task highly influenced by personal sensitivity and encyclopedic knowledge. In the IMPAQTS project, the protocol includes the study of the relevant literature, an ad hoc vademecum, and training by the project manager.

In the pragmatic annotation, not only are the strings of text marked with the tags corresponding to the implicit strategy and its pragmatic function, but an explicit version of the implicit content is made available (a procedure whose importance was highlighted by Sbisà 2021). The IMPAQTS protocol leads to extremely comprehensive explicitation, avoiding anaphorics and deictics to untie the implicit content completely.

Each speech is annotated by three independent annotators, one of which subsequently adopts the role of curator, comparing the three annotated versions and validating the definitive one.

To this aim, a WebAnno-MM[5] instance has been set up on a local server. WebAnno-MM is the multimodal version of the WebAnno[6] cooperative annotation tool: in addition to providing an online user-friendly annotation environment, it allows playing the video/audio segments during annotation. Submission of the transcription into HIAT-TEI format (Rehbein et al., 2004) is necessary to upload text and video for multimodal annotation. Annotation analysis and curation are also performed through the WebAnno-MM platform. At the end of this process, annotated files are exported to XMI, [7]tagged with parts of speech and lemmas with TreeTagger, [8] and converted to VRT to be further inserted in the search engine platform. After the annotation is finished, all the VRT files will be indexed and, together with the corresponding video or audio source, loaded into EMMAcorp (Cominetti et al., 2022) for linguistic searches.

Although the inter-annotator agreement has not been evaluated yet, a few main issues can be mentioned. Curators noticed that less expert annotators tend to go through a phase of "hyper-annotation", in particular when wrongly tagging as implicatures merely re-elaborated content and logical implications, or on the contrary full deductions. Implicit strategies with clear linguistic triggers (including some kinds of presupposition and vagueness and conventional implicatures) tend to show larger agreement, even if hyper-annotation may still be present due to the sometimes uncertain recognition of bona fide true content. The most difficult implicit strategy to manage seems to be topicalization, especially when only activated by prosodic cues.

The whole corpus with implicit annotation is stored in XML format. Figure 1 shows the annotation of the implicature of example 2 in section 2.2. The implicature is annotated with the tag <impl>,

[3] https://cloud.google.com/speech-to-text/

[4] https://github.com/readbeyond/aeneas

[5] https://github.com/webanno/webanno-mm

[6] https://webanno.github.io/webanno/

[7] https://www.omg.org/spec/XMI/

[8] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Sp. Type	Words	Implicits	/100Kw
Parliam.	887,965	19,538	2,200
Rally	479,053	11,602	2,422
Party ass.	229,379	4,583	1,998
Statements	462,277	8,053	1,742
Total	*2,058,674*	*43,776*	*2,126*

Table 5: Number of implicits per speech type.

Sp. Type	IMPL	PPP	VAG	TOP
Parliam.	845	873	419	221
Rally	895	868	557	191
Party ass.	616	805	515	169
Statements	609	717	386	209

Table 6: Relative frequency of implicit strategies per speech type (number of implicits per 100Kw).

along with its classification (type), its communicative function (function), and an explanation of the implicit content (comment).

3.4. Preliminary Results on Implicit Strategies and Types of Speech

Table 5 shows the number of words and implicit per speech type, along with the relative frequency of implicits, estimated per 100,000 words (last column). Table 6 reports, for each speech type, the relative frequency of the different implicit strategies: implicature (IMPL), presupposition (PPP), vagueness (VAG), and topicalization (TOP). All these numbers refer to 93.5% of the whole IMPAQTS corpus.

As Tables 5 and 6 show, IMPAQTS_PARL is above the average political discourse for global implicitness. This is due to a relatively high presence of the two most common implicit categories: implicatures (a trait shared with rallies) and presuppositions (a trait shared with rallies and party assemblies). If compared with the single most implicit political genre, rallies, parliamentary speeches prove to be significantly less vague but higher in topicalizations. On one side, this may be linked to the tendency of rallies to include many promises (a linguistic act often tending to vagueness). On the other, parliamentary speeches are the most carefully planned type of political speech, and accordingly often show elaborate syntax, in which subordinates and other circumstantial phrases may be framed as topics.

Certainly, this is merely a preliminary outline of an analysis of such data, and further elaboration would be necessary for comprehensive development.

4. Further Research

Subsections of the IMPAQTS corpus and its pragmatic annotation have already been used for the description of pragmatic phenomena, including the in-depth analysis of under-described linguistic implicit triggers (Lombardi Vallauri et al., 2021; Cominetti and Giunta, 2022), and the interaction of linguistic implicitness and different aspects of grammar (Cominetti, 2023; Cimmino and Cominetti, 2023). The large collection of texts has allowed us to extend to political discourse a kind of study that was limited so far to other text types. Once made available to the scientific community, the large size of the corpus, its diachronic and multimodal nature, and the unprecedented pragmatic annotation will certainly be useful for an array of research in all the fields of linguistics.

5. Acknowledgements

The IMPAQTS project was funded by the Italian Ministry of University and Research within the Italian national research program PRIN "IMPAQTS: Implicit Manipulation in Politics – Quantitatively Assessing the Tendentiousness of Speeches" (2020-2023, PRIN code: 2017STJCE9), coordinated by Edoardo Lombardi Vallauri from Università Roma Tre, and involving Università di Firenze as computational linguistics unit.

We are grateful to Radio Radicale[9] for allowing the offline processing of many of their audio-video documents.

6. References

Marcella Bertuccelli Papi. 2009. Implicitness. *Key notions in pragmatics*, pages 139–162.

Nicola Brocca, Davide Garassino, and Viviana Masia. 2016. Politici nella rete o nella rete dei politici? l'implicito nella comunicazione politica italiana su twitter. *PhiN-Beiheift*, 11(2016):66.

Patrick Charaudeau. 2005. *Le discours politique: les masques du pouvoir*. Vuibert.

Paul A Chilton. 2005. Manipulation, memes and metaphors. *Manipulation and ideologies in the twentieth century*, pages 15–43.

Doriana Cimmino and Federica Cominetti. 2023. Italian davvero ('really') as a trigger of implicit contents in persuasive discourse. *Journal of Pragmatics*, 211:84–95.

Federica Cominetti. 2023. Nominalization as an enhancer of linguistic implicitness in political discourse. *Lingue e Linguaggi*, 56:69–88.

[9] https://www.radioradicale.it/

```
<doc id='WVEL11-A1' parlante='Walter Veltroni'  [...]>
[...]
Dobbiamo uscire da questa crisi e dobbiamo uscirne più forti come italiani.
Tutti sapete che la corsa non di un governo, ma di una lunga fase politica,
    durata quindici anni, è finita.
<impl type='cvrs' comment='Implica che il governo [...]'  function='TT'>
Lo dicono quei sondaggi che un tempo venivano tanto citati e oggi tanto
    nascosti.
</impl>
[...]
</doc>
```

Figure 1: XML annotation of the implicit phenomenon reported in the example (2) (implicature).

Federica Cominetti, Doriana Cimmino, Claudia Coppola, Giorgia Mannaioli, and Viviana Masia. 2023. Manipulative impact of implicit communication: A comparative analysis of French, Italian and German political speeches. *Linguistik online*, 120(2):41–64.

Federica Cominetti and Giulia Giunta. 2022. Change of state and factive nominals and nominalizations as presupposition triggers. *Italian Journal of Linguistics*, 34:59–102.

Federica Cominetti, Lorenzo Gregori, Edoardo Lombardi Vallauri, Alessandro Panunzi, et al. 2022. IMPAQTS: un corpus di discorsi politici italiani annotato per gli impliciti linguistici. In *Corpora e Studi linguistici. Atti del LIV Congresso della Società di Linguistica Italiana (Online, 8–10 settembre 2021), a cura di Emanuela Cresti e Massimo Moneglia. Milano, Officinaventuno*, pages 151–164.

Emanuela Cresti and Massimo Moneglia. 2005. *C-ORAL-ROM: integrated reference corpora for spoken romance languages*. John Benjamins Publishing.

Paul Danler. 2005. Morpho-syntactic and textual realizations as deliberate pragmatic argumentative linguistic tools. *Manipulation and ideologies in the twentieth century*, pages 45–60.

Daria Dayter. 2014. Self-praise in microblogging. *Journal of Pragmatics*, 61:91–102.

Ash Evans. 2016. Stance and identity in Twitter hashtags. *Language@internet*, 13(1).

Charles J Fillmore. 1969. Verbs of judging: An exercise in semantic description. *Research on Language & Social Interaction*, 1(1):91–117.

Gottlob Frege. 1892. Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100(1):25–50.

Davide Garassino, Nicola Brocca, and Viviana Masia. 2022. Is implicit communication quantifiable? a corpus-based analysis of British and Italian political tweets. *Journal of Pragmatics*, 194:9–22.

Davide Garassino, Viviana Masia, and Nicola Brocca. 2019. Tweet as you speak: the role of implicit strategies and pragmatic functions in political communication: Data from a diamesic comparison. *RILA: Rassegna Italiana di Linguistica Applicata: 2/3*, pages 187–208.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Lauri Karttunen. 1971. Some observations on factivity. *Research on Language & Social Interaction*, 4(1):55–69.

Lauri Karttunen. 1973. Presuppositions of compound sentences. *Linguistic inquiry*, 4(2):169–193.

Paul Kiparsky and Carol Kiparsky. 1971. Fact'in semantics dd steinberg and la jakobovits, eds. *Semantics (1 971)*, pages 345–69.

Jayeon Lee and Weiai Xu. 2018. The more attacks, the more retweets: Trump's and clinton's agenda setting on Twitter. *Public Relations Review*, 44(2):201–213.

Stephen C Levinson. 1983. *Pragmatics*. Cambridge university press.

Edoardo Lombardi Vallauri. 2000. *Grammatica funzionale delle avverbiali italiane*. Carocci.

Edoardo Lombardi Vallauri. 2009. La struttura informativa: forma e funzione negli enunciati linguistici. *(No Title)*.

Edoardo Lombardi Vallauri. 2016a. Implicits as evolved persuaders. *Pragmemes and Theories of Language Use*, pages 725–748.

Edoardo Lombardi Vallauri. 2016b. The "exaptation" of linguistic implicit strategies. *SpringerPlus*, 5(1):1106.

Edoardo Lombardi Vallauri. 2019. La lingua disonesta: Contenuti impliciti e strategie di persuasione, il mulino.

Edoardo Lombardi Vallauri, Laura Baranzini, Doriana Cimmino, Federica Cominetti, Claudia Coppola, and Giorgia Mannaioli. 2020. Implicit argumentation and persuasion: A measuring model. *Journal of Argumentation in Context*, 9(1):95–123.

Edoardo Lombardi Vallauri, Federica Cominetti, and Laura Baranzini. 2021. Presupposing indefinite descriptions. *Journal of Pragmatics*, 180:173–186.

Edoardo Lombardi Vallauri and Viviana Masia. 2014. Implicitness impact: measuring texts. *Journal of Pragmatics*, 61:161–184.

Patrick Morency, Steve Oswald, and Louis De Saussure. 2008. Explicitness, implicitness and commitment attribution: A cognitive pragmatic approach. *Belgian journal of linguistics*, 22(1):197–219.

Jochen Rehbein, Thomas Schmidt, Bernd Meyer, Franziska Watzke, and Annette Herkenrath. 2004. Handbuch fur das computergestutzte transkribieren nach hiat. *Working papers in multilingualism*, 56.

Martin Reisigl. 2008. 11. rhetoric of political speeches. *Handbook of communication in the public sphere*, 4:243.

Andrea Rocci. 2002. Are manipulative texts coherent? In *New perspectives on manipulation and ideologies: theoretical aspects, Amsterdam: John Benjamins (selected papers from the conference "Manipulation in the totalitarian ideologies of the twentieth century", Monte Verità/Ascona*.

Marina Sbisà. 2021. Presupposition and implicature: Varieties of implicit meaning in explicitation practices. *Journal of Pragmatics*, 182:176–188.

Marina Sbisà et al. 1999. Ideology and the persuasive use of presupposition. In *Language and ideology. Selected papers from the 6th International Pragmatics Conference*, volume 1, pages 492–509. International Pragmatics Association Antwerp.

Wilfrid Sellars. 1954. Presupposing. *The Philosophical Review*, 63(2):197–215.

Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.

Teun A Van Dijk. 1992. Discourse and the denial of racism. *Discourse & society*, 3(1):87–118.

Teun A Van Dijk. 2011. Discourse and ideology. *Discourse studies: A multidisciplinary introduction*, pages 379–407.

Teun A Van Dijk et al. 1997. What is political discourse analysis. *Belgian journal of linguistics*, 11(1):11–52.

Teun A Van Dijk et al. 2000. New (s) racism: A discourse analytical approach. *Ethnic minorities and the media*, 37:33–49.

Teun Adrianus Van Dijk. 1997. *Discourse as social interaction*, volume 2. Sage.

ParlaMint Ngram Viewer:
Multilingual Comparative Diachronic Search Across 26 Parliaments

Asher de Jong, Taja Kuzman, Maik Larooij, Maarten Marx
Information Retrieval Lab, Informatics Institute, University of Amsterdam,
Department of Knowledge Technologies, Jožef Stefan Institute, Slovenia
asher2912@gmail.com, taja.kuzman@ijs.si, {larooij|maartenmarx}@uva.nl

Abstract

We demonstrate the multilingual search engine and Ngram viewer that was built on top of the Parlamint dataset (Erjavec et al., 2023), using the recently available translations (Kuzman et al., 2023). The user interface and SERP are carefully designed for querying parliamentary proceedings and for the intended use by citizens, journalists and political scholars. **Demo:** `https://debateabase.wooverheid.nl/`

Keywords: Multilingual Search, Parliamentary Proceedings, Ngram Viewer, Machine Translation

1. Introduction

The ParlaMint collection contains the complete parliamentary proceedings of 26 European national and regional parliaments, all in the same XML format, from the period 2015–2022 (Erjavec et al., 2023; Kuzman et al., 2023). Strong analysis tools like the Sketch Engine concordancer are available for (corpus) linguists, but access to this valuable dataset for social scientists and the general public has been lacking. So we decided to build a dedicated parliamentary search engine for ParlaMint. The availability of good quality automatic translations of all corpora to English (Kuzman et al., 2023) made it possible to develop a multilingual search and analysis tool, allowing both scholars and ordinary citizens to compare stances, opinions, and policies about a topic across different nations. We developed two integrated information systems for this data. The first entry after a query is a diachronic comparative saliency analysis tool, reminiscent of Google's Ngram viewer (Mann et al., 2014), that provides a fast and clear overview of the development of topics through time and across nations. From this in essence unordered faceted presentation of search results, the user can enter the vertical search engine yielding relevance ranked speeches given in various parliaments.

This paper describes the broad technical details, zooms in on the design choices made for the user interaction, and provides details of the automatic translation process. Our demo is available at `https://debateabase.wooverheid.nl/`, the raw data at `http://hdl.handle.net/11356/1810`, and the code for creating the demo at `https://github.com/AsherIDE/Debate-a-Base`.

Related Work With more and more easy to process parliamentary corpora becoming available, we saw several non-governmental initiatives to open up the proceedings to the general public with specialized vertical search engines e.g., Marx (2009); Beelen et al. (2017); Kaptein and Marx (2010), a process that started in 2003 with TheyWorkForYou.com in the UK. The proceedings of the European Parliament were multilingual from the early beginning, and the EuroParl corpus (Koehn, 2005) kickstarted the field of statistical machine translation. Cross-language information retrieval is an active research field since the late 1990's (Oard and Diekema, 1998) and is still very relevant today (Nie, 2022). Ngram viewers have been used to visualize and analyse temporal and comparative trends in multilingual corpus linguistics (Lin et al., 2012), psychology (Pettit, 2016), geosciences (Brandt, 2018), and political speech (de Goede et al., 2013).

2. The ParlaMint Dataset

The search engine uses the Parlamint.ana 3.0 dataset[1] (Erjavec et al., 2023) and its machine-translated English version, ParlaMint-en.ana 3.0[2] (Kuzman et al., 2023). The corpora were collected in the ParlaMint II project[3], which focused on creation and curation of parliamentary corpora from different countries in a harmonised and uniform format (Erjavec et al., 2023).

The ParlaMint 3.0 corpora include parliamentary sessions from 26 national and regional parliaments with a total of over 1.2 billion words (Erjavec et al., 2023). All corpora encompass the sessions held in the 8 years between 2015 and 2022, with many also including earlier sessions. The corpus collection consists of 27 languages; 24 in the Latin alphabet, 2 in Cyrillic (Bulgarian and Ukrainian corpus) and 1 in

[1] `http://hdl.handle.net/11356/1488`
[2] `http://hdl.handle.net/11356/1810`
[3] `https://www.clarin.eu/parlamint`

Country	Years	Speeches	EN tokens	Tokens	Speakers	Parties	Languages
Austria	27	228K	67M	66M	853	9	German
Bosnia-Hz.	25	126K	22M	18M	603	40	Bosnian
Belgium	9	199K	43M	43M	787	66	Dutch, French
Bulgaria	9	210K	30M	27M	1,033	19	Bulgarian
Czech Republic	10	181K	34M	28M	592	19	Czech
Denmark	9	399K	43M	41M	383	19	Danish
Estonia	12	228K	32M	23M	488	6	Estonian
Spain: Catalonia	8	50K	16M	16M	364	21	Catalan, Spanish
Spain: Galicia	8	83K	19M	18M	227	7	Galician
France	6	715K	47M	49M	908	26	French
Great Britain	8	671K	-	126M	1,951	2	English
Greece	8	342K	53M	50M	635	13	Greek
Croatia	20	504K	103M	88M	1,036	45	Croatian
Hungary	9	105K	35M	28M	426	9	Hungarian
Iceland	8	95K	33M	31M	261	9	Icelandic
Italy	10	173K	34M	31M	771	45	Italian
Latvia	9	163K	13M	9M	234	11	Latvian
Netherlands	9	609K	68M	68M	586	35	Dutch
Norway	25	399K	99M	89M	1,106	13	Norwegian
Poland	8	228K	44M	36M	1,223	9	Polish
Portugal	8	171K	18M	18M	723	10	Portuguese
Serbia	26	316K	99M	85M	1,724	71	Serbian
Sweden	8	85K	33M	29M	650	13	Swedish
Slovenia	23	311K	83M	70M	973	27	Slovenian
Turkiye	12	681K	63M	45M	1,346	5	Turkish
Ukraine	12	196K	23M	19M	2,192	48	Ukrainian, Russian
Total	-	7.5M	1.2B	1.2B	22K	597	27 langs

Table 1: For each corpus in the Parlamint collection: number of years, speeches, tokens in English and in the original language, number of different speakers, parties, and the languages of the proceedings. *Note*: Total English tokens represent the number of tokens in the corpora that were machine-translated into English. As the British parliamentary corpus is originally in English, it was not included in the machine-translated ParlaMint-en.ana corpus.

the Greek alphabet. Certain corpora are bilingual, such as the Belgian and Catalan corpus. The sizes of the corpora are presented in Table 1.

While the ParlaMint corpora in original languages are a very rich source of information, most users would be able to search only a small part of the corpus that is in the languages which they understand. That is why we included in the search engine the translated version as well – the ParlaMint-en.ana 3.0[4] corpus (Kuzman et al., 2023), which allows the users to browse through the entire dataset at once in one language.

The ParlaMint-en.ana 3.0 corpora (Kuzman et al., 2023) provide the English translations obtained with machine translation using the pre-trained OPUS-MT models (Tiedemann and Thottingal, 2020). These freely-available[5] Transformer-based models are based on the MarianNMT neural machine translation toolbox (Junczys-Dowmunt et al., 2018) and were trained on parallel corpora from the OPUS repository (Tiedemann, 2012). For each language, a manual evaluation of a translated sample was conducted to determine the most suitable model. The evaluations confirmed that the translations exhibited satisfactory quality. However, it is important for users of the search engine to be aware that the translations contain errors. The manual evaluation revealed incorrect translations of proper names, terms, and multi-word expressions, as well as repetitions, insertions, and incorrect translations that are unrelated to the source sentence (commonly referred to as "hallucinations" of MT systems). The search engine's interface allows users to verify the accuracy of the translations by toggling between the translated and the source text.

[4] http://hdl.handle.net/11356/1810

[5] https://github.com/Helsinki-NLP/Opus-MT

3. The Demo

The aim of the Ngram viewer (Figure 1) is to provide insight into the relative use of a phrase (ngram) through time, and to compare these temporal developments across countries: it is a diachronic comparison tool. Users can temporally zoom in on the visualization and view the relative counts also in months and even days. If the user is interested in the debates that were held at a certain day, she simply clicks in the ngram graph and is redirected to the search engine result page listing all speeches of that day relevant for the given phrase.

Users can search, read and compare debates on the debates page (Figure 2). Through this page a user can also gain insights into the actual statements that politicians made, by only having to provide the topic they are interested in. It is possible to filter on country, person, political party and date.

3.1. Interface Design

The interface design of the search engine was based on the SERP (Search Engine Result Page) design principles laid out by Hearst (Hearst, 2009). It features two main screens, the Ngram viewer (Figure 1) and the SERP combined with a document inspector (Figure 2). An important design choice was to use all the non-linguistic metadata in the collection (like name, gender, party affiliation of speakers) in the SERP. The added numbers 1–7 in Figure 2 highlight some of the design choices specially made for multilingual parliamentary search: 1: the ranked list of speeches that match the query; 2: inspection of a user-opened speech shown in the context of the surrounding debate; 3: filters for querying with the values of the used filters highlighted; 4: Debate file identifier (same #tag identifier means same debate); 5: move to next and previous speeches in the debate which are hits for the query; 6: highlighting of used search terms; 7: button to switch between the original language and English translation of the debate.

The design of the Ngram viewer is standard. To normalize counts across parliaments, it shows the fraction of speeches containing the N-gram. To reduce clutter, parliaments with few hits for a Ngram are ignored in the viewer, and the user can remove more. Users can temporally zoom in, and clicking on a line brings the user to the SERP for the Ngram as query restricted to the parliament connected to the clicked line.

3.2. Back-End Framework

The website is built with the Python based[6] Flask web framework. The search engine is built in Elasticsearch (ES) and uses the default BM25 ranking[7], but with slight tweaks to return exact string matches for Ngram queries. Normal debate speech queries only have to contain the queried word. Both the website and ES are placed in a Docker container. Following https://www.theyworkforyou.com/, we took the individual speeches as the objects, which are indexed and returned after a query.

All XML files were extracted with a variety of Python scripts that can be found on our Github repo. The complete corpus contains 7.5 million speeches. For the debates overview, all data was uploaded to ES, where one row contained one speech. In ES, one can respond to Ngram queries using phrase-queries but this turned out to be much too slow. So, in line with other Ngram viewer architectures, we simply precomputed the number of hits for each Ngram (N between 1 and 5) for each parliament, and for each day, month and year and stored these as documents in ES. With 5.8 billion different Ngrams, this did not fit into a regular ES index that has a limit of about 2.1 billion[8], so we created an index for each N.

3.3. Search Engine Evaluation

After the creation of the website, it was tested by 10 participants, with a mean age of 30 and varying educational backgrounds. Participants had to answer 7 questions using our Ngram viewer and search engine. An intervention only occurred if the participant got stuck on a question. During the experiment, participants were encouraged to think aloud constantly. From the results it became apparent that the debates page was not clear enough about which search results (speeches) belonged to the same debate. We improved the design by adding a document number next to each search result. Some participants did not realize they could jump to the speeches from the Ngrams page. To solve this, we added instructions below the visualization.

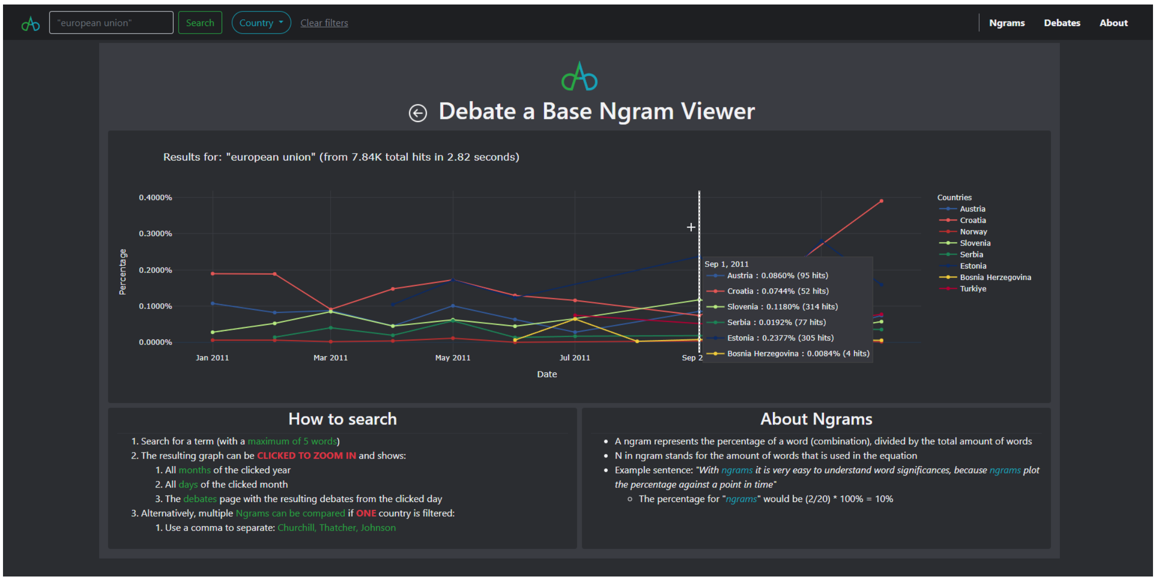
4. Conclusion

Our goal was to make the ParlaMint corpus easily available to a much wider audience, in particular to people with very little technical background. As the main aim of ParlaMint is the ability to *compare* speeches through time and across nations, we designed our interface based on that principle. The

[6] https://www.fullstackpython.com/flask.html

[7] https://www.elastic.co/guide/en/elasticsearch/reference/7.17/similarity.html

[8] https://issues.apache.org/jira/browse/LUCENE-5843



Debate a Base Ngram Viewer

Results for: "european union" (from 7.84K total hits in 2.82 seconds)

Sep 1, 2011
Austria : 0.0860% (95 hits)
Croatia : 0.0744% (52 hits)
Slovenia : 0.1180% (314 hits)
Serbia : 0.0192% (77 hits)
Estonia : 0.2377% (305 hits)
Bosnia Herzegovina : 0.0084% (4 hits)

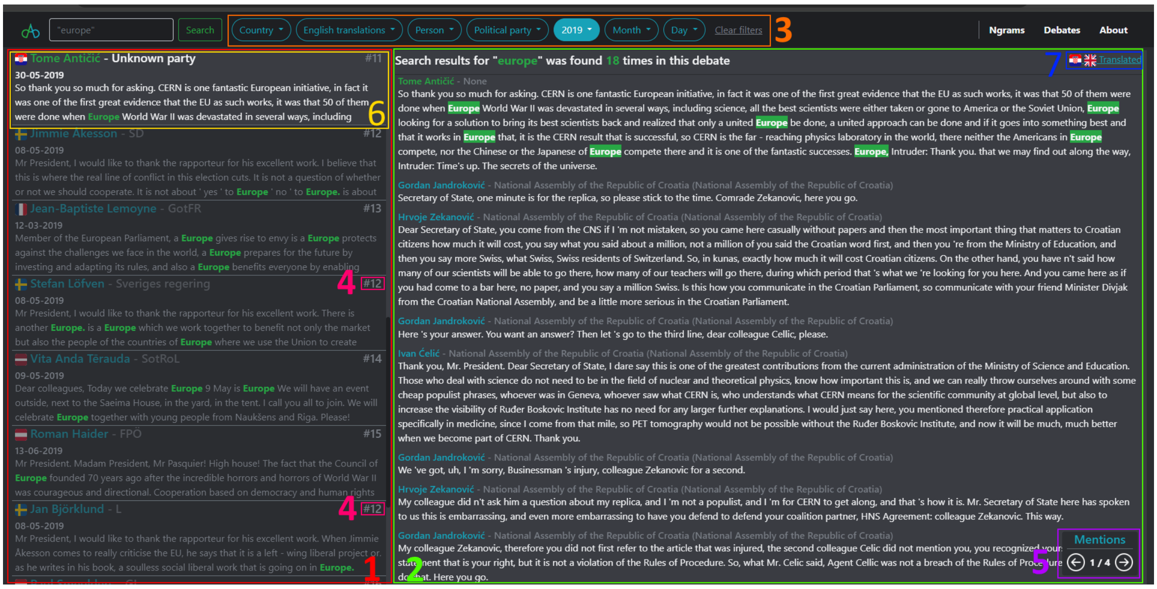
How to search

1. Search for a term (with a maximum of 5 words)
2. The resulting graph can be CLICKED TO ZOOM IN and shows:
   1. All months of the clicked year
   2. All days of the clicked month
   3. The debates page with the resulting debates from the clicked day
3. Alternatively, multiple Ngrams can be compared if ONE country is filtered:
   1. Use a comma to separate: Churchill, Thatcher, Johnson

About Ngrams

- A ngram represents the percentage of a word (combination), divided by the total amount of words
- N in ngram stands for the amount of words that is used in the equation
- Example sentence: "*With ngrams it is very easy to understand word significances, because ngrams plot the percentage against a point in time*"
  - The percentage for "*ngrams*" would be (2/20) * 100% = 10%

Figure 1: Debateabase Ngram Viewer



Tome Antičić - Unknown party #11
30-05-2019
So thank you so much for asking. CERN is one fantastic European initiative, in fact it was one of the first great evidence that the EU as such works, it was that 50 of them were done when Europe World War II was devastated in several ways, including

Jimmie Åkesson - SD #12
08-05-2019
Mr President, I would like to thank the rapporteur for his excellent work. I believe that this is where the real line of conflict in this election cuts. It is not a question of whether or not we should cooperate. It is not about ' yes ' to Europe ' no ' to Europe. is about

Jean-Baptiste Lemoyne - GotFR #13
12-03-2019
Member of the European Parliament, a Europe gives rise to envy is a Europe protects against the challenges we face in the world, a Europe prepares for the future by investing and adapting its rules, and also a Europe benefits everyone by enabling

Stefan Löfven - Sveriges regering #12
08-05-2019
Mr President, I would like to thank the rapporteur for his excellent work. There is another Europe. is a Europe which we work together to benefit not only the market but also the people of the countries of Europe where we use the Union to create

Vita Anda Tērauda - SotRoL #14
09-05-2019
Dear colleagues, Today we celebrate Europe 9 May is Europe We will have an event outside, next to the Saeima House, in the yard, in the tent. I call you all to join. We will celebrate Europe together with young people from Naukšens and Riga. Please!

Roman Haider - FPÖ #15
13-06-2019
Mr President. Madam President, Mr Pasquier! High house! The fact that the Council of Europe founded 70 years ago after the incredible horrors and horrors of World War II was courageous and directional. Cooperation based on democracy and human rights

Jan Björklund - L #12
08-05-2019
Mr President, I would like to thank the rapporteur for his excellent work. When Jimmie Åkesson comes to really criticise the EU, he says that it is a left - wing liberal project or. as he writes in his book, a soulless social liberal work that is going on in Europe.

Search results for "europe" was found 18 times in this debate

Tome Antičić - None
So thank you so much for asking. CERN is one fantastic European initiative, in fact it was one of the first great evidence that the EU as such works, it was that 50 of them were done when Europe World War II was devastated in several ways, including science, all the best scientists were either taken or gone to America or the Soviet Union, Europe looking for a solution to bring its best scientists back and realized that only a united Europe be done, a united approach can be done and if it goes into something best and that it works in Europe that, it is the CERN result that is successful, so CERN is the far - reaching physics laboratory in the world, there neither the Americans in Europe compete, nor the Chinese or the Japanese of Europe compete there and it is one of the fantastic successes. Europe, Intruder: Thank you. that we may find out along the way, Intruder: Time's up. The secrets of the universe.

Gordan Jandroković - National Assembly of the Republic of Croatia (National Assembly of the Republic of Croatia)
Secretary of State, one minute is for the replica, so please stick to the time. Comrade Zekanovic, here you go.

Hrvoje Zekanović - National Assembly of the Republic of Croatia (National Assembly of the Republic of Croatia)
Dear Secretary of State, you come from the CNS if I 'm not mistaken, so you came here casually without papers and then the most important thing that matters to Croatian citizens how much it will cost, you say what you said about a million, not a million of you said the Croatian word first, and then you 're from the Ministry of Education, and then you say more Swiss, what Swiss, Swiss residents of Switzerland. So, in kunas, exactly how much it will cost Croatian citizens. On the other hand, you have n't said how many of our scientists will be able to go there, how many of our teachers will go there, during which period that 's what we 're looking for you here. And you came here as if you had come to a bar here, no paper, and you say a million Swiss. Is this how you communicate in the Croatian Parliament, so communicate with your friend Minister Divjak from the Croatian National Assembly, and be a little more serious in the Croatian Parliament.

Gordan Jandroković - National Assembly of the Republic of Croatia (National Assembly of the Republic of Croatia)
Here 's your answer. You want an answer? Then let 's go to the third line, dear colleague Cellic, please.

Ivan Ćelić - National Assembly of the Republic of Croatia (National Assembly of the Republic of Croatia)
Thank you, Mr. President. Dear Secretary of State, I dare say this is one of the greatest contributions from the current administration of the Ministry of Science and Education. Those who deal with science do not need to be in the field of nuclear and theoretical physics, know how important this is, and we can really throw ourselves around with some cheap populist phrases, whoever was in Geneva, whoever saw what CERN is, who understands what CERN means for the scientific community at global level, but also to increase the visibility of Ruđer Boskovic Institute has no need for any larger further explanations. I would just say here, you mentioned therefore practical application specifically in medicine, since I come from that mile, so PET tomography would not be possible without the Ruđer Boskovic Institute, and now it will be much, much better when we become part of CERN. Thank you.

Gordan Jandroković - National Assembly of the Republic of Croatia (National Assembly of the Republic of Croatia)
We 've got, uh, I 'm sorry, Businessman 's injury, colleague Zekanovic for a second.

Hrvoje Zekanović - National Assembly of the Republic of Croatia (National Assembly of the Republic of Croatia)
My colleague did n't ask him a question about my replica, and I 'm not a populist, and I 'm for CERN to get along, and that 's how it is. Mr. Secretary of State here has spoken to us this is embarrassing, and even more embarrassing to have you defend to defend your coalition partner, HNS Agreement: colleague Zekanovic. This way.

Gordan Jandroković - National Assembly of the Republic of Croatia (National Assembly of the Republic of Croatia)
My colleague Zekanovic, therefore you did not first refer to the article that was injured, the second colleague Celic did not mention you, you recognized your statement that is your right, but it is not a violation of the Rules of Procedure. So, what Mr. Celic said, Agent Cellic was not a breach of the Rules of Procedure do that. Here you go.

Figure 2: Debateabase SERP; design choices highlighted by numerals 1-7.

ParlaMint corpus made corpus linguistic comparisons possible by standardizing the technical format of all debates, but it is the availability of the translations into one language that makes *comparisons on content* possible. This also opens up the corpus to a far wider group of users.

We designed our system using time-tested examples: a Google style search engine, speeches as the unit of retrieval and counting, as initiated by TheyWorkForYou, the Ngram viewer in which we can compare normalized saliency timelines across parliaments, and proven-to-work interface choices for dealing with facets and multilinguality.

The goal of the demo is really to show the richness of the ParlaMint corpus (that is why we also included e.g., the political party of a speaker and more information), and to provide a somewhat familiar manner to explore its vast possibilities. Our hope is that (the idea of) this demo is taken up by a party which can sustain it and hopefully also keep the whole corpus up to date. The demo shows that with limited computing resources and freely available software a strong prototype covering all of ParlaMint can indeed be created. The value of a corpus lies in its use. Our aim with the demo is to widen that use both to a new audience and to new types of questions asked to the ParlaMint corpus.

5. Acknowledgements

This research was supported in part by the Netherlands Organization for Scientific Research (NWO) through the ACCESS project grant CISC.CC.016 and an NWO Open Science Fund grant nr 01607400. The creation of the ParlaMint corpora was funded by CLARIN ERIC and by in-kind contributions of partners from ParlaMint I, (https://www.clarin.eu/parlamint#Partners), including POIR.04.02.00-00C002/19 and the Polish Ministry of Education and Science 2022/WK/09, programs Language resources and technologies for Slovene P6-0411 and Digital Humanities P6-0436, projects LiLaH N6-0099 and MEZZANINE J7-4642, all funded by the Slovenian Research Agency; CLaDA-BG DO01-301/17.12.21, funded by the Bulgarian Ministry of Education and Science; the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ.

6. Bibliographical References

Kaspar Beelen, Timothy Alberdingk Thijm, Christopher Cochrane, Kees Halvemaan, Graeme Hirst, Michael Kimmins, Sander Lijbrink, Maarten Marx, Nona Naderi, Ludovic Rheault, et al. 2017. Digitization of the Canadian parliamentary debates. *Canadian Journal of Political Science/Revue Canadienne de Science Politique*, 50(3):849–864.

Danita S Brandt. 2018. Charting the geosciences with Google Ngram Viewer. *GSA Today*, 5:66–67.

Bart de Goede, Justin van Wees, Maarten Marx, and Ridho Reinanda. 2013. PoliticalMashup Ngramviewer. In *Research and Advanced Technology for Digital Libraries*, pages 446–449. Springer.

Tomaz Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubesic, Kiril Simov, Andrej Pancur, Michal Rudolf, Matyás Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çagri Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevicius, Tomas Krilavicius, Roberts Dargis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fiser. 2023. The ParlaMint corpora of parliamentary proceedings. volume 57, pages 415–448.

Marti Hearst. 2009. *Search User Interfaces*. Cambridge University Press.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121.

Rianne Kaptein and Maarten Marx. 2010. Focused retrieval and result aggregation with political data. *Information retrieval*, 13:412–433.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit X: Papers*, pages 79–86.

Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the Google books Ngram corpus. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 169–174.

Jason Mann, David Zhang, Lu Yang, Dipanjan Das, and Slav Petrov. 2014. Enhanced search with wildcards and morphological inflections in the google books ngram viewer. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 115–120.

Maarten Marx. 2009. Advanced information access to parliamentary debates. *Journal of Digital Information*, 10(6).

Maarten Marx, Nelleke Aders, and Anne Schuth. 2010. Digital sustainable publication of legacy parliamentary proceedings. In *Proceedings of the 11th Annual International Digital Government Research Conference on Public Administration Online: Challenges and Opportunities*, dg.o '10, page 99–104. Digital Government Society of North America.

Jian-Yun Nie. 2022. *Cross-language information retrieval*. Springer Nature.

Douglas W Oard and Anne R Diekema. 1998. Cross-language information retrieval. *Annual Review of Information Science and Technology (ARIST)*, 33:223–56.

Michael Pettit. 2016. Historical time in the age of big data: Cultural psychology, historical change, and the Google Books Ngram Viewer. *History of psychology*, 19(2):141.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *LREC*, volume 2012, pages 2214–2218. Citeseer.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT–Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

7. Language Resource References

Erjavec, Tomaž and Kopp, Matyáš and Ogrodniczuk, Maciej and Osenova, Petya and Fišer, Darja et al. 2023. *Multilingual comparable corpora of parliamentary debates ParlaMint 3.0*. Slovenian language resource repository CLARIN.SI.

Kuzman, Taja and Ljubešić, Nikola and Erjavec, Tomaž and Kopp, Matyáš and Ogrodniczuk, Maciej et al. 2023. *Linguistically annotated multilingual comparable corpora of parliamentary debates in English ParlaMint-en.ana 3.0*. Slovenian language resource repository CLARIN.SI.

Investigating Political Ideologies through the Greek ParlaMint corpus

Maria Gavriilidou, Dimitris Gkoumas, Stelios Piperidis, Prokopis Prokopidis
ILSP / Athena RC
Artemidos 6, 15125 Marousi, Greece
{maria, dgkoumas, spip, prokopis}@athenarc.gr

Abstract

This paper has two objectives: to present (a) the creation of ParlaMint-GR, the Greek part of the ParlaMint corpora of debates in the parliaments of Europe, and (b) preliminary results on its comparison with a corpus of Greek party manifestos, aiming at the investigation of the ideologies of the Greek political parties and members of the Parliament. Additionally, a gender related comparison is explored. The creation of the ParlaMint-GR corpus is discussed, together with the solutions adopted for various challenges faced. The corpus of party manifestos, available through CLARIN:EL, serves for a comparative study with the corpus of speeches delivered by the members of the Greek Parliament, with the aim to identify the ideological positions of parties and politicians.

Keywords: parliamentary corpora, party manifestos, ideology identification

1. Introduction

Parliamentary data is considered extremely important as it contains rich linguistic content corresponding to local and international events, on political, social, economic, environmental and health issues, among others. In addition to the significance of the content, rich metadata (e.g., speaker, party affiliation, gender, role) as well as additional clues (interruptions, voting results) can often be obtained. In the field of political science, the study of political ideology and position (left-right) of members of parliament (MPs) is of great importance. For this reason, this paper aims to describe the process of assembling and encoding the ParlaMint-GR corpus as part of the ParlaMint project, and to investigate the ideology of politicians in the Greek parliament, based on two sets of corpora: the *ParlaMint-GR corpus*, on the one hand, and the *Party manifestos of Greek Parliamentary Parties corpus*, on the other. Both corpora are available through the CLARIN:EL infrastructure, while the ParlaMint-GR corpus is also available through the CLARIN.SI repository.

Section 2 describes the ParlaMint project, which provided the framework within which the ParlaMint-GR corpus was created; Section 3 elaborates on the creation of the ParlaMint-GR corpus, the solutions adopted for data acquisition, encoding and annotation; Section 4 describes the Party manifestos corpus; Section 5 presents the creation of ParlaMint-GR specific word embeddings; Section 6 discusses the experiments on the two corpora with the aim to investigate various aspects of political ideology declaration, and Section 6 concludes with future steps.

2. The ParlaMint Project

The objective of the ParlaMint project (Erjavec et al., 2022) was the creation of multilingual, comparable, and uniformly annotated corpora, following uniform collection principles, adhering to common structural and linguistic annotation principles and to a common metadata model. The first phase of the project (ParlaMint I: 2020 – 2021) produced 17 corpora, while the second phase (ParlaMint II: 2022 – 2023) resulted in corpora in 29 languages, one of them being Greek. All corpora were automatically translated into English for comparability purposes. They are hosted at the Slovenian CLARIN repository[1], accompanied by tools for querying the data (such as concordancers, corpus analysis and statistical tools, etc.) (Erjavec et al. 2023).

3. The Corpora

3.1 The ParlaMint-GR Corpus

3.1.1 Data Acquisition

The source for Greek parliamentary data acquisition was the Hellenic Parliament official site[2], the scraping of which yielded 1,263 files in total (approximately 50 MWs), which consist in approximately 350,000 speeches made by 634 members of the Parliament and corresponding to proceedings from January 2015 to February 2022[3]. Besides the speeches, the proceedings also contain transcribers' notes, which record various incidents happening during the Parliamentary meetings (e.g., notes related to time *"the meeting started at 10:00 am"* or recording voting results *"80 voted Yes and 57 voted No"* etc.), vocal non-lexicalized sounds such as shouts, laughter, etc., clapping, or any other incident affecting communication.

The Greek parliament is a unicameral parliament with a multi-party political system. The proceedings are organized in Parliamentary terms (a term is the period between two general elections). Each Parliamentary term is divided into Sessions; a parliamentary term has regular Sessions, while extraordinary and special Sessions are also foreseen. Each Session is divided

[1] https://www.clarin.si/repository/xmlui/handle/11356/1859
[2] https://www.hellenicparliament.gr/en/
[3] https://www.hellenicparliament.gr/Praktika/Synedriaseis-Olomeleias

into Meetings, and each Meeting into Sittings (multiple sittings are possible, e.g., morning/afternoon sittings).

3.1.2 Data Encoding and Metadata

The collected speeches made by members of Parliament and recorded in the proceedings were automatically processed. For every speech external metadata were provided, as well as structural and linguistic annotation. These tasks were preceded by a necessary phase of data and metadata curation, given that the minutes were not always free of typographical errors, spelling mistakes or discrepancies (e.g., in the names of members of parliament).

Dedicated metadata obtained from various sources were added for all relevant entities, i.e., government, political parties, and members of parliament (MPs). Metadata for each government, i.e. the starting and ending date of governance, Prime Minister and ministers of each government, their corresponding ministries with the relevant dates, as well as any resignations or suspensions, were obtained from the Secretariat General for Legal and Parliamentary Affairs[4], where this information is provided for all Greek governments since 1909.

For each party, the metadata include its name, its acronym, its leader, the year of establishment and the year it ceased to exist (where appropriate), whether it forms part of the government or the opposition, a link to its Wikipedia page (where additional information can be found), and finally the party's position as regards ideology and policy issues (based on Chapel Hill Expert Survey[5]). The Chapel Hill expert surveys estimate party positioning on European integration, ideology and policy issues for national parties in a variety of European countries. The first survey was conducted in 1999 (14 Western European countries), and the latest in 2019 (31 countries). The survey includes questions on various issues such as Ideology, EU integration, Specific EU Policy Questions (Agriculture, Environment, Economics, etc.), Tax policy, Welfare, immigration, position on civil rights, human rights etc., and places each party on a spectrum from far left to far right.

Information on the political party each MP belonged to during each period was acquired by scraping the Hellenic parliament official site, where a dedicated page[6] lists all Members of Parliament from 1974 until today, with their political affiliations. For additional information about each MP, such as their gender, their parliamentary roles (i.e., Prime Minister, Minister, party president, parliament president, vice-president, etc.), their government positions, and electoral districts (Dritsa, 2020)[7] was deployed, with modifications and adjustments to the original code.

For each of the 1,263 proceedings' files the following types of information were identified and annotated: term, number and date for each session, beginning and ending of each speech, speaker and his/her role (Chairperson, Regular speaker, Guest speaker). Each speech that was extracted from the Hellenic Parliament proceedings files was encoded as utterance. Each utterance was classified as being a proper Speech by an MP, or as Vocal, i.e., non-lexical vocal sounds by other MPs (shouts, laughter, etc.), as recorded in the minutes, and annotated accordingly.

The association of each speaker with the collected metadata (role, gender, political party and period) was based on Jaro-Winkler distance calculation between the speaker in question and all possible MPs in our list. The similarity threshold was set at 0.95, to avoid false positives.

3.1.3 Linguistic Annotation

Besides structural annotation and the relevant metadata, all proceedings files were automatically processed and linguistically annotated. For the linguistic processing we used the ILSP Neural NLP toolkit (Prokopidis and Piperidis, 2020), available through CLARIN:EL[8]. The toolkit integrates modules, models and lexical resources for sentence splitting, tokenization, part of speech tagging, lemmatization, dependency parsing (Universal Dependencies) and Named entity recognition, recognizing PERSON, LOCATION, ORGANIZATION, FACILITY, and GPE (Geopolitical entity). The output of the toolkit is in conllu format and underwent appropriate conversions rendering it compatible with the ParlaMint guidelines.

3.1.4 Availability

The ParlaMint-GR corpus (v4.0) is freely available, together with all the ParlaMint corpora, through the Slovenian CLARIN node[9], and through CLARIN:EL, the Greek infrastructure for Language Resources and Technologies (v3.0, 2023)[10]. A detailed description of ParlaMint-GR is found in (Gavriilidou et al. 2023).

4. The Party Manifestos Corpus

This corpus consists of 5 sub-corpora, available through the CLARIN:EL infrastructure, collected, curated and deposited by Panteion University, member of the Greek CLARIN national network[11]. These are collections of electoral manifestos, involving programmatic stances and policy positions, as stated officially by the Greek Parliamentary Parties, in the occasions of five consecutive

[4] https://gslegal.gov.gr/?page_id=776&sort=time

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

[6] https://www.hellenicparliament.gr/Vouleftes/Diatelesantes-Vouleftes-Apo-Ti-Metapolitefsi-Os-Simera/

[7] https://github.com/iMEdD-Lab/Greek_Parliament_Proceedings

[8] http://hdl.handle.net/11500/CLARIN-EL-0000-0000-67B2-3

[9] https://www.clarin.si/info/about/

[10] http://hdl.handle.net/11500/CLARIN-EL-0000-0000-7603-8

[11] https://inventory.clarin.gr/search/party%20manifestos?repository__term=Panteion%20University%20Repository

Parliamentary elections: in 2009, 2012, January and September 2015, and 2019.

The five corpora add up to a total of approximately 1,4Mb of monolingual Greek texts, in plain txt UTF-8 format, with no annotation.

5. ParlaMint-GR Embeddings

Using the open-source fastText library and the ParlaMint-GR corpus we obtained ParlaMint-GR specific embeddings. During training, and in order to get the 100-dimensional vectors, we kept all parameters to their default values. To evaluate the quality of our embeddings we queried our model for the nearest neighbours of different words.

Table 1 shows that the 3 closest words to *Mitsotakis* (the Greek Prime Minister from 2019 till now) are *Prime Minister*, *Kyriakos* (his first name) and *Tsipras* (the previous Greek Prime Minister). Respectively, for the word *Prime Minister* the closest words are *Mitsotakis and Tsipras*. Interestingly, for the word *woman* the closest ones are *man* and *mother* and *mom*. Finally, for the word KKE, acronym of a left-wing party, the most similar words are *communist, movement*, and *comunist* (wrongly spelled).

Query word	Top 3 similar words
μητσοτάκης (mitsotakis)	πρωθυπουργός (prime minister)
	κυριάκος (kyriakos)
	τσίπρας (tsipras)
γυναίκα (woman)	άντρας (man)
	μητέρα (mother)
	μάνα (mom)
πρωθυπουργός (prime minister)	πρωθυπουργός (prime minister)
	μητσοτάκης (mitsotakis)
	τσίπρας (tsipras)
κκε (kke)	κομμουνιστικό (communist)
	κίνημα (movement)
	κομουνιστικό (comunist)

Table 1: The nearest neighbours of given words as provided by word embeddings

An additional evaluation step for the produced embeddings is to assess their ability to capture analogies between words. For this, we tested our model by seeding it with the following word triplet: *PASOK* (socialist party), *Gennimata* (president of PASOK 2015-21), and *SYRIZA* (left-wing party). Our model successfully captured the hidden analogy and returned as the most probable word the term *Tsipras,* who was indeed the president of SYRIZA.

6. Experimental Investigations of Political Ideologies

Utilizing the described datasets, we conducted a number of experiments focusing on the year 2015. This was a year of special interest for Greece, due to the political turbulence: there were 2 parliamentary elections, and also the bailout referendum, the first after many decades in the country, which was due to the financial crisis and the strict economic measures imposed by the country's creditors. Finally, this year was the first time a left-wing party (SYRIZA) came into power by forming a coalition with ANEL, a right-wing party.

6.1 Similarity of Manifestos across Parties

First, we used the party manifestos dataset and calculated the cosine similarity of the texts. Using cosine similarity to retrieve similar documents is widely used in computer science and information retrieval (Lahitani et al., 2016), (Ramya et al., 2018), (Gunawan et al. 2018). Initially, we represented each text as a vector with features the frequency of each word (bag of words). By doing this, we expect to have a quantitative measure of how similar or dissimilar the programs of the various parties are. We investigate the similarity of the following parties (Table 2): ANEL and New Democracy (ND) which are both right-wing, Golden Dawn (fascist), KKE and SYRIZA (left-wing), PASOK (socialist) and POTAMI (center-left), through their manifestos for the January 2015 elections. One quite interesting observation is that SYRIZA seems to have the lowest similarity with ANEL, the party they formed a coalition with twice during this period, i.e. despite the coalition, each party kept its ideology. Among the various explanations for this coalition, the dominant one seems to be that SYRIZA considered this coalition as the only way to form a government, despite their wide ideological differences. Cosine similarity is used to determine how similar the party's manifestos are to each other, not their ideological placement. Therefore, the fact that the similarity score of ND (right-wing) and ANEL (right-wing) is lower than the one between ND (right-wing) and KKE (left-wing) suggests that ND is using a vocabulary more similar to KKE than to ANEL in their party manifesto.

Manifesto 201501	ANEL	GD	KKE	ND	PASOK	POTAMI	SYRIZA
ANEL	1	0.821	0.751	0.781	0.8	0.833	0.776
Golden Dawn	0.821	1	0.853	0.882	0.928	0.936	0.895
KKE	0.751	0.853	1	0.84	0.856	0.887	0.824
ND	0.781	0.882	0.84	1	0.858	0.889	0.867
PASOK	0.8	0.928	0.856	0.858	1	0.96	0.903
POTAMI	0.833	0.936	0.887	0.889	0.96	1	0.882
SYRIZA	0.776	0.895	0.824	0.867	0.903	0.882	1

Table 2: Party manifestos similarity using bag-of-words

Apart from the traditional bag of words representation technique, we also computed the cosine similarity of the manifestos using the centroids of their words' embeddings. Using this more recent and advanced representation method we aim to study if the initial results still hold or if the semantic representations of the words instead of the words themselves, give another insight.

Table 3 depicts the results. One initial observation is the very high similarity scores between all manifests. This denotes that all parties are topically very close to each other when it comes to their pre-election programmes (irrespectively of the solutions promised). Secondly, regarding the similarity between the manifestos of SYRIZA and ANEL (coalition government), we observe that now they have the second lowest similarity (with POTAMI being the most similar).

Manifesto 201501	ANEL	GD	KKE	ND	PASOK	POTAMI	SYRIZA
ANEL	1.000	0.983	0.963	0.941	0.963	0.974	0.953
Golden Dawn	0.983	1.000	0.968	0.966	0.966	0.981	0.970
KKE	0.963	0.968	1.000	0.932	0.930	0.945	0.927
ND	0.941	0.966	0.932	1.000	0.957	0.971	0.973
PASOK	0.963	0.966	0.930	0.957	1.000	0.992	0.990
POTAMI	0.974	0.981	0.945	0.971	0.992	1.000	0.988
SYRIZA	**0.953**	0.970	0.927	0.973	0.990	**0.988**	1.000

Table 3: Party manifestos similarity using word embeddings

6.2 MPs Speeches vs Party Manifestos

Aiming to compare the pre-election party manifestos with the post-election speeches in the Parliament, we placed each party manifesto on an ideological scale and studied the members of Parliament speeches against that scale. In order to achieve this, we use the unsupervised text scaling method wordfish (Slapin and Proksch 2008 & 2010), which has been widely used in political science to estimate party positions.

The objective of the investigation was to identify where the party members speeches in the Parliament are positioned compared to the party manifestos; in other words, whether speakers follow their parties' political stance(s) when addressing the Parliament. The first, most obvious observation (Figure 1), is that the Golden Dawn (GD) party is indeed placed on the far right. This is valid conceptually, as the Golden Dawn party is a neo-Nazi party with extremely racist discourse.

Worthwhile noticing is that, between the elections of January and September 2015, there was an ideological movement of the ANEL right-wing party towards the left, due to the government coalition with left-wing SYRIZA, possibly in an attempt to exhibit political homogeneity.

6.3 Gender-Related Observations

Investigating possible gender differences in ideology, we distinguished speeches made by male (M) and female (F) MPs of the right-wing ND party (see bottom lines in Figure 1). We see that their speeches (either M or F) are placed on the center-left ideological range, and certainly more to the left than their party's manifesto (Figure 1, 3rd line from the bottom). An interpretation for this might be that MPs, when delivering their speeches in Parliament, do not feel compelled to adhere to the right-wing discourse of their party, as attested in the respective manifesto.

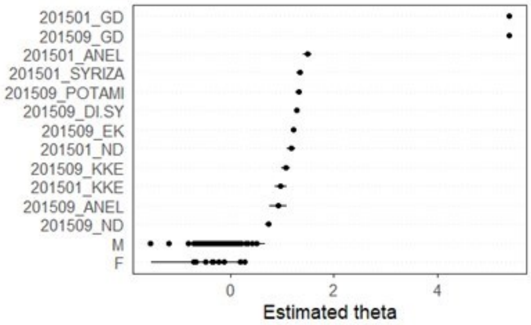


Figure 1: MPs speeches and party ideology. The Y-axis contains (from top to bottom) the parties' manifestos for the 2 elections of 2015 and at the bottom with F and M we denote the female and male MP's of ND party. The X-axis contains the ideological positions of all parties, as estimated by the wordfish method.

In the above Figure, we observe that the positioning of male and female MPs does not differ significantly, and consequently specific conclusions cannot be drawn from the specific data, concerning the position of the speeches on the ideological dimension in relation to the gender of the speaker. Since the wordfish scaling method relies on word occurrences and lexical overlap, if men and women are using similar vocabularies in their speeches, their position scores on the ideological scale will be very similar. This coincides with recent studies (Hargrave & Blumenau, 2022) reporting that the gender gap in most dimensions has narrowed in recent years.

However, it is evident from the data that the dominance of male over female MPs is still true, in numbers of MPs, and, consequently, in number of speeches. Extracting the MPs' speeches from the Parlamint-GR dataset, the descriptive statistics for the year 2015 show a clear dominance (by approximately 81%) of the Parliamentary floor by male MPs (Table 4).

Number of Speeches	
All parties	39,123
All parties Male MPs	31,604
All parties Female MPs	7,519

Table 4: Total and gender specific speech statistics for 2015

An additional snapshot of the number of speeches given by MPs of the most important political parties of the year 2015 confirms the above observation: as shown in Table 5, female MPs are drastically less heard than their male colleagues, irrespectively of political ideology. Whether fascist (F), right-wing (R), socialist (S), or left-wing (L), women stand much less on the Parliament’s podium than men: specifically, women talk 3.5 times less than men in the Greek Parliament.

Party	Gender	No speeches
ANEL (R)	F	31
	M	1011
GoldenDawn (F)	F	208
	M	1271
KKE (L)	F	321
	M	2978
ND (R)	F	1051
	M	7346
PASOK (S)	F	131
	M	1354
SYRIZA (L)	F	5363
	M	11443

Table 5: Speeches of most significant parties by gender for year 2015

7. Conclusions and Future Steps

We have presented the ParlaMint-GR corpus (its creation, the metadata used, its encoding and annotation) and the Greek party manifestos corpus. We presented some preliminary results of experiments we conducted, aiming to comparatively investigate political ideologies of parties and members of the Parliament, as expressed in these two corpora. In the immediate future we intend to broaden the scope of this study and to deal with further research questions related to political ideology as expressed in these corpora.

8. Bibliographical References

Erjavec, T. et al. (2022). The ParlaMint corpora of parliamentary proceedings. Language Resources and Evaluation. https://doi.org/10.1007/s10579-021-09574-0

Erjavec, T. et al. (2023). Multilingual comparable corpora of parliamentary debates ParlaMint 4.0. http://hdl.handle.net/11356/1859

Gavriilidou, M., Gkoumas, D., Prokopidis P., Papavassiliou, V., and Piperidis S. (2023). The ParlaMint-GR corpus: Annotated Greek Parliamentary Proceedings, in *Proceedings of the 16th International Conference on Greek Linguistics*, 14-17 December 2023, Thessaloniki, Greece.

Gunawan, Dani, C. A. Sembiring, and Mohammad Andri Budiman. (2018). "The implementation of cosine similarity to calculate text relevance between two documents." *Journal of physics: conference series*. Vol. 978. IOP Publishing.

Hargrave, L., & Blumenau, J. (2022). No longer conforming to stereotypes? Gender, political style and parliamentary debate in the UK. *British Journal of Political Science*, *52*(4), 1584-1601.

Hjorth, F. et al. (2015). Computers, coders, and voters: Comparing automated methods for estimating party positions. In *Research & Politics* 2.2 (2015): 2053168015580476.

Lahitani, Alfirna Rizqi, Adhistya Erna Permanasari, and Noor Akhmad Setiawan. (2016). Cosine similarity to determine similarity measure: Study case in online essay assessment. 4th International Conference on Cyber and IT Service Management. IEEE, 2016.

Prokopidis, P, and Piperidis, S. (2020). A Neural NLP toolkit for Greek. In 11th Hellenic Conference on Artificial Intelligence (SETN 2020).

Slapin, J.B., and Proksch S-O. (2008). A scaling model for estimating time-series party positions from texts. In *American Journal of Political Science* 52.3 (2008): 705-722.

Proksch, S-O., and Slapin, J.B. (2010). Position taking in European Parliament speeches. British Journal of Political Science 40.3 (2010): 587-611.

Ramya, R. S. et al. (2018). DRDLC: discovering relevant documents using latent dirichlet allocation and cosine similarity. In *Proceedings of the 2018 VII International Conference on Network, Communication and Computing.*

9. Language Resources References

Multilingual comparable corpora of parliamentary debates ParlaMint 3.0. CLARIN:EL http://hdl.handle.net/11500/CLARIN-EL-0000-0000-7603-8.

Party manifestos of Greek Parliamentary Parties - double Parliamentary elections 2012. CLARIN:EL http://hdl.handle.net/11500/PANTEION-0000-0000-5DF1-8.

Party manifestos of Greek Parliamentary Parties - Parliamentary elections January 2015. CLARIN:EL http://hdl.handle.net/11500/PANTEION-0000-0000-5DFE-B.

Party manifestos of Greek Parliamentary Parties - Parliamentary elections Sept. 2015. CLARIN:EL http://hdl.handle.net/11500/PANTEION-0000-0000-5E17-E.

Party manifestos of Greek Parliamentary Parties - Parliamentary elections July 2019. CLARIN:EL http://hdl.handle.net/11500/PANTEION-0000-0000-5E26-D.

ParlaMint in TEITOK

Maarten Janssen and Matyáš Kopp
Charles University, Faculty of Mathematics and Physics
Prague, Czechia
janssen,kopp@ufal.mff.cuni.cz

Abstract

This paper describes the ParlaMint 4.0 parliamentary corpora as made available in TEITOK at LINDAT. The TEITOK interface makes it possible to search through the corpus, to view each session in a readable manner, and to explore the names in the corpus. The interface does not present any new data, but provides an access point to the ParlaMint corpus that is less oriented to linguistic use only, and more accessible for the general public or researchers from other fields.

Keywords: ParlaMint, TEITOK, Document visualization

1. Introduction

ParlaMint (Erjavec et al., 2022) is "a CLARIN Flagship project which focuses on the creation of comparable and uniformly annotated corpora of parliamentary debates in Europe"[1]. The current ParlaMint 4.0 (Erjavec et al., 2023) release contains parliamentary sessions from 29 European countries and autonomous regions, with over a billion words in total. All the texts have been linguistically annotated, and adorned with bibliographical information about all speakers in all the documents. All information is encoded using the TEI/XML standard, and made publicly available via the CLARIN.SI repository[2], fully following the FAIR principles and making the data accessible for research purposes.

The data in ParlaMint are relevant for more than just linguistic research, and one could even argue that linguistic investigation is only a minor use case for the data in the repository. Yet despite being fully available in theory, making use of the ParlaMint data is not trivial. The sheer amount of data makes it difficult to get started. The linguistic annotation in the source code makes it hard to grasp the structure of the data. The TEI standard allows to encode a vast array of different information. The ParlaMint schema[3] reduces the number of elements significantly, but it still may mean that many of the structures used will not be familiar to everyone. And there are many cross-links between different files, making it even more complex to figure out what all the various elements stand for.

Apart from the repository itself, the data are also made searchable via NoSketchEngine[4] (henceforth NSE), as well as Kontext[5], which makes access a lot easier. However, NSE and Kontext are both very much designed for linguistic research. There are rich metadata about the speaker, the party he/she belongs to, etc. that make it possible to search for words, and get statistical differences in language use between parties, periods, genders, etc. But it is not that easy to just read the texts, or to see which parties are in the parliament at any given time, what topics are being discussed, or who is speaking. So for much of the potential audience of ParlaMint, NSE and Kontext are not optimal interfaces.

We attempted to provide a more generally accessible version of ParlaMint by creating a corpus out of it in the TEITOK (Janssen, 2016) corpus system. TEITOK is a corpus management system that provides linguistic search options in much the same way as NSE, but furthermore provides a document visualization system that provides an easy to read version of the documents. And TEITOK is a modular system that is maintained within LINDAT[6] (the Czech node of CLARIAH), allowing us to add dedicated functions to the system designed specifically to make various data in ParlaMint accessible. In this article, we first give a short overview of TEITOK, then describe how the ParlaMint data were put into TEITOK, and finally describe the functionality of the interface of the TEITOK version of ParlaMint, as is available at LINDAT: `https://lindat.mff.cuni.cz/services/teitok/parlamint-40/`

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

[2] `https://www.clarin.si/repository/xmlui/handle/11356/1860`

[3] `https://clarin-eric.github.io/ParlaMint/`

[4] `https://www.clarin.si/ske/`

[5] `https://www.clarin.si/kontext/`

[6] `https://lindat.mff.cuni.cz/`

2. TEITOK

TEITOK is an online corpus platform that combines various corpus tasks into a single platform. Each document in TEITOK is a TEI/XML file. There are various modules that can visualize these TEI/XML files depending on their content. There are modules to edit the content of the TEI/XML files by running NLP tasks over them or performing manual annotations and corrections. And the system can create a searchable corpus out of the set of TEI/XML files. Searching the corpus will render an XML fragment, that will be linked back to the source XML.

TEITOK is an open source repository[7], that is designed to be installed locally. It is intended as a non-intrusive tool that can be customized to the style of the organization or project where it is used rather than impose its own style. Each TEITOK corpus is an independent folder, and can be fully customized. It has been used in a wide range of different projects with installations in various universities around the world[8].

TEITOK has a modular design that makes it easy to create additional modules for custom visualization of documents, or for providing additional information taken from sources other than the corpus documents. There is an ever growing number of modules to work with different types of corpus documents. There are for instance modules to work with manuscript corpora with alignment to facsimile images (Janssen, 2018a). There are modules to work with audio or video corpora with alignment between the audio and the transcription (Janssen, 2021). And there are modules to work with dependency parsed corpora that can visualize dependency trees (Janssen, 2018b).

For the corpus search, by default TEITOK uses the Corpus WorkBench (CWB) (Evert and Hardie, 2011), but it is also possible to use other search engines, including dependency based search languages such as PML-TQ[9] or Grew (Guillaume, 2019), or it is possible to let external tools like Kontext (Machálek, 2020) handle the search (Janssen, 2020). For NLP tasks, the default in TEITOK is to use UDPIPE[10].

TEITOK has been used in many different projects in different universities. At LINDAT, it is used to gradually provide a search interface to all data in the repository, and it is the primary tool for creation and deployment for many new projects. One of the projects made available in TEITOK at LINDAT is ParCzech (Kopp et al., 2021; Kopp, 2024), the Czech parliamentary corpus that forms the basis of the Czech subcorpus of ParlaMint. The experience with ParCzech was one of the main motivations for creating the version of ParlaMint in TEITOK.

3. Converting ParlaMint to TEITOK

TEITOK documents are stored in tokenized TEI/XML format, and so are the files of ParlaMint. So in principle, creating a TEITOK version of the ParlaMint corpus is easy. However, there are differences in the way TEI is used in the two projects. ParlaMint uses an adaptation of the Parla-CLARIN guidelines[11], while TEITOK is designed to work with almost any kind of TEI, but with a limited number of constructions that cannot be used, some deviation from pure TEI, and some prefered constructions that differ from those used in Parla-CLARIN. Therefore, the documents cannot be used directly, but some minor conversions are needed.

Because of ParCzech in TEITOK, much of the conversion was already in place, but still needed to be adapted for ParlaMint. Firstly, ParCzech contains not only the transcription but also the audio recording. Secondly, because there are differences between the different ParlaMint subcorpora that were not accounted for by the scripts. And thirdly, because we needed the TEITOK version of ParlaMint to follow some of the decisions made in the NSE version for consistency.

The conversion consists, apart from some trivial naming differences, in providing information locally as much as possible, rather than distributed as it is in Parla-CLARIN. ParlaMint uses a central repository of names (per subcorpus) and each utterance is linked to a person. People can have multiple affiliations over time, and even multiple names. This is very good for complex cases and for political correctness, but not very helpful for giving a unique answer about the name of a person in a search or in mouse-over information. And the same holds to a lesser extent for information like the dependency relations and the (chronological) order of the transcriptions.

For the conversion, the data corresponding to the repository record were downloaded onto the server, where a local script did all the necessary conversions, subcorpus by subcorpus and one file at a time[12]. The conversion was done with pre-final

[7] https://gitlab.com/maartenes/TEITOK

[8] http://www.teitok.org/index.php?action=projects

[9] https://ufal.mff.cuni.cz/pmltq

[10] https://lindat.mff.cuni.cz/services/udpipe/

[11] https://clarin-eric.github.io/parla-clarin/

[12] Conversions and editing in TEITOK are typically done via the interface, or via the API for larger conversions. But in the case of ParlaMint, no editing is needed since the TEITOK corpus is used just as an interface for existing data and not the primary data source itself. And given the sheer size of the ParlaMint corpus, even the API is too slow for the amount of processing needed.

releases of the corpus, so that the TEITOK corpus could be ready at the time of the launch of ParlaMint. This meant that the conversion had to be rerun due to some last minute corrections in ParlaMint, but also that inconsistencies could be communicated back to the project. And it means that the scripts are streamlined and can be easily used to convert possible future updates of the corpus.

The script compiles the person data from the tab separated text file included in the repository, which was compiled for the NSE version of ParlaMint. The reason for using this compiled file instead of the raw source data is not only that it avoids having to account for complex cases and possible inconsistencies since they have already been dealt with; but also because that way, all decisions will coincide with those taken for the NSE corpus, so that people doing the same search in the different versions of the corpus will (as far as possible) get the same answers. The script also places the dependency information directly on the tokens following the style of CoNLL-U, and introduces pagination markers to be able to display reasonably sized parts of the transcription files in the browser.

4. The TEITOK ParlaMint Interface

4.1. Subcorpora

The ParlaMint corpus in TEITOK is divided into a separate subcorpus for each of the parliamentary sub-parts of ParlaMint. Therefore, the user first has to select the subcorpus he wants to consult. For convenience, the sub-corpora are not only presented as a list, but also shown on a map of Europe, following the CLARIN map style. The subcorpus select page is shown in Figure 1, which is not only convenient, but also gives a quick view of which European countries are included in the ParlaMint release.

Selecting a subcorpus brings you to the landing page of that subcorpus, which is a static page that combines a number of different data about the subcorpus. Let us use CZ as an example:

- It provides the description of the subcorpus, as included in the repository as the README for that subcorpus, converted from MD to HTML - so the contents on README-CZ.md

- It lists the source(s) used for the compilation of the subcorpus, as listed in the `<sourceDesc>` of ParlaMint-CZ.ana.xml, with the title, the link to the source, and the begin and end date used for ParlaMint

- It lists the location(s) where the parliamentary sittings took place, as listed in the `<settingsDesc>` of ParlaMint-CZ.ana.xml

- It lists all the people listed as responsible for the creation of the subcorpus with their role, as listed in the `<respStmt>` of ParlaMint-CZ.ana.xml

This way, the interface makes various types of information that should be pertinent for the use and attribution of the corpus more visible than they are in the repository or the project site.

4.2. Search and Browse

The search provided in the ParlaMint project in TEITOK uses CWB. The search and statistics options are very similar to other interfaces based on the CWB Query Language (CQL) with mostly visualation differences. That inludes the NSE and Kontext interfaces to ParlaMint mentioned earlier, but also the Polish Parliamentary Corpus (PPC)[13], the Plenary Sessions of the Parliament of Finland[14] and many others.

Where the TEITOK interface differs is that from the search result (KWIC list) it brings you to the full document visualization with the matching word highlighted. Many search engine, like Kontext, provide a limited, mostly plain text context or do not provide any larger context at all (often intentionally). And for instance PPC does provide a link to the full context for each search result, but the context is provided as a PDF document with no indication where in the text the matching segment of text can be found.

Another option in the TEITOK interface of ParlaMint is that it allows visitors to browse through the transcriptions, and find transcriptions based by sitting or by date. Browsing by date uses a feature that was introduced for the ParCzech corpus in which a calendar is presented with dates for which transcriptions are available highlighted, and those for which there are not greyed out.

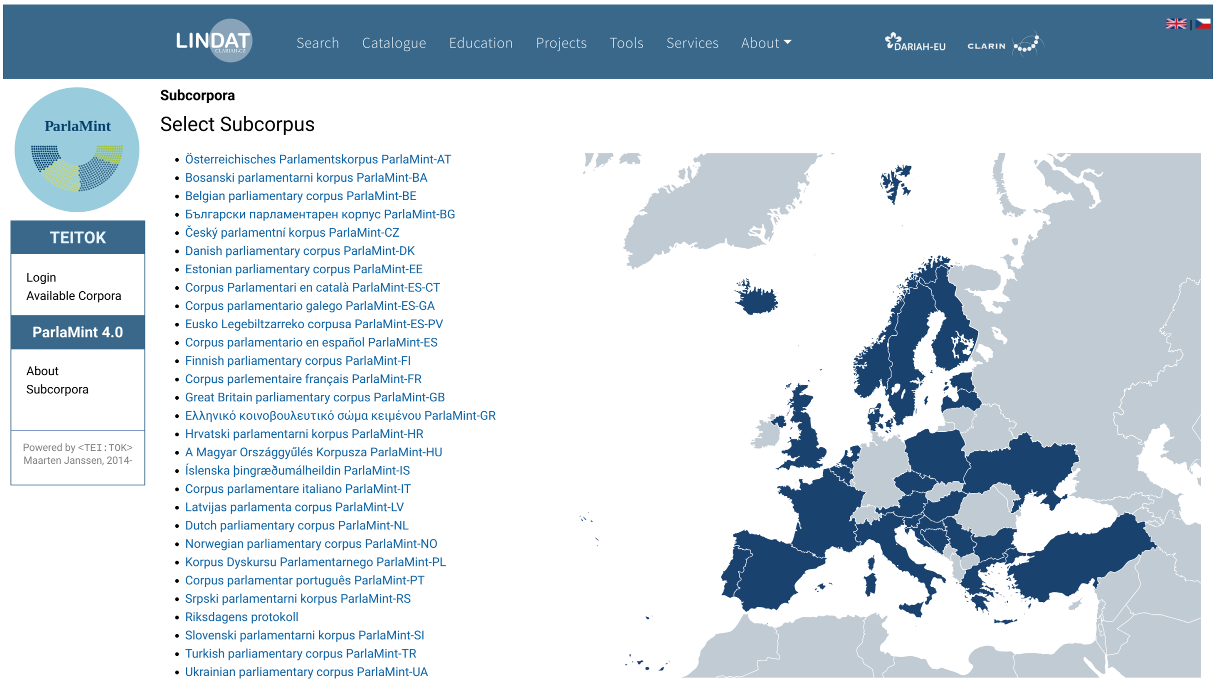
4.3. Document Visualization

The interface can display individual transcription files with various types of information assembled in the interface, as can be seen in Figure 2.

The header of the page contains the pertinent metadata of the file: the country of the parliament; the source it was taken from with a direct link where available; the information about the session - the date, term, meeting, sitting and agenda; and a link to the previous and the next document in the corpus.

[13] https://kdp.nlp.ipipan.waw.pl/query_corpus/

[14] https://www.kielipankki.fi/korp/#corpus=eduskunta&cqp=%5B%5D

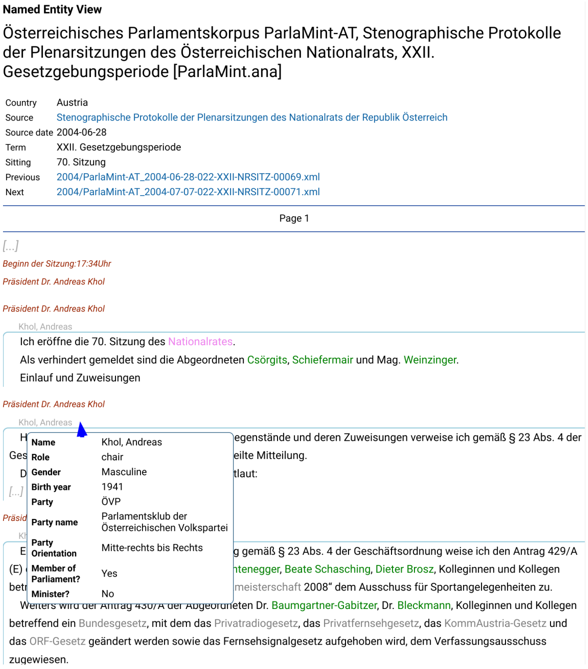


Subcorpora

Select Subcorpus

- Österreichisches Parlamentskorpus ParlaMint-AT
- Bosanski parlamentarni korpus ParlaMint-BA
- Belgian parliamentary corpus ParlaMint-BE
- Български парламентарен корпус ParlaMint-BG
- Český parlamentní korpus ParlaMint-CZ
- Danish parliamentary corpus ParlaMint-DK
- Estonian parliamentary corpus ParlaMint-EE
- Corpus Parlamentari en català ParlaMint-ES-CT
- Corpus parlamentario galego ParlaMint-ES-GA
- Eusko Legebiltzarreko corpusa ParlaMint-ES-PV
- Corpus parlamentario en español ParlaMint-ES
- Finnish parliamentary corpus ParlaMint-FI
- Corpus parlementaire français ParlaMint-FR
- Great Britain parliamentary corpus ParlaMint-GB
- Ελληνικό κοινοβουλευτικό σώμα κειμένου ParlaMint-GR
- Hrvatski parlamentarni korpus ParlaMint-HR
- A Magyar Országgyűlés Korpusza ParlaMint-HU
- Íslenska þingræðumálheildin ParlaMint-IS
- Corpus parlamentare italiano ParlaMint-IT
- Latvijas parlamenta corpus ParlaMint-LV
- Dutch parliamentary corpus ParlaMint-NL
- Norwegian parliamentary corpus ParlaMint-NO
- Korpus Dyskursu Parlamentarnego ParlaMint-PL
- Corpus parlamentar português ParlaMint-PT
- Srpski parlamentarni korpus ParlaMint-RS
- Riksdagens protokoll
- Slovenski parlamentarni korpus ParlaMint-SI
- Turkish parliamentary corpus ParlaMint-TR
- Ukrainian parliamentary corpus ParlaMint-UA

Figure 1: The subcorpus select page


Named Entity View

Österreichisches Parlamentskorpus ParlaMint-AT, Stenographische Protokolle der Plenarsitzungen des Österreichischen Nationalrats, XXII. Gesetzgebungsperiode [ParlaMint.ana]

Country	Austria
Source	Stenographische Protokolle der Plenarsitzungen des Nationalrats der Republik Österreich
Source date	2004-06-28
Term	XXII. Gesetzgebungsperiode
Sitting	70. Sitzung
Previous	2004/ParlaMint-AT_2004-06-28-022-XXII-NRSITZ-00069.xml
Next	2004/ParlaMint-AT_2004-07-07-022-XXII-NRSITZ-00071.xml

Page 1

[...]

Beginn der Sitzung:17:34Uhr

Präsident Dr. Andreas Khol

Präsident Dr. Andreas Khol

Khol, Andreas

Ich eröffne die 70. Sitzung des Nationalrates.

Als verhindert gemeldet sind die Abgeordneten Csörgits, Schiefermair und Mag. Weinzinger.

Einlauf und Zuweisungen

Präsident Dr. Andreas Khol

Khol, Andreas

H egenstände und deren Zuweisungen verweise ich gemäß § 23 Abs. 4 der Ges eilte Mitteilung.

D tlaut:

[...]

Name	Khol, Andreas
Role	chair
Gender	Masculine
Birth year	1941
Party	ÖVP
Party name	Parlamentsklub der Österreichischen Volkspartei
Party Orientation	Mitte-rechts bis Rechts
Member of Parliament?	Yes
Minister?	No

Präsid

Kh

E g gemäß § 23 Abs. 4 der Geschäftsordnung weise ich den Antrag 429/A (E) ntenegger, Beate Schasching, Dieter Brosz, Kolleginnen und Kollegen betr meisterschaft 2008" dem Ausschuss für Sportangelegenheiten zu.

Weiters wird der Antrag 430/A der Abgeordneten Dr. Baumgartner-Gabitzer, Dr. Bleckmann, Kolleginnen und Kollegen betreffend ein Bundesgesetz, mit dem das Privatradiogesetz, das Privatfernsehgesetz, das KommAustria-Gesetz und das ORF-Gesetz geändert werden sowie das Fernsehsignalgesetz aufgehoben wird, dem Verfassungsausschuss zugewiesen.

Figure 2: An indidivual transcription

The text itself is a direct visualization of the source TEI/XML file, and hence does not only contain the utterances in the transcription, but also all the comments and other information in the source, which are not in the searchable corpus.

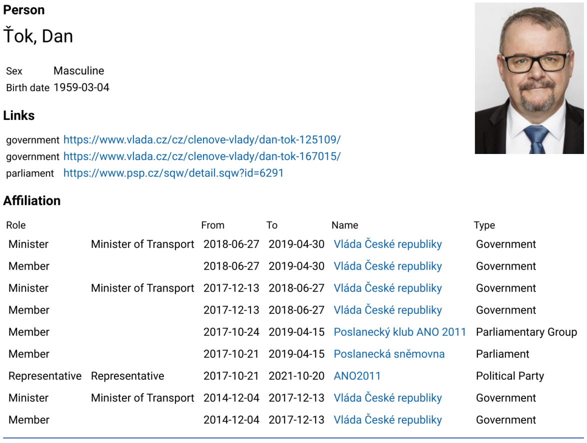
The default visualization in TEITOK is a linguistic view that shows all linguistic annotations about the tokens on mouse-over. Since the expectation is that the majority of visitors of ParlaMint in TEITOK does not have a linguistic background, in the ParlaMint project the linguistic view is instead linked on the bottom of the text, with the default view highlighting all named entities in the text, with information about the type of entity on mouse-over.

And each speaker is identified on top of the transcription of their speech act, with all available information about the speaker shown on mouse-over: full name, gender, birth year, and the name and political orientation of the party he/she belongs to. As mentioned in the previous section, this information is taken from the tabular data compiled for NSE to make sure that the data are consistent across the different interfaces, and all data should reflect the status of the person and the party at the time of the sitting.

4.4. People and Organizations

Apart from the transcriptions themselves, the TEITOK interface also provides a visualization of the people and organizations in the ParlaMint sources. For each subcorpus, it presents a searchable list of all people in the metadata file. Each person is listed along with its sex and birth date. Clicking on a name will open up a window about that person, with on top all biographical data available, such as name, sex, birth date or photographs. Below that all links to external pages related to that person as present in the ParlaMint sources. And then a list of all the organizations that that person has been a member of, with the name, the period, and the role of the person in the organization. An example is given in Figure 3.

Similarly, you can also start from the organizations, where each organization provides all available information such as political orientation of the organization, external links, and a list of all people that were a member of that organization, with name, period, and sex.



Person

Ťok, Dan

Sex Masculine
Birth date 1959-03-04

Links

government https://www.vlada.cz/cz/clenove-vlady/dan-tok-125109/
government https://www.vlada.cz/cz/clenove-vlady/dan-tok-167015/
parliament https://www.psp.cz/sqw/detail.sqw?id=6291

Affiliation

Role		From	To	Name	Type
Minister	Minister of Transport	2018-06-27	2019-04-30	Vláda České republiky	Government
Member		2018-06-27	2019-04-30	Vláda České republiky	Government
Minister	Minister of Transport	2017-12-13	2018-06-27	Vláda České republiky	Government
Member		2017-12-13	2018-06-27	Vláda České republiky	Government
Member		2017-10-24	2019-04-15	Poslanecký klub ANO 2011	Parliamentary Group
Member		2017-10-21	2019-04-15	Poslanecká sněmovna	Parliament
Representative	Representative	2017-10-21	2021-10-20	ANO2011	Political Party
Minister	Minister of Transport	2014-12-04	2017-12-13	Vláda České republiky	Government
Member		2014-12-04	2017-12-13	Vláda České republiky	Government

Figure 3: The person record visualization

5. Conclusion

The TEITOK interface for the ParlaMint data provides a more accessible entry point for people with a background other than linguistics. It attempts to focus on those data that the average user is expected to be most interested in, accessing data by people, dates, and organizations. The interface makes all the information in the ParlaMint corpus easy to view and browse. Of course the interface still presents the data as a corpus of text with additional data - so dedicated research for instance in the field of political sciences would likely still need to start from the raw data, but for more cursory access, we believe the interface makes the data accessible to a much wider audience.

The interface could be improved in future versions of ParlaMint. Due to the modular set-up of TEITOK it is easy to add more dedicated functionality over time. For instance, if apart from named entity recognition the names in the transcriptions would also be entity linked. This would make it possible to create an interface for all topics discussed in the parliamentary sitting, which is probably one of the most interesting issues for many people. But named entity recognition alone, especially with many languages in the corpus where names are inflected, does not give satisfactory results.

We are currently working on creating a live version of ParlaMint alongside the static version of ParlaMint 4.0 currently provided. New versions of subcorpora are sometimes released, and for most users, the most pertinent version is the most recent version of the documents. But replacing the searchable corpus would break the reproducibility of published results based on ParlaMint 4.0. So the intention is to keep the version of ParlaMint 4.0 unmodified, while at the same time having a separate version that always contains the most recent version of all subcorpora.

Another direction of work is to make the english version of the corpus accessible[15]. The idea is to leverage those translations in a number of different ways – to provide searches in both the original and the translation, or combinations of those. To display all metadata both in the original and the translations, and to display the two versions next to each other. At this point in time, there are still several issues to be resolved before this can be fully implemented.

6. Acknowledgements

This work has been supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ and CLARIN ERIC, ParlaMint: Towards Comparable Parliamentary Corpora.

7. Bibliographical References

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darundefinedis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2022. The parlamint corpora of parliamentary proceedings. *Lang. Resour. Eval.*, 57(1):415–448.

Tomaž Erjavec, Matyáš Kopp, Maciej Ogrodniczuk, Petya Osenova, Rodrigo Agerri, Manex Agirrezabal, Tommaso Agnoloni, José Aires, Monica Albini, Jon Alkorta, Iván Antiba-Cartazo, Ekain Arrieta, Mario Barcala, Daniel Bardanca, Starkaður Barkarson, Roberto Bartolini, Roberto Battistoni, Nuria Bel, Maria del Mar Bonet Ramos, María Calzada Pérez, Aida Cardoso, Çağrı Çöltekin, Matthew Coole, Roberts Darģis, Jesse de Does, Ruben de Libano, Griet Depoorter, Katrien Depuydt, Sascha Diwersy, Réka Dodé,

[15] http://hdl.handle.net/11356/1864

Kike Fernandez, Elisa Fernández Rei, Francesca Frontini, Marcos Garcia, Noelia García Díaz, Pedro García Louzao, Maria Gavriilidou, Dimitris Gkoumas, Ilko Grigorov, Vladislava Grigorova, Dorte Haltrup Hansen, Mikel Iruskieta, Johan Jarlbrink, Kinga Jelencsik-Mátyus, Bart Jongejan, Neeme Kahusk, Martin Kirnbauer, Anna Kryvenko, Noémi Ligeti-Nagy, Nikola Ljubešić, Giancarlo Luxardo, Carmen Magariños, Måns Magnusson, Carlo Marchetti, Maarten Marx, Katja Meden, Amália Mendes, Michal Mochtak, Martin Mölder, Simonetta Montemagni, Costanza Navarretta, Bartłomiej Nitoń, Fredrik Mohammadi Norén, Amanda Nwadukwe, Mihael Ojsteršek, Andrej Pančur, Vassilis Papavassiliou, Rui Pereira, María Pérez Lago, Stelios Piperidis, Hannes Pirker, Marilina Pisani, Henk van der Pol, Prokopis Prokopidis, Valeria Quochi, Paul Rayson, Xosé Luís Regueira, Michał Rudolf, Manuela Ruisi, Peter Rupnik, Daniel Schopper, Kiril Simov, Laura Sinikallio, Jure Skubic, Minna Tamper, Lars Magne Tungland, Jouni Tuominen, Ruben van Heusden, Zsófia Varga, Marta Vázquez Abuín, Giulia Venturi, Adrián Vidal Miguéns, Kadri Vider, Ainhoa Vivel Couso, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, Rodolfo Zevallos, and Darja Fišer. 2023. Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 4.0. http://hdl.handle.net/11356/1860.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Corpus Linguistics 2011*.

Bruno Guillaume. 2019. Graph Matching for Corpora Exploration. In *JLC 2019 - 10èmes Journées Internationales de la Linguistique de corpus*, Grenoble, France.

Maarten Janssen. 2016. TEITOK: Text-faithful annotated corpora. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 4037–4043.

Maarten Janssen. 2018a. Adding words to manuscripts: From pagesxml to TEITOK. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11057 LNCS:152–157.

Maarten Janssen. 2018b. TEITOK as a tool for dependency grammar. *Procesamiento del Lenguaje Natural*, 61:185–188.

Maarten Janssen. 2020. Integrating TEITOK and Kontext at LINDAT. In *Proceedings of CLARIN Annual Conference 2020*, Madrid, Spain. CLARIN.

Maarten Janssen. 2021. A corpus with Wavesurfer and TEI: Speech and video in TEITOK. In *Text, Speech, and Dialogue: 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings*, page 261–268, Berlin, Heidelberg. Springer-Verlag.

Matyáš Kopp. 2024. ParCzech 4.0. http://hdl.handle.net/11234/1-5360.

Matyáš Kopp, Vladislav Stankov, Jan Oldřich Krůza, Pavel Straňák, and Ondřej Bojar. 2021. Parczech 3.0: A large czech speech corpus with rich metadata. In *24th International Conference on Text, Speech and Dialogue*, pages 293–304, Cham, Switzerland. Springer.

Tomáš Machálek. 2020. KonText: Advanced and flexible corpus query interface. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7003–7008, Marseille, France. European Language Resources Association.

Historical Parliamentary Corpora Viewer

Alenka Kavčič, Martin Stojanoski, Matija Marolt
University of Ljubljana, Faculty of Computer and Information Science
Večna pot 113, 1000 Ljubljana, Slovenia
alenka.kavcic@fri.uni-lj.si, ms7072@student.uni-lj.si, matija.marolt@fri.uni-lj.si

Abstract

Historical parliamentary debates offer a window into the past and provide valuable insights for academic research and historical analysis. This paper presents a novel web application tailored to the exploration of historical parliamentary corpora in the context of Slovenian national identity. The developed web viewer enables advanced search functions within collections of historical parliamentary records and has an intuitive and user-friendly interface. Users can enter search terms and apply filters to refine their search results. The search function allows keyword and phrase searching, including the ability to search by delegate and place names. It is also possible to search for translations of the text by selecting the desired languages. The search results are displayed with a preview of the proceedings and highlighted phrases that match the search query. To review a specific record, the full PDF document can be displayed in a separate view, allowing the user to scroll through the PDF document and search the content. In addition, the two corpora of Slovenian historical records integrated into the viewer—the Carniolan Provincial Assembly Corpus and the Parliamentary Corpus of the First Yugoslavia—are described and an insight into the corresponding preparation processes is provided.

Keywords: parliamentary corpora, web application, Parla-CLARIN, Carniolan Provincial Assembly, National Representation of the First Yugoslavia

1. Introduction

Parliamentary debates have long been a valuable resource of research data, as they are systematically prepared and reflect the state of society at the time. They cover a wide range of topics and are an important source for historians as well as for scholars from diverse fields such as political science, sociology, economics and linguistics.

While contemporary parliamentary corpora, such as those produced in projects like ParlaMint (Erjavec et al., 2022), are widely available and well-structured, historical records are often available in less accessible forms, such as PDFs or unstructured text. This poses a major challenge for effective queries and large-scale analysis and limits the use of these resources.

Although digitization has made it easier for researchers to access the historical records by enabling keyword searches and remote access to the data, advanced search and analysis tools are needed to enhance research capabilities and facilitate exploration of the data. These tools include keyword search, advanced search filters, text mining algorithms, sentiment analysis, topic modeling, and visualization techniques.

2. Historical Parliamentary Corpora

Historical parliamentary records in digital format are still a rarity and online access and exploration even rarer. However, there are notable exceptions that provide better access to transcriptions of past parliamentary debates. A good example of this is the digitized collection of lower house parliamentary debates in the French parliament from 1881 to 1940, accessible via the digital repository of the France National Library (Gallica[1]). This corpus, carefully curated as part of the AGODA project (Puren et al., 2022), has undergone extensive processing, including OCR, annotation and semantic enrichment, making it easily accessible for scholarly research.

Another notable example is the Historical Hansard[2] corpus, which contains transcriptions of speeches and debates in the UK Houses of Lords and the Commons from 1803 to the present day (Coole et al., 2020). Hansard has traditionally been published in printed form since the early 19th century. The older volumes of the collection have been digitized and enhanced with metadata, including tokenization, part-of-speech tagging and semantic annotations, to form a comprehensive corpus. In addition, a web-based search interface has been developed that allows linguistic queries to be performed on the corpus while offering visualizations that provide a deeper understanding of the evolution of political discourse over time.

Additional example is Congress.gov[3], which offers numerous ways to access the US Congressional Records from 1873 (43rd Congress) to the present day. Users can select records by specific

[1] https://gallica.bnf.fr/ark:/12148/cb328020951/date.item

[2] https://hansard.parliament.uk/

[3] https://www.congress.gov/

dates and examine complete issues and all sections. The platform also facilitates keyword and phrase searches, with the option to refine results using various filters.

2.1. Slovenian Parliamentary History

The beginnings of Slovenian parliamentary history date back to the mid-19th century, and a large part of the parliamentary debates from this period have been digitized and made available in PDF format. While these digitized archives are invaluable repositories of historical knowledge, they also present significant challenges in terms of readability (due to archaic language or poor reproduction quality), completeness (due to gaps or omissions in the records), and biases associated with the reporting and recording processes, so a critical eye is required when interpreting these materials.

Beyond mere digital preservation, the efficiency of data exploration depends on the enrichment of resources with metadata and the provision of structured, annotated content. Enriching these digitized archives with comprehensive metadata facilitates efficient retrieval and categorization, while structured and annotated content improves interpretability and enables a more complex and systematic study of Slovenian parliamentary history.

In this section, we present two newly created historical corpora from the present-day territory of Slovenia.

2.2. Carniolan Provincial Assembly

The Carniolan Provincial Assembly (*Kranjski deželni zbor* in Slovenian or *Krainer Landtag* in German) was the highest legislative body of Duchy of Carniola, which was a hereditary land of the Habsburg monarchy and a part of Austrian Empire (from 1867 Austro-Hungarian Empire). The Carniolan Provincial Assembly was introduced with the February patent, a constitution of the Austrian Empire proclaimed on 26 February 1861. After 12 parliamentary terms, it ended with the onset of the First World War. A unicameral assembly consisted of 37 members (in 1908 the number was increased to 50) and was chaired by the provincial governor (*deželni glavar* or *Landeshauptmann*) who was appointed by the Emperor from among the members. The Carniolan Provincial Assembly passed laws that were within the province's jurisdiction, including educational, municipal, ecclesiastical and military matters, and issues of provincial importance (e.g. agriculture, culture, public buildings, public construction works, various economic matters, charity institutes).

The parliamentary meeting proceedings from 1861 to 1913 are available in the Carniolan Provincial Assembly corpus Kranjska 1.0 (Kavčič et al., 2023a). The corpus covers 694 sessions, with two documents for each parliamentary session: one in Parla-CLARIN compliant TEI XML format (see section 2.4) and a corresponding facsimile in PDF format (an example of the PDF facsimile is shown in Figure 2).

The documents are mostly bilingual; 58% of sentences are in Slovenian and 42% in German language. The XML documents together include over 44 thousand utterances, over 540 thousand sentences and approximately 10 million words that are also linguistically annotated (tokenisation, part-of-speech tagging and lemmatisation were used).

2.3. National Representation of the First Yugoslavia

First Yugoslavia refers to the Yugoslav state between the two world wars: the Kingdom of Serbs, Croats, and Slovenes, established after the collapse of the Austro-Hungarian Empire in 1918, and renamed Kingdom of Yugoslavia in 1929. In the newly formed Kingdom of Serbs, Croats, and Slovenes, a joint government and a Temporary National Representation were established. The latter performed the functions of Parliament from March 1919 to October 1920, when the elections to the Constituent Assembly were called. Its 296 delegates were not elected, but appointed. First parliamentary elections were held in November 1920, when 419 delegates were elected to the Constituent Assembly. There have been seven elections altogether in that time (i.e. between the two world wars): besides 1920, also in 1923, 1925 and 1927 (because of political instability every 2 years, although the terms lasted 4 years), and in 1931, 1935 and 1938. There were no elections during the dictatorship from January 1929 to September 1931, as the National Assembly was abolished at that time. In 1931, a bicameral system was introduced with the new constitution: National Representation consisted of Senate and National Assembly. Parliament's autonomy was very limited due to the great authority of the King, who according to the constitution controlled all three branches of government, including the legislative. The parliament had general passive and active suffrage, decided on laws, and on amendments to the constitution, while the King had the power, among the others, to suspend the law, and to summon and dissolve the Parliament.

The parliamentary meeting proceedings from 1919 to 1939 are available in the Parliamentary corpus of first Yugoslavia yu1Parl 1.0 (Kavčič et al., 2023b), covering proceedings in the three periods:

- Temporary National Representation of the Kingdom of Serbs, Croats, and Slovenes (1919-1920);

- Legislative Committee of National Assembly of the Kingdom of Serbs, Croats, and Slovenes (1921-1922);
- National Representation (National Assembly and Senate) of the Kingdom of Yugoslavia (1931-1939).

The meeting proceedings of the National Assembly of the Kingdom of Serbs, Croats, and Slovenes between years 1923 and 1928 are not (yet) available in digital form and therefore not included in the corpus.

The corpus comprises 714 sessions, where each session is available in two documents of different formats: Parla-CLARIN compliant TEI XML and a corresponding facsimile in PDF.

The documents are multilingual, in Slovenian (3% of sentences) and Serbo-Croatian. The latter is typeset in the Cyrillic (Serbian, 59% of sentences) or in the Latin (Croatian, 38% of sentences) alphabet. The XML documents together include over 34 thousand utterances, 578 thousand sentences and approximately 13 million linguistically annotated words, where words in Cyrillic script (Serbian) have lemmas in Latin script.

2.4. Parla-CLARIN TEI Format

There are various ways to annotate the content, but the TEI guidelines are the de facto standard for encoding text in the digital humanities (TEI Consortium, 2019). For parliamentary corpora, the Parla-CLARIN Guidelines (Erjavec and Pančur, 2021) were developed as a common TEI-based annotation scheme.

The preparation of the corpus started with the scanned images of the meeting proceedings. The scanned documents were OCR processed and automatically parsed with rule-based Python scripts to extract the metadata and annotate the speeches.

For structuring the session data, we used a subset of the Parla-CLARIN tags (Erjavec and Pančur, 2022) that were best suited for our content, the multilingual historical parliamentary debates.

Each meeting proceeding has a unique identifier based on the content of the proceeding and the date of the session. An XML file consists of a header and a body. The header contains the metadata of the file: title in several languages (English, Slovenian and other main languages used in the proceedings), information about the publisher, publication date, link to the related PDF files, etc. The body is parsed from the content: it starts with the title of the session, information about the delegates present, the agenda and the starting time of the session. This is followed by the individual sections dealing with the speeches of the individual delegates. Each section is labelled with the name of the speaker, followed by the content of the speech, which may also include comments or events during the speech (i.e. described events in the room such as "laughter from the left" or "reads"). The speech is divided into sentences, and these in turn are divided into words and punctuation. As the transcripts are multilingual, the language is marked for each sentence. The words are also linguistically annotated. The end time of the session is noted at the end.

The linguistic annotation in both corpora included tokenisation, MSD tagging and lemmatisation. Since the corpora contain different languages, different tools were used for the linguistic annotation in each corpus. The languages of the Carniolan Provincial Assembly were German and Slovene, so Trankit[4] was used as it works well for both languages. The National Representation of the First Yugoslavia, on the other hand, includes Slavic languages (Slovenian and Serbo-Croatian), which is why we opted for CLASSLA[5] (Ljubešić and Dobrovoljc, 2019; Terčon and Ljubešić, 2023).

3. Historical Parliamentary Corpora Viewer

To make the historical parliamentary session proceedings accessible to a wider audience that may not be proficient in parsing TEI encoded files, we developed the Historical Parliamentary Corpora Viewer. The Viewer is a web application that supports searching over collections of multilingual proceedings of historical parliamentary sessions. It allows the user to search for texts across languages and limit the results to certain speakers or place names. It is also possible to filter the results by language, date of sessions and to sort the results by date or relevance.

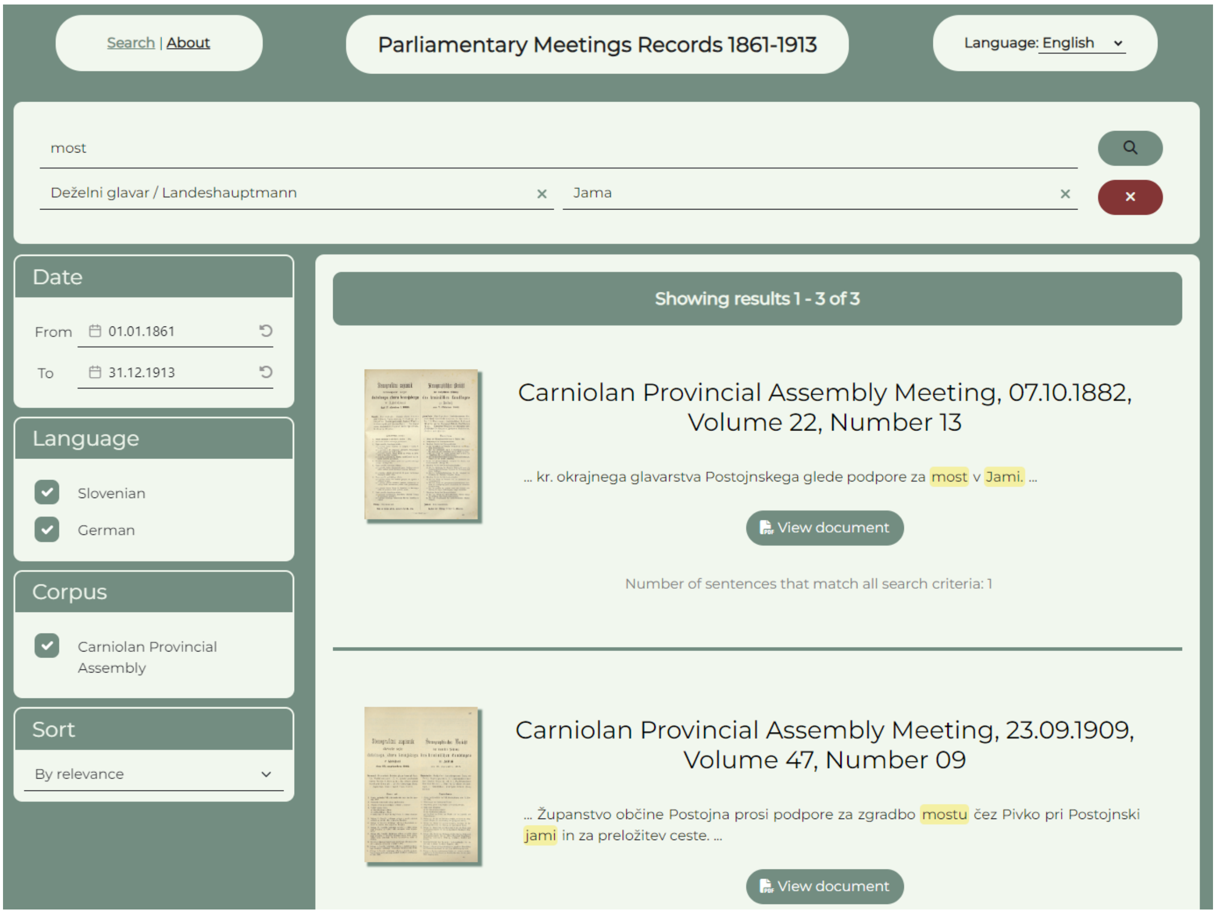
The two corpora described in sections 2.2 and 2.3 are currently included in the Viewer. If further parliamentary proceedings are scanned and prepared so that they are available as PDF and TEI XML documents, they can be added to the Viewer as an additional corpus.

3.1. Technical Details

The web application consists of three parts: frontend, backend and database. The frontend runs on a client device and implements the user interface. It sends requests to the backend and receives the data to be displayed to the user. The frontend was developed in the Vue.js JavaScript framework and uses the HTTP protocol for communication between the client and the server.

[4] https://github.com/nlp-uoregon/trankit

[5] https://github.com/clarinsi/classla



Parliamentary Meetings Records 1861-1913

most

Deželni glavar / Landeshauptmann ×

Jama ×

From 01.01.1861

To 31.12.1913

☑ Slovenian

☑ German

☑ Carniolan Provincial Assembly

By relevance

Showing results 1 - 3 of 3

Carniolan Provincial Assembly Meeting, 07.10.1882, Volume 22, Number 13

... kr. okrajnega glavarstva Postojnskega glede podpore za most v Jami. ...

Number of sentences that match all search criteria: 1

Carniolan Provincial Assembly Meeting, 23.09.1909, Volume 47, Number 09

... Županstvo občine Postojna prosi podpore za zgradbo mostu čez Pivko pri Postojnski jami in za preložitev ceste. ...

Figure 1: Searching the corpora. The search fields are at the top, the filters for narrowing down the results are on the left, while the search results are displayed in the main part of the page.

The backend runs on a Node.js server was implemented with Express.js. It offers a RESTful API and thus complies with the specifications of the REST architecture. The backend communicates with the database via the HTTP protocol. Elasticsearch[6], a RESTful search and analytics engine was used as the database, enabling fast unstructured text searches.

3.2. User Interface

The user interface is minimalist, intuitive and easy to use. It is divided into two parts: a page for searching and browsing the parliamentary proceedings and a page that displays a facsimile of the selected proceedings and allows search within the proceedings, as well as the display of PDF (OCRed) text transcription and its translations into target languages.

3.3. Searching the Corpora

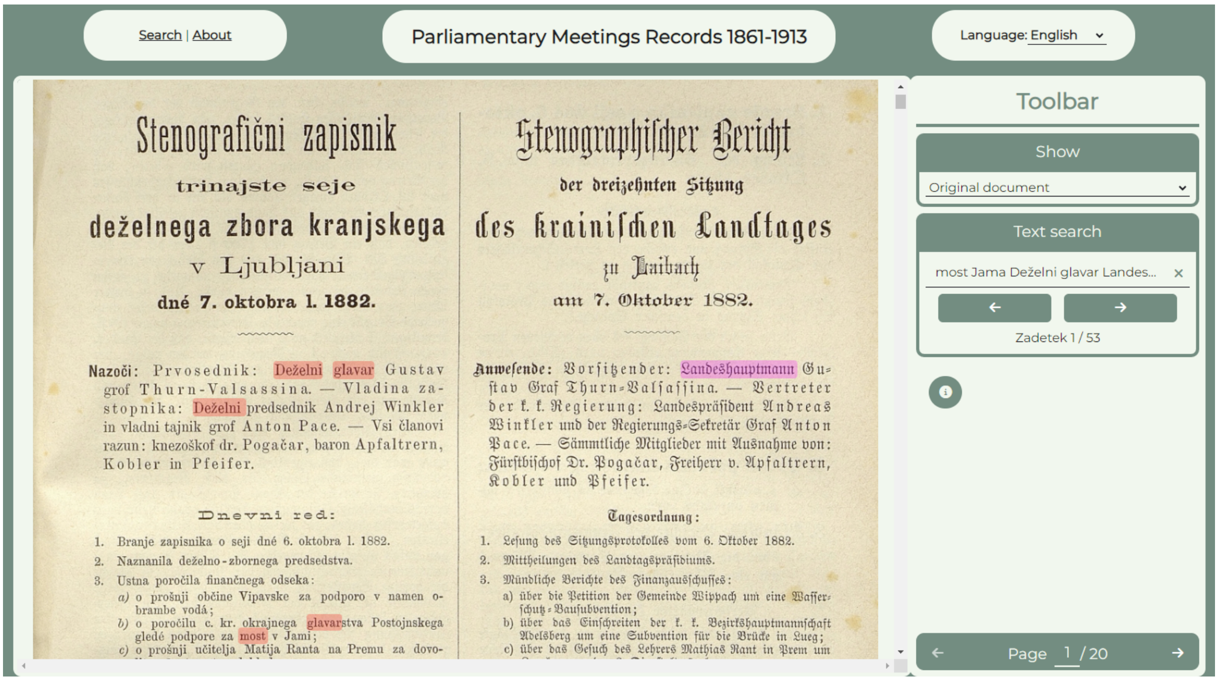
The page for searching in the corpora is shown in Figure 1. The user can enter search terms and/or set specific filters for the search results.

If no search parameters are entered, all the proceedings in the corpus are displayed, so that the user can browse the collection.

By default, the keyword search finds all proceedings that contain all the words in the search query. The search terms can also be separated with OR operator to search for documents that contain at least one of the specified search words. It is also possible to search for a phrase by enclosing the phrase in quotation marks. The search in translations of the text can be activated by selecting the desired languages in the filters. If the search word is entered in its basic form (lemma), the search will also find all other forms of the word in the text.

Several filters are available to limit the search results: date, language and corpus (shown on the left in Figure 1). Restricting the search results by date is an important filter for historical documents, as it limits the search results to the parliamentary sessions within a selected time period. Without this filter, the search is applied to all documents contained in the selected corpora. As the documents are multilingual, it is also possible to use a language filter and search for keywords in all languages (i.e.

[6] https://www.elastic.co/



Language: English

Original document

most Jama Deželni glavar Landes...

Zadetek 1 / 53

Page 1 /20

Stenografični zapisnik

trinajste seje

deželnega zbora kranjskega

v Ljubljani

dné 7. oktobra l. 1882.

Nazoči: Prvosednik: Deželni glavar Gustav grof Thurn-Valsassina. — Vladina zastopnika: Deželni predsednik Andrej Winkler in vladni tajnik grof Anton Pace. — Vsi članovi razun: knezoškof dr. Pogačar, baron Apfaltrern, Kobler in Pfeifer.

Dnevni red:

1. Branje zapisnika o seji dné 6. oktobra l. 1882.
2. Naznanila deželno-zbornega predsedstva.
3. Ustna poročila finančnega odseka:
   a) o prošnji občine Vipavske za podporo v namen obrambe vodá;
   b) o poročilu c. kr. okrajnega glavarstva Postojnskega gledé podpore za most v Jami;
   c) o prošnji učitelja Matija Ranta na Premu za dovo-

Stenographiſcher Bericht

der dreizehnten Sitzung

des krainiſchen Landtages

zu Laibach

am 7. Oktober 1882.

Anweſende: Vorſitzender: Landeshauptmann Guſtav Graf Thurn-Valſaſſina. — Vertreter der k. k. Regierung: Landespräſident Andreas Winkler und der Regierungs-Sekretär Graf Anton Pace. — Sämmtliche Mitglieder mit Ausnahme von: Fürſtbiſchof Dr. Pogačar, Freiherr v. Apfaltrern, Kobler und Pfeifer.

Tagesordnung:

1. Leſung des Sitzungsprotokolles vom 6. Oktober 1882.
2. Mittheilungen des Landtagspräſidiums.
3. Mündliche Berichte des Finanzausſchuſſes:
   a) über die Petition der Gemeinde Wippach um eine Waſſerſchutz-Bauſubvention;
   b) über das Einſchreiten der k. k. Bezirkshauptmannſchaft Adelsberg um eine Subvention für die Brücke in Lueg;
   c) über das Geſuch des Lehrers Mathias Rant in Prem um

Figure 2: The viewer for PDF documents, showing an excerpt from the PDF facsimile of the Carniolan Provincial Assembly session dated 7.10.1882. Words from the query are highlighted.

also in the translations of the texts).

The search results are displayed with the proceedings preview and some sentences with the highlighted words corresponding to the query. The “Show document” button displays the PDF facsimile of the proceedings in an integrated PDF viewer.

3.4. Searching the PDFs and Transcriptions

If users want to inspect specific parliamentary proceedings, they can display the document in a separate view that allows them to scroll through the PDF and search within the PDF content. Figure 2 shows the PDF viewer with the search bar on the right.

Instead of a facsimile of the document, the user can display a transcript in all the main languages used in the parliamentary proceedings of a particular corpus (e.g. German and Slovene in the case of the records of the Carniolan Provincial Assembly), which also supports the content search.

4. Conclusion

The development of the web viewer for exploring the Slovenian historical parliamentary corpora represents a significant step forward in terms of the availability of historical resources for researchers and enthusiasts alike. The intuitive user interface caters to diverse users, from students to experienced scholars, and enables seamless navigation and exploration of the invaluable historical documents. By bridging the gap between academia and the public, this application not only enhances scholarly research, but also promotes a deeper understanding and appreciation of Slovenian parliamentary history.

Since the web viewer is designed to display corpora in a Parla-CLARIN TEI compatible format, the integration of new corpora is a straightforward process. In the future, we plan to gradually expand our dataset by integrating additional corpora. We are aiming for comprehensive coverage of parliamentary debates from the mid-19th century to the present day.

5. Acknowledgements

The work was supported by the Slovenian Research Agency research programme P6-0436: Digital Humanities: resources, tools and methods (2022–2027).

6. Bibliographical References

Matthew Coole, Paul Rayson, and John Mariani. 2020. Unfinished business: Construction and maintenance of a semantically tagged historical parliamentary corpus, UK Hansard from 1803 to the present day. In *Proceedings of the Second ParlaCLARIN Workshop*, pages 23–27, Marseille, France. European Language Resources Association.

Tomaž Erjavec and Andrej Pančur. 2021. The parla-clarin recommendations for encoding corpora of parliamentary proceedings. *Journal of the Text Encoding Initiative*, 14.

Tomaž Erjavec and Andrej Pančur. 2022. Parla-clarin: A tei schema for corpora of parliamentary proceedings. https://clarin-eric.github.io/parla-clarin/. Accessed: 2024-02-08.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darģis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2022. The parlamint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 57(1):415–448.

Nikola Ljubešić and Kaja Dobrovoljc. 2019. What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 29–34, Florence, Italy. Association for Computational Linguistics.

Marie Puren, Pierre Vernus, Aurélien Pellet, Nicolas Bourgeois, and Fanny Lebreton. 2022. Extracting and providing online access to annotated and semantically enriched historical data. The AGODA project. In *DH Benelux 2022*, Luxembourg, Luxembourg.

TEI Consortium. 2019. TEI P5: Guidelines for electronic text encoding and interchange. http://www.tei-c.org/Guidelines/P5/. Accessed: 2024-02-08.

Luka Terčon and Nikola Ljubešić. 2023. Classla-stanza: The next step for linguistic processing of south slavic languages.

7. Language Resource References

Kavčič, Alenka and Mundjar, Aleksander and Marolt, Matija. 2023a. *Carniolan Provincial Assembly corpus Kranjska 1.0*. Faculty of Computer and Information Science, University of Ljubljana. PID http://hdl.handle.net/11356/1824. Slovenian language resource repository CLARIN.SI.

Kavčič, Alenka and Mundjar, Aleksander and Marolt, Matija. 2023b. *Parliamentary corpus of first Yugoslavia (1919-1939) yu1Parl 1.0*. Faculty of Computer and Information Science, University of Ljubljana. PID http://hdl.handle.net/11356/1845. Slovenian language resource repository CLARIN.SI.

The dbpedia R Package: An Integrated Workflow for Entity Linking (for ParlaMint Corpora)

Christoph Leonhardt and Andreas Blätte
University of Duisburg-Essen
{christoph.leonhardt, andreas.blaette}@uni-due.de

Abstract

Entity Linking is a powerful approach for linking textual data to established structured data such as survey data or adminstrative data. However, in the realm of social science, the approach is not widely adopted. We argue that this is, at least in part, due to specific setup requirements which constitute high barriers for usage and workflows which are not well integrated into analyitical scenarios commonly deployed in social science research. We introduce the `dbpedia` R package to make the approach more accessible. It has a focus on functionality that is easily adoptable to the needs of social scientists working with textual data, including the support of different input formats, limited setup costs and various output formats. Using a ParlaMint corpus, we show the applicability and flexibility of the approach for parliamentary debates.

Keywords: Entity Linking, ParlaMint, Parliamentary Data

1. Introduction

Recent innovations such as transformer-based machine learning and large language models come with huge promises and great potential for scholars of different disciplines (Linegar et al., 2023). The unprecedented wealth of available data and tools continuously inspires new research questions and innovative methodological approaches. At the same time, the analysis of well-established types of structured data such as survey data or administrative data is methodologically mature and advanced at the same time, and continues to provide invaluable insights into social processes. In consequence, the possibility to combine findings from both textual data and structured data constitutes an important perspective for innovative research.

In the field of parliamentary research, these potentials are particularly apparent. Given the efforts of projects such as ParlaMint (Erjavec et al., 2023a) to create interoperable corpora of parliamentary debates and the large variety of data sets which can enrich these collections such as the Chapel Hill Expert Survey (Bakker et al., 2015), the Manifesto Project (Budge and Bara, 2001) or other statistical or administrative data sets, the combination of different types of data opens up novel perspectives for research questions which previously would be impossible or hard to address due to a lack of data and integrated analyses.

A central way to link textual data with structured data is the method of Entity Linking. Entity Linking is both an established but also actively researched area of study in the field of Natural Language Processing and Information Retrieval. In a nutshell, it comprises the disambiguation and assignment of entities in a document – often representing persons, organizations and locations – to corresponding entities in an external knowledge graph (Linhares Pontes et al., 2020, p. 218; Möller et al., 2022, p. 925). This way, the text can be represented in a "computer-processable form" (Al-Moslmi et al., 2020, p. 32862), thus potentially facilitating integrated analyses by shared unique identifiers and access to other data sets.

However, realizing this potential can be challenging. While the large number of analyses using all kinds of approaches to text analysis illustrates the interest of social scientists and beyond to apply innovative approaches in their research, Entity Linking is – until now – adopted only sporadically. We argue that this is, at least in part, due to a lack of integrated workflows and established best practices. Existing approaches do not necessarily provide guidance for social science applications and often constitute individual use cases which do not necessarily generalize well enough or are poorly maintained. Improving the accessibility of such innovative methods by approaching them from a perspective of social science and the humanities may thus be an important driver of progress.

To address this, this contribution introduces the `dbpedia` R package which is currently developed by the authors of this paper. `dbpedia` constitutes a wrapper for the statistical programming language R (R Core Team, 2023) for the Entity Linking service DBpedia Spotlight (Mendes et al., 2011; Daiber et al., 2013). In particular, it integrates the communication with the service into an R-based analysis workflow which makes Entity Linking available for existing text analysis pipelines.

This contribution proceeds as follows: First, existing applications of Entity Linking with a focus on

ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora, pages 133–144
May 20, 2024.

parliamentary textual data are presented. This is followed by a discussion of requirements for the adoption of the approach. In the third section, the `dbpedia` R package is presented, using a sample of the UK corpus of the ParlaMint project (Erjavec et al., 2023a) as a show case to illustrate input formats and enrichment. This contribution concludes with a discussion of limitations and necessary next steps to contribute to the adoption of Entity Linking in social science research.

2. Related Work

2.1. Entity Linking in Social Science and Parliamentary Research

There is a number of approaches and services to facilitate the linking of entities to knowledge graphs (for a comprehensive overview, see Al-Moslmi et al., 2020). Prominent proponents of the approach are DBpedia Spotlight (Mendes et al., 2011; Daiber et al., 2013) which includes the identification, disambiguation and linking of entities in text and targets the DBpedia knowledge graph (Auer et al., 2007) and Wikidata (Vrandečić and Krötzsch, 2014) which can be used to link previously identified entities to the knowledge graph of the same name.

Focusing on parliamentary debates, Olieman et al. (2015) evaluate the performance of DBpedia Spotlight and discuss associated challenges when deploying such Entity Linking systems in domain-specific settings. Using DBpedia Spotlight as a baseline to perform Entity Linking on Dutch parliamentary proceedings, they show that the tool provides links with a precision of 0.69 and a recall of 0.40 (Olieman et al., 2015; see also Olieman et al., 2017). The authors show that these measures vary between targeted entity types and provide further suggestions on improving the approach. Similarly, van Heusden et al. (2022) compare the Entity Linking solutions of DBpedia Spotlight, YAGO and Wikidata. Using corpora of ParlaMint (Erjavec et al., 2023a), they show that while feasible, the approach can be challenging when confronted with different languages and alphabets as well as with "real world data" (van Heusden et al., 2022, p. 47). The performance of the approach varies between languages and deployed systems, but all in all they "found that the WikiData [sic!] system performed the best overall for the local politicians, although all systems performed relatively well" (van Heusden et al., 2022, p. 53).

DBpedia Spotlight is also used by Glaser et al. (2022) who provide a very illustrative example on how to use Entity Linking with DBpedia Spotlight to facilitate a substantive analysis in the realm of debates of the United Nations Security Council. They discuss the method and potential limitations, thus providing some guidance on how to deploy the approach in general. Using DBpedia Spotlight instead of other Entity Linking solutions was, among other things, also informed by its relative ease-of-use and the possibility to run the service locally (Glaser et al., 2022, pp. 54-55).

For the `dbpedia` R package, we follow similar motivations when selecting DBpedia Spotlight as the service of choice. Aside from the relatively low effort to use the service (as discussed below), DBpedia Spotlight is also considered because it can be configured depending on the use case. The authors of DBpedia Spotlight describe the service as a "comprehensive and flexible solution" to annotate mentions of entities in a text with resources from the DBpedia knowledge graph (Mendes et al., 2011, p. 1). As it performs the identification of entities and uses the ontology of the underlying knowledge graph, it is not limited to pre-annotated entity types or to specific classes (Mendes et al., 2011, p. 1). The flexibility and architecture of DBpedia Spotlight are also discussed in Olieman et al. (2014, pp. 14-16).

2.2. The Need for a Package

While the projects discussed above provide great insight into the potential of the approach, a broader adoption requires that the cost of setup is minimized; workflows must integrate well into those commonly deployed in the social sciences and humanities. Accordingly, we argue that a software solution which can provide a robust framework for analyses, is easy enough to use and which provides a code base which can be maintained easier than, for example, a stand-alone script is thus an essential building block towards this goal.

In this vein, there are software implementations and wrappers for DBpedia Spotlight which address a part of the problem. As an interesting example, `spacy-dbpedia-spotlight`[1] is a library implemented in Python for users who are familiar with the popular spaCy NLP suite.[2] It seems to be well maintained and is comprehensively documented. However, its main focus is to extend the NLP pipeline of spaCy which, by itself, is not directly integrated into common social science workflows. This is true for many packages which provide the core functionality to query DBpedia Spotlight but do not provide easy paths, clear guidelines and best practices on how to use the approach in substantive analyses.

Accordingly, the `dbpedia` R package should be both robustly developed – providing options with

[1] `https://github.com/MartinoMensio/spacy-dbpedia-spotlight` (2024-02-13).

[2] `https://spacy.io` (2024-02-14).

useful default values, telling error messages, etc. – and flexible enough to be deployed in different scenarios. By providing an integrated workflow for different input types, a condensed but configurable set of commands, and including the possibility to add the enriched data to the initial input structure, the package should address some common issues when adopting the approach and equip researchers of various fields with a tool which enables them to focus on substantive research.

3. The `dbpedia` R Package

3.1. At a Glance

Currently only available on GitHub, the installation of `dbpedia` is described in some detail in the online documentation.[3] In principle, it can be run like any ordinary R package. Without any additional setup, it only needs a few lines of code to query the English public endpoint and receive Uniform Resource Identifiers (URIs) from the DBpedia knowledge graph for identified entities in a document. At the time of writing, this endpoint is provided by the maintainers of DBpedia Spotlight and can be used for minimal setup. Being a public endpoint, rate limits might apply and availability might not be guaranteed.[4]

Running the following chunk of code will result in the output similar to that shown in table 1. The results will include character offsets describing the start positions of tokens, the entities itself as well as the identified URI of the entity.

```
library(dbpedia) # v0.1.2.9004 or higher

annotations <- get_dbpedia_uris(
  x = "The city of Turin is located
  at the river Po."
)
```

3.2. Advanced Setup

As described above, one of the advantages of DBpedia Spotlight is the easy local deployment which improves performance, avoids potential rate limits and saves resources of the publicly available endpoint. Accordingly, for our experiments and examples, we run the service locally in a Docker container. This is described by its maintainers in the corresponding online documentation.[5] Necessary computational resources depend on the language model used, but should, in general, be manageable for most modern systems.

3.3. Advanced Scenario

In the example above, we simply sent a `character vector` to the service. In this instance, the `get_dbpedia_uris()` method is somewhat limited to a wrapper which sends and receives HTTP requests and parses results. This is realized using established R packages such as `httr` (Wickham, 2023) and `jsonlite` (Ooms, 2014). However usually, challenges occur in more advanced scenarios. They include the preparation of different input formats and the presentation of results in a useful way, for example by mapping identified entities back to the input data. In the following, we present the functionality of `dbpedia` to adopt Entity Linking in a plausible social science scenario.

3.3.1. Input Data

Textual data comes in different shapes and forms. While sometimes, it is provided as a single continuous string, other times it is already separated into individual tokens. Sometimes the data is available in a tabular representation and other times it is represented in more complex formats such as XML or the Corpus Workbench format (Evert and Hardie, 2011). The `dbpedia` package is designed to account for this variety of input formats and provides workflows for different data types such `character vectors`, `quanteda` corpora (Benoit et al., 2018), `Corpus Workbench` subcorpora and `XML`.

As discussed before, parliamentary debates are an attractive subject for Entity Linking. Accordingly, this contribution focuses on an emerging standard for encoding this type of textual data and presents the workflow of `dbpedia` for corpora represented in the XML schema of the ParlaMint project (Erjavec et al., 2023a). The corpora of ParlaMint follow strict encoding guidelines for parliamentary data in the XML data format, thus ensuring interoperability and comparability. The corpora include different levels of structural and linguistic annotation. Named entities are identified, but not linked to an external knowledge base.

The interoperable format of ParlaMint also benefits the development of tools such as the `dbpedia` R package, as it increases the number of potential use cases. While DBpedia Spotlight supports many languages out of the box,[6] ParlaMint also

[3] The package is available on `https://github.com/PolMine/dbpedia` (2024-03-30).

[4] Also see the presentation of the tool here: `https://www.dbpedia-spotlight.org` (2024-02-14).

[5] `https://github.com/dbpedia-spotlight/spotlight-docker` (2024-02-13).

[6] The documentation of the `spacy-dbpedia-spotlight` Python library referred to above provides a useful overview regarding supported languages.

start	text	dbpedia_uri
5	city	http://dbpedia.org/resource/City
13	Turin	http://dbpedia.org/resource/Turin
45	Po	http://dbpedia.org/resource/Po_(river)

Note: Entity types annotated by DBpedia Spotlight are omitted for legibility.

Table 1: Entities returned by DBpedia Spotlight

provides a machine-translated English version of all corpora, further broadening the applicability of the approach. Realizing a robust Entity Linking workflow for ParlaMint thus opens up avenues for a host of corpora in the realm of parliamentary research, facilitating both longitudinal and comparative research by enriching the textual data with URIs (see also van Heusden et al., 2022).

The data used in this example application is taken from the linguistically annotated sample of ParlaMint for Great Britain provided in the ParlaMint GitHub repository. The chosen single sample file is based on the corpus prepared by Matthew Coole as part of the ParlaMint 4.0 release (Erjavec et al., 2023b).[7] Since the following steps only illustrate the Entity Linking process in general, the specific file has been chosen rather arbitrarily after it became apparent that the document contained substantive speech and, in consequence, entities which could be linked to a knowledge graph.

With ParlaMint being well-formed XML, the data is first read into R using the `xml2` R package (Wickham et al., 2023).

3.3.2. Entity Linking and Parsing

To start the Entity Linking process, the package is loaded.

```
library(dbpedia)
```

When the package is first loaded, setup messages inform the user about the endpoint of the service and the chosen language. It will also indicate whether DBpedia Spotlight is running locally in a Docker container. While the endpoint indicates where the queries are sent to, the language parameter indicates the chosen language model and is used to select a list of stop words which are excluded from the Entity Linking process. Both the endpoint and the language parameters are used as arguments in the main function of the package presented below.

`dbpedia` provides the `get_dbpedia_uris()` method which takes care of pre-processing the data, interaction with DBpedia Spotlight as well as the parsing of the linking results into a format which is appropriate for different analysis scenarios. The method can handle different input formats such as tokenized XML.

In keeping with the motivation to streamline the process of Entity Linking when working with textual data, the set of commands and parameters was carefully chosen to limit the number of confusing and potentially overwhelming options. Nevertheless, the process should also be transparent and open for configuration. As such, a number of parameters can be set. The package, while still in development, provides documentation for a number of basic scenarios. The most important arguments specific for XML input are the following:

- `x`: the input XML
- `feature_tag`: a `character vector` containing the name of XML elements which should be considered for Entity Linking. Can be used to select pre-annotated named entities.
- `segment`: a `character vector` describing segments into which the document should be split (e.g. paragraphs), to account for the maximum length of documents supported by DBpedia Spotlight.
- `token_tags`: a `character vector` containing the names of XML tags representing tokens

Setting these parameters requires some knowledge about the input data. For ParlaMint it seems reasonable to segment the input using the `<seg>` tag provided in the data. Assuming that these nodes represent paragraphs, this segmentation should provide sufficient context for the entity linking approach (see also Glaser et al., 2022, p. 55). This is also related to the `max_len` parameter which indicates the maximum length of segments of text to be sent to the server in one query. The default is mainly informed by the maximum length of

[7] The file was downloaded from https://github.com/clarin-eric/ParlaMint/blob/main/Samples/ParlaMint-GB/ParlaMint-GB_2022-07-21-commons.ana.xml on 2024-02-06. As stated in this example file, the corpus is licensed under the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/) and the Open Parliament Licence v3.0 (https://www.parliament.uk/site-information/copyright-parliament/open-parliament-licence/).

characters which can be reliably processed in one query before DBpedia Spotlight starts to return errors. The feature_tag parameter can be useful when data is already pre-annotated with named entities and the envisioned analysis focuses on specific elements such as persons, organizations and locations. In this case, dbpedia will limit the output to entity links which exactly match the pre-annotated entities. Otherwise, the method will return a large number of entities of all kinds of types. The parameters confidence, support and types are described by Mendes et al. (2011, pp. 3-4). All arguments are also documented in the package.

```r
annotations <- get_dbpedia_uris(
  x = xml_doc,
  language = getOption("dbpedia.lang"),
  feature_tag = NULL,
  segment = "seg",
  token_tags = c("w", "pc"),
  text_tag = "text",
  max_len = 5600L,
  confidence = 0.7,
  api = getOption("dbpedia.endpoint"),
  types = character(),
  support = 20,
  expand_to_token = FALSE,
  drop_inexact_annotations = TRUE,
  verbose = TRUE
)
```

After this call, the method creates a token stream using the elements (“token_tags”) of each segment (“seg”) and sends it to the DBpedia Spotlight service. DBpedia Spotlight identifies token spans representing entities and assigns types as well as URIs of entries in the DBpedia knowledge graph to these spans.

3.3.3. Working with the Output

get_dbpedia_uris() returns a tabular representation of identified entities and additional information such as individual entity types for many entities. Depending on the input format, character offsets or token IDs describing the position of the enriched entity are returned as well. Table 1 above illustrates this for the input of character vectors, while table 2 shows the output for the ParlaMint XML format. Table 3 visualizes retrieved entities for a single segment.

3.3.4. Enrichment with SPARQL

Often, the addition of DBpedia URIs is not the final objective of the approach but only an intermediate step to enrich entities with information available in external knowledge graphs such as DBpedia itself or Wikidata. Since the community can directly add information to the latter, Wikidata can be particularly interesting as a target knowledge graph to enrich textual data with additional information via Entity Linking (Möller et al., 2022, pp. 936-938).

In line with the aspiration to provide a cohesive workflow, dbpedia integrates the functionality to query DBpedia as well as Wikidata using the SPARQL query language. The respective functions dbpedia_get_wikidata_uris() and wikidata_query() facilitate this enrichment. Both functions work as wrappers included to alleviate some of the burden to construct valid SPARQL queries for specific endpoints of the knowledge graphs. In a nutshell, both functions take URIs as an input, prepare a SPARQL query using a template and send it to the respective SPARQL endpoints. The main functionality of dbpedia_get_wikidata_uris() is the retrieval of Wikidata IDs based on the owl:sameAs property provided by the knowledge graph. If desired, additional information – e.g. the ISO code of countries – could be retrieved. In this example, we focus only on the retrieval of Wikidata IDs. Note that rate limits and other limitations apply for the public endpoint.[8]

```r
endpnt <- "https://dbpedia.org/sparql/"

wd_uris <- dbpedia_get_wikidata_uris(
  annotations[["dbpedia_uri"]],
  endpoint = endpnt,
  wait = 5,
  chunksize = 100,
  progress = TRUE
)
```

The returned values suggest that mapping DBpedia URIs to Wikidata IDs is not without challenges. owl:sameAs often contains multiple Wikidata IDs for a single DBpedia URI. For example, for the entity “United_Kingdom”, three Wikidata IDs are returned by DBpedia which describe the “United Kingdom” as a “country in northwest Europe” (Q145), the “United Kingdom of Great Britain and Ireland” as a “historical sovereign state (1801–1922)” (Q174193) and “Great Britain” as an “island in the North Atlantic Ocean off the northwest coast of continental Europe” (Q23666).[9]

This observation is already described by Glaser et al. (2022, p. 55). To address this, Glaser et al. (2022) compare the labels of both knowledge graphs to identify the correct Wikidata ID for each item. van Heusden et al. (2022, p. 49) suggest an approach to identify missing Wikidata IDs by retrieving the Wikipedia page the DBpedia item is based on. This allows them to gather the Wikidata

[8] See the documentation here https://www.dbpedia.org/resources/sparql/ (2024-02-26).

[9] Cited passages refer to the entity labels of the three items on Wikidata as of 2024-02-16.

original_id	dbpedia_uri	text
ParlaMint-GB_2022-07-21-commons.seg5.2.10 ParlaMint-GB_2022-07-21-commons.seg5.2.11	http://dbpedia.org/resource/Free_trade	free trade
ParlaMint-GB_2022-07-21-commons.seg5.2.14	http://dbpedia.org/resource/India	India
ParlaMint-GB_2022-07-21-commons.seg870.1.10	http://dbpedia.org/resource/Glasgow	Glasgow
ParlaMint-GB_2022-07-21-commons.seg870.1.13	http://dbpedia.org/resource/Scotland	Scotland
ParlaMint-GB_2022-07-21-commons.seg870.1.17 ParlaMint-GB_2022-07-21-commons.seg870.1.18	http://dbpedia.org/resource/United_Kingdom	United Kingdom
ParlaMint-GB_2022-07-21-commons.seg870.5.15	http://dbpedia.org/resource/Christmas	Christmas

Note: Two illustrative segments of the sample document. Removed columns 'segment_id' and 'types' for improved legibility. Additional line breaks for Token IDs in column 'original_id'.

Table 2: Entities returned by DBpedia Spotlight - Tabular Overview

segment_id	text	entities
ParlaMint-GB_2022-07-21-commons.seg5	1. What progress her Department has made on securing a free trade agreement with India.	free trade (http://dbpedia.org/resource/Free_trade) | India (http://dbpedia.org/resource/India)

Table 3: Entities returned by DBpedia Spotlight - In Segments

ID indirectly. Following this approach could make it possible to identify a suitable ID if more than one Wikidata ID is provided for an entity in the DBpedia knowledge graph. In this case, instead of using the `owl:sameAs` property, the Wikipedia page would be queried and mapped to its corresponding Wikidata ID.

The ontology of Wikidata could also be used to distinguish different entities. Using the example above, the different versions of "United Kingdom" could be queried on Wikidata to retrieve the instances they are a part of (property P31) such as "sovereign state" (Q3624078), "island" (Q23442) or "historical country" (Q3024240). `dbpedia` includes the functionality for this to make this step easier via the `wikidata_query()` function which uses the `WikidataQueryServiceR` R package (Popov, 2020) and queries the Wikidata Query Service SPARQL endpoint.[10] As above, rate limits apply.

```
wd_ids <- c("Q145", "Q174193", "Q23666")

wd_props <- wd_ids |>
  wikidata_query(
    id = "P31",
    progress = TRUE)
```

However, when using Wikidata in this way, the assignment relies on the specific configuration of the knowledge graph. For instance, while this would allow to select only items which describe "sovereign states", both item Q145 (which we likely would keep as the appropriate Wikidata ID) and item Q174193 (the "historical sovereign state") are instances of this class in the knowledge graph. For the latter item, the instance of "sovereign state" is not returned by the SPARQL query above because in this specific query the returned value is limited to the highest ranked value in the statement.[11]

In consequence, while the integration of querying additional knowledge graphs seems useful for the scope and purpose of the package, there are limits to its current implementation. Addressing more complex applications is still to be tested. Ultimately, what `dbpedia_get_wikidata_uris()` and `wikidata_query()` facilitate are basic queries and the enrichment of DBpedia URIs with plausible Wikidata IDs and some additional data. More complex scenarios which also require some knowledge about the underlying knowledge graph and its ontology and structure can still be addressed with SPARQL queries regardless of the features of this package, however.

[10] https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service (2024-02-21).

[11] This is described in the documentation of Wikidata: https://www.wikidata.org/wiki/Help:Ranking (2024-02-16).

3.3.5. Enrichment of XML

A crucial feature of dbpedia is the possibility to map identified entities back to the tokens in the input data. After the DBpedia URIs, Wikidata IDs and additional properties are retrieved, the function xml_enrich() takes care of this. It extracts entities from the annotation table and maps them onto the input data via their relative IDs. For ParlaMint this technically comprises of either adding parent nodes to tokens which are identified as entities by DBpedia Spotlight or enriching existing entity annotations with additional attributes describing the type and URI of the entity. Regarding the enrichment of existing entity annotations, it has to be noted that the alignment of pre-existing and newly added entity spans can be challenging and is under development in the current version of the dbpedia R package. As discussed above, DBpedia Spotlight returns types for many entities. These often include references to types in a number of different knowledge graphs and ontologies. Since the encoding guidelines of ParlaMint limit possible values for the “type” attribute, types returned by DBpedia Spotlight can be mapped onto this allowed set of values to adhere to specific guidelines or applications.[12]

Aside from the annotation table created by the get_dbpedia_uris() method, the arguments of the function account for the name of nodes which potentially contain entities, a name for the entity nodes to be added or enriched as well as the names of columns in the annotation table which should be added as XML attributes. For a visualization of these modifications, please see the listings in the appendix (A and B) which represent a single sentence of the document.

```
xml_enrich(
  xml = xml_doc,
  annotation_dt = annotations,
  entity_name = "name",
  token_tags = c("w", "pc"),
  feature_tag = "name",
  ref = "dbpedia_uri",
  type = "category"
)
```

4. Limitations and Next Steps

The R package dbpedia provides an intuitive and cohesive workflow to perform Entity Linking using the DBpedia Spotlight Entity Linking tool with a variety of input formats. In its current state there are some limitations concerning Entity Linking in social science research as a whole and the design principles and applicability of the R package dbpedia in particular.

Regarding Entity Linking with DBpedia Spotlight, we currently lack benchmarks on the actual performance of DBpedia Spotlight when applied to parliamentary research and beyond. While benchmarks provided by the developers of DBpedia Spotlight (Daiber et al., 2013) and others indicate the usefulness of the approach, given the specificities of research scenarios in social science research, further steps of quality control should be taken. This is of particular relevance as the importance of the specific domain of textual data for approaches and corresponding benchmarks for Entity Linking is subject of some discussion and challenges (van Erp et al., 2016, pp. 4377-4378). As discussed above, the study by Olieman et al. (2015) presents some crucial insights into the performance of DBpedia Spotlight concerning Dutch parliamentary proceedings and van Heusden et al. (2022) provide some valuable perspectives on the general performance of different approaches for parliamentary debates across different languages. However, further evaluation would be crucial for substantive research. When does the approach work and when does it fail? Which accuracy can be expected? How does this affect substantive downstream tasks? In comparative parliamentary research, for example, the applicability of the approach might not only depend on the language or genre of a text but also on other aspects such as time. If the reliability of results varies over time, substantive results might depend on whether the performance of Entity Linking is worse on older documents than on more recent ones or vice versa.

Focusing on making the approach easier to use as a necessary starting point, this contribution does not yet add to these perspectives on the performance of DBpedia Spotlight. However, despite potential challenges when evaluating Entity Linking systems and creating reliable gold standard annotation (Olieman et al., 2017), given its relevance for the applicability of the approach in the envisioned scenarios and its broader adoption, the estimation of its performance as well as accompanying guidelines and advice on how to best facilitate reliable research is a crucial next step.

DBpedia Spotlight was purposefully chosen as the backbone of the package. Given its relatively easy deployment in particular, the implementation of Entity Linking with this tool provides a great baseline to address questions of usefulness, accessibility and the actual usage of the approach in social science research. This also means that the current approach relies on the DBpedia knowledge graph. However, considering the recent promi-

[12] According to https://clarin-eric.github.io/ParlaMint/#sec-ner allowed types in ParlaMint are PER (person), LOC (location), ORG (organization) and MISC (miscellaneous) (2024-03-31).

nence of Wikidata which could also be used as a direct target for Entity Linking (Möller et al., 2022) and the challenges of mapping DBpedia URIs to Wikidata, finding better solutions to access other knowledge graphs is worth pursuing.

5. Conclusion

`dbpedia` aims to make innovations in Natural Language Processing and Information Retrieval accessible in the social science community in order to facilitate the combination of unstructured data such as textual data and data such as survey data and administrative data. This contribution illustrated the possibilities of an integrated workflow for parliamentary debates in the form of corpora in the ParlaMint encoding schema. The package allows to create immediate representations of extracted and disambiguated entities, but also facilitates the addition of the enriched data to the initial corpora. This, in turn, makes it possible to use this additional information – for example statistical data added via extracted URIs – in workflows scholars working with corpora are already familiar with, for example by creating relevant subsets of documents or deploying common methods of corpus analysis.

Since it is work-in-progress, the functionality of the package is subject to future changes. The current focus of `dbpedia` is on the development of a slim set of functions and commands which apply in different scenarios.

As indicated in the previous section, there are some obvious next steps: We neither discussed the substantive performance of the approach, nor is DBpedia Spotlight the only Entity Linking solution worth considering. While the local deployment and performance are advantages, there are more recent developments which should be evaluated. For future research, this might entail complementing `dbpedia` with other components which build on the functionality and API of the presented package and facilitate the integration of different approaches. In consequence, a modular design of tools and workflows might be needed which can handle different but standardized input formats such as the ParlaMint corpora and beyond.

6. Acknowledgement

The development of the `dbpedia` R package is made possible by funding from KonsortSWD (project number 442494171) within the German National Research Data Infrastructure (Nationale Forschungsdateninfrastruktur/NFDI). This funding is gratefully acknowledged. We also want to thank the anonymous reviewers for their insightful comments and useful suggestions.

7. Bibliographical References

Tareq Al-Moslmi, Marc Gallofré Ocaña, Andreas L. Opdahl, and Csaba Veres. 2020. Named Entity Extraction for Knowledge Graphs: A Literature Overview. *IEEE Access*, 8:32862–32881.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ryan Bakker, Catherine de Vries, Erica Edwards, Liesbet Hooghe, Seth Jolly, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen, and Milada Anna Vachudova. 2015. Measuring party positions in Europe: The Chapel Hill expert survey trend file, 1999–2010. *Party Politics*, 21(1):143–152.

Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774.

Ian Budge and Judith Bara. 2001. Introduction: Content Analysis and Political Texts. In Ian Budge, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, and Eric Tanenbaum, editors, *Mapping Policy Preferences. Estimates for Parties, Electors, and Governments 1945-1998*, pages 1–16. Oxford University Press, Oxford; New York.

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, I-SEMANTICS '13, pages 121–124, New York, NY, USA. Association for Computing Machinery.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darģis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2023a. The ParlaMint corpora of parliamentary proceedings.

Language Resources and Evaluation, 57:415–448.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millenium. In *Proceedings of the Corpus Linguistics 2011 conference*, University of Birmingham, UK.

Luis Glaser, Ronny Patz, and Manfred Stede. 2022. UNSC-NE: A Named Entity Extension to the UN Security Council Debates Corpus. *Journal for Language Technology and Computational Linguistics*, 35(2):51–67.

Mitchell Linegar, Rafal Kocielnik, and R. Michael Alvarez. 2023. Large language models and political science. *Frontiers in Political Science*, 5.

Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G. Moreno, Emanuela Boros, Ahmed Hamdi, Nicolas Sidère, Mickaël Coustaty, and Antoine Doucet. 2020. Entity Linking for Historical Documents: Challenges and Solutions. In *Digital Libraries at Times of Massive Societal Transition*, volume 12504 of *Lecture Notes in Computer Science*, pages 215–231, Cham. Springer International Publishing.

Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia Spotlight: Shedding Light on the Web of Documents. In *Proceedings of the 7th International Conference on Semantic Systems*, Graz, Austria.

Cedric Möller, Jens Lehmann, and Ricardo Usbeck. 2022. Survey on English Entity Linking on Wikidata. Datasets and approaches. *Semantic Web*, 13:925–966.

Alex Olieman, Hosein Azarbonyad, Mostafa Dehghani, Jaap Kamps, and Maarten Marx. 2014. Entity linking by focusing DBpedia candidate entities. In *Proceedings of the First International Workshop on Entity Recognition & Disambiguation*, ERD '14, pages 13–24, New York, NY, USA. Association for Computing Machinery.

Alex Olieman, Kaspar Beelen, Milan van Lange, Jaap Kamps, and Maarten Marx. 2017. Good Applications for Crummy Entity Linkers? The Case of Corpus Selection in Digital Humanities. In *Proceedings of the 13th International Conference on Semantic Systems*, Amsterdam, Netherlands.

Alex Olieman, Jaap Kamps, Maarten Marx, and Arjan Nusselder. 2015. A Hybrid Approach to Domain-Specific Entity Linking. *arXiv:1509.01865*.

Jeroen Ooms. 2014. The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. *arXiv:1403.2805*.

Mikhail Popov. 2020. *WikidataQueryServiceR: API Client Library for 'Wikidata Query Service'*. R package version 1.0.0.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Marieke van Erp, Pablo N. Mendes, Heiko Paulheim, Filip Ilievski, Julien Plu, Giuseppe Rizzo, and Joerg Waitelonis. 2016. Evaluating Entity Linking: An Analysis of Current Benchmark Datasets and a Roadmap for Doing a Better Job. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4373–4379, Portorož, Slovenia. European Language Resources Association (ELRA).

Ruben van Heusden, Maarten Marx, and Jaap Kamps. 2022. Entity Linking in the ParlaMint Corpus. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 47–55, Marseille, France. European Language Resources Association.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85. Publisher: ACM New York, NY, USA.

Hadley Wickham. 2023. *httr: Tools for Working with URLs and HTTP*. R package version 1.4.7.

Hadley Wickham, Jim Hester, and Jeroen Ooms. 2023. *xml2: Parse XML*. R package version 1.3.5.

8. Language Resource References

Erjavec, Tomaž and Kopp, Matyáš and Ogrodniczuk, Maciej and Osenova, Petya and Agerri, Rodrigo and Agirrezabal, Manex and Agnoloni, Tommaso and Aires, José and Albini, Monica and Alkorta, Jon and Antiba-Cartazo, Iván and Arrieta, Ekain and Barcala, Mario and Bardanca, Daniel and Barkarson, Starkaður and Bartolini, Roberto and Battistoni, Roberto and Bel, Nuria and Bonet Ramos, Maria del Mar and Calzada Pérez, María and Cardoso, Aida and Çöltekin, Çağrı and Coole, Matthew and Darģis, Roberts and de Does, Jesse and de Libano, Ruben and

Depoorter, Griet and Depuydt, Katrien and Diwersy, Sascha and Dodé, Réka and Fernandez, Kike and Fernández Rei, Elisa and Frontini, Francesca and Garcia, Marcos and García Díaz, Noelia and García Louzao, Pedro and Gavriilidou, Maria and Gkoumas, Dimitris and Grigorov, Ilko and Grigorova, Vladislava and Haltrup Hansen, Dorte and Iruskieta, Mikel and Jarlbrink, Johan and Jelencsik-Mátyus, Kinga and Jongejan, Bart and Kahusk, Neeme and Kirnbauer, Martin and Kryvenko, Anna and Ligeti-Nagy, Noémi and Ljubešić, Nikola and Luxardo, Giancarlo and Magariños, Carmen and Magnusson, Måns and Marchetti, Carlo and Marx, Maarten and Meden, Katja and Mendes, Amália and Mochtak, Michal and Mölder, Martin and Montemagni, Simonetta and Navarretta, Costanza and Nitoń, Bartłomiej and Norén, Fredrik Mohammadi and Nwadukwe, Amanda and Ojsteršek, Mihael and Pančur, Andrej and Papavassiliou, Vassilis and Pereira, Rui and Pérez Lago, María and Piperidis, Stelios and Pirker, Hannes and Pisani, Marilina and van der Pol, Henk and Prokopidis, Prokopis and Quochi, Valeria and Rayson, Paul and Regueira, Xosé Luís and Rudolf, Michał and Ruisi, Manuela and Rupnik, Peter and Schopper, Daniel and Simov, Kiril and Sinikallio, Laura and Skubic, Jure and Tamper, Minna and Tungland, Lars Magne and Tuominen, Jouni and van Heusden, Ruben and Varga, Zsófia and Vázquez Abuín, Marta and Venturi, Giulia and Vidal Miguéns, Adrián and Vider, Kadri and Vivel Couso, Ainhoa and Vladu, Adina Ioana and Wissik, Tanja and Yrjänäinen, Väinö and Zevallos, Rodolfo and Fišer, Darja. 2023b. *Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 4.0*. ISSN: 2820-4042.

Appendices

A. ParlaMint Example XML output before Entity Linking

```
<s xml:id=”ParlaMint-GB_2022-07-21-commons.seg5.2”>
  <w lemma=”what” msd=”UPosTag=DET|PronType=Int” pos=”WDT” xml:id=”
      ParlaMint-GB_2022-07-21-commons.seg5.2.1”>What</w>
  <w lemma=”progress” msd=”UPosTag=NOUN|Number=Sing” pos=”NN” xml:id=”
      ParlaMint-GB_2022-07-21-commons.seg5.2.2”>progress</w>
  <w lemma=”she” msd=”UPosTag=DET|Gender=Fem|Number=Sing|Person=3|Poss=Yes|
      PronType=Prs” pos=”PRP$” xml:id=”ParlaMint-GB_2022-07-21-commons.seg5
      .2.3”>her</w>
  <w lemma=”Department” msd=”UPosTag=PROPN|Number=Sing” pos=”NNP” xml:id=”
      ParlaMint-GB_2022-07-21-commons.seg5.2.4”>Department</w>
  <w lemma=”have” msd=”UPosTag=VERB|Mood=Ind|Number=Sing|Person=3|Tense=
      Pres|VerbForm=Fin” pos=”VBZ” xml:id=”ParlaMint-GB_2022-07-21-commons.
      seg5.2.5”>has</w>
  <w lemma=”make” msd=”UPosTag=VERB|Tense=Past|VerbForm=Part” pos=”VBN” xml
      :id=”ParlaMint-GB_2022-07-21-commons.seg5.2.6”>made</w>
  <w lemma=”on” msd=”UPosTag=ADP” pos=”IN” xml:id=”ParlaMint-GB_2022-07-21-
      commons.seg5.2.7”>on</w>
  <w lemma=”secure” msd=”UPosTag=VERB|VerbForm=Ger” pos=”VBG” xml:id=”
      ParlaMint-GB_2022-07-21-commons.seg5.2.8”>securing</w>
  <w lemma=”a” msd=”UPosTag=DET|Definite=Ind|PronType=Art” pos=”DT” xml:id
      =”ParlaMint-GB_2022-07-21-commons.seg5.2.9”>a</w>
  <w lemma=”free” msd=”UPosTag=ADJ|Degree=Pos” pos=”JJ” xml:id=”ParlaMint-
      GB_2022-07-21-commons.seg5.2.10”>free</w>
  <w lemma=”trade” msd=”UPosTag=NOUN|Number=Sing” pos=”NN” xml:id=”
      ParlaMint-GB_2022-07-21-commons.seg5.2.11”>trade</w>
  <w lemma=”agreement” msd=”UPosTag=NOUN|Number=Sing” pos=”NN” xml:id=”
      ParlaMint-GB_2022-07-21-commons.seg5.2.12”>agreement</w>
  <w lemma=”with” msd=”UPosTag=ADP” pos=”IN” xml:id=”ParlaMint-GB_2022
      -07-21-commons.seg5.2.13”>with</w>
  <w join=”right” lemma=”India” msd=”UPosTag=PROPN|Number=Sing” pos=”NNP”
      xml:id=”ParlaMint-GB_2022-07-21-commons.seg5.2.14”>India</w>
  <pc msd=”UPosTag=PUNCT” pos=”.” xml:id=”ParlaMint-GB_2022-07-21-commons.
      seg5.2.15”>.</pc>
</s>
```

Listing 1: XML before Entity Linking

Note: A single sentence based on sample data for the ParlaMint 4.0 corpora (Erjavec et al., 2023b). Removed syntactic information for better legibility. See footnote 7 regarding the source of the data.

B. ParlaMint Example XML output after Entity Linking

```
<s xml:id=”ParlaMint-GB_2022-07-21-commons.seg5.2”>
  <w lemma=”what” msd=”UPosTag=DET|PronType=Int” pos=”WDT” xml:id=”
      ParlaMint-GB_2022-07-21-commons.seg5.2.1”>What</w>
  <w lemma=”progress” msd=”UPosTag=NOUN|Number=Sing” pos=”NN” xml:id=”
      ParlaMint-GB_2022-07-21-commons.seg5.2.2”>progress</w>
  <w lemma=”she” msd=”UPosTag=DET|Gender=Fem|Number=Sing|Person=3|Poss=Yes|
      PronType=Prs” pos=”PRP$” xml:id=”ParlaMint-GB_2022-07-21-commons.seg5
      .2.3”>her</w>
  <w lemma=”Department” msd=”UPosTag=PROPN|Number=Sing” pos=”NNP” xml:id=”
      ParlaMint-GB_2022-07-21-commons.seg5.2.4”>Department</w>
  <w lemma=”have” msd=”UPosTag=VERB|Mood=Ind|Number=Sing|Person=3|Tense=
      Pres|VerbForm=Fin” pos=”VBZ” xml:id=”ParlaMint-GB_2022-07-21-commons.
      seg5.2.5”>has</w>
  <w lemma=”make” msd=”UPosTag=VERB|Tense=Past|VerbForm=Part” pos=”VBN” xml
      :id=”ParlaMint-GB_2022-07-21-commons.seg5.2.6”>made</w>
  <w lemma=”on” msd=”UPosTag=ADP” pos=”IN” xml:id=”ParlaMint-GB_2022-07-21-
      commons.seg5.2.7”>on</w>
  <w lemma=”secure” msd=”UPosTag=VERB|VerbForm=Ger” pos=”VBG” xml:id=”
      ParlaMint-GB_2022-07-21-commons.seg5.2.8”>securing</w>
  <w lemma=”a” msd=”UPosTag=DET|Definite=Ind|PronType=Art” pos=”DT” xml:id
      =”ParlaMint-GB_2022-07-21-commons.seg5.2.9”>a</w>
  <name type=”MISC” ref=”http://dbpedia.org/resource/Free_trade”>
    <w lemma=”free” msd=”UPosTag=ADJ|Degree=Pos” pos=”JJ” xml:id=”ParlaMint
        -GB_2022-07-21-commons.seg5.2.10”>free</w>
    <w lemma=”trade” msd=”UPosTag=NOUN|Number=Sing” pos=”NN” xml:id=”
        ParlaMint-GB_2022-07-21-commons.seg5.2.11”>trade</w>
  </name>
  <w lemma=”agreement” msd=”UPosTag=NOUN|Number=Sing” pos=”NN” xml:id=”
      ParlaMint-GB_2022-07-21-commons.seg5.2.12”>agreement</w>
  <w lemma=”with” msd=”UPosTag=ADP” pos=”IN” xml:id=”ParlaMint-GB_2022
      -07-21-commons.seg5.2.13”>with</w>
  <name type=”LOC” ref=”http://dbpedia.org/resource/India”>
    <w join=”right” lemma=”India” msd=”UPosTag=PROPN|Number=Sing” pos=”NNP”
        xml:id=”ParlaMint-GB_2022-07-21-commons.seg5.2.14”>India</w>
  </name>
  <pc msd=”UPosTag=PUNCT” pos=”.” xml:id=”ParlaMint-GB_2022-07-21-commons.
      seg5.2.15”>.</pc>
</s>
```

Listing 2: XML after Entity Linking

Note: A single sentence based on sample data for the ParlaMint 4.0 corpora (Erjavec et al., 2023b). Removed syntactic information for better legibility. See footnote 7 regarding the source of the data.

Video Retrieval System Using Automatic Speech Recognition for the Japanese Diet

Mikitaka Masuyama[1], Tatsuya Kawahara[2], Kenjiro Matsuda[3]
National Graduate Institute for Policy Studies[1], Kyoto University[2], Kobe Shoin Women's University[3]
Minato-ku, Tokyo[1], Sakyo-ku, Kyoto[2], Nada-ku, Kobe[3], Japan
mmasuyama@grips.ac.jp[1], kawahara@i.kyoto-u.ac.jp[2], kenjiro@shoin.ac.jp[3]

Abstract

The Japanese House of Representatives, one of the two houses of the Diet, has adopted an Automatic Speech Recognition (ASR) system, which directly transcribes parliamentary speech with an accuracy of 95 percent. The ASR system also provides a timestamp for every word, which enables retrieval of the video segments of the Parliamentary meetings. The video retrieval system we have developed allows one to pinpoint and play the parliamentary video clips corresponding to the meeting minutes by keyword search. In this paper, we provide its overview and suggest various ways we can utilize the system. The system is currently extended to cover meetings of local governments, which will allow us to investigate dialectal linguistic variations.

Keywords: speech recognition, video retrieval, keyword search

1. Introduction

In recent times, there has been a surge in the development of analytical tools and techniques for analyzing the textual data of parliamentary proceedings. However, with the growing trend of parliamentary video streaming, there is a pressing need for similar tools to be developed for audio-visual data. While visual data offers a clear advantage over textual data for a more comprehensive analysis of parliamentary debates, it can be challenging to pinpoint the exact scene of a particular utterance by a specific speaker in lengthy video recordings that can span for hours.

To remedy this situation, we have launched an Internet video retrieval system for the Japanese Diet. Using the speech recognition system dedicated to Parliamentary speech which creates timestamp data to match parliamentary video feeds and the minutes of proceedings, it can pinpoint and play the parliamentary video clips corresponding to the minutes of proceedings through keyword search.

2. Video Retrieval System for Diet Deliberations

One of the authors has developed automatic speech recognition (ASR) technology, which the Japanese House of Representatives has deployed in the transcription system since 2011. The ASR system was trained with a large amount of parliamentary speech data, which covers terms and expressions used in the Parliament (Kawahara 2012, Kawahara 2017). It introduced an efficient lightly-supervised training based on statistical language model transformation, which fills the gap between faithful transcripts of spoken utterances and final texts for documentation. Once the mapping is trained, faithful transcripts for training acoustic and language models are no longer needed. The ASR system has consistently achieved character accuracy of over 90% since 2011, which helps streamline the transcription process. The accuracy rate currently has improved to 95 percent.

The Diet Library currently provides digitized minutes of parliamentary meetings via the Internet. Although these are not “official” records, they are amenable to keyword searching. On the other hand, we can watch the online live streaming at each house’s secretariat website. We can also search the video library and watch videos of parliamentary meetings. The House of Representatives has made the parliamentary videos available since 2010, while the House of Councillors, the other house of the Diet, makes the videos available only one year after the meetings.

https://www.shugiintv.go.jp/index.php

https://www.webtv.sangiin.go.jp/webtv/index.php

Diet deliberation videos can be searched by meeting date, meeting title, subject, and speaker, although the English interface only offers the first two search options. Even if we successfully retrieve the desired deliberation video, we must watch the video from the beginning to the speech or debate segment we are interested in. It is not uncommon for a committee meeting to last more than 7 hours. While the video breakdown by questioner is available in the Japanese interface, video segmentation is usually 30 to 60 minutes long. No such breakdown is available in the English interface. Moreover, replies to parliamentary questions are arranged by the questioner, and we cannot search prime and cabinet ministers’ deliberation videos answering parliamentary questions.

By linking the Diet Library’s proceedings database and the Diet secretariats’ deliberation video libraries, our “Video Retrieval System for Diet Deliberations (VRS)” makes it possible to retrieve the video clips corresponding to the minutes of the parliamentary meetings through keyword searching:

https://gclip1.grips.ac.jp/video/

With our system, we can directly retrieve the portion of the video feed we are interested in. We can instantly gain a visual understanding of the flow of parliamentary debate and check the facial

May 20, 2024.

expressions and body language of the speaker, all of which are not possible from a simple reading of the minutes of parliamentary meetings.

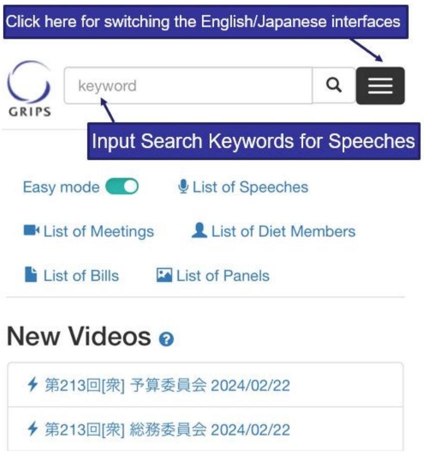
Our video retrieval system consists of two sub-systems. One uses the latest speech recognition techniques to create timestamp data to match the Diet Library's proceedings database and the Diet secretariats' deliberation video databases. The second sub-system uses the timestamp data to search the parliamentary minutes stored in our system and retrieve the Diet deliberation videos corresponding to the minute in question by keyword search. The results of keyword searches are deliberation video links, and the portion of the video we are interested in can be played partially by clicking the URL link for the deliberation video available in the Diet secretariats' databases (not stored in our system).

The system has been in operation and publicly available since November 2012. It is possible to keyword search all the plenary and committee meetings in the House of Representatives since January 2010 and those in the House of Councillors since December 2012[1].

Below, we briefly describe how our video retrieval system works. Figure 1 shows the top page of our web-based search interface, allowing us to search for deliberation video segments by typing keywords. The Japanese interface will appear when the user clicks "Japanese" in the upper right-hand corner.

One can type English keywords separated by spaces in the search field, and they will be translated automatically into Japanese and used in keyword searching. For instance, if one types "Kishida Fumio" (the name of the current Prime Minister of Japan) and "tax increase" in the search field and hits the search button, a list of the search results will appear in ascending order of date (Figure 2). As the default setting, our system searches the database for the past year, although it can be extended or shortened by calendar and filtered by other factors in the search results interface. Then, one can click one of the video links, and our system will instantly play the portion of the video corresponding to the speech, including the keywords (Figure 3).

The video-playing interface shows subtitles under the video and the speeches at the meeting on the right side, highlighting the current speech (not shown in Figure 3). By default, the video will play for one minute or three speeches. Alternatively, one can keep playing the video by clicking the play button in the toolbar under the video. Double-clicking any speech in the speech list allows one to instantly watch the video portion of the speeches before and after the speech found by keyword search. Once the user has moved on to another speech, the original speech found by keyword search remains highlighted in yellow.

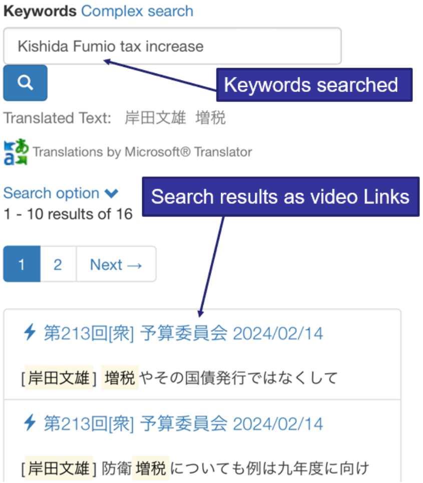


New Videos

第213回[衆] 予算委員会 2024/02/22

第213回[衆] 総務委員会 2024/02/22

Figure 1: Keyword Search Interface in English



Kishida Fumio tax increase

Translated Text: 岸田文雄 増税

Translations by Microsoft® Translator

1 - 10 results of 16

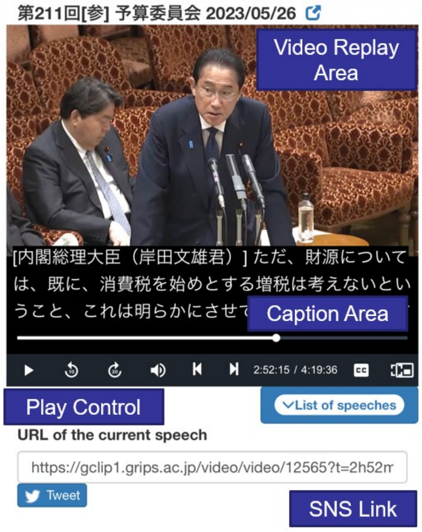
第213回[衆] 予算委員会 2024/02/14

[岸田文雄] 増税 やその国債発行ではなくして

第213回[衆] 予算委員会 2024/02/14

[岸田文雄] 防衛 増税 についても例は九年度に向け

Figure 2: Search Results Interface



第211回[参] 予算委員会 2023/05/26

[内閣総理大臣（岸田文雄君）] ただ、財源については、既に、消費税を始めとする増税は考えないということ、これは明らかにさせて

URL of the current speech

https://gclip1.grips.ac.jp/video/video/12565?t=2h52m

Figure 3: Video Replay Interface

[1] At the time of Feb. 22, 2024, our database covers the time period since Jan. 18, 2010, which includes a total of 14,156 hours, 12,685 meetings, 8,245,621 speeches, 13,282 speakers, and 554,634,003 Japanese characters.

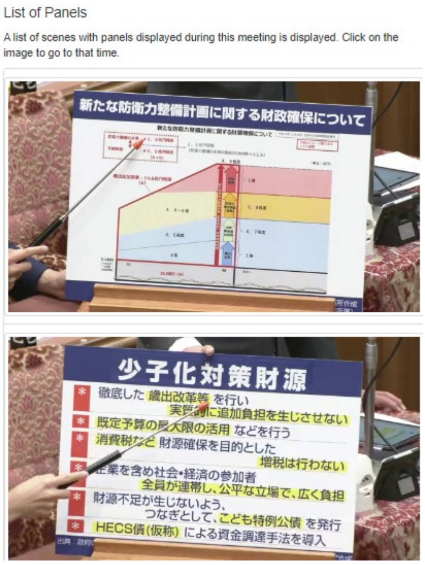
The video-playing interface shows subtitles under the video, highlighting the current speech. By default, the video will play for one minute or three utterances. Alternatively, one can keep playing the video by clicking the play button in the toolbar under the video. Double-clicking any speech in the speech list (not shown in Figure 3) allows one to instantly watch the video portion of the speeches before and after the speech found by keyword search. Once the user has moved on to another speech, the original speech found by keyword search remains highlighted in yellow.

Moreover, the video-playing interface shows the URL for the corresponding video portion, and one can easily share the URL via SNS by clicking the tweet button while the video stream is playing. The text of the speech and the URL will immediately appear in the tweet box. Moreover, the bottom of the page offers information about the speaker, followed by a list of agendas and the Diet members attending the meeting (not shown in Figure 3).

3. Usage beyond Keyword Search

We can utilize our video retrieval system in various ways. For instance, it allows us to obtain the URL for a moment of video streaming and to create and share a list of video links without downloading and editing the video files. Another way of utilizing the interfaces for keyword searching and partial replay is to post deliberation video links to internet news.

The minutes are essential for parliamentary discussion but do not tell the whole story. For instance, supplementary materials often used in committee meetings are graphic materials such as figures and tables, which concisely summarize the discussion points but are not usually included in the minutes. Thus, we combined speech and pattern recognition techniques to distinguish between the portions of videos that focus on the speaker and automatically extract video clips, including the moments focusing on supplementary materials used in committee meetings. Furthermore, we have developed an automatic text recognition system for these clips to extract and store text information in the database to be amenable to keyword search so that our system searches the video portion, focusing on the supplementary materials by keyword search through their content (Figure 4). The minutes are silent regarding non-verbal communication, and we are developing a web-based program to automatically extract and analyze the speaker's facial expressions and body language[2].



List of Panels

A list of scenes with panels displayed during this meeting is displayed. Click on the image to go to that time.

新たな防衛力整備計画に関する財政確保について
少子化対策財源
徹底した歳出改革等を行い
実質的に追加負担を生じさせない
既定予算の最大限の活用などを行う
消費税など財源確保を目的とした
増税は行わない
企業を含め社会・経済の参加者
全員が連帯し、公平な立場で、広く負担
財源不足が生じないよう、
つなぎとして、こども特例公債を発行
HECS債(仮称)による資金調達手法を導入

Figure 4: Supplementary Materials

The speech recognition output may contain irrelevant fillers and incorrectly recognized errors. The minutes become "official" by eliminating fillers, correcting inappropriate wording, and adding commas and periods. With our web-based program, we can systematically analyze the correspondence between the official minutes and speech recognition results. We can also check how we pronounce heteronyms, common in Japanese. While it is impossible to detect by reading the minutes, we can utilize our video retrieval system to analyze how parliamentary members pronounce heteronyms through keyword searching[3].

4. Conclusion

Our video retrieval system has great potential to boost the usage of parliamentary information. The speech recognition techniques for creating timestamp data for matching video and text information can apply to various meetings, including local assemblies, international conferences, and other less formal public and private meetings. For instance, many local assemblies in Japan increasingly use YouTube to disseminate deliberation videos. By extending the video retrieval system to such local assemblies, we can expect to improve speech recognition for dialectal diversity. Also, since some parliaments use multiple languages, we can develop multi-linguistic speech recognition by extending our system to such parliaments. Furthermore, international conferences like the United Nations Commission on Human Rights have stopped producing conference proceedings and recently disseminated meeting videos. A video retrieval system like ours may become the only way to search the content of such meetings.

[2] There are studies extracting emotions from minutes and videos (Rheault et al. 2016, Werlen et al. 2021, Rheault & Borwein 2019) and comparing verbal and non-verbal emotions (Werlen et al. 2018).

[3] Linguistic scholars focus on how politicians pronounce Iraq and figure out their diplomatic stance (Hall-Lew et al. 2010). Political scientists try to unravel politicians' gender differences in discussing women's issues by analyzing pitch (Dietrich 2019).

5. Bibliographical References

Kawahara. T. (2012). Transcription system using automatic speech recognition for the Japanese Parliament (Diet). In Proc. AAAI/IAAI, pp.2224—2228.

Kawahara T. (2017). Automatic meeting transcription system for the Japanese Parliament (Diet). In Proc. APSIPA ASC.

Dietrich et al. (2019). “Pitch Perfect: Vocal Pitch and the Emotional Intensity of Congressional Speech” *American Political Science Review* 113(4) 941-962.

Hall-Lew. (2010). “Indexing Political Persuasion: Variation in the Iraq Vowels” *American Speech*. 85: 91-102.

Rheault et al. (2016). “Measuring Emotion in Parliamentary Debates with Automated Textual Analysis” *PLOS ONE* 11(12) 1-18.

Rheault & Borwein. (2019). “Multimodal Techniques for the Study of Affect in Political Videos” *Prepared for the PolMeth Conference*, MIT, Cambridge, MA, July 18-20, 2019.

Werlen et al. (2018). “Is reading mirrored in the face? A comparison of linguistic parameters and emotional facial expressions” *CEUR-WS.org*, 1-2226, paper2.

Werlen et al. (2021). “Emotions in the parliament: Lexical emotion analysis of parliamentarian speech transcriptions” *CEUR-WS.org*, 1-2957, paper10.

One Year of Continuous and Automatic Data Gathering from Parliaments of European Union Member States

Ota Mikušek
Lexical Computing, Brno, Czech Republic
ota.mikusek@sketchengine.eu

Abstract

This paper provides insight into automatic parliamentary corpora development. One year ago, I created a simple set of tools designed to continuously and automatically download, process, and create corpora from speeches in the parliaments of European Union member states. Despite the existence of numerous corpora providing speeches from European Union parliaments, the tools are more focused on collecting and building such corpora with minimal human interaction. These tools have been operating continuously for over a year, gathering parliamentary data and extending corpora, which together have more than one billion words. However, the process of maintaining these tools has brought unforeseen challenges, including issues such as being blocked by some parliaments due to overloading the parliament with requests, the inability to access the most recent data of a parliament, and effectively managing interrupted connections. Additionally, potential problems that may arise in the future are provided, along with possible solutions. These include problems with data loss prevention and adaptation to changes in the sources from which speeches are downloaded.

Keywords: parliamentary protocols, continuous downloading, corpus processing, automatic tools, corpus development, maintenance of tools

1. European Parliamentary Corpora

Between July 2020 and May 2021, the ParlaMint I (Erjavec et al., 2022) project aimed to create corpora of transcriptions from the sessions of 17 European Union parliaments from 2015 to October 2019. ParlaMint I was the largest project of its kind for European parliamentary corpora at the time. Each parliamentary corpus had a dedicated lead developer.

In December 2021, the ParlaMint II (Erjavec et al., 2021) project extended the work of ParlaMint I by including parliamentary transcriptions up to July 2022. This project also involved updates to the schema, validation, and enhancement of corpora with additional metadata.

In July 2023 ParlaMint 3.0 (Erjavec et al., 2023b) and in October 2023 ParlaMint 4.0 (Erjavec et al., 2023a) follows ParlaMint II and extend it. Currently, ParlaMint 4.0 provides 29 corpora, namely for Bulgarian, Croatian, Polish, Slovenian, Czech, Icelandic, Belgian, Danish, Spanish, Dutch, Turkish, Italian, Hungarian, Latvian, French, Bosnian, Catalonian, Galician, Greek, Norwegian, Serbian, Swedish, Ukrainian, Finnish, Estonian, Basque, United Kingdom, Portuguese and Austrian parliament.

For all corpora, ParlaMint 4.0 provides unified metadata, including timestamps, speaker details, transcriber notes, and source URLs for documents. Expanding coverage to include other parliaments is a future objective for the ParlaMint project.

In addition, there are other initiatives to create parliamentary corpora, such as the Polish Parliamentary Corpus (Ogrodniczuk, 2018), which covers debates from 1919 to the present, and the German Parliamentary Corpus (GerParCor) (Abrami et al., 2022), which includes transcripts from Germany, Liechtenstein, Austria, and Switzerland up to 2021, with plans for continuous development. The Czech Parliamentary Corpus (CzechParl) (Jakubíček and Kovář, 2010) is based on Czech parliament stenographic protocols from the 1990s. The Dutch Parliamentary Corpus (DutchParl) (Marx et al., 2010) aims to collect Dutch-language parliamentary documents and has different sized corpora for Belgium, Flanders, and the Netherlands, with ongoing development efforts.

2. Automatic Tools

A year ago, I created a toolset written in Python language providing continuous automatic development of corpora from transcriptions of parliamentary chambers from selected members of the EU. From suitable sources of parliamentary protocols on chamber websites, created scripts are gathering protocols in different formats and unifying their format as preverticals[1]. The prevertical format is a file format containing plain text and structures. The structures enclose the text and provide metadata about the text. An example of a document in prevertical format, created by the tools, is shown in Figure 1.

Created scripts are independent of each other and work autonomously, automatically, and atom-

[1] https://www.sketchengine.eu/my_keywords/prevertical/

May 20, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

icly. Each script consists of three parts: shared code, a tool for discovering and downloading new protocols, and a tool for processing downloaded protocols into prevertical files. In case of any error, scripts are able to log this error, notify the script administrator, and roll back to the last consistent state.

2.1. Downloading of Data

Reliable sources of protocols were searched on parliamentary official websites. For a source to be considered reliable, it must come directly from the parliament, it has to provide an option to discover newly added protocols, and it must not rely on website-provided scripts (mainly javascript).

The reason why script execution to access or discover new protocols is unwanted is that user-side scripts can change over time, and these changes may cause errors during the automatic download process. Such dependency is unwanted because it increases maintenance difficulty.

Found sources provided data in plain text, HTML, JSON, CSV, XML, XLSX, and DOCX format. PDF file format was also available. However, PDF format introduced problems with the ordering of the paragraphs, and text extraction, when words were split at the end of the line by “-" character. In cases when the source was not found on the parliament website, the parliament was contacted via email.

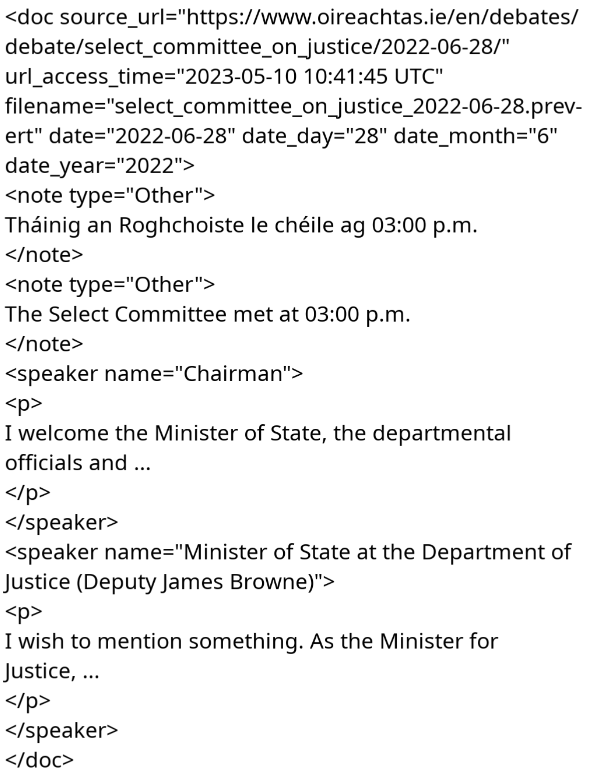
Created scripts are downloading protocols from sources automatically and atomically. If the downloading of a protocol fails, this information is logged, and the download will be retried during the next script execution.

2.2. Processing of Protocols

A script that processes downloaded protocols called prevertbuilder was created for each chamber website. The prevertbuilder is responsible for metadata extraction and unifying downloaded protocols into prevertical format. Common metadata across all corpora are the speaker name, the date, the source URL, the URL access time, and the filename where prevertical is stored. More metadata, like notes of transcriber, are also provided for some corpora.

The prevertbuilder works like a pipe. It contains the initialization, writing, and finalization methods, which process downloaded protocols linearly and do not require the whole protocol to be loaded in memory. This capability is used, for example, in the Swedish parliament, where one downloaded document consists of protocols from a month period.

A protocol is marked as successfully processed only when prevertbuilder process the protocol without an error. Prevertbuilders are capable of detecting the presence of new information (for example, new tags or attributes) in processed protocols. By default, in these cases, protocols are processed without these new elements. However, their occurrence is logged as a warning in the script log.



```
<doc source_url="https://www.oireachtas.ie/en/debates/
debate/select_committee_on_justice/2022-06-28/"
url_access_time="2023-05-10 10:41:45 UTC"
filename="select_committee_on_justice_2022-06-28.prev-
ert" date="2022-06-28" date_day="28" date_month="6"
date_year="2022">
<note type="Other">
Tháinig an Roghchoiste le chéile ag 03:00 p.m.
</note>
<note type="Other">
The Select Committee met at 03:00 p.m.
</note>
<speaker name="Chairman">
<p>
I welcome the Minister of State, the departmental
officials and ...
</p>
</speaker>
<speaker name="Minister of State at the Department of
Justice (Deputy James Browne)">
<p>
I wish to mention something. As the Minister for
Justice, ...
</p>
</speaker>
</doc>
```

Figure 1: Example of prevertical format from the upper chamber of the Irish parliament (modified)

The final corpora is created using (No)Sketch Engine (Kilgarriff et al., 2014) infrastructure and are available on Sketch Engine[2] under the name Parliament debates.

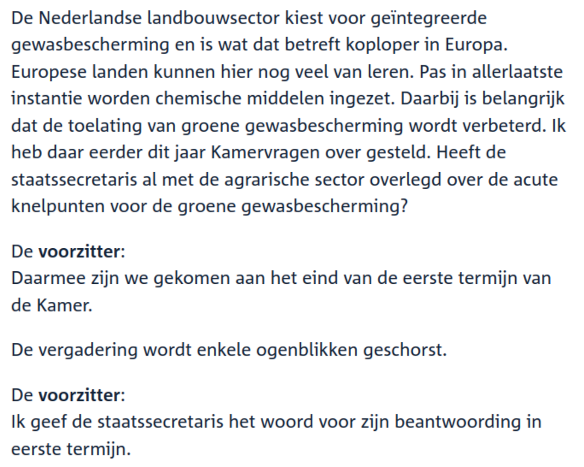
3. Flaws of Current Design

The original toolset was the first of its kind. Some of the original goals, like zero human interaction and the ability to have the most recent data could be considered naive after running them for over one year. During the maintenance of these tools, several problems were encountered.

3.1. Speaker name attribute detection

In some cases, sources do not provide the name of a speaker but just their role. This can be seen in Figure 1, where in one case, only the speaker role "Chairman" is provided, without the actual name of the speaker. This information can be acquired elsewhere and could be resolved at a possible cost

[2] https://app.sketchengine.eu/



De Nederlandse landbouwsector kiest voor geïntegreerde gewasbescherming en is wat dat betreft koploper in Europa. Europese landen kunnen hier nog veel van leren. Pas in allerlaatste instantie worden chemische middelen ingezet. Daarbij is belangrijk dat de toelating van groene gewasbescherming wordt verbeterd. Ik heb daar eerder dit jaar Kamervragen over gesteld. Heeft de staatssecretaris al met de agrarische sector overlegd over de acute knelpunten voor de groene gewasbescherming?

De **voorzitter**:
Daarmee zijn we gekomen aan het eind van de eerste termijn van de Kamer.

De vergadering wordt enkele ogenblikken geschorst.

De **voorzitter**:
Ik geef de staatssecretaris het woord voor zijn beantwoording in eerste termijn.

Figure 2: Lower parliament chamber of the Netherlans

of more dependencies and, therefore, higher maintenance.

3.2. Notes of transcriber detection

Notes of transcriber are hard to detect in the lower parliament chamber of the Netherlands. In Figure 2 is a sample of discussion[3]. All sentences were spoken except the penultimate sentence, which is a note from the transcriber saying that the sitting is suspended for a few moments.

The format of the note is indistinguishable from the rest of the spoken text. Currently, these notes remain undetected and are added as spoken text to the current speaker.

3.3. Overloading of Parliaments

In the original release of the tools, none of the tools were using delays between requests to parliamentary source. The Parliament of Denmark started to require human verification to access its website two weeks after the first run of the original tools. The Parliament of the Netherlands banned the IP address of the server where the tools were originally running. This led to a quick fix by adding random time delays between requests to each source. No more problems that could be related to overloading the parliaments were encountered since the fix.

3.4. Delay in Data Source

During the selection of a suitable source for the Finnish Parliament, I was unable to find any reliable source for the Finnish Parliament website. I contacted the Finnish Parliament via email to ask for such a reliable source. The Finnish Parliament responded with webpage[4] where, according to the Finnish Parliament, new data should be available only twice a year.

However, it seems that data are not updated two times per year but only once a year. Unfortunately, in both cases, this is not ideal since one of the core ideas was to have up-to-date parliamentary transcripts with just a little delay from the time they are published on the source.

3.5. Connection Errors

Whenever tools encounter a problem, the problem is classified as a warning if the tool can continue or an error when the tool cannot continue. In both cases, an email is sent to the tool administrator to resolve the issue.

The most common type of error encountered during tools execution are connection errors when connection with sources is interrupted. The correct reaction to this error is waiting until another day when tools are automatically executed again. However, email is sent anyway, which leads to spamming the tool administrator's inbox with errors that require no action.

The collection of errors and warnings frequency is important. If connection errors become frequent in some tools, action may be required. I recommend solving this issue by creating an email filter that automatically archives this specific error. In cases when the connection errors would persist for a longer period, other tools safeguards will inform about the error, like checking if the tools data were recently compiled.

4. Future Work

As I present my tools, others also express their concerns about them. Currently, these concerns center on two main issues.

4.1. No Backups of Downloaded Sources

Downloaded sources are processed in memory. Only the output of the prevertbuilder is stored. This means that in the case when it would be found out that some part of the tools was working incorrectly, currently, the only way to reprocess incorrectly processed transcripts is to rely on their presence in parliamentary sources.

This situation may require redownloading a bigger portion of transcripts from the source, which could be a problem since, in the past, some parliaments were actively blocking the toolset because it was gathering too much data. Because of that, the current main focus is on fixing this issue.

[3] https://www.tweedekamer.nl/kamerstukken/plenaire_verslagen/detail/2016-2017/85

[4] https://avoindata.eduskunta.fi/#/fi/dataset-search

4.2. Major Change of the Source

One of the main ideas was the toolset's ability to adapt to changes in parliamentary transcript sources. Most of the time, only new elements or segments are added to the parliament source, which provides no more information to the gathered transcripts. In other cases, useful information is added to the overall structure of the transcripts, which does not interrupt the continuous and automatic download process. Fortunately, there was never a change that would require a complete rewrite of the downloading tool.

This means that the toolset currently could be run on sources that were available at the time of toolset creation and still work correctly. Problems may arise when parliamentary sources would undergo complete renewal. This would mean that the ability to go back to older transcripts would be lost.

5. Conclusion

Currently, tools are running for over one year and have collected over 1,200 million words, as can be seen in Table 1. Development and maintenance of the automatic parliamentary corpora toolset have revealed several flaws in its original design.

The most important flaws are attribute detection and connection errors. The connection errors show the importance of atomicity. Problems with attribute detection still remain to be solved.

Still, the tools are working and doing what was expected from them. Small human interaction is still required, but those interactions are not critical for tools correct function.

The source code of all the tools is licensed under GNU Lesser General Public License 3.0 and available in a GitLab repository.[5]

corpus name	words	words now
bg_deputies	5.40M	5.86M
cz_deputies	18.41M	20.79M
cz_senate	11.32M	11.58M
dk_deputies	79.00M	79.59M
nl_deputies	71.20M	80.25M
nl_senate	9.99M	11.01M
ir_deputies	40.70M	87.31M
ee_deputies	9.04M	10.49M
fi_deputies	21.09M	21.11M
be_deputies	54.94M	56.77M
be_senate	0.06M	0.69M
fr_deputies	21.09M	59.57M
fr_senate	169.08M	173.53M
at_deputies	6.94M	7.21M
at_senate	2.73M	2.88M
de_deputies	125.03M	125.53M
gr_deputies	58.31M	59.48M
hu_deputies	3.08M	3.93M
it_deputies	3.32M	5.16M
it_senate	13.31M	14.62M
pl_senate	20.08M	20.26M
pt_deputies	141.10M	154.37M
ro_deputies	14.02M	14.86M
ro_senate	26.36M	26.88M
sk_deputies	6.76M	8.74M
si_deputies	15.49M	23.70M
es_deputies	66.66M	68.73M
se_deputies	131.74M	131.75M
sum	1,146.25M	1,286.65M

Table 1: Comparison of processed data from May 2023 to April 2024 (word count from final corpora)

6. Bibliographical References

Giuseppe Abrami, Mevlüt Bagci, Leon Hammerla, and Alexander Mehler. 2022. German parliamentary corpus (gerparcor). In *Proceedings of the Language Resources and Evaluation Conference*, pages 1900–1906, Marseille, France. European Language Resources Association.

Tomaž Erjavec, Matyáš Kopp, Maciej Ogrodniczuk, Petya Osenova, Rodrigo Agerri, Manex Agirrezabal, Tommaso Agnoloni, José Aires, Monica Albini, Jon Alkorta, Iván Antiba-Cartazo, Ekain Arrieta, Mario Barcala, Daniel Bardanca, Starkaður Barkarson, Roberto Bartolini, Roberto Battistoni, Nuria Bel, Maria del Mar Bonet Ramos, María Calzada Pérez, Aida Cardoso, Çağrı Çöltekin, Matthew Coole, Roberts Darģis, Jesse de Does, Ruben de Libano, Griet Depoorter, Katrien Depuydt, Sascha Diwersy, Réka Dodé, Kike Fernandez, Elisa Fernández Rei, Francesca Frontini, Marcos Garcia, Noelia García Díaz, Pedro García Louzao, Maria Gavriilidou, Dimitris Gkoumas, Ilko Grigorov, Vladislava Grigorova, Dorte Haltrup Hansen, Mikel Iruskieta, Johan Jarlbrink, Kinga Jelencsik-Mátyus, Bart Jongejan, Neeme Kahusk, Martin Kirnbauer, Anna Kryvenko, Noémi Ligeti-Nagy, Nikola Ljubešić, Giancarlo Luxardo, Carmen Magariños, Måns Magnusson, Carlo Marchetti, Maarten Marx, Katja Meden, Amália Mendes, Michal Mochtak, Martin Mölder, Simonetta Montemagni, Costanza Navarretta, Bartłomiej Nitoń, Fredrik Mohammadi Norén, Amanda Nwadukwe, Mihael Ojsteršek, Andrej Pančur, Vassilis Papavassiliou, Rui Pereira, María Pérez Lago, Stelios Piperidis, Hannes Pirker, Marilina Pisani, Henk van der Pol, Prokopis Prokopidis, Vale-

[5] https://gitlab.com/Atom194/european-parliamentary-protocols

ria Quochi, Paul Rayson, Xosé Luís Regueira, Michał Rudolf, Manuela Ruisi, Peter Rupnik, Daniel Schopper, Kiril Simov, Laura Sinikallio, Jure Skubic, Minna Tamper, Lars Magne Tungland, Jouni Tuominen, Ruben van Heusden, Zsófia Varga, Marta Vázquez Abuín, Giulia Venturi, Adrián Vidal Miguéns, Kadri Vider, Ainhoa Vivel Couso, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, Rodolfo Zevallos, and Darja Fišer. 2023a. Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 4.0. Slovenian language resource repository CLARIN.SI.

Tomaž Erjavec, Matyáš Kopp, Maciej Ogrodniczuk, Petya Osenova, Darja Fišer, Hannes Pirker, Tanja Wissik, Daniel Schopper, Martin Kirnbauer, Michal Mochtak, Nikola Ljubešić, Peter Rupnik, Henk van der Pol, Griet Depoorter, Jesse de Does, Kiril Simov, Vladislava Grigorova, Ilko Grigorov, Bart Jongejan, Dorte Haltrup Hansen, Costanza Navarretta, Martin Mölder, Neeme Kahusk, Kadri Vider, Nuria Bel, Iván Antiba-Cartazo, Marilina Pisani, Rodolfo Zevallos, Xosé Luís Regueira, Adina Ioana Vladu, Carmen Magariños, Daniel Bardanca, Mario Barcala, Marcos Garcia, María Pérez Lago, Pedro García Louzao, Ainhoa Vivel Couso, Marta Vázquez Abuín, Noelia García Díaz, Adrián Vidal Miguéns, Elisa Fernández Rei, Sascha Diwersy, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Amanda Nwadukwe, Dimitris Gkoumas, Vassilis Papavassiliou, Prokopis Prokopidis, Maria Gavriilidou, Stelios Piperidis, Noémi Ligeti-Nagy, Kinga Jelencsik-Mátyus, Zsófia Varga, Réka Dodé, Starkaður Barkarson, Tommaso Agnoloni, Roberto Bartolini, Francesca Frontini, Simonetta Montemagni, Valeria Quochi, Giulia Venturi, Manuela Ruisi, Carlo Marchetti, Roberto Battistoni, Roberts Darģis, Ruben van Heusden, Maarten Marx, Katrien Depuydt, Lars Magne Tungland, Michał Rudolf, Bartłomiej Nitoń, José Aires, Amália Mendes, Aida Cardoso, Rui Pereira, Väinö Yrjänäinen, Fredrik Mohammadi Norén, Måns Magnusson, Johan Jarlbrink, Katja Meden, Andrej Pančur, Mihael Ojsteršek, Çağrı Çöltekin, and Anna Kryvenko. 2023b. Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 3.0. Slovenian language resource repository CLARIN.SI.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Vladislava Grigorova, Michał Rudolf, Andrej Pančur, Matyáš Kopp, Starkaður Barkarson, Steinĥór Steingrímsson, Henk van der Pol, Griet Depoorter, Jesse de Does, Bart Jongejan, Dorte Haltrup Hansen, Costanza Navarretta, María Calzada Pérez, Luciana D. de Macedo, Ruben van Heusden, Maarten Marx, Çağrı Çöltekin, Matthew Coole, Tommaso Agnoloni, Francesca Frontini, Simonetta Montemagni, Valeria Quochi, Giulia Venturi, Manuela Ruisi, Carlo Marchetti, Roberto Battistoni, Miklós Sebők, Orsolya Ring, Roberts Darģis, Andrius Utka, Mindaugas Petkevičius, Monika Briedienė, Tomas Krilavičius, Vaidas Morkevičius, Roberto Bartolini, Andrea Cimino, Sascha Diwersy, Giancarlo Luxardo, and Paul Rayson. 2021. Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1. Slovenian language resource repository CLARIN.SI.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darģis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2022. The parlamint corpora of parliamentary proceedings. *Language Resources and Evaluation*.

Miloš Jakubíček and Vojtěch Kovář. 2010. Czechparl: Corpus of stenographic protocols from czech parliament. In *Proceedings of Recent Advances in Slavonic Natural Language Processing 2010*, pages 41–46, Brno. Masaryk University.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, pages 7–36.

Maarten Marx, A Schuth, et al. 2010. Dutchparl. a corpus of parliamentary documents in dutch. *Proceedings Language Resources and Evaluation (LREC)*, pages 3670–3677.

Maciej Ogrodniczuk. 2018. Polish parliamentary corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Government and Opposition in Danish Parliamentary Debates

Costanza Navarretta, Dorte Haltrup Hansen
University of Copenhagen
Emil Holms Kanal 2, 2300 Copenhagen, Denmark
costanza@hum.ku.dk, dorteh@hum.ku.dk
https://nors.ku.dk/english/staff/?pure=en/persons/53191,
https://nors.ku.dk/english/staff/?pure=en/persons/127790

Abstract

In this paper, we address government and opposition speeches made by the Danish Parliament's members from 2014 to 2022. We use the linguistic annotations and metadata in ParlaMint-DK, one of the ParlaMint corpora, to investigate some characteristics of the transcribed speeches made by government and opposition and test how well classifiers can identify the speeches delivered by these groups. Our analyses confirm that there are differences in the speeches made by government and opposition e.g., in the frequency of some modality expressions. In our study, we also include parties, which do not directly support or are against the government, the *other* group. The best performing classifier for identifying speeches made by parties in government, in opposition or in *other* is a transformer with a pre-trained Danish BERT model which gave an F1-score of 0.64. The same classifier obtained an F1-score of 0.77 on the binary identification of speeches made by government or opposition parties.

Keywords: Parliamentary Speeches, Classification, Government/Opposition

1. Introduction

This paper addresses the parliamentary speeches delivered by Danish politicians in government, in opposition or in a group called *other*, which comprises parties neither supporting directly the government nor being against it. More precisely, we want to investigate whether the speeches by the three groups are different in some linguistic aspects, and then we apply classifiers to their transcriptions in order to automatically identify which of the three groups produced the speeches.

The data we use are extracted from ParlaMint-DK, one of the 29 corpora in the ParlaMint v. 4.0[1]. The corpora were collected and annotated under the ParlaMint project[2], which was initiated and partially funded by the European CLARIN infrastructure [3] (Erjavec et al., 2022).

ParlaMint-DK covers the debates of the Danish parliament, *Folketinget*, in the period 2014-2022. As all the other ParlaMint corpora, ParlaMint-DK contains various information types, hereunder the party, gender and age of the speaker, as well as information on whether the party of the speaker at that time is in government or opposition. Moreover, parties that are in neither group (*other*) can be identified.

The ParlaMint corpora also contain automatically produced linguistic annotations in the same theoretic framework. Furthermore, all corpora are encoded in the same TEI format and contain the same type of metadata (Erjavec et al., 2022). The corpora are both available as texts[4] and in a linguistically annotated version[5].

Recently, many of the ParlaMint corpora have been automatically translated into English[6].

The paper is organized as follows. In section 2, we shortly present some background studies, and in section 3 we describe the data and account for some linguistic differences in the speeches made by government, opposition and *other*. In section 4 we outline related work on automatic text classification, and in section 5, we present our classification experiments. Finally, in section 6, we discuss our results and in section 7 we conclude and outline future work.

2. Background Studies

Various researchers have addressed the speeches made by government and opposition parties. Many of these studies focus on different aspects related to the sentiment expressed in the speeches by the two groups, see e.g. the overview in (Abercrombie and Batista-Navarro, 2020).

For example, Sawhney et al. (2020) address the automatic identification of the political stance in speeches by government and opposition par-

[1] All the corpora are freely available at `https://www.clarin.si/repository/xmlui/handle/11356/1864`.

[2] `https://www.clarin.eu/parlamint`

[3] `https:\clarin.eu`

[4] `https://www.clarin.si/repository/xmlui/handle/11356/1859`

[5] `https://www.clarin.si/repository/xmlui/handle/11356/1860`

[6] `https://www.clarin.si/repository/xmlui/handle/11356/1864`

ties, while Curini et al. (2020) analyse government and opposition in the Japanese parliament over sixty years using Wordfish, a method which uses a scaling technique for predicting positions based on word frequencies in political texts (Slapin and Proksch, 2008).

Izumi and Medeiros (2022) apply a Naive Bayes Classifier to the speeches in the Brazilian Senate in order to classify the positive or negative sentiment presented by the speakers when talking on different issues. The authors annotated a number of speeches manually in order to train and test the classifier. They find that the differences in sentiment between the speeches do not correspond to the left-wing and right-wing dichotomy as they expected, but they reflect much more the government and opposition division. In their opinion, this result indicates that the politicians in the government use a more sentiment rich language in order to influence the politicians in the Senate to vote in favor of their bills.

In our study, we were partly inspired by the findings in (Izumi and Medeiros, 2022). Differing from their work, however, we do not look at the sentiment expressed by politicians , but we use linguistic features of the transcriptions of the Danish speeches in order to determine whether the speeches made by the three groups *government*, *opposition* and *other* differ and can, therefore, be automatically identified, even if the policy stances of many Danish parties are common in many cases, at least with respect to how they vote in the parliament. More precisely, most Danish parties collaborate during the various legislative periods, and many laws are therefore supported by both parties in government and parties outside it. In fact, counting the votes in the parliament in the investigated period, we found that in approx. 22% of the cases the votes were unanimous. Moreover, the two parties, *The Social Democratic Party* and *The Liberal Party*, which belonged to opposite wings and chaired each two of the governments in this period, also expressed the same votes in additionally 8.2% of the cases. This means that in more than 30% of the cases, the politicians of the two main parties voted in the same way independently on whether they were in government or opposition.

3. The Data

The data in our studies was extracted from the annotated ParlaMint-DK, one of the annotated ParlaMint v. 4.0 corpora[7]. The ParlaMint-DK corpus covers the transcriptions of the speeches from the 7 October 2014 to the 7 June 2022. The transcriptions and some of the metadata included in ParlaMint-DK were downloaded from the Danish Parliament website[8], while other metadata and the linguistic annotations of the corpus were made by researchers from CLARIN-DK (Jongejan et al., 2021).

Speech Group	Speeches	Tokens
Government	56,369	14,039,122
Opposition	74,922	16,919,425
Other	59,403	13,542,700
Speaker	207,915	3,139,068
Total	398,609	47,640,315

Table 1: Number of speeches and tokens in the corpus.

ParlaMint-DK contains 47,640,315 tokens, 3,139,068 uttered by the Speaker (the chair) and 44,501,247 tokens uttered by the members of the parliament and the ministers. Only the latter speeches are relevant for this work. All these speeches are marked as either belonging to the government, the opposition or none of the two (the *other* group).

The government can comprise one or more parties; the opposition always consists of more parties from the opposite political wing in the studied period. The group *other* is more heterogeneous. It consists of both parties which give parliamentary support to the government, without being part of it, and parties which are not in direct opposition to the government. *other* also comprises small independent parties, e.g., the parliament members from Greenland and the Faroe Islands, which have acted as parliamentary support to various governments.

The distribution of the speeches in the three groups, and the number of tokens in them are in Table 1.

The Speaker often takes the floor, but does not speak for a long time since the Speaker's role is to chair the meetings and ensure that the formal rules are followed (average number of tokens per speech is 11). The largest number of speeches comes from the opposition parties, followed by the *other* parties. The government parties take the floor less often than the parties in the two other groups, but their speeches are longer (in average 249 words per speech) than the speeches made by the *other* parties (228 words per speeches) and the opposition parties (226 words per speeches). The fact that members of the government parties speak for a longer time than those in the other groups is not surprising since ministers often present the bills.

There were 20 parties in the Danish parliament in the investigated time span. Table 2 shows the positions of at the time largest 11 left and right-wing

[7] The corpus is available a from `https://www.clarin.si/repository/xmlui/handle/11356/1864` as the other ParlaMint annotated corpora.

[8] `ftp://oda.ft.dk`

	Government	Opposition	Other
Left w.		EL	EL
		SF	SF
		ALT	ALT
	S	S	
	RV	RV	RV
Right w.			M
	V	V	
	KF	KF	KF
	LA	LA	LA
			DF
			NB

Table 2: Largest parties' positions in the investigated period.

parties in the various legislative periods., that is some parties were always in the *other* groups, while some parties in some legislation periods were in government, while in other ones were in opposition. The remaining 9 parties, all part of the group *other* are smaller, they have never been in a government, and their members seldom take the floor. They are not shown in Table 2. The 11 parties shown in Table 2 from the left to the right are the following:

EL The Red-Green Unity List (*Enhedslisten*)

SF Socialist People's Party (*Socialistik Folkeparti*)

ALT The Alternative (*Alternativet*)

S The Social Democratic Party (*Socialdemokratiet*) has been leading two governments in the investigated period (2014-2016, and 2019-)

RV Danish Social Liberal Party (*Radikale Venstre*)

V The Liberal Party (*Venstre*) has been leading two right-wing governments in the investigated time (2009-2014, 2016-2019)

K Conservative People's Party (*Konservative Folkeparti*)

LA The Liberal Alliance (*Liberal Alliance*)

DF Danish People's Party (*Dansk Folkeparti*)

NB New Right (*Nye Borgerlige*)

In the period covered by the ParlaMint-DK data, the Social Democrats (S) and the Liberals (V) are always either in government or in opposition, while other parties like The Red/Green Alliance (EL) or Danish People's Party (DF), are never part of the government. In the Lars Løkke Rasmussen II Cabinet (28.06.2015 - 28.11.2016), the government consisted of only one party, The Liberal Party (V), while Danish People's Party (DF), Liberal Alliance (LA) and Conservative People's Party (KF) were the parliamentary support. From 28.11.2016 to 27.06.2019, the liberals (V) were at the government with the Liberal Alliance (LA) and Conservative People's Party (KF). The opposition consisted of the left-wing parties, which also comprised a centre party, the Danish Social Liberal Party. From 2014 to 28.06.2015 the social democrats (S) headed a left-wing government which also comprised ministers from the Danish Social Liberal Party (RV). After the election in 2019, in the Mette Frederiksen I Cabinet (27. 06 2019 til 15. 12 2022), the social democrats alone formed the government with the other "left-wing" parties as parliamentary support. During these governments, the right-wing parties were the opposition.

3.1. Analysis of the Speeches

The data from the ParlaMint-DK annotated corpus, which we use in the present research are the following: the tokenised transcriptions, the lemmatised transcriptions and, for each speech, information about whether it was delivered by a speaker whose party was in government (GOV), in opposition (OPPN) or in the *other* group.

In our first study, we looked into whether there is an overlap of the lemmas in the three groups of speeches, and we found that 60,989 lemmas only occurred in the government speeches, 13,225 only occurred in the opposition speeches and 34,333 lemmas only occurred in the speeches by the *other* parties. Thus, we found that the government speeches contained the largest number of lemmas which did not appear in the speeches of the other groups, followed by the speeches of the *other* group. A first analysis of the lemmas that only occur in each of the three groups indicates that they mostly consist of compounds, such as *affaldshåndteringsgebyr* (waste management fee), which only occurs in the speeches made by parties in government and *affaldsforbrændingskapacitet* (waste incineration capacity) which only occur in the speeches by opposition parties. This indicates that even if the topics discussed in the parliament by the parties in the three groups are the same, the politicians can address different details about the same topics. Moreover, the data shows the great amounts of compounds which characterise Danish as other Germanic languages.

In the second study, we wanted to investigate the speaker's attitudes to what is said by looking into some of the ways of expressing modality in the speeches. The use of modality in political speeches has been addressed in several studies since through modality speakers can express their attitudinal state towards what they say or others have expressed, see e.g., (Simon-Vandenbergen, 1996; Lillian, 2008; Sharififar and Rahimi, 2015).

The most frequent way of expressing modality in Danish is with modal auxiliaries and modal adverbs. More specifically, *mood* in verbs usually expressed the speaker's or another person's attitude towards an utterance, e.g., (Allan et al., 2015),. The modal auxiliaries in Danish are *kunne* (could), *skulle* (should), *ville* (would), *måtte* (had to), *turde* (dare), *burde* (ought to), *gide* (bother). When they are used in past tense, they often indicate a non-factual (hypothetical) attitude to what is said, while when they are used in present tense, they often indicate a firmer and more factual attitude.

Examples of the modal auxiliary *skulle* in 1) present tense and 2) past tense, are the following:

1. S: *jeg* ***skal*** *som med de foregående dobbeltbeskatningsoverenskomster også meddele at Socialdemokratiet støtter dette lovforslag*
   (I **must** also announce like with the previous double taxation agreements that the Social Democracy supports this bill)

2. V: *det var bare lige for at notere at vi også gerne stadig væk* ***skulle*** *have en positiv stemning i frikommunerne*
   (it was just to note that we still *would like* to maintain a positive atmosphere in the free municipalities)

In the first example, a social democrat in government presents the position of its party with respect to the existing double taxation agreements (a fact), while in the second example, a liberal in the opposition express a desire.

Danish modal adverbs are divided by Jensen (1997) into epistemic and factual adverbs. As for the modal auxiliaries, the distinction between the two groups is that the epistemic adverbs can indicate a more hesitant attitude, while the factual adverbs show more firmness. The epistemic adverbs listed in (Jensen, 1997) are the following: *måske* (maybe), *nok* (probably), *muligvis* (possibly), *dog* (though), *vist* (possibly), *formodentlig* (probably), *åbenbart* (apparently), *tilsyneladende* (seemingly), *egentlig* (actually), *vel* (I guess), while the factual adverbs are *desværre* (unfortunately), *uheldigvis* (unfortunately), and *heldigvis* (fortunately).

Examples of 1) a factive adverb and b) an epistemic adverb are in what follows:

1. V: ***heldigvis*** *er der flere unge med minoritetsbaggrund, der blander sig i debatten og siger fra*
   (**fortunately**, there are several young people with minority backgrounds, who are getting involved in the debate and put their foot down)

2. EL: *hvis ikke det her lovforslag, som* ***muligvis*** *krænker menneskerettighederne, og som i hvert fald træder på retssikkerheden, blev vedtaget*
   (if this bill, which **possibly** violates human rights and certainly undermines legal certainty, was not adopted)

Group	**Modal pres**	**Modal past**
Government	480,687	60,492
Opposition	552,544	90,020
Other	453,628	77,532
Group	**Factive adv**	**Epistemic adv**
Government	5,412	57,171
Opposition	5,903	80,466
Other	4,658	69,340

Table 3: Occurrences of modal auxiliaries and modal adverbs

Group	**Modal pres**	**Modal past**
Government	3.58	0.44
Opposition	3.49	0.57
Other	3.49	0.6
Group	**Factive adv**	**Epistemic adv**
Government	30.39	0.42
Opposition	0.37	0.51
Other	0.36	0.53

Table 4: Relative frequency of modal auxiliaries and modal adverbs

In the first example a liberal expresses a fact, while in the second a example member of the Red-green Union list expresses a possibility regarding a bill, which might violate human rights. We extracted the two types of modal auxiliary verb (present vs. past form) and the factual vs. epistemic adverbs in the parliamentary speeches by government, opposition and *other* group in order to determine whether the parties in government use more confident and factual expressions, and the parties in the other two groups express less confidence when they speak as e.g., was noted in the sentiment analysis of the speeches made by the politicians in the Brazilian senate (Izumi and Medeiros, 2022).

In table 3, the number of each type of modal auxiliary and clausal adverb in each group of speeches is shown, while table 4 shows their relative frequency.

There are no statistically significant differences in the occurrences of modal auxiliaries in present tense and of factual adverbs in the speeches by the three groups. On the contrary, we found significant differences (chi-square's $p < 0.0001$, $df = 1$) in the use of both past tense modal auxiliaries and epistemic adverbs in the speeches by politicians in government and politicians in the other two groups. The politicians in government use significantly less epistemic adverbs and non-factual modal auxiliaries than the politicians in opposition or in the *other* group, thus the politicians in government express themselves in a more confident way.

We also investigated whether we could find the same differences in the speeches of politicians belonging to the two parties that chaired left-wing and right-wing governments (the Social Democrats and the Liberals) comparing cases when they were chairing the government and when they were in opposition, and the above differences in the use of modal auxiliaries in past tense and of epistemic adverbs were confirmed with the same significance values.

These results show that politicians in government use less hypothetical constructions than the politicians that are not in government.

Concluding, our first quantitative study indicates that there are differences in the speeches by the three groups' politicians, and the second study shows differences between speeches made by government parties and parties not in the government. These results are promising for applying text classification to the transcriptions.

4. Text Classification: Related Work

Automatic text classification is one of the main applications of natural language processing. It aims to assign pre-defined labels to whole texts or parts of them. Machine learning based approaches use annotated data to identify the labels in non-annotated data.

The features and algorithm that have been tested the past decades are many, see e.g., (Kowsari et al., 2019; Minaee et al., 2021). The most frequently used features are n-grams, word vectors, TF*IDF (Term Frequency * Inverse Document Frequency)[9] vectors, word embeddings (Kowsari et al., 2019).

Traditional machine learning classifiers comprise e.g., Naïve Bayes and Logistic regression, while examples of deep learning methods used for classification are Multilayer Perceptrons (MLP), Recurrent Neural Networks (RNN), and Long-Short Term Memory systems (LSTM). More recently transformers and pre-trained large language models have improved the state-of-the-art results on some of the most common classification tasks such as sentiment analysis and classification of news articles (Minaee et al., 2021).

Text classification has also been applied to political data, and specifically to parliamentary debates. Many of these studies have addressed the classification of opinions in the debates, inter alia (Abercrombie and Batista-Navarro, 2018; Sawhney et al., 2020), but also the automatic identification of ideology or position in the speeches (Proksch and Slapin, 2012; Riabinin, 2009), the automatic identification of policy domains (Ristilä and Elo, 2023; Navarretta and Hansen, 2022) and of parties (Kapočiūtė-Dzikienė and Krupavičius, 2014; Navarretta and Hansen, 2020).

In our classification experiments, we follow this line of research with the aim of identifying speeches by government, opposition and *other* parties. We test traditional machine learning classifiers and a neural network classifier training them on the most frequently used representations of the ParlaMint-DK transcriptions. In the final experiments, we applied a transformer and a pre-trained Danish BERT model to our data.

5. Classification Experiments

The aims of our classification experiments were to test to which extent various feature types and machine learning classifiers can predict if speeches are delivered by politicians in *Government*, *Opposition* or *other*.

The data we used were the tokenised and lemmatised transcriptions of the speeches, as well as information about whether the speaker's party was in government, opposition or in the *other* group.

The experiments were run in python 3 and the main libraries used are Pandas, Numpy, and Scikit-learn[10]. For the final experiments with a transformer and pre-trained BERT model, pytorch[11] was used.

Firstly, we ran a number of classifiers on the word vectors and TF*IDF vectors of tokens and lemmas in order to determine whether the former or the latter dataset performed best for this task. All classifiers gave the best results with lemma based features. The classifiers we tested were a stratified classifier[12], which is our first baseline, a Multinomial Naïve Bayes, our second baseline, Logistic Regression[13], and a Multilayer Perceptron Classifier[14]. All classifiers' implementations were those provided in Scikit-learn.

We also ran these classifiers with vector and TF*IDF vector representations of the data's bigrams and trigrams.

Secondly, we ran the classifiers and the unigrams features from the first experiments on speeches from only government and opposition (binary classification), since the government vs opposition distinction is used in many political studies. Moreover, many of the parties in the *other* group only

[9] TF*IDF is a technique proposed in (Luhn, 1958) and then adopted by both information retrieval and NLP. It allows to identify documents on the basis of the frequency of their words relative to the words' frequency in the whole dataset.

[10] https://scikit-learn.org/stable/

[11] https://pytorch.org/

[12] The classifier generates predictions by following the training set's class distribution.

[13] Logistic Regression was run with the *lbfgs* solver.

[14] Mulilayer Perceptron was run with the *sgd* solver, *tanh* activation, $alpha = 0.001$, 3 hidden layers, and *constant* learning rate.

belonged to it in the investigated period, and we wanted to address especially parties that have been part of different groups in different periods in order to be sure that linguistic differences in the speeches are not exclusively party dependent.

Finally, we run on the lemmas of the speeches a hugging face transformer[15] with a pre-trained Danish BERT model[16], which has been trained and distributed by the Danish company Certainly[17].

In the first group of experiments, we tested word vectors and the TF*IDF vectors with 15,000 to 19,000 features. The results of classification improved when going from 15,000 features to 17,000 and then decreased. Therefore, we only report the results obtained with the two vectorized datasets and $max\_features = 17000$. The same number of vector features were then also used in the second group of experiments. 10-fold cross validation was performed and Precision (P), Recall (R) and weighted F1-score (F1) are given as evaluation measures.

The results when the classifiers were trained on unigrams, bigrams and trigrams vector representations are in Table 5.

Naïve Bayes classifier outperforms the stratified baseline (F1-score 0.34 vs. 0.47) and is the only algorithm that performs slightly better when trained on vectors of bigrams and unigrams. Both Logistic Regression and Multilayer Perceptron outperform the second baseline, that is the results of the Naïve Bayes classifier. The best results are also produced by Logistic Regression trained on TF*IDF vectors of lemmas with $F1 = 0.61$. Multilayer Perceptron also performs best when trained on TF*IDF unigrams' vectors. The results of the two classifiers decrease slightly when they were run on the vectorized bigrams, and their performance decreases even more when they were trained on the two types of vectorized trigrams.

The confusion matrix from Logistic Regression trained on the TF*IDF lemma vectors is in figure 1[18].

Classifier	**P**	**R**	**F1**
Stratified	0.34	, 0.34	0.34
	Lemma vectorized		
NaïveBayes	0.52	0.49	0.47
LogisticR	0.58	0.58	0.58
MultilayerP.	0.6	0.593	0.594
	TF*IDF		
NaïveBayes	0.541	0.49	0.0.47
LogisticR	0.61	0.61	**0.61**
MultilayerP.	0.6	0.6	0.6
	Bigrams vectorized		
NaïveBayes	0.52	0.5	0.48
Logistic	0.57	0.57	0.57
MultilayerP.	0.51	0.51	0.51
	TF*IDF bigrams		
NaïveBayes	0.53	0.502	0.483
LogisticR	0.6	0.6	0.6
MultilayerP.	0.584	0.584	0.584
	Trigrams vectorized		
NaïveBayes	0.513	0.51	0.50
LogisticR	0.533	0.534	0.533
MultilayerP.	0.472	0.472	0.472
	TF*IDF trigrams		
NaïveBayes	0.52	0.5	0.48
Logistic	0.553	0.554	0.552
MultilayerP.	0.541	0.542	0.541

Table 5: Results of the first classification experiments

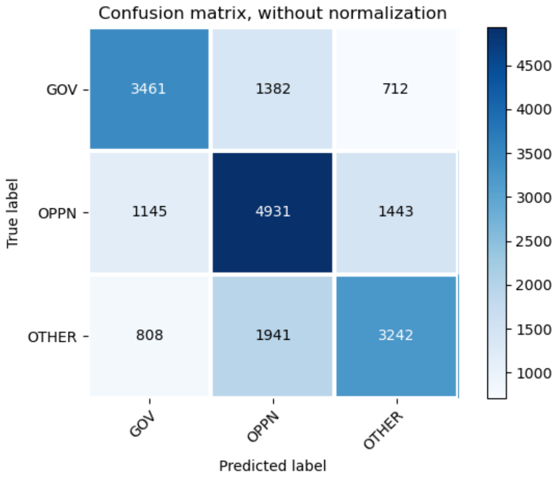


Figure 1: Confusion matrix for ternary classification with Logistic Regression

The classes that are most often confused with each other are *Opposition* and *other*, and this could be expected since they both consist of speeches made by parties that are not in government. We also analyzed some of the erroneously classified speeches and found that some were short, and/or they did not address a specific political issue, as it is shown in the following speech examples:

- *ja* (yes)
- *tak* (thank you)
- *nej det har jeg ikke* (no, I have not)

[15] https://huggingface.co/docs/transformers/index

[16] Version 2, https://github.com/certainlyio/nordic_bert

[17] https://certainly.io/

[18] In the two confusion matrices in the paper, GOV, stands for government, OPPN for opposition, and OTHER for *other* since these were the labels used in the dataset.

Classifier	P	R	F1
Stratified	0.5	0.5	0.5
Lemma vectors			
NaiveBayes	0.654	0.66	0.65
Logistic	0.733	0.734	0.732
MultilayerP.	0.735	0.736	0.737
TFIDF vectors			
NaiveBayes	0.69	0.68	0.66
Logistic	0.752	0.753	**0.751**
MultilayerP.	0.741	0.742	0.741

Table 6: Results of the binary classification experiments

- *jamen så kan man rejse et civilt søgsmål* (well then you can bring a civil action)
- *jeg tror ikke at jeg har yderligere kommentarer* (I do not think that I have further comments)

In the second group of experiments, we applied the same classifiers and used the same unigrams features as in the first group of experiments, but in this case we only addressed the speeches made by parties in government and opposition (binary classification). The results of these experiments are in Table 6.

Also in these experiments, the Multinomial Naïve Bayes classifier outperforms the stratified classifier, and both Logistic Regression and Multilayer Perceptron give better results than the Naïve Bayes classifier, which also performs quite well on this task. Also in these experiments, the best results were achieved by Logistic Regression trained on TF*IDF lemma vectors (F1-score$= 0.751$). The F1-score of Logistic Regression outperforms the F1-score of the Stratified classifier with more than 0.25. Multilayer Perceptron gave slightly better results than Logistic Regression when the two classifiers were trained on word vector representations, while it gave slightly worse results when trained on TF*IDF vectors.

The confusion matrix from the binary classification performed by Logistic Regression trained on the TF*IDF lemma vectors is in figure 2. The confusion matrix shows that speeches made by the government are more often classified as speeches made by the opposition than the contrary. Also in this case, part of the wrongly classified speeches were short and/or did not address a specific political issue.

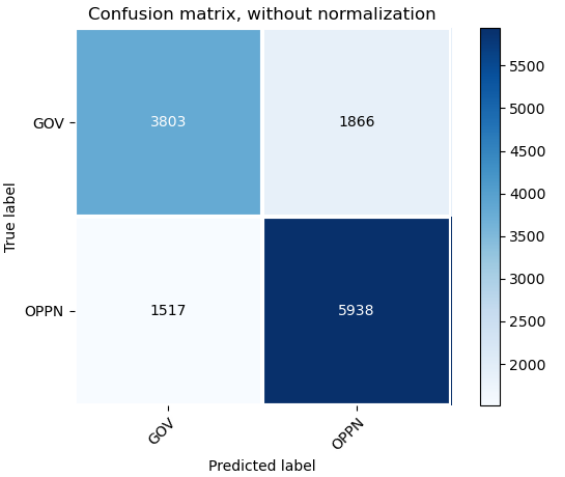


Figure 2: Confusion matrix for binary classification

In the third group of experiments, a bidirectional Encoder Representation from Transformers was run using the pre-trained Danish BERT model[19]. The results for the ternary classification were the following: $Precision = 0.68$, $Recall = 0.64$ and F1-score$= 0.64$. The results of the transformer improve especially precision compared to the best results obtained with the more traditional classifiers, but also recall gets better.

Using even larger language models would probably give even better results. However, environmental sustainability issues should be considered, since it took much more time to fine tune and train the transformer on this data than training Logistic Regression (48 hours vs. half an hour) even if we used a much stronger processor when running the transformer than when training Logistic Regression.

Finally, we run the transformer and the pretrained Danish BERT model on the data consisting only of speeches made by government and opposition parties. The data was so that 80% was used for fine tuning the pre-trained model and 10% were used for testing and 10% for validation.The results for the binary classification were the following: $Precision = 0.79$, $Recall = 0.77$ and F1-score$= 0.77$. Also in this case, the transformer gives the best results.

6. Discussion

Our first quantitative analyses of the parliamentary speeches in the ParliaMint-DK corpus show that there are differences in the speeches delivered by government, opposition or the *other* group.

The politicians in the government use less hypothetical constructions than politicians in the other two groups. Moreover, the fact that a number of lemmas in the speeches of each group do not occur in the speeches produced by politicians belonging

[19] The experiment was run on an Intel Xeon gold processor with 64 cores and 364 GB memory provided by https:cloud.sdu.dk. Optimization was performed with the pytorch implementation of the AdamW optimizer. The learning rate was $5e-5$ and $eps = 1e-8$ (the default). 16 batches and 4 epochs were used.

to the other two groups might indicate that there are issues, which are addressed more by one group or that politicians in government, opposition and *other* parties use some particular words depending on their party's current position.

This aspect should be examined further. In future, we could also investigate whether the differences between the three groups are more evident when they address specific policy areas.

The results of our ternary classification experiments (F1-score$= 0.64$) confirm that identifying the speeches of politicians in government, opposition and outside the two groups are quite good given the type of data. The best results were obtained with a transformer trained on a BERT, but also a traditional ML classifier, Logistic Regression, trained on TF*IDF vectors of lemmas gave a good F-score (0.61).

The results of ternary classification when traditional ML classifiers were trained on bigrams and trigrams vector representations gave different results depending on the classifier and the type of vector, but in general the results decreased slightly when going from unigrams to bigrams, and even more when trigrams were used.

In our binary classification experiments, we again obtained the best results using the transformer and the pre-trained Danish BERT model, with an F1-score of 0.77. This result is also good when compared to the results obtained by other researchers on different text classification tasks (Minaee et al., 2021). The second best result was again obtained by Logistic Regression on TF*IDF vectors of lemmas (best results with 17,000 features: $F1 - score = 0.754$). The analysis of randomly selected speeches, which were wrongly classified, showed that some of them were short and did not address a specific policy domain. Many of these examples, in fact, had a communication management function (Bunt et al., 2010).

7. Conclusions and Future Work

In this paper, we have presented quantitative analyses of the transcriptions of Danish parliamentary speeches as well classification experiments aimed to determine whether the speeches were produced by politicians in government, opposition or *other* parties. Both the results of our preliminary analyses of the speeches and our ternary and binary classification experiments show that there are differences between the speeches of government parties and parties outside it. These differences were also found within parties taking either the role of chairing the government or being in opposition in different years of the investigated period.

The results of this study also confirm some of the observations by Izumi and Medeiros (2022) who classified sentiment in Brazilian Senate speeches delivered by left-wing and right-wing parties.

Future extensions of our work are many, such as a) making further analyses of the linguistic characteristics of the speeches of government parties and parties outside the government, b) investigating whether there are policy domains which are more often addressed by each of the three groups, c) reducing the classification experiments to the speeches of one of the two large parties which have been in government and in opposition in different periods, and d) comparing the results from this study with similar studies of the speeches from other ParlaMint corpora. Since all the ParlaMint corpora have the same metadata and linguistic annotation types (Erjavec et al., 2022), it should be possible to extend this kind of study to other parliamentary data also comparing language specific characteristics of e.g., speeches made by government and opposition parties. Moreover, the English translation of ParlaMint-DK could be used in a replication study in order to evaluate the quality of the automatic translation.

Finally, more Large Language Models could be tested for classification, but environmental sustainability should be considered given the larger amount of resource they require compared with traditional machine learning classifiers.

8. Acknowledgements

We would thank the CLARIN infrastructure for supporting the ParlaMint project and making the ParlaMint data available.

Part of the computation done for this article was performed on the UCloud interactive HPC system, which is managed by the eScience Center at the University of Southern Denmark.

9. Bibliographical References

Gavin Abercrombie and Riza Batista-Navarro. 2020. Sentiment and position-taking analysis of parliamentary debates: a systematic literature review. *Journal of Computational Social Science*, 3(1):245–270.

Gavin Abercrombie and Riza Theresa Batista-Navarro. 2018. Identifying opinion-topics and polarity of parliamentary debate motions. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 280–285.

Robin Allan, Tom Lundskaer-Nielsen, and Philip Holmes. 2015. *Danish: A Comprehensive Grammar*, first edition. Routledge, London.

H. Bunt, J. Alexandersson, J. Carletta, J.-W. Choe, A. Chengyu Fang, K. Hasida, K. Lee, V. Petukhova, A. Popescu-Belis, L. Romary, C. Soria, and D. Traum. 2010. Towards and iso standard for dialogue act annotation. In *Proceedings 7th international conference on language resources and evaluation (LREC 2010)*, pages 2548–2555.

Luigi Curini, Aito Hino, and Atsushi Osaka. 2020. The Intensity of Government–Opposition Divide as Measured through Legislative Speeches and What We Can Learn from It: Analyses of Japanese Parliamentary Debates, 1953–2013. *Government and Opposition*, 55(2):184–201.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešic, Kiril Simov, Andrej Pancur, Michał Rudolf, Matyáš Kopp, Starkadhur Barkarson, Steinthór Steingrímsson, Çagrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevicius, Tomas Krilavicius, Roberts Dargis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fiser. 2022. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*.

Mauricio Y Izumi and Danilo B Medeiros. 2022. Government and opposition in legislative speechmaking: Using text-as-data to estimate brazilian political parties' policy positions–corrigendum. *Latin American Politics and Society*, 64(1):174–175.

Eva Skafte Jensen. 1997. Modalitet og dansk. *NyS, Nydanske Sprogstudier*, 23(23):9–24.

Bart Jongejan, Dorte Haltrup Hansen, and Costanza Navarretta. 2021. Enhancing CLARIN-DK Resources While Building the Danish ParlaMint Corpus. In *CLARIN Annual Conference 2021 Proceedings*, pages 70–73. CLARIN ERIC.

Jurgita Kapočiūtė-Dzikienė and Algis Krupavičius. 2014. Predicting party group from the Lithuanian parliamentary speeches. *Information Technology and Control*, 43(3):321–332.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text classification algorithms: A survey. *Information*, 10(4):150.

Donna L Lillian. 2008. Modality, persuasion and manipulation in canadian conservative discourse. *Critical Approaches to Discourse Analysis across Disciplines*, 2(1):1–16.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3):1–40.

Costanza Navarretta and Dorte Haltrup Hansen. 2020. Identifying parties in manifestos and parliament speeches. In *Proceedings of the Second ParlaCLARIN Workshop*, pages 51–57.

Costanza Navarretta and Dorte Haltrup Hansen. 2022. The Subject Annotations of the Danish Parliament Corpus (2009-2017) - Evaluated with Automatic Multi-label Classification. In *Proceedings of LREC 2022*. ELRA.

Sven-Oliver Proksch and Jonathan B. Slapin. 2012. Institutional foundations of legislative speech. *American Journal of Political Science*, 56(3):520–537.

Yaroslav Riabinin. 2009. Computational identification of ideology in text: A study of canadian parliamentary debates. *MSc paper, Department of Computer Science, University of Toronto*.

Anna Ristilä and Kimmo Elo. 2023. Observing political and societal changes in Finnish parliamentary speech data, 1980–2010, with topic modelling. *Parliaments, Estates and Representation*, pages 1–28.

Ramit Sawhney, Arnav Wadhwa, Shivam Agarwal, and Rajiv Ratn Shah. 2020. GPolS: A contextual graph-based language model for analyzing parliamentary debates and political cohesion. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4847–4859, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Massoud Sharififar and Elahe Rahimi. 2015. Critical discourse analysis of political speeches: A case study of obama's and rouhani's speeches at un. *Theory and Practice in Language studies*, 5(2):343.

Anne-Marie Simon-Vandenbergen. 1996. Image-building through modality: the case of political interviews. *Discourse & Society*, 7(3):389–415.

Jonathan B Slapin and Sven-Oliver Proksch. 2008. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722.

A New Resource and Baselines for Opinion Role Labelling in German Parliamentary Debates

Ines Rehbein, Simone Paolo Ponzetto
University of Mannheim
ines.rehbein@uni-mannheim.de, simone.ponzetto@uni-mannheim.de

Abstract

Detecting opinions, their holders and targets in parliamentary debates provides an interesting layer of analysis, for example, to identify frequent targets of opinions for specific topics, actors or parties. In the paper, we present GePaDe-orl, a new dataset for German parliamentary debates where subjective expressions, their opinion holders and targets have been annotated. We describe the annotation process and report baselines for predicting those annotations in our new dataset.

Keywords: Opinion Role Labelling, Political Text Analysis, Holder and Target Extraction

1. Introduction

Recent work in the area of political text analysis has seen an increasing interest in using NLP methods to investigate the sentiment and positions of political actors in parliamentary debates (see Abercrombie and Batista-Navarro (2020) for an overview). Most work, however, sticks to rather coarse-grained analyses like the prediction of sentiment (positive, neutral, negative) at the level of sentences or documents (Proksch et al., 2019; Abercrombie and Batista-Navarro, 2018) or the prediction or scaling of ideology on a binary scale (*left–right*) (Laver et al., 2003; Slapin and Proksch, 2008).

We thus argue that more work is needed to enable analyses of political text on a more fine-grained level. One possible approach is Opinion Role Labelling (ORL), i.e., the extraction of opinion holders and their targets from text. ORL offers an interesting layer of analysis by distinguishing different perspectives expressed in a text. For illustration, see Fig. 1 and the examples below.

Ex. 1.1 *The German government regrets sending the wrong message to authoritarian leaders.*

Ex. 1.2 *The German government risks sending the wrong message to authoritarian leaders.*

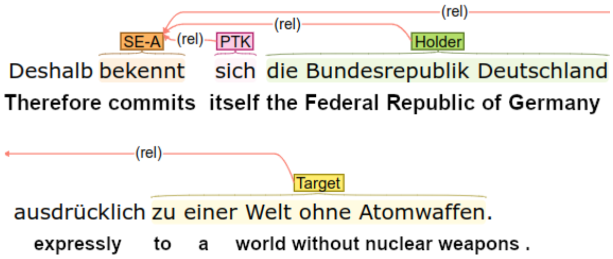

(rel)
(rel)
SE-A
(rel)
PTK
Holder
Deshalb bekennt sich die Bundesrepublik Deutschland
Therefore commits itself the Federal Republic of Germany
(rel)
Target
ausdrücklich zu einer Welt ohne Atomwaffen.
expressly to a world without nuclear weapons .

Figure 1: Example annotation from our corpus (SE-A: Subjective Expression, Agent-view; PTK: particles and reflexive pronouns).

While both sentences express negative sentiment, the first one is written from the point of view of the German government, while the second sentence reflects the speaker's perspective. This subtle but crucial difference results in very different analyses. Instead of classiying both sentences as *negative*, a more informative analysis should capture that the first example expresses the regrets of an opinion holder (*the German government*) about an action (*sending the wrong message to authoritarian leaders*), where we can infer that the stance of the holder towards the target is negative. For the second example, we would like to know that the German government is *not* the opinion holder but the target of the opinion, while the holder is not stated explicitly but can be inferred as the speaker of the utterance.

In the paper, we present a new dataset of parliamentary debates from the German Bundestag where such differences are encoded on the level of subjective expressions (SEs) and their opinion roles. Our annotation follows a lexico-semantic approach to the identification of opinions and their holders and targets (Wiegand and Ruppenhofer, 2015), based on the detection of subjective expressions for *agent, patient* and *speaker view* verbs (for details, see Section 3). We then use our new dataset to train a state of the art Semantic Role Labelling (SRL) system that can automatically predict subjective expressions and opinion roles in text and present baselines for our new corpus.

The paper is structured as follows. We start with a short review of related work on sentiment and stance detection in political communication (§ 2) and present our lexico-semantic approach to opinion role labelling (§ 3). Section 4 describes

ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora, pages 163–170
May 20, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

our new dataset and annotation process, and we report baselines for the automatic prediction of opinion roles in Section 5. Section 6 concludes and outlines future work.

2. Related Work

Detecting politicians' positions towards certain policy issues is an active field of research in the computational political science community (Subramanian et al., 2017; Rauh, 2018; Abercrombie and Batista-Navarro, 2018; Abercrombie et al., 2019; Koh et al., 2021; Abercrombie and Batista-Navarro, 2022).[1] However, due to a lack of resources for fine-grained analyses of the sources and targets of opinions in political debates, many works have tried to approximate the stances of political actors with sentiment predictions, assuming that the concepts are sufficiently correlated (Jose and Choorali, 2015; Murthy, 2015; Rezapour et al., 2017; Uthirapathy and Sandanam, 2023).

Bestvater and Monroe (2023) address this issue and present three case studies showing that approximating stance with sentiment introduces noise and can thus have a negative impact on the validity of the results. Therefore, they discourage the use of sentiment dictionaries and classifiers for modelling stance and, instead, recommend to train in-domain stance classifiers for the task at hand. Below, we explain the difference between stance detection and opinion role labelling and shortly overview relevant work in each field.

Stance detection for political text analysis In contrast to sentiment classifiers that label a text as either *positive, negative* or *neutral* without specifying the target of the sentiment, a stance detection classifier takes a text and a given target and tries to determine the stance of the text toward that target as either *in favour, against* or *neither*.[2]

Work on the intersection of NLP and political science often tries to predict political preferences for a large set of fine-grained issues (Subramanian et al., 2017; Abercrombie and Batista-Navarro, 2018; Abercrombie et al., 2019; Koh et al., 2021; Abercrombie and Batista-Navarro, 2022), *inter alia*. Most notably is the Manifesto Project[3] which has created a large, multilingual collection of political manifestos across countries, where policy issues and preferences are coded on the sentence level.

Vamvas and Sennrich (2020) present a multilingual, multi-target dataset for online political debates. Mascarell et al. (2021) release a corpus of German news articles with stance annotations for a set of 91 target issues. Barriere et al. (2022a,b) create a multilingual, multi-target dataset of online debates with self-rated comments in 26 European languages. They augment the data with around 1,200 comments in 6 languages, manually annotated for stance. Göhring et al. (2021) present the German deInStance corpus, including 1,000 answers by politicians taken from the X-Stance corpus of Vamvas and Sennrich (2020), focussing on the challenging task of inferring *implicit* stances from text.

Opinion Role Labelling is the task of identifying subjective expressions in text, together with their holders and targets. Previous work has used the term "fine-grained entity or aspect-level sentiment analysis" for identifying the sentiment (*positive, negative*) of a text toward the target of an opinion (Liu, 2012), which is very similar to our goal. However, unlike aspect-level sentiment analysis and stance detection, ORL does not require any prior knowledge of the target(s), but attempts to identify them "on the fly", together with their sources.

Following the seminal work of Stoyanov et al. (2004) and Wiebe et al. (2005a) for English, Ruppenhofer et al. (2014, 2016) have presented a corpus of Swiss-German parliamentary debates annotated with subjective expressions, their opinion holders and targets. The data set has been used in two shared tasks.[4] While being similar in spirit to our work, their data is substantially smaller with around 26,500 tokens compared to over 200,000 tokens in our data. However, due to the full text annotation approach where all subjective verbs, nouns, adjectives and multi-word expressions have been coded, the density of annotated SEs in the shared task data is much higher than in our corpus.

Other work from the area of Argumentation Mining has focussed on German newswire, presenting a dataset of German newspaper articles, manually annotated for claims about the migration crisis (Lapesa et al., 2020). The authors identify and code claims, together with their holders (the ones who stated the claim), and also annotate the polarity of the claim. This results in a high-quality dataset for this particular topic. However, the approach offers limited generalisability, as the data is tailored toward one particular policy issue.

Instead, the ORL approach is more generalisable as it can be used on any text, without a predefined topic or target. This, however, comes at the cost of interpretability. While stance detection asks what stance a text conveys towards the target (e.g.,

[1] Also see Abercrombie and Batista-Navarro (2020) for a survey of recent work on sentiment and stance detection in parliamentary debates.

[2] Often the label *neutral* is also included.

[3] https://manifestoproject.wzb.eu

[4] See the IGGSA 2014 shared task: https://sites.google.com/site/iggsasharedtask/task-1 and for 2016: https://iggsasharedtask2016.github.io.

A	(Wir)$_{Holder}$ lehnen (diesen Antrag)$_{Target}$ ab$_{Ptc}$ (We) Agent reject (this motion) Patient
P	(Die USA)$_{Target}$ haben (mich)$_{Holder}$ enttäuscht (The USA) Agent disappointed (me) Patient
S	(Deutschland)$_{Target}$ verfehlt (seine Ziele)$_{Other}$ (Germany) Agent fails to meet (its targets)

Table 1: Examples for agent (A), patient (P) and speaker (S) view verbs and the mapping to opinion holder and target (A: agent=holder, patient=target; P: agent=target, patient=holder; S: agent=target, holder=speaker).

a political actor like *Trump* or *Obama* or a topic like *abortion, death penalty*), the targets identified in ORL can be very heterogeneous, making it hard to map them to a predefined topic (e.g., *sending the wrong message to authoritarian leaders*). In addition, ORL does not encode the polarity of the subjective expression. The different approaches are therefore not equally suitable for all types of analyses, but should be carefully selected depending on the research question.

3. Agent, Patient and Speaker Views

To create a corpus annotated for subjective expressions, their holders and targets, we follow the lexico-semantic approach described in Wiegand and Ruppenhofer (2015). The authors show that semantic roles like *agent* and *patient* are not sufficient for distinguishing opinion *holders* from their *targets* and propose to categorise opinion verbs into three distinct views: (i) agent view, (ii) patient view, and (iii) speaker view verbs.[5]

The three views specify how the opinion holder is mapped to high-level semantic roles on the syntax-semantics interface: In the agent view, the opinion holder is the syntactic subject of the clause and is linked to the semantic role of the agent. For patient view, the holder of the opinion is not the subject but the direct object of the clause and can be mapped to the semantic role of the patient while the agent role encodes the opinion target (see Table 1). For speaker view, the semantic agent role again encodes the opinion target while the opinion holder is implicit and can be inferred as the speaker of the utterance.

Therefore, determining the correct view of the subjective expression should help us to identify the correct target as either the grammatical subject or the object of the utterance. We use this schema to create a dataset of German parliamentary debates where we annotate subjective expressions, their holders and targets and some additional roles (see Section 4). In the next section, we present our new dataset and describe the annotation process.

4. Data and Annotation

Our dataset, GePaDe-ORL, includes German parliamentary debates, manually annotated for verbal subjective expressions and their opinion roles, i.e., their opinion *holders* and *targets*. The speeches are taken from the 19th legislative term of the German Bundestag, however, the distribution of topics in GePaDe-ORL is not representative of the larger data but has been sampled to cover a more diverse range of topics, with contributions from all parties distributed over the whole legislative term. Below, we describe the sampling procedure in more detail.

Sampling procedure We extracted a sample of parliamentary debates from the German Bundestag, covering all speeches from the 19th legislative term (2017–2021). The sample includes speeches by 807 different speakers, with over 900,000 sentences and over 16 mio tokens. From this corpus, we selected individual speeches for annotation, controlled for topic and including speeches for each of the political parties. In addition, we wanted the texts to be evenly distributed over the time span of the legislative term. To achieve this goal, we selected specific agenda items that covered a range of topics, and then sampled all speeches that belong to this specific agenda item, to increase the comparability of the contributions made by the different speakers.

We based our topic selection on the coding scheme developed in the Comparative Agendas Project (CAP) (Bevan, 2019). The CAP scheme includes 21 major topics and more than 200 fine-grained subtopics. We used a topic classifier to select speeches for eight of the major CAP topics for annotation (*Cultural Policy Issues, Defense, Domestic Macroeconomic Issues, Education, Environment, Health, Immigration and Refugee Issues, Law, Crime, Family Issues*) and manually validated the results.[6]

Annotation Our annotation follows a lexicographic approach, based on the automatically created German opinion verb lexicon of Wiegand and Ruppenhofer (2015). The lexicon includes 1,416 verbal subjective expressions, categorised as either *agent* (533), *patient* (141) or *speaker view* verbs (742). Our annotation setup proceeds as follows. We mark all verbs from the lexicon in

[5] Speaker view verbs have previously been described by Wiebe et al. (2005b) as *expressive subjectivity* and by Maks and Vossen (2011) as *speaker subjectivity*, see Wiegand and Ruppenhofer (2015).

[6] For more detailed information, please refer to the data sheet in our github repository: `https://github.com/umanlp/GePaDe-ORL`.

our data for annotation and ask our annotators to disambiguate the view as either *agent, patient* or *speaker* view. If the verb can not be interpreted as a subjective expression in this particular context, then we assign the label *none*. After disambiguating the subjective expressions, the annotators are instructed to identify the holder and target for this subjective expression.[7]

In addition to *holder* and *target*, we annotate the *effect* role for patient and speaker view (see Ex. (1) below). We use the label *other* to encode a set of verb-specific roles (such as Cause, Theme, Goal) for speaker view verbs (see examples in Table 1).

(1) (Der Fall Susanna)$_{Target}$ zeigt beispielhaft (den Maximalschaden der Durchwinkekultur)$_{Effect}$.
(The Susanna case)$_{Target}$ shows (the maximum damage caused by the wave-through culture)$_{Effect}$.

The *particle* role (**PTC**) marks separated verb particles, as shown in Ex. (2) where the verb form “verlorengehen” (be lost) has a meaning very different from “gehen” (go) alone without the verb particle. To encode the actual meaning of the verb, we mark the separated verb particle as PTC. In addition, we use this label for obligatory reflexive pronouns (see Fig. 1).

(2) Über viele Jahrhunderte gewachsenes kulturelles Kapital geht hier (verloren)$_{Ptc}$.
Cultural capital that has grown over many centuries is being lost here.

We use the label **SVC** to indicate the nominal component of a support verb construction where the meaning is largely shifted from the verb to the noun, as illustrated in Ex. (3).

(3) Zeigen Sie endlich (Rückgrat)$_{SVC}$.
Finally show some (backbone)$_{SVC}$.

Our annotated dataset has a size of 214,229 tokens and 13,222 clauses.[8] The number of annotated subjective expressions and their roles is shown in Table 2. The numbers refer to SE and role counts where each role can consist of multiple tokens.

The annotation has been done independently by two trained student assistants. Throughout the annotation, we had weekly meetings to discuss open questions and difficult cases. After the coding has been completed, all disagreements have been resolved by a trained linguist and further consistency checks have been made to assure the quality of the data. We computed inter-annotator agreement (IAA) between the two students for role assignment as precision, recall and f-score on the token level. We first considered Annotator1 as the ground truth and evaluated Annotator2’s predictions against A1. Then we switched roles and report the averaged agreement as prec: 74.83%, recall: 74.90%, and F1: 74.27%.

	Agent	Patient	Speaker	Total
SE	2,325	138	859	3,322
Roles (all)	4,594	278	1,503	6,375
Target	2,422	109	752	3,283
Holder	1,998	116	12	2,126
Other	1	0	643	644
PTC	142	4	53	199
SVC	31	5	38	74
Effect	0	44	5	49

Table 2: Distribution of roles and views in our new data set. The numbers refer to counts on the **SE/role** level. PTC: separated verb prefixes and obligatory reflexive pronouns; SVC: support verb constructions.

Error analysis One frequent error concerns roles where one annotator assigned a specific label and the other coder also marked the same span but forgot to select a label for this span. Another frequent source of disagreements regards the selection of the role spans. Our student annotators had a background in political and social sciences and therefore sometimes struggled to identify the correct syntactic phrase for role annotation, as illustrated below. Here, A1 correctly chose the relative pronoun for target annotation while A2 assigned the target label to the head of the relative clause.

A1: die Frostschäden, (unter denen)$_{Target}$ (die Obstbauern)$_{Holder}$ zu leiden hatten

A2: (die Frostschäden)$_{Target}$, unter denen (die Obstbauern)$_{Holder}$ zu leiden hatten

Gloss: (the frost damage)$_{A2}$, (under which)$_{A1}$ (the fruit_growers)$_{Holder}$ to suffer had

Translation: *the frost damage suffered by fruit growers*

In a similar vein, we observed cases where one annotator had marked the whole noun phrase (as specified in the annotation guidelines) while A2 marked only the head of the noun phrase but left out modifier phrases or complement clauses attached to the head. This shows that for this type of annotation, linguistic training is more important than a background in political or social sciences.

5. Evaluation

We now present an evaluation where we assess how well an automatic system can predict the subjective expressions and opinion roles in our new dataset.

[7] While the lexicon specifies the view of each verb, some of the verbs also have other senses that belong to a different view and thus need to be disambiguated.

[8] We used spacy for sentence splitting which results in segments at the clause level, with an average size of around 16 tokens/clause.

5.1. Experimental Setting

We split our data into training, development and test sets with 9,298/927/3,067 sentences, respectively. We ensure that none of the agenda items in the test set are included in the training set which results in a more challenging and realistic setting compared to distributing speeches from the same agenda item into training and test sets. This amounts to 177/18/72 (train/dev/test) different speeches, with 2,302 (train), 257 (dev) and 763 (test) annotated subjective expressions.

Baseline system The structure of our data is similar to semantic roles (see Fig. 1), which allows us to train a state of the art Semantic Role Labelling (SRL) system on our data. We chose the SRL system of Conia and Navigli (2020), a language- and syntax-agnostic model that jointly learns to predict the predicates, their senses and arguments (i.e., opinion roles). The model combines a predicate-aware word encoder with a predicate-argument encoder. The first component yields contextualised word representations with respect to the predicate of the sentence, while the second encoder learns predicate-aware argument representations. We initialise the model with the pretrained gbert-large[9] language model (Chan et al., 2020) and select the best fine-tuned model on the development set.[10]

Evaluation metric We report precision, recall and F1 (micro) for the prediction of subjective expressions and roles. Note that, due to our lexicographic approach, the position of all potential SEs are given (hence recall for SE prediction is 100%) and the system only has to decide whether a given verb form at position i is a subjective expression (SE) or not (none).[11] As the role labels can cover more than one token, they are therefore represented as sets of (possibly discontinuous) tokens. The annotation scheme assumes that a given verb can bear at most one SE annotation, that is, it can evoke at most one instance of subjective expression. For roles this is not true: a set of tokens could bear multiple role labels, usually in relation to different SEs. According to our annotation guidelines, roles are dependent on SEs and so system roles can match gold roles only if they are related to the same SE. In line with this, the evaluation first checks how system SEs and gold SEs align. System SEs that cannot be aligned to gold SEs produce false positives, including for their associated roles. In symmetric fashion, gold SEs that cannot be aligned to a system SE result in false negatives. For roles, alignment requires non-zero overlap with the tokens covered by a label of the same type on the other side. Each component token of aligned labels is counted as a true or false positive, or as a false negative. This means that longer spans contribute more to the overall score than shorter labels.

	Prec	Rec	F1	Prec	Rec	F1
SE	93.0	100	96.3	93.0	100	96.4
Roles	74.0	74.8	74.4	70.7	76.8	73.6
Target	77.7	78.2	77.9	71.2	82.0	76.3
Holder	76.0	78.9	77.4	75.9	84.7	80.0
Other	57.7	59.6	58.5	69.1	55.9	61.6
PTC	41.1	54.5	46.2	68.9	56.8	60.0
SVC	33.7	14.6	20.3	20.5	14.3	16.7
Effect	42.0	47.1	44.4	22.5	85.0	35.5

Table 3: Precision, recall and F1 (micro) for SE prediction and roles (token overlap). Results are averaged over three runs with different initialisations (white: dev set, gray: test set).

Results Table 3 shows precision, recall and F1 (micro) for SEs and roles on the development (white) and test set (gray). Precision for the prediction of subjective expressions is around 93% for all runs, with a standard deviation of 0.33/0.68% on the dev/test set. This shows that the system has no problem to distinguish between subjective and non-subjective uses in our data.

Results for roles are substantially lower with around 73% micro-F1 for all roles. However, results for holders and targets, which are at the center of our interest, are still high with an F1 in the range of 77-80%. The gap in results between holders and targets can be explained by their length. Holders in our corpus have an average length of 1.5 tokens while targets are much longer with 5.5 tokens on average, making them more challenging to predict. For the other, less frequent roles, however, results are much lower as there are not enough annotations for the model to learn.

6. Conclusions

In the paper, we presented a new dataset for German political debates, with manual annotations for subjective expressions and their opinion roles. We showed that we can use an SRL system to identify subjective expressions and their holders and targets in text with good prediction accuracy. In future work, we plan to apply our system to predict opinion holders and their targets in a large corpus of parliamentary debates, to study the sources and targets of opinions for specific topics across speakers and parties.

[9] https://huggingface.co/deepset/gbert-large

[10] To ensure replicability, we will release the configuration files together with the train/dev/test splits.

[11] The data includes 3,322 subjective expressions and 1,167 non-subjective uses of those verb forms, i.e., 26% of the candidate expressions have the label None.

7. Limitations

An important limitation of our work is that our corpus only includes annotations for one language (German) and text type (parliamentary debates). However, we expect that our approach can be easily extended to similar text types such as party press releases, manifestos or newspaper articles and plan to investigate this in future work. Another weakness of our work are the low results for the low-frequency labels. As we are mostly interested in the identification of holders and targets, this is not a severe problem but we strongly recommend users who apply our model not to rely on the predictions for the other labels.

Acknowledgements

The work presented in this paper was funded in part by a research grant from the Ministry of Science, Research and the Arts (MWK) Baden-Württemberg and by the German Research Foundation (DFG) under the UNCOVER project (RE3536/3-1). We would also like to thank our student annotators for their dedicated work.

8. Bibliographical References

Gavin Abercrombie and Riza Batista-Navarro. 2018. ‘Aye’ or ‘No’? Speech-level Sentiment Analysis of Hansard UK Parliamentary Debate Transcripts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Gavin Abercrombie and Riza Batista-Navarro. 2020. Sentiment and position-taking analysis of parliamentary debates: A systematic literature review. *Journal of Computational Social Science*, 3(1):245–270.

Gavin Abercrombie and Riza Batista-Navarro. 2022. Policy-focused Stance Detection in Parliamentary Debate Speeches. *Northern European Journal of Language Technology*, 8(1).

Gavin Abercrombie and Riza Theresa Batista-Navarro. 2018. Identifying Opinion-Topics and Polarity of Parliamentary Debate Motions. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 280–285, Brussels, Belgium. Association for Computational Linguistics.

Gavin Abercrombie, Federico Nanni, Riza Batista-Navarro, and Simone Paolo Ponzetto. 2019. Policy Preference Detection in Parliamentary Debate Motions. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 249–259, Hong Kong, China. Association for Computational Linguistics.

Valentin Barriere, Alexandra Balahur, and Brian Ravenet. 2022a. Debating Europe: A multilingual multi-target stance classification dataset of online debates. In *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*, pages 16–21, Marseille, France. European Language Resources Association.

Valentin Barriere, Guillaume Guillaume Jacquet, and Leo Hemamou. 2022b. CoFE: A new dataset of intra-multilingual multi-target stance classification from an online European participatory democracy platform. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 418–422, Online only. Association for Computational Linguistics.

Samuel E. Bestvater and Burt L. Monroe. 2023. Sentiment is not stance: Target-aware opinion classification for political text analysis. *Political Analysis*, 31(2):235–256.

Shaun Bevan. 2019. Gone Fishing: The Creation of the Comparative Agendas Project Master Codebook. In Frank R. Baumgartner, Christian Breunig, and Emiliano Grossman, editors, *Comparative Policy Agendas: Theory, Tools, Data*. Oxford: Oxford University Press.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German’s next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Simone Conia and Roberto Navigli. 2020. Bridging the gap in multilingual semantic role labeling: a language-agnostic approach. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1396–1410, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Anne Göhring, Manfred Klenner, and Sophia Conrad. 2021. DeInStance: Creating and evaluating a German corpus for fine-grained inferred stance detection. In *Proceedings of the 17th Conference on Natural Language Processing*

(KONVENS 2021), pages 213–217, Düsseldorf, Germany. KONVENS 2021 Organizers.

Rincy Jose and Varghese S Chooralil. 2015. Prediction of election result by enhanced sentiment analysis on twitter data using word sense disambiguation. In *2015 International Conference on Control Communication & Computing India (ICCC)*, pages 638–641.

Allison Koh, Daniel Kai Sheng Boey, and Hannah Béchara. 2021. Predicting policy domains from party manifestos with BERT and convolutional neural networks. In *Proceedings of the 1st Workshop on Computational Linguistics for Political Text Analysis (CPSS-2021)*, pages 67–77, Düsseldorf, Germany.

Gabriella Lapesa, Andre Blessing, Nico Blokker, Erenay Dayanik, Sebastian Haunss, Jonas Kuhn, and Sebastian Padó. 2020. DEbateNet-mig15:tracing the 2015 immigration debate in Germany over time. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 919–927, Marseille, France. European Language Resources Association.

Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review*, 02(97):311–331.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Isa Maks and Piek Vossen. 2011. A verb lexicon model for deep sentiment analysis and opinion mining applications. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 10–18, Portland, Oregon. Association for Computational Linguistics.

Laura Mascarell, Tatyana Ruzsics, Christian Schneebeli, Philippe Schlattner, Luca Campanella, Severin Klingler, and Cristina Kadar. 2021. Stance detection in German news articles. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 66–77, Dominican Republic. Association for Computational Linguistics.

Dhiraj Murthy. 2015. Twitter and elections: Are tweets, predictive, reactive, or a form of buzz? *Information, Communication & Society*, 18(7):816–831.

Sven-Oliver Proksch, Will Lowe, Jens Wäckerle, and Stuart Soroka. 2019. Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches. *Legislative Studies Quarterly*, 44(1):97–131.

Christian Rauh. 2018. Validating a sentiment dictionary for German political language—a workbench note. *Journal of Information Technology & Politics*, 15(4):319–343.

Rezvaneh Rezapour, Lufan Wang, Omid Abdar, and Jana Diesner. 2017. Identifying the Overlap between Election Result and Candidates' Ranking Based on Hashtag-Enhanced, Lexicon-Based Sentiment Analysis. In *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, pages 93–96, San Diego, CA, USA. IEEE.

Josef Ruppenhofer, Julia Maria Struß, Jonathan Sonntag, and Stefan Gindl. 2014. Iggsa-steps: Shared task on source and target extraction from political speeches. *Journal for Language Technology and Computational Linguistics*, 29(1):33 – 46.

Josef Ruppenhofer, Julia Maria Struss, and Michael Wiegand. 2016. Overview of the iggsa 2016 shared task on source and target extraction from political speeches. In *Proceedings of IGGSA Shared Task 2016 Workshop*, pages 1–9.

Jonathan B Slapin and Sven-Oliver Proksch. 2008. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52:705–722.

Veselin Stoyanov, Claire Cardie, Diane Litman, and Janyce Wiebe. 2004. Evaluating an opinion annotation scheme using a new multi-perspective question and answer corpus. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*.

Shivashankar Subramanian, Trevor Cohn, Timothy Baldwin, and Julian Brooke. 2017. Joint sentence-document model for manifesto text analysis. In *Proceedings of the Australasian Language Technology Association Workshop, ALTA 2017, Brisbane, Australia, December 6-8, 2017*, pages 25–33.

Samson Ebenezar Uthirapathy and Domnic Sandanam. 2023. Topic Modelling and Opinion Analysis On Climate Change Twitter Data Using LDA And BERT Model. *Procedia Computer Science*, 218:908–917.

Jannis Vamvas and Rico Sennrich. 2020. X-stance: A multilingual multi-target dataset for stance detection. In *5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*. CEUR Workshop Proceedings.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005a. Annotating expressions of opinions and emotions in language. *Lang. Resour. Evaluation*, 39(2-3):165–210.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005b. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210.

Michael Wiegand and Josef Ruppenhofer. 2015. Opinion holder and target extraction based on the induction of verbal categories. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 215–225, Beijing, China. Association for Computational Linguistics.

ParlaMint Widened: a European Dataset of Freedom of Information Act Documents (Position Paper)

Maarten Marx, Maik Larooij, Gerda Viira
Information Retrieval Lab, Informatics Institute, University of Amsterdam.
{maartenmarx|larooij}@uva.nl, gerda.viira@student.uva.nl

Abstract

This position paper makes an argument for creating a corpus similar to that of ParlaMint, not consisting of parliamentary proceedings, but of documents released under Freedom of Information Acts. Over 100 countries have such an act, and almost all European countries. Bringing these now dispersed document collections together in a uniform format into one portal will result in a valuable language resource. Besides that, our Dutch experience shows that such new larger exposure of these documents leads to efforts to improve their quality at the sources.

Keywords: Freedom of Information Act, ParlaMint, Government Data

1. ParlaMint

The ParlaMint corpus of Parliamentary proceedings in 27 languages from 26 European parliaments covering at least 10 years of data for each parliament enables diachronic comparative research done by corpus linguists but also by social and political scientists (Erjavec et al., 2023). With the recently released translations into English (Kuzman et al., 2023), it is easy to conduct large scale comparative research on European and global topics like immigration, climate change, the pandemic, the War in Ukraine, or European integration.

The ParlaMint corpus shows that such a huge corpus in a tightly controlled format can be created with a decentralized approach with independent groups taking care of "their own data", and together creating an archive which derives its strength from the fact that it is an integrated data warehouse covering so many nations and languages.

The parliamentary proceedings are just one example of a resource which has the needed properties for such a huge socio-linguistic data collection and harmonization project. The key properties are:

- resources are built on top of a shared data model (for the parliamentary proceedings this is the Hansard model);
- the resources mean more or less the same in each country (what they represent is very similar: speech acts in parliament);
- there is enough overlap in context among the different resources.

There are other types of resources with these properties for which it is useful and desirable to collect and harmonize them. For instance, notes of cabinet meetings, Supreme Court rulings, and Addresses to the Nation (e.g., State of the Unions).

2. Freedom of Information Act

We are advocating in this position paper to bring together resources which are made public after a request based on the local Freedom of Information Act (FOIA). According to Wikipedia 102 nations have FOIA legislation by which citizens can request the public release of government documents on a certain topic. By 2018, every European country, except Luxembourg, has implemented some form of Freedom of Information law (Mokrosinska, 2021).

Also these FOIA resources share the desired properties needed to bring them together into a ParlaMint-like corpus.

We have created a data model and a corpus for Dutch Freedom of Information Requests, and extensively tested it with examples from very different sources: ministries, provinces, municipalities, the police, some universities, and regulating bodies like the gambling, the financial and the consumer authorities. This yielded a daily updated corpus of over 10K requests coming from 50 different governing bodies, consisting of 87K documents and 1.6M pages, all in a uniform format, accessible via a search engine called Woogle (the Dutch FOIA is abbreviated as Woo), and via datadumps in csv format (Marx, 2023).

We tested whether our data model could also fit FOIA documents from another country, and proved that it did with a corpus of 720K linked FOIA documents originating from 57 different Estonian governing bodies: the Estonian Woogle

3. Building the Corpus

Creating the corpus came with new challenges that we did not encounter when creating the Dutch ParlaMint corpus. As documents released under FOIA may contain sensitive information they often contain text redaction (pieces of the text made unreadable). This redaction process is often done by scanning

Type of Institution	Count
Government Agency	22
Local Government	15
Constitutional Institution	10
Other State Agencies	8
Educational Institution	1
State Held Companies	1
Total	57

Table 1: Institutions in the Estonian FOIA corpus.

the documents, thereby effectively removing (destroying is a better term) all the textual and structural content of the documents. Afterwards, often no OCR is applied, and if it is, it is usually of poor quality. Thus we had to OCR all documents ourselves (van Heusden et al., 2023). Besides this, the Dutch government has the habit of concatenating all released documents into one huge PDF document, *without clearly indicating the borders between the original documents*. To recover the original separate documents we had to use *Page Stream Segmentation* techniques (Wiedemann and Heyer, 2021). Thus much more low level document analysis was needed than we expected beforehand. Besides this, as the provided metadata was hardly existing, we needed to do document classification and information extraction (Bakker et al., 2024).

4. FAIR Data

As indicated above, "raw" FOIA documents are far from being FAIR research data, as defined in (Wilkinson et al., 2016). In fact the Dutch FOIA law stipulates that all documents released under this law have to be machine readable, contain all relevant metadata, and have to comply to European accessibility and re-use guidelines, covering exactly the four FAIR principles: data should be findable, accessible, interoperable and reusable.

Being rather frustrated that we had to use documents of such poor quality, we widely published about this in Dutch media directed to civil servants, and information professionals. The fact that these documents were being collected for scientific purposes and brought together in a convenient search platform like our Woogle, and thus could also be compared to documents from other publishers had a positive effect on the awareness by stakeholders of this problem. We already see the first changes and improved quality of data released under the Dutch FOIA.

By reusing data and exposing it, data will become more FAIR. We have seen this in the early 2000's with TheyWorkForYou.com and PoliticalMashup, two precursors of ParliaMint, and now we see the same happening with Woogle. The same process is known from self-organizing systems like Wikipedia: infoboxes have become so much more standardized after the advent of large knowledge graphs like DBPedia and Yago based on them.

5. Call for Action

Our goal with this position paper is to start an incentive to collect FOIA documents on a European scale, using a similar setup as ParlaMint. We believe that with Woogle we have already a strong foundation, in terms of a proven well fitting data model, a proven data processing methodology with reliable software, and a stable initial infrastructure for collecting and storing the data.

If you have FOIA data that you want to add to our collection, make contact with us, and we are happy to help.

6. Bibliographical References

Femke Bakker, Ruben van Heusden, and Maarten Marx. 2024. Timeline extraction from decision letters using ChatGPT. In *Proc. CASE 2024 Colocated with EACL*, St. Julians, Malta.

Tomaz Erjavec, Maciej Ogrodniczuk, Petya Osenova ..., and Darja Fiser. 2023. The ParlaMint corpora of parliamentary proceedings. *Lang. Resour. Evaluation*, 57(1):415–448.

Maarten Marx. 2023. Woogle dump. Technical report, DANS. doi.org/10.17026/dans-zau-e3rk.

Dorota Mokrosinska. 2021. *Transparency and secrecy in European democracies: contested tradeoffs*. Routledge.

Ruben van Heusden, Hazel Ling, Lars Nelissen, and Maarten Marx. 2023. Making PDFs accessible for visually impaired users (and findable for everybody else). In *Proc. TPDL 2023*.

Gregor Wiedemann and Gerhard Heyer. 2021. Multi-modal page stream segmentation with convolutional neural networks. *Lang. Resour. and Evaluation*, 55(1):127–150.

Mark D Wilkinson et al. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific data*, 3(1):1–9.

7. Language Resource References

Kuzman, Taja and Ljubešić, Nikola and Erjavec, Tomaž and Kopp, Matyáš and Ogrodniczuk, Maciej et al. 2023. *Linguistically annotated multilingual comparable corpora of parliamentary debates in English ParlaMint-en.ana 3.0*. Slovenian language resource repository CLARIN.SI.

Author Index