LREC 2026

ParlaCLARIN V:
Workshop on Interoperability, Multilinguality, and Multimodality in Parliamentary Corpora

Workshop Proceedings

Editors
Maria Eskevich, Vincent Vandeghinste, David Bordon

16 May 2026

Proceedings of ParlaCLARIN V Workshop on Interoperability, Multilinguality, and Multimodality in Parliamentary Corpora @ LREC 2026

ISBN 978-2-493814-81-4

ParlaCLARIN V @ LREC2026: Preface

April 22, 2026

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 fifth ParlaCLARIN workshop is a continuation of the 2018[1], 2020[2], 2022[3], and 2024[4] 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[5] and data formats such as ParlaCLARIN[6] .

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

[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/ParlaCLARIN-IV
[5] https://www.clarin.eu/parlamint
[6] https://github.com/clarin-eric/parla-clarin

- 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.)

In 2026 the following special themes were also brought for discussion at the workshop:

- Multimodal corpora, containing video, speech, speech transcripts, official proceeding and research using such data
- Multilingual corpora, possibly containing live interpretation into spoken or signed languages and research using such data
- Position papers on the status and the future of the field of parliamentary corpus collection and research

The half-day workshop programme is composed of a keynote talk by Nikola Ljubešić from Jožef Stefan Institute in Ljubljana, Slovenia and 8 peer-reviewed papers by 19 authors from 6 countries.

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

The ParlaCLARIN V workshop was held in person with the a possibility of hybrid attendance in Palma de Mallorca (Spain), as part of the 2026 International Confrence on Language Resources and Evaluation (LREC2026).

M. Eskevich, V. Vandeghinste, D. Bordon May 2026

Organizing Committee

- Maria Eskevich, Huygens Institute, KNAW
- Vincent Vandeghinste, Dutch Language Institute & KU Leuven
- David Bordon, University of Ljubljana

Programme Committee

- Robert Borges, Uppsala University, SE
- Çağrı Çöltekin, University of Tübingen, DE
- Tomaž Erjavec, Jožef Stefan Institute, SI
- Maria Gavriilidou, Athena Research Center, GR
- Francesca Frontini, CNR-ILC, IT
- Pasi Ihalainen, University of Jyväskylä, FI
- Cristina Lastres-López, Universidad de Sevilla, ES
- Bente Maegaard, University of Copenhagen, DK
- Christian Mair, University of Freiburg, DE
- Maarten Marx, University of Amsterdam, NL
- Michal Mochtak, Radboud University, NL
- Jan Odijk, Utrecht University, NL
- Maciej Ogrodniczuk, Polish Academy of Sciences, PL
- Petya Osenova, Sofia University, BG
- Paul Rayson, Lancaster University, UK
- Hugo Sanjurjo-González, University of Deusto, ES
- Sara Tonelli, Fondazione Bruno Kessler, IT

Table of Contents

Workshop Program

16 May 2026

14:00–14:40 **Keynote**

From Concordance to Inference: ParlaCAP Helps ParlaMint Escape the Linguistics Lab
Nikola Ljubešić

14:40–15:40 **Linguistic and Discourse Analysis**

Quantifying Code-Switching in a Ukrainian Parliamentary Dataset 1990-2021
Olha Kanishcheva and Maria Shvedova

Ours and Yours: A Discourse Analysis of Political Identity Markers in Slovenian Parliamentary Discourse
Meden Katja

Representations of Europe and the European Union in Parliamentary Discourse from a Corpus-Assisted Perspective
Anna Kryvenko

15:40–16:00 **Corpus Creation I**

Towards ParlaMint-DE: Improving the Interoperability of the GermaParl Corpus of Plenary Protocols of the German Bundestag
Christoph Leonhardt and Andreas Blätte

16 May 2026 (continued)

16:40–17:00 **Corpus Creation II**

From Transcripts to Insights: A Digital Corpus and Interactive Speech Analysis Platform for Turkish Parliamentary Records
Basak Tepe, Irem Nur Yildirim, Onur Gungor and Susan Uskudarli

17:00–18:00 **Processing of Parliamentary Data**

Transcription and Recognition of Italian Parliamentary Speeches Using Vision-Language Models
Luigi Curini, Alfio Ferrara, Giovanni Pagano and Sergio Picascia

Beyond OCR: Structural Segmentation and Speaker Attribution in Historical Italian Parliamentary Debates
Claudia Corbetta, Samuele Mazzei and Alessio Palmero Aprosio

Computational Political Landscape of the Netherlands and Prime Minister Schoof's Position
Wessel Ledder and Iris Hendrickx

From Concordance to Inference: ParlaCAP Helps ParlaMint Escape the Linguistics Lab

Nikola Ljubešić
Jožef Stefan Institute, University of Ljubljana,
Institute of Contemporary History, Ljubljana, Slovenia
nikola.ljubesic@ijs.si

Abstract

ParlaCAP is an OSCARS Open Science cascading grant project aimed at extending the use of the ParlaMint parliamentary corpora beyond corpus linguistics into the wider Social Sciences and Humanities (SSH). While ParlaMint provides a rich, comparable collection of parliamentary debates and accompanying metadata, its broader uptake has been limited. ParlaCAP addresses this by enriching the data with automatically derived political agendas and sentiment, enabling new forms of comparative political analysis. Using recent advances in multilingual transformer models, the project annotates over 8 million speeches from 28 European parliaments in more than 20 languages. By integrating ParlaMint with the Comparative Agendas Project (CAP) coding schema, ParlaCAP produces a FAIR dataset suitable for cross-national research on interaction of policy, sentiment, and political identity. The enrichments rely on two models, XLM-R-ParlaSent and XLM-R-ParlaCAP, both performing comparably to human annotators. The latter is trained using a teacher–student approach, where GPT-4o-generated labels are used to fine-tune a scalable classifier. The dataset is available via the CROSSDA repository and a user-friendly API. The talk concludes with a series of use cases demonstrating how meaningful insights can be obtained with minimal technical effort.

1. Summary of the Talk

I will present the results of ParlaCAP[1], an OSCARS Open Science cascading grant initiative aimed at extending the usability of the ParlaMint corpora beyond their traditional audience of corpus linguists. While ParlaMint[2] offers a rich and comparable collection of parliamentary debates across Europe, its uptake in broader Social Sciences and Humanities (SSH) fields, such as political science and sociology, has remained limited. ParlaCAP addresses this gap by transforming ParlaMint into a semantically enriched, analysis-ready dataset for comparative political research.

The project leverages recent advances in natural language processing and artificial intelligence to automatically identify political agendas and sentiments in debates from 28 European parliaments. The dataset comprises more than 8 million speeches in over 20 languages, making manual annotation infeasible. However, multilingual transformer models now enable highly consistent and accurate large-scale coding across languages and contexts.

A central contribution of ParlaCAP is the integration of ParlaMint with the Comparative Agendas Project (CAP) coding scheme. This allows for the automatic assignment of policy topics to parliamentary speeches, effectively bridging linguistic resources and political science methodologies. The result is a FAIR dataset that supports cross-national and longitudinal analyses of political agendas and enhances transparency in legislative discourse.

The enrichment is driven by two models: XLM-R-ParlaSent for sentiment analysis and XLM-R-ParlaCAP for agenda classification. Both are based on the XLM-RoBERTa architecture and perform comparably to human annotators. Notably, XLM-R-ParlaCAP is developed using a teacher–student approach, where the GPT-4o teacher generated labels that are then used to fine-tune the XLM-R student, combining the strengths of large language models with scalable deployment.

The resulting dataset is openly available via the Croatian CESSDA repository CROSSDA[3] and through a user-friendly API[4], that simplifies flexible data selection and followup analysis. Users can filter data by country, time, speaker attributes, agenda categories, and sentiment, supporting both exploratory and reproducible research.

The talk concludes with a series of use cases demonstrating how meaningful insights can be obtained with minimal technical effort, often in just a few lines of Python. These examples highlight shifts in policy attention, cross-country differences in political sentiment, and new opportunities for interdisciplinary research.

ParlaCAP thus helps ParlaMint "escape the linguistics lab", making parliamentary corpora accessible, interpretable, and valuable for a much broader SSH community.

[1] https://clarinsi.github.io/parlacap/
[2] https://www.clarin.eu/parlamint
[3] https://doi.org/10.23669/1ZTELP
[4] https://parlacap.ipipan.waw.pl

Proceedings of ParlaCLARIN V Workshop on Interoperability, Multilinguality, and Multimodality in Parliamentary Corpora @ LREC 2026, page 1

Quantifying Code-Switching in a Ukrainian Parliamentary Dataset 1990-2021

Olha Kanishcheva[1,2] Maria Shvedova[3,4]
[1]Heidelberg University, [2]SET University,
[3]National Technical University "Kharkiv Polytechnic Institute", [4]Friedrich Schiller University Jena
Grabengasse 1, 69117 Heidelberg; Mykoly Shpaka St. 3, 03113, Kyiv;
Kyrpychova str. 2, 61002, Kharkiv; Fürstengraben 1 07743, Jena
kanichshevaolga@gmail.com, o.kanishcheva@setuniversity.edu.ua, mariia.shvedova@khpi.edu.ua

Abstract

Analyzing code-switching – the practice of mixing multiple languages in one discourse – remains a significant task in natural language processing (NLP). This study examines the Ukrainian-Russian bilingual context, focusing on quantifying language alternation in a multilingual dataset. We introduce metrics to assess linguistic boundaries and patterns, specifically addressing the complexities of processing texts where Ukrainian and Russian are used interchangeably, including word-level hybridization. Using a corpus of approximately 200,000 tokens derived from parliamentary transcripts (1990-2021), we apply code-switching metrics to identify frequency and patterns of language use. Our findings provide insights into bilingual communication dynamics and can be used to improve language identification models for mixed-language data.

Keywords: code-switching, code-mixing, dataset, Ukrainian, Russian, code-mixing metrics

1. Introduction

Code-switching or code-mixing, the alternating use of multiple languages within discourse, is a widespread phenomenon in multilingual communities. Understanding the patterns and triggers of code-switching is crucial for fields ranging from linguistics and sociolinguistics to NLP (Winata et al., 2023; Doğruöz et al., 2021), where the goal is to model and understand the use of human language. The terms *code-switching* and *code-mixing* are used in studies of multilingual discourse, with distinctions based either on structural differences (seamless mixing vs. distinct switching) or speaker intentionality (intentional switching vs. unintentional mixing) (Hakimov, 2021), but in this paper, we use *code-switching* as an umbrella term encompassing both phenomena, focusing specifically on intra-sentential instances as our dataset consists of isolated sentences.

The structural properties of code-switching have been extensively studied since the 1980s. Poplack (1980) demonstrated that code-switching is not random, but is governed by grammatical constraints, most notably the *Equivalence Constraint*, which predicts that switches occur at points where the surface syntax of both languages is congruent. The *Matrix Language Frame* model (Myers-Scotton, 1993) further argued that one language serves as the grammatical matrix, supplying morphosyntactic structure, while the other contributes lexical insertions – a distinction that proves particularly relevant for morphological mixing. Muysken (2000) proposed a typological framework distinguishing *insertion*, *alternation*, and *congruent lexicalization* as the three fundamental strategies of code-mixing. The latter strategy, in which two languages share grammatical structure that can be filled lexically by either language, is especially pertinent to typologically close language pairs such as Ukrainian and Russian.

The Ukrainian-Russian bilingual community presents a valuable case study for code-switching analysis. As two closely related Slavic languages, Ukrainian and Russian share significant lexical, syntactic, and morphological similarities, making code-switching between them particularly fluid and frequent. Additionally, both languages use the Cyrillic alphabet, which further contributes to the complexity of distinguishing between them, as homographs – words that are spelled identically but differ in pronunciation and sometimes meaning – are common.

In this paper, we aim to fill this gap by presenting a comprehensive analysis of code-switching in a Ukrainian-Russian bilingual dataset from an NLP perspective. We focus on quantifying code-switching through the use of various metrics and conduct an in-depth analysis to uncover patterns at both the lexical and syntactic levels. By leveraging tools from natural language processing, such as symbol n-gram analysis, part-of-speech tagging, etc., we seek to identify the linguistic and contextual factors that influence when and how speakers switch between Ukrainian and Russian.

Our contributions include the calculation of different metrics for quantifying code-switching in bilingual text and the application of these metrics to a dataset comprising spoken Ukrainian-Russian

Proceedings of ParlaCLARIN V Workshop on Interoperability, Multilinguality, and Multimodality in Parliamentary Corpora @ LREC 2026, pages 2–12

bilingual data.

The structure of our article is as follows: In Section 2, we provide a detailed description of the dataset, including the source, selection process, and annotation methodology. Section 3 outlines the metrics employed for evaluating code-switching, along with the results obtained from these evaluations. Section 4 presents an analysis of key linguistic features at code-switching points, such as n-grams and parts of speech. Section 5 analyzes collocations that occur at language boundaries and provides detailed examples of code-switching instances. Finally, Section 6 offers the conclusions drawn from our study.

2. Data Description and Statistics

In this study, we compiled a dataset of sentences from Ukrainian parliamentary session transcripts that exhibit code-switching between Ukrainian and Russian. Parliamentary transcripts provide a large volume of contemporary texts published in the public domain and thus often serve as the basis for corpus linguistic research (Erjavec et al., 2024). The transcripts of the Ukrainian parliament's sessions published on the official website[1] also have additional value as linguistic material, as the texts are transcribed verbatim, preserving colloquial syntax, language errors, language switching, etc.

Different transcription nuances in different years need to be taken into account when working with the dataset. Most of the texts were transcribed manually and the work was done quickly, so the texts contain typos, particularly in sentences with code-switching, where the transcriber did not always manage to switch the keyboard layout in time. In the 1990s, transcripts were at least partially edited (in particular, vocative forms were normalized (Shvedova and Lukashevskyi, 2025)). Since 2023, we have noticed signs of automatic transcription, which reduces the value of the material for linguistic research, as the program normalizes the text (up to replacing colloquial words with literary synonyms); this was discovered by comparing transcriptions from recent years with audio recordings.

The primary language of Ukrainian parliamentary transcripts is standard Ukrainian, with a minor presence of Russian that declined annually, with substantial Russian fragments becoming rare after 2017 (Kanishcheva et al., 2023). Nevertheless, code-switching instances, including brief lexical insertions (1-2 words), occur throughout the corpus and are more frequent in earlier transcripts, providing material for creating a bilingual dataset.

To obtain code-switching content, we excluded sentences that were entirely or predominantly in Russian using CleanText.groovy[2]. The remaining sentences were lemmatized with the dictionary-based TagText parser[3] and filtered to retain only those containing more than two out-of-vocabulary words, which typically indicate a mixture of Ukrainian and Russian. This approach yielded a dataset of approximately 150,000 tokens of identified code-switching content. To ensure a balanced distribution for the language identification task, we supplemented this data with an additional 50,000 tokens from previously excluded Russian sentences.

Figure 1 presents the annual distribution of the collected data alongside the frequency of language transitions from 1990 to 2021. The blue bars represent the text frequency (the volume of unique sentences containing code-switching events), while the orange bars illustrate the code-switching frequency (the total count of language switches identified). A higher count of switches relative to the number of sentences, as observed in peak years like 2019, indicates more frequent transitions between Ukrainian and Russian within individual utterances, suggesting a higher density of mixing in those specific periods.

We processed the sentences by tokenizing them and manually labeling the language of each token. The tokens were categorized into five distinct classes: Ukrainian (UK), Russian (RU), Ukrainian-Russian hybridized words (MIX), Others (OTH), Numbers (NUM), and Punctuation (PUNCT). The resulting corpus comprises over 200,000 tokens (Table 1).

While the full annotated corpus comprises over 200,000 tokens to support the classification task, all subsequent statistical calculations and linguistic analyses in this study are conducted on the core code-switching subset of 150,000 tokens.

The dataset was annotated by three native bilingual speakers of Ukrainian and Russian. One annotator (a graduate student) performed the initial annotation with access to expert consultation for difficult cases. Upon completion of this first pass, a systematic review revealed that the task was more nuanced than anticipated. Consequently, we implemented a validation stage in which two expert annotators collaboratively reviewed a substantial portion of the dataset, identified problematic cases, and developed explicit annotation guidelines. Challenging annotation cases are analyzed in detail in (Kanishcheva et al., 2026), such as orthographically identical words at code-switching boundaries, hybrid and morphologically adapted forms, syntactic calques, and the distinction between borrowings, colloquial variants, and dialectal forms.

[1] https://www.rada.gov.ua/documents/Stenbul_pz

[2] https://github.com/brown-uk/nlp_uk/blob/master/doc/README_other.md

[3] https://github.com/brown-uk/nlp_uk

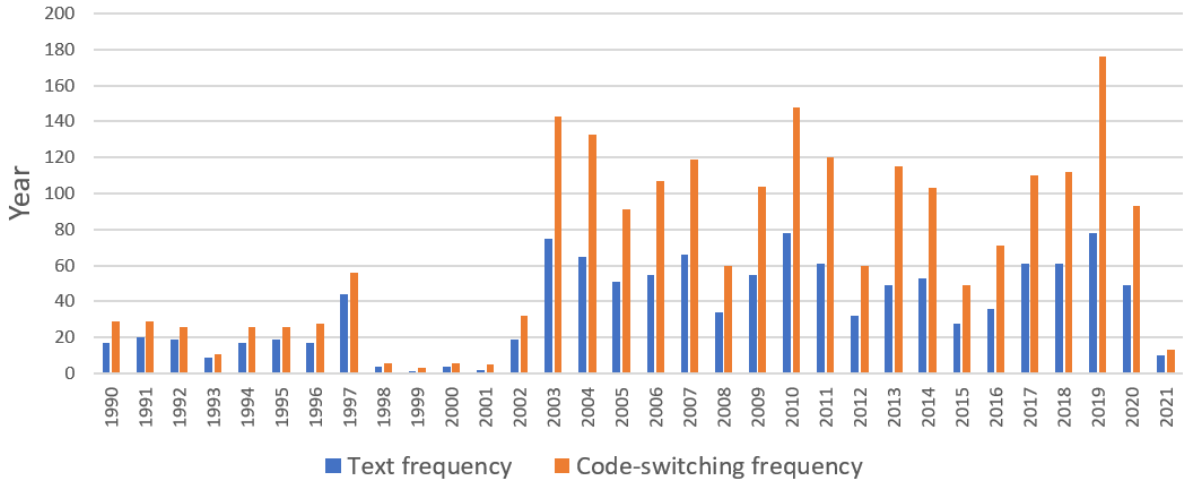


Figure 1: Quantity of sentences containing code-switching instances per year.

The dataset has been published on Zenodo[4] under the Creative Commons Attribution 4.0 International license.

3. Code-Switching Metrics

Code-switching metrics are quantitative measures used to analyze and evaluate patterns of code-switching in bilingual or multilingual discourse (Mave et al., 2018; Guzmán et al., 2017). These metrics provide insights into the frequency, distribution, and structure of code-switching phenomena. Researchers employ various metrics to assess aspects such as the number of code-switching instances, the types of switches (e.g., intra-sentential or inter-sentential), the linguistic levels involved (e.g., lexical, syntactic), and the languages or varieties being switched between. Additionally, code-switching metrics may include measures of proficiency, fluency, and sociolinguistic factors to capture the complexity of code-switching behavior accurately. Overall, the use of code-switching metrics facilitates systematic analysis and comparison of language mixing phenomena across different contexts and populations.

When estimating a code switch in a dataset, the following metrics are usually used:

The **Multilingual Index (M-index)**, developed by Barnett et al. (Barnett et al., 2000) from the Gini coefficient, is a word-count-based measure that quantifies the inequality of the distribution of language tags in a corpus of at least two languages. The M-index is calculated as follows, where $k > 1$ is the total number of languages represented in the corpus, p_j is the total number of words in the language j over the total number of words in the corpus, and j ranges over the languages present in the corpus:

$$M - index = \frac{1 - \sum_{j=1}^{k} p_j^2}{(k-1)\sum_{j=1}^{k} p_j^2}. \quad (1)$$

The index is bounded between 0 (monolingual corpus) and 1 (each language in the corpus is represented by an equal number of tokens).

The **Integration Index** is the approximate probability that any given token in the corpus is a switch point (Guzmán et al., 2017; Guzman et al., 2016). Given a corpus composed of tokens tagged by language $\{l_j\}$, where i ranges from 1 to $n-1$, the corpus size. The I-index is computed as follows:

$$I - index = \frac{1}{n-1} \sum_{1 \le i = j-1 \le n-1} S(l_i, l_j), \quad (2)$$

where $S(l_i, l_j) = 1$ if $l_i \neq l_j$ and 0 otherwise. For a corpus with n tokens, there are $n-1$ possible switch points. It quantifies the frequency of code-switching in a corpus.

The **Code-Mixing Index** is calculated at the utterance level, by finding the most frequent language in the utterance and then counting the frequency of the words belonging to all other languages present in the dataset as illustrated in equation (3) (Das and Gambäck, 2014; Gambäck and Das, 2016).

$$CMI = \frac{\sum_{i=1}^{n}(w_i) - \max\{w_i\}}{n - u}, \quad (3)$$

where $\sum_{i=1}^{n}(w_i)$ is the sum over N languages in the utterance, $\max\{w_i\}$ the highest number of words present from any language, N the number of languages in the utterance, n the number of tokens, and u the number of language-independent tokens. The range of CMI value is [0, 100). If an

[4] https://zenodo.org/records/14724542

Labels	Description	Count	%
UK	Ukrainian words	91,592	41.85
RU	Russian words	82,375	37.65
MIX	Ukrainian-Russian hybridized words	652	0.29
NUM	Numbers	2,123	0.97
OTH	Words in other languages	234	0.1
PUNCT	Punctuation	41,832	19.11

Table 1: Dataset statistics for the language pair Ukr-Rus.

utterance has language-independent tokens or only monolingual tokens, then the corresponding CMI value is 0. A higher value of CMI indicates a higher level of mixing between the languages. CMI-all is an average over all utterances in the corpus and CMI-mixed is an average over only code-switched instances.

Language Entropy (LE): An information-theoretic alternative to the M-index. Measures the number of bits required to describe the distribution of language tags.

$$LE = -\sum_{i=1}^{k} p_i \log_2(p_i), \tag{4}$$

where, k – number of languages, p_i – number of words in language j divided by the total number of words. This metric is 0 for a monolingual corpus and is bounded by equally distributed k languages. Both LE and M-index can be derived from one another.

As a result of applying the above-considered metrics to our data, the values indicated in Table 2 were obtained.

Our dataset shows a high M-Index (58.14%), indicating a balanced distribution of words between the two languages (Table 2). This balance is consistent with the language distribution data presented in Table 1. The high values for both CMI (33.84) and the M-Index confirm a high frequency of code-switching points throughout the corpus.

Furthermore, the token entropy values ($\sim$ 11.5 for both languages) suggest high lexical diversity and an absence of dominant tokens, reflecting a complex and varied vocabulary. When compared to established Hindi-English and Spanish-English datasets (Table 3), our corpus demonstrates a significantly higher CMI. This suggests that our data is not only comparable to but potentially more complex than many existing benchmarks in terms of switching density.

4. N-gram Analysis of Code-Mixing Data

Beyond basic quantitative metrics, we examine the structural characteristics of the code-switching data through character n-grams. The degree of similarity in n-gram distributions between languages directly correlates with the complexity of language identification and sequence labeling tasks. For this analysis, we extracted character n-grams of lengths $n \in \{2, \ldots, 6\}$ from the respective language vocabularies. Figure 2 illustrates the overlap between Ukrainian and Russian n-grams. As is typical for closely related languages, the overlap decreases significantly as the n-gram length increases. A higher overlap probability (e.g., approaching 60%) signifies greater lexical ambiguity, increasing the challenge for computational models to distinguish between the two languages based on sub-word features.

The following analysis highlights the most significant frequency discrepancies between Ukrainian and Russian for 2-grams and 3-grams within our dataset (see Figures 3 and 4). Table 4 presents a comprehensive overview of n-grams ($n \in \{2, \ldots, 6\}$), detailing the total count for each language and the extent of their overlap.

Certain n-gram combinations are orthographically distinct to Ukrainian, as they contain the Cyrillic letters і and ї, which are absent from the Russian alphabet. Many high-frequency n-grams represent characteristic morphemes, such as the Ukrainian -ння *-nnja*, -ськ *-s'k*, and -ти *-ty*, or the Russian пре- *pre-*, -ени(е) *-eni(je)*, and -ть *-t'*. Furthermore, combinations such as -то *-to* (a component of Russian pronouns like то *to*, что *čto*, это *ėto* 'that', 'this') and -ого *-ogo* (the genitive singular masculine ending) highlight differences in frequent word forms. Notably, while -ого *-ogo (RU)/-oho (UK)* exists in both languages, in Ukrainian it appears more frequently in pronominal forms such as його *joho*, нього *n'oho* 'his', 'him', чого *čoho* 'that', нічого *ničoho* 'nothing', and свого *svoho* 'one's own'.

N-gram analysis serves not only to quantify the similarities and differences between Ukrainian and Russian but also as a robust feature for developing token-level language identification models.

5. Code-Switching in Collocations

Beyond static analysis, it is crucial to examine which n-grams emerge at the points of code-switching. Consequently, our study investigates n-gram pat-

Dataset	M-index (%)	I-index	CMI	LE (UK)	LE (RU)
Our Dataset	58.14	7.99	33.84	11.55	11.48

Table 2: Quantitative metrics and language entropy (H) for the Uk-Ru code-switching dataset.

Language pair	CMI index	Source article
English-Bengali	22.48	(Das and Gambäck, 2014)
Dutch-Turkish	22.65	(Nguyen and Doğruöz, 2013)
Spanish-English	22.11	(Mave et al., 2018)
Hindi-English	22.22	(Mave et al., 2018)
Nepali-English	20.32	(Solorio et al., 2014)
Magahi-Hindi-English	51.54	(Rani et al., 2022)

Table 3: Code-mixing index for the different language pair datasets.

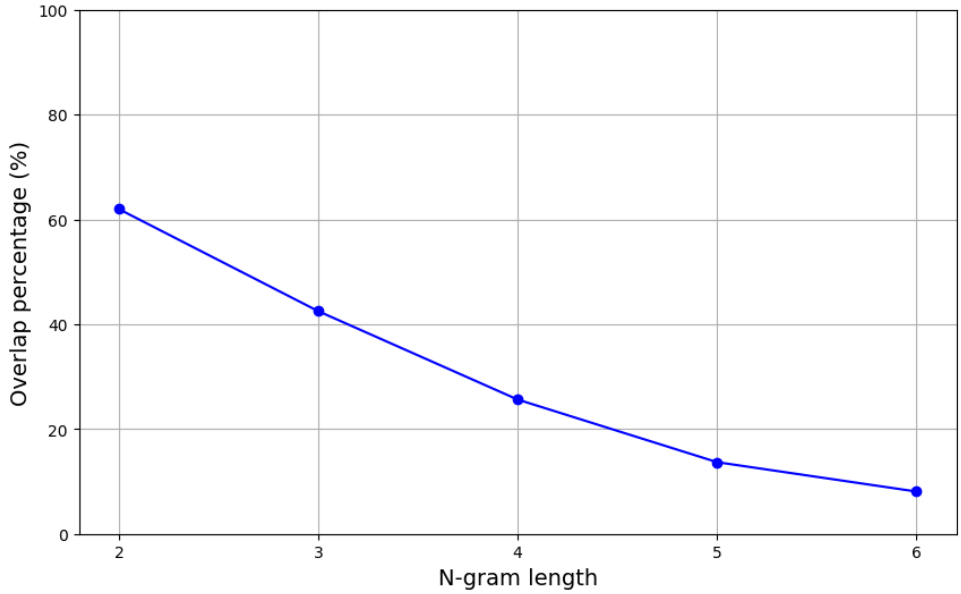


Figure 2: Plot of character N-grams overlap between the Ukrainian and Russian languages, $n \in \{2, \ldots, 6\}$.

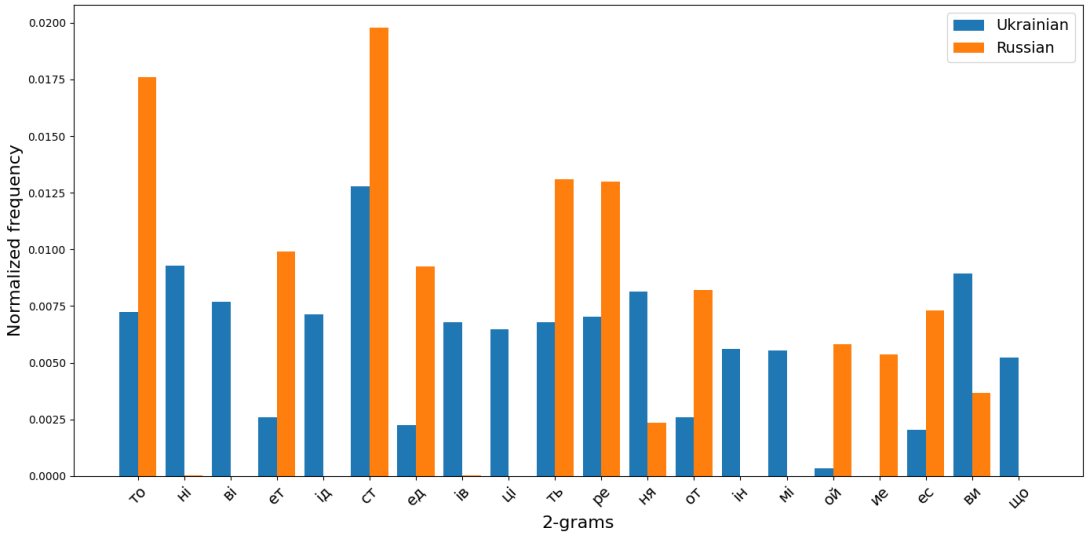


Figure 3: Top-20 2-grams with the greatest frequency discrepancy between Ukrainian and Russian.

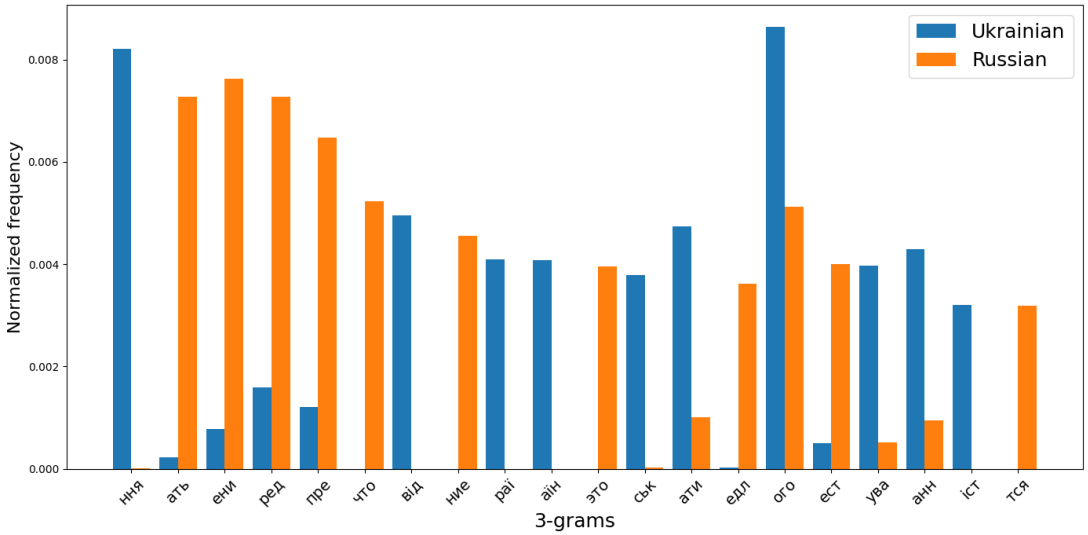


Figure 4: Top-20 3-grams with the greatest frequency discrepancy between Ukrainian and Russian.

N-grams	n=2	n=3	n=4	n=5	n=6
Unique n-grams in Ukrainian	256	2,929	12,248	23,531	28,565
Unique n-grams in Russian	194	2,121	9,438	18,341	22,415
Common n-grams	33	3,730	7,471	6,636	4,478

Table 4: Information about n-grams (n=2..6) of Uk-Ru code-switching dataset.

terns in these transitions, alongside the parts of speech (POS) most frequently involved in switching events (see Figure 5). For morphological analysis, we utilized the *spaCy* model[5].

Most often, the code-switching boundary occurs in syntactically related phrases between an adjective and a noun. This finding aligns with syntactic constraints on code-switching established in earlier linguistics research: the distribution of switching points around nouns and adjectives is consistent with the Equivalence Constraint (Poplack, 1980), which predicts switches at positions where surface syntax is shared across both languages. We analyzed 150 such collocations of the ADJ+NOUN type manually. Most of them are cases where a speaker inserts one or more words in another language. Almost all collocations under consideration are syntactically related, switching can occur in both directions.

5.1. Lexical Code-Switching

This section presents examples of lexical code-switching, where individual words from one language are inserted into an utterance otherwise belonging to the other. The examples are grouped by the direction of the switch: from Ukrainian to Russian ($uk \Rightarrow ru$) and vice versa ($ru \Rightarrow uk$).

[5] https://spacy.io/models/uk

5.1.1. Ukrainian to Russian switching (uk=>ru)

(1) більшу часть
bigger.ACC.F(UK) part.ACC.F(RU)
'the bigger part'

(2) політичний кризис
political.NOM.M(UK) crisis.NOM.M(RU)
'political crisis'

(3) ганебна возня
shameful.NOM.F(UK) fuss.NOM.F(RU)
'shameful fuss'

5.1.2. Russian to Ukrainian switching (ru=>uk)

(4) следующий рік
next.NOM.M(RU) year.NOM.M(UK)
'next year'

(5) остальних продуктів
other.GEN.PL(RU) product.GEN.PL(UK)
'of the other products'

(6) второго зауваження
second.GEN.N(RU) remark.GEN.N(UK)
'of the second remark'

5.2. Idiomatic expressions and calques

Some expressions combining Ukrainian and Russian elements may represent calques of idioms

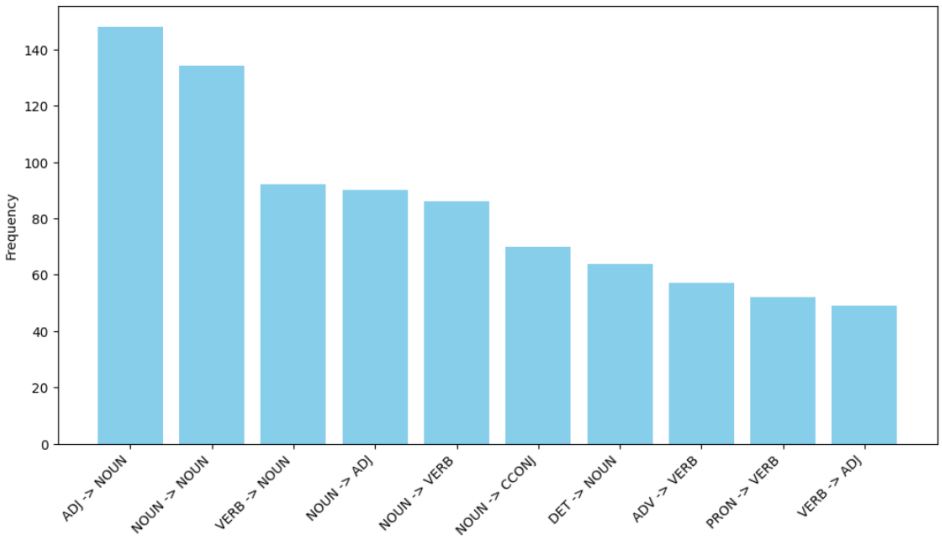


Figure 5: POS switching distribution for the code-switch boundaries.

or frequent collocations. The most common are Ukrainianized calques of Russian expressions:

(7) поетичного слогу
poetic.GEN.M style.GEN.M
'of poetic style'
(UK) (UK)

[calque of RU: поэтического слога; expected UK: поетичного стилю]

(8) железной дорозі
iron.DAT.F road.DAT.F
'railway'
(RU) (UK)

[calque of RU: железной дороге; expected UK: залізниці]

(9) борзими щенками
borzoi.INSTR.PL puppy.INSTR.PL
'with borzoi puppies'
(UK) (RU)

[calque of RU idiom from Gogol's "The Government Inspector": брать взятки борзыми щенками,
lit. 'take bribes with greyhound puppies', 'take bribes in kind rather than money']

5.3. Morphological Code-Switching

Beyond purely lexical switching, morphological patterns reveal more subtle forms of code-mixing where shared lexemes take grammatical forms from different languages. Such patterns are predicted by the Matrix Language Frame model (Myers-Scotton, 1993), which posits that inflectional morphology tends to follow the dominant (matrix) language of the speaker, even when lexical items are borrowed from the other language.

5.3.1. Ukrainian or shared lexemes with Russian morphology

The following collocation could be considered fully Ukrainian, except for the Russian grammatical form:

(10) місцевих бюджетов
local.GEN.PL(UK) budget.GEN.PL(RU,-ов)
'of local budgets'
[expected UK: бюджетів -ів]

In some cases, non-standard Ukrainian endings coincide with standard Russian ones. Such cases may be considered colloquial Ukrainian rather than actual language mixing. The most common cases of this type are:

- **Uncontracted adjective forms**, which may be either Russian or dialectal Ukrainian (northern dialects, see (Zales'kyj and Matviyas, 2001), p. III, map No. 36):

(11) уникальную турботу
unique.ACC.F(RU,-ую) care.ACC.F(UK)
'unique care'
[expected UK: унікальну -у]

(12) новую частину
new.ACC.F(RU,-ую) part.ACC.F(UK)
'new part'
[expected UK: нову -у]

- **Genitive masculine endings** -а , which may be either Russian or colloquial Ukrainian. In Russian, -а is the only possible ending in most

cases (except for a limited number of words that can also have the -y *-u* ending in partitive meaning, but such cases are not considered in our data), while in Ukrainian, part of the nouns have the ending -a and part -y, depending on semantics. This literary norm is often not strictly observed in spoken Ukrainian.

(13) другого етапа
second.GEN.M(UK) stage.GEN.M(RU,-a)
'of the second stage'
[expected UK: етапу -y]

(14) основного капитала
main.GEN.M(UK) capital.GEN.M(RU,-a)
'of the main capital'
[expected UK: капіталу -y]

5.3.2. Russian or shared lexemes with Ukrainian morphology

The following collocations could be considered fully Russian, except for the Ukrainian grammatical form:

(15) досадна ошибка
regrettable.NOM.F(MIX) mistake.NOM.F(RU)
'regrettable mistake'
[expected RU: досадная -ая]

(16) игорного бізнесу
gambling.GEN.M(RU) business.GEN.M(UK)
'of gambling business'
[expected RU: бизнеса -a]

5.3.3. Bidirectional morphological switching

Some collocations can be morphologically normalized, both towards Ukrainian and Russian:

(17) новой редакції
new.GEN.F(RU) edition.GEN.F(UK)
'of the new edition'
[normalized UK: нової редакції]
[normalized RU: новой редакции]

5.4. Orthographic Code-Switching

A controversial issue in annotation is whether cases where the difference between Ukrainian and Russian is only orthographical should be marked as code-switching, since it does not come from the original speaker but is the decision of the transcriber. It can be assumed that the speaker's phonetics influenced the transcriber in such cases, so perhaps they should be considered as well.

5.4.1. Ukrainian phrases with shared words in Russian orthography

Collocations that could be Ukrainian but contain elements of Russian orthography:

(18) слідчої комиссии
investigative.GEN.F(UK)

commission.GEN.F(RU|RU.ORTH?)
'of the investigative commission'
[UK spelling: комісії]

(19) транспортных засобів
transport.GEN.PL(RU|RU.ORTH?)

vehicle.GEN.PL(UK)
'of transport vehicles'
[UK spelling: транспортних]

5.4.2. Russian phrases with shared words in Ukrainian orthography

Collocations that could be Russian but contain elements of Ukrainian orthography:

(20) Огромный дефіцит
huge.NOM.M(RU)

deficit.NOM.M(UK|UK.ORTH?)
'huge deficit' [RU spelling: дефицит]

(21) номінальний прирост
nominal.NOM.M(UK|UK.ORTH?)

increase.NOM.M(RU)
'nominal increase' [RU spelling: номинальный]

5.5. Syntactic-Level Code-Switching

Clearly, in some cases, code-switching cannot be described at the token level only.

5.5.1. Ambiguous prepositions

The combination ADP+NOUN is also frequent, but not always informative, because most common prepositions in Ukrainian and Russian are orthographically identical: на *na* 'on', в *v* 'in', у *u* 'in/near', до *do* 'to', за *za* 'by', про *pro* 'about', для *dl'a* 'for', etc. Determining the language of such prepositions is primarily a matter of annotation principles: whether to consider the preposition at the code-switching boundary as part of the grammatical form of the noun and attribute it the language of the noun, or to consider it a separate item, and in this case the boundary can be between the preposition and the noun.

Examples of such questionable case:

(22) їм плескали в ладони
they.DAT clap.PST in palm.ACC.PL
'[people] clapped for them'
(UK) (UK) (UK/RU?) (RU)

Verbal government as a disambiguation cue

In some cases, syntactic patterns can help determine the language of ambiguous prepositions:

(23) Присягаю на верность Украине
swear.1SG on loyalty Ukraine.DAT
'I swear loyalty to Ukraine'
(UK) (UK/RU?) (RU) (RU)

In this example, although the preposition на 'on' is orthographically identical in Ukrainian and Russian, it belongs to the Ukrainian verbal government pattern присягати на + Accusative (swear on + Accusative), whereas Russian uses a different construction клясться в + Locative (swear in + Locative). This suggests that despite the ambiguous preposition, the syntactic structure follows Ukrainian grammar, with only the lexical items верность 'loyalty' and Украине 'Ukraine' being Russian.

Clear cross-language preposition-noun combinations

There are cases when a preposition and a noun definitely belong to different languages, e.g.:

(24) у кавычках
in quotation.mark.LOC.PL
'in quotation marks'
(UK) (RU)

(25) к діям
to action.DAT.PL
'to actions'
(RU) (UK)

5.5.2. Context-dependent ambiguity

Finally, there are collocations where one of the words is orthographically completely identical in the two languages, and the code-switching can only be determined based on the wider context, not always with absolute certainty:

(26) з секретного соглашения
from secret.GEN.N agreement.GEN.N
'from the secret agreement'
(UK) (UK/RU?) (RU)

Full context: Я (UK) хотів (UK) би (UK) звернути (UK) увагу (UK), що (UK) все (UK) це (UK) почалося (UK) з (UK) секретного (UK/RU?) соглашения (RU) в (RU) отношении (RU) уровня (RU) заработной (RU) платы (RU) Тимошенко (RU)
'I would like to draw attention that all this began with a secret agreement regarding Tymoshenko's salary level'

The collocation analysis shows that taking syntactic relations and idiomatic expressions into account is a promising approach for further study of code-switching. The results of the POS switching distribution (Figure 5) provide insights into the syntactic structures where code-switching is most likely to occur. These results highlight that code-switching is not random but rather occurs in specific syntactic environments, particularly around nouns and their modifiers or connected verbs. This is consistent with established findings that certain parts of speech, particularly nouns and their modifiers, are especially susceptible to code-switching. Understanding these patterns can provide deeper insights into the linguistic mechanisms of code-switching in the Ukrainian-Russian bilingual context.

6. Conclusion

In this paper, we evaluated a comprehensive set of token-level metrics on a Ukrainian-Russian code-switching dataset. Our results demonstrate that this corpus exhibits a high switching density, surpassing several established benchmarks in the field. Through a detailed n-gram analysis ($n \in \{1, \ldots, 6\}$), we quantified the degree of cross-linguistic overlap and identified language-specific sub-word features that are critical for robust language identification.

Furthermore, our investigation into the morphological patterns at switching points reveals that code-switching is not random but follows specific part-of-speech distributions, providing empirical, corpus-based support for established theoretical frameworks in the linguistic literature that posit morphosyntactic structure as a key constraint on code-switching behavior (Poplack, 1980; Myers-Scotton, 1993; Muysken, 2000). These findings suggest that integrating morphological and syntactic features can significantly enhance the performance of automated code-switching detection systems. This work provides both a validated dataset and a methodological foundation for future research in Slavic-centric multilingual NLP.

Acknowledgments

We would like to thank the reviewers for their time and effort in reviewing this manuscript. We sincerely appreciate their valuable comments and suggestions, which greatly helped us improve the quality of the work. This research was partially funded by the Alexander von Humboldt Foundation, and this work received support from the COST Action CA21167 'UniDive'[6] (European Cooperation in Science and Technology). The authors are also grateful to Friedrich Schiller University Jena for providing the research facilities and support that made this work possible.

[6] https://unidive.lisn.upsaclay.fr/

7. Bibliographical References

References

B. Barnett, E. Codo, E. Eppler, M. Forcadell, P. Gardner-Chloros, R. van Hout, M. Moyer, M. Torras, M. Turell, M. Sebba, M. Starren, and S. Wensink. 2000. The LIDES coding manual - A document for preparing and analyzing language interaction data. Version 1.1, July 1999. *International Journal of Bilingualism*, 4(2):131–270.

Amitava Das and Björn Gambäck. 2014. Identifying languages at the word level in code-mixed Indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 378–387, Goa, India. NLP Association of India.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online. Association for Computational Linguistics.

Tomaž Erjavec, Matyáš Kopp, Nikola Ljubešić, Taja Kuzman, Paul Rayson, Petya Osenova, Maciej Ogrodniczuk, Çağrı Çöltekin, Danijel Koržinek, Katja Meden, Jure Skubic, Peter Rupnik, Tommaso Agnoloni, José Aires, Starkaður Barkarson, Roberto Bartolini, Núria Bel, María Calzada Pérez, Roberts Darģis, Sascha Diwersy, Maria Gavriilidou, Ruben van Heusden, Mikel Iruskieta, Neeme Kahusk, Anna Kryvenko, Noémi Ligeti-Nagy, Carmen Magariños, Martin Mölder, Costanza Navarretta, Kiril Simov, Lars Magne Tungland, Jouni Tuominen, John Vidler, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, and Darja Fišer. 2024. ParlaMint II: advancing comparable parliamentary corpora across Europe. *Language Resources and Evaluation*.

Björn Gambäck and Amitava Das. 2016. Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1850–1855, Portorož, Slovenia. European Language Resources Association (ELRA).

G. Guzmán, J. Ricard, J. Serigos, B. E. Bullock, and A. J. Toribio. 2017. Metrics for modeling code-switching across corpora. In *Proceedings of Interspeech 2017*, pages 67–71, Stockholm, Sweden. ISCA.

Gualberto A. Guzman, Jacqueline Serigos, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2016. Simple tools for exploring variation in code-switching for linguists. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 12–20, Austin, Texas. Association for Computational Linguistics.

Nikolay Hakimov. 2021. *Explaining Russian-German code-mixing*. Number 3 in Contact and Multilingualism. Language Science Press, Berlin.

Olha Kanishcheva, Tetiana Kovalova, Maria Shvedova, and Ruprecht von Waldenfels. 2023. The parliamentary code-switching corpus: Bilingualism in the Ukrainian parliament in the 1990s-2020s. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 79–90, Dubrovnik, Croatia. Association for Computational Linguistics.

Olha Kanishcheva, Maria Shvedova, Liudmyla Dyka, and Kristina Husenko. 2026. Study of language identification task on the token level for Ukrainian-Russian code-switching dataset. *Northern European Journal of Language Technology*, 12(1).

Deepthi Mave, Suraj Maharjan, and Thamar Solorio. 2018. Language identification and analysis of code-switched social media text. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 51–61, Melbourne, Australia. Association for Computational Linguistics.

Pieter Muysken. 2000. *Bilingual Speech: A Typology of Code-Mixing*. Cambridge University Press, Cambridge.

Carol Myers-Scotton. 1993. *Duelling Languages: Grammatical Structure in Codeswitching*. Clarendon Press, Oxford.

Dong Nguyen and A. Seza Doğruöz. 2013. Word level language identification in online multilingual communication. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 857–862, Seattle, Washington, USA. Association for Computational Linguistics.

Shana Poplack. 1980. Sometimes I'll start a sentence in Spanish y termino en español: Toward a typology of code-switching. *Linguistics*, 18(7–8):581–618.

Priya Rani, John P. McCrae, and Theodorus Fransen. 2022. MHE: Code-mixed corpora for similar language identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3425–3433, Marseille,

France. European Language Resources Association.

Maria Shvedova and Arsenii Lukashevskyi. 2025. Case choice in Ukrainian vocative expressions: A study of parliamentary transcripts (1990–2024) annotated with Universal Dependencies. In *Grammar and Corpora: 10th International Conference, Book of Abstracts*, pages 106–108, Riga. University of Latvia Press.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar. Association for Computational Linguistics.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023. The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics.

Antin Zales'kyj and Ivan Matviyas, editors. 2001. *Atlas of the Ukrainian Language*, volume 3. Naukova Dumka, Kyiv. 206 maps.

Ours and Yours: A Discourse Analysis of Political Identity Markers in Slovenian Parliamentary Discourse

Katja Meden
Department of Knowledge Technologies, Jožef Stefan Institute;
Jožef Stefan International Postgraduate School;
Institute of Contemporary History,
Ljubljana, Slovenia
katja.meden@ijs.si

Abstract

With recent enrichments of the ParlaMint corpora, new opportunities have emerged for examining a range of political and discursive phenomena. This paper utilises the Slovenian ParlaMint corpus to investigate the construction of political identities through the possessive pronouns “our"" (slv. *naš*) and “your" (slv. *vaš*) in Slovenian parliamentary discourse. The analysis uses a corpus-assisted approach, combining text-type, collocation, and keyword analyses of the lemmas *naš* and *vaš*. The data are drawn from three subcorpora compiled from speeches of Members of Parliament: (1) *Our*, containing speeches in which *naš* occurs; (2) *Your*, containing speeches in which *vaš* occurs; and (3) *Our&Your*, containing speeches in which both lemmas co-occur within the same sentence. The results indicate a clear alignment with established patterns of positive self-representation and negative other-representation. Occurrences of "our" are predominantly associated with positively evaluative discourse, with *Norms & Values* emerging as a central category. In contrast, occurrences of “your" are more typically linked to negative sentiment and keywords, with *Activities/Discourse* identified as the most prominent category. These findings suggest that these possessive pronouns function as markers of ideological positioning and discursive polarisation in Slovenian parliamentary debates.

Keywords: parliamentary discourse, political identities, ParlaMint

1. Introduction

Political identities (PI), a shared constructs of members of social collectivities, are expressed through language and discourse and defined within the realm of political power (Van Dijk, 2010). One such realm is Parliament, where discourse occurs in a controlled environment and can be investigated using parliamentary corpora, which contain transcriptions of essentially spoken language (Fišer and Pahor de Maiti, 2021). Parliamentary discourse is shaped not only by its procedural turn-taking but also by the additional context and relationships between actors on the parliamentary podium (Ilie, 2005; Modrijan, 2007).

One aspect of expressing political identity is the Us versus Them concept, which frames political struggles between two camps and underpins politically charged discursive strategies of defence and attack (Wirth-Koliba, 2016; Van Dijk, 2010). The concepts of identity and the "Us versus Them" distinction are therefore closely linked to political division (or polarisation). One such pattern is the use of possessive contrasts (e.g., “ours" vs. “yours"), which signal group alignment and introduce a division between groups with different political constituencies (e.g., “your side" vs. “our side"). In Slovenian discourse, this is commonly expressed through the pronouns “our" and “your" (slv. *naši in vaši*) which have historically appeared in various (political) distinctions (Zajc and Polajnar, 2012). In this way, language not only reflects but also actively shapes political identities.

The study therefore aims to answer how political identities are expressed within Slovenian parliamentary debates through the use of the terms “our" and “your", using corpus-assisted discourse analysis. Specifically, we are interested in the categories involved in the construction of political identities and how these relate to parliamentary debates.

The paper is organised as follows: Section 2 introduces the concept of political identities and its connection to parliamentary debates through the Us versus Them concept. Section 3 briefly outlines the approaches and methods used to analyse identity markers in discourse. The results of the analyses are presented in Section 4, where we specify the findings identified within each individual approach. Finally, we discuss the role of ‘our’ and ‘your’ as identity markers in Slovenian parliamentary debates and outline future research directions.

2. Political Identities and Related Work

The concept of “identity" is a complex notion in the humanities and social sciences that resists explicit definition (Van Dijk, 2010) and encompasses various expressions of the “self".

Social identities are shared cognitive constructs expressed through discourse and interaction, and,

Proceedings of ParlaCLARIN V Workshop on Interoperability, Multilinguality, and Multimodality in Parliamentary Corpora @ LREC 2026, pages 13–21

while generally stable, may evolve gradually over time. They are shared by members of social groups, whose members are typically aware of them (e.g., "I am a woman/man/citizen...") and are self-attributed, with individuals sometimes affiliating with multiple identities simultaneously (Stets and Burke, 2000; Van Dijk, 2010). Political identities, a subtype of social identity, share many characteristics with social identities. However, because they are grounded in political functions, they also possess specific additional characteristics (Van Dijk, 2010). Like social identities, they are relatively stable, but are formed later in an individual's life and are mostly ideological in nature. As they are defined within the domain of power, they inherently focus more on political in-groups (e.g. membership of a specific country, party, etc.) than on political out-groups, and are associated with different world views (such as the distinction between Left and Right). In discourse, these identities can be expressed through various descriptors, one of which are also personal and possessive pronouns (Van Dijk, 2010, 2018).

While identities are often related to the mental representations that people hold (which are inaccessible for analysis), there are specific characteristics or base categories that allow us to capture and articulate such identities. Van Dijk (2010) proposes the following basic categories: *Membership* (what we are), *Activities/Discourse* (what we do), *Aims* (what we want to achieve – politically), *Norms & Values* (what is politically good or bad), *Ideology* (what we believe in), *Group relations* (who are our political friends or enemies), and *Power resources* (what are our political affordances/what we are able to do).

Parliamentary discourse is closely tied to the construction of political and national identities (Skubic and Fišer, 2022; Riihimäki, 2019). Within this highly regulated setting, political identities are shaped not only by social markers (such as gender, age, ethnicity, or nationality), but also by political functions and institutional (parliamentary) roles (such as MP, minister, chairperson, or member of the coalition or opposition) (Van Dijk, 2004, 2018; Ilie, 2010). Additionally, parliamentary procedures can influence the use of identities, which are regulated through different pronouns, rhetorical strategies (such as positive self-representation or negative other-representation) (Van Dijk, 2004), or polite forms of address to parliamentary speakers (Modrijan, 2007; Ilie, 2005).

One of the most notable strategies in parliamentary discourse, which is closely connected with both political ideology and political identities, is the concept of Us versus Them, a direct expression of political identities through language (Van Dijk, 2010). This concept of political activity always involves an opposition camp (Them), as well as us and our allies (Us), who are continually engaged in a struggle for political power and domination (Wirth-Koliba, 2016; Van Dijk, 2010). The struggles between in-group and out-group separations or polarisations follow the ideological square: emphasise Our good qualities, emphasise Their bad qualities; and conversely, de-emphasise Our bad qualities and de-emphasise Their good qualities (Van Dijk, 2010; Al Maani et al., 2022).

Research on the Us vs. Them construct in parliamentary debates has received increasing attention in recent years, particularly when using corpus-based methods. However, much of this research has focused on specific thematic domains and primarily examines personal pronoun use to explore notions of belonging, for example, the pronouns "we" and "us" (Räikkönen, 2024; Kryvenko, 2024). Some studies in Croatian have examined the possessive pronouns "naši i vaši" (which roughly correspond to the Slovenian terms "naši in vaši") through critical discourse analysis (Blagus Bartolec, 2024).

In Slovenian public and political discourse, the opposition between "naši" (ours) and "vaši" (yours) has developed into an implicit, (and at times explicit) marker of political division, which remains the greatest constant in the Slovenian parliamentary (Gašparič and Kustec, 2020) and political sphere (Mahmutović and Lovec, 2024), which is often conveyed through linguistic patterns. In relation to "naši" and "vaši", as demonstrated by Zajc and Polajnar (2012) in their analysis of historical newspaper discourse, these terms have historically appeared in diverse contexts of social differentiation, including racial, national, and ideological distinctions.

3. Methodology

The corpus-assisted discourse analysis comprises text type, collocation and keyword analysis of the speeches, containing lemmas "our" and "yours" within Slovenian parliamentary debates in ParlaMint-SI 5.1 corpus[1] (Erjavec et al., 2025a,b). The corpus covers the minutes of the National Assembly of the Republic of Slovenia from 3^{rd} to the 8^{th} legislative period (2000 – 2022). The corpus contains 81.683.385 tokens, 69.032.700 words and 311.347 utterances (i.e. speeches). The corpus also includes additional enrichments (Erjavec et al., 2025b), relevant to this analysis, specifically the information on utterance-level sentiment and political orientation of parliamentary groups and political parties. The analysis was done via NoSketch En-

[1]The beta corpus includes corrected transcriptions and updated metadata, which will be incorporated in the next official CLARIN.SI ParlaMint release. Replication with version 5.0 yields minor frequency differences but does not affect the overall conclusions.

gine (Kilgarriff et al., 2014)(beta)[2] concordancer, which allows for subcorpus creation.

To conduct the analysis we created three subcorpora: 1) subcorpus *Our*: speeches containing occurrences of lemma *naš* (our), 2) Subcorpus *Your*: speeches containing occurrences of lemma *vaš* (your), 3) Subcorpus *Our&Your*: subcorpus containing speeches in which both lemmas co-occur within the same sentence.

As the analysis focuses on the discourse of Members of Parliament (MPs) and by extension, their political affiliations with parliamentary groups/political parties, we restricted our subcorpus creation by removing Chairperson, Minister and non-MP (guest) speakers speeches[3]. The basic distributions of the created subcorpora are shown in Table 1.

	Our	Your	Our&Your
Hits	119,440	38,362	2,578
% in Corpus	0.15	0.047	0.0032
Freq./mil. tokens	1462.23	469.64	31.56
Words	37,383,574	14,473,892	1,829,072
Tokens	31,593,806	12,232,2534	1,545,795

Table 1: Basic distributions of the subcorpora.

As the *Our&Your* subcorpora contain speeches, within which both “our" and “your" lemmas appear in a single sentence, this causes a slight overlap of speeches in the *Our* and *Your* subcorpora. Effectively, the overlapping speeches in *Our&Your* subcorpus represent only 3.72% within the *Our* subcorpus, and 9.47% of the *Your* subcorpus.

3.1. Text type Analysis

In the text-type analysis, we examined distributions across identity-relevant dimensions, specifically, party status (coalition vs opposition), political orientation (ideological positioning), and sentiment (polarity stance). These dimensions were selected because they represent institutional, ideological, and affective axes through which political identities are discursively constructed in parliamentary debates.

To support our interpretation, we used metrics available in NoSketch Engine: (raw) frequency (F), relative frequency in text type (RF), and relative density (RD), each conveying distinct information relevant to the analysis. Of these, relative density was highlighted as the primary metric, as it compares how frequent an item is in a specific text type relative to its frequency in the entire corpus, adjusted for the size of that text type. Values below 100% indicate that the item is less typical of that text type; values around 100% indicate equal typicality; and values above 100% indicate that the item is more typical or characteristic of that text type (Pahor de Maiti Tekavčič and Kryvenko, 2026; Relative text type frequency).

3.2. Collocation Analysis

We also investigated the lemma pairings for each subcorpus through collocation analysis. Specifically, we examined the lowercase lemmas within an L0R1 window (i.e. words that occur directly after our keyword in context (KWIC)) to capture direct references to “our" and “your" (e.g., “our government" and “your ministers"). We analysed the 20 most relevant collocational patterns, as identified by the LogDice metric (Brezina, 2018; Rychlỳ, 2008; Statistic measure LogDice), which measures the strength of association between collocates. While strong collocations are typically defined as those with LogDice values of 7 or above (Pahor de Maiti Tekavčič and Kryvenko, 2026; Jaworska and Kinloch, 2018), we also included weaker collocations in our analysis to account for the more constrained contextual window.

Collocates outside the top 20 with LogDice ≥ 5 were included if potentially relevant to political identity and subjected to manual review. For the *Our&Your* subcorpus, the context window was expanded to L3R3 to capture multi-word KWIC. Relevant collocates were examined in parliamentary context and annotated with a primary PI category, with secondary categories assigned where context-dependent variation was observed.

The annotation procedure can be further illustrated with an example: “*You, too, should think about whether it is worth ruining the future of your children and ours*"[4]. The reference to “children” invokes a widely shared value, framing their care as desirable and their neglect as undesirable; it is therefore assigned the primary PI-building category *Norms & Values*. The sentence also contains an implicit group distinction between “your” and “our,” which, in a parliamentary context, reflects divides such as coalition versus opposition and is captured as the secondary category *Group relation*.

Given the large volume of speeches, categorisation was supported by systematic manual review. Each collocate was assessed in its recurrent usage to determine alignment with PI categories prior to annotation. These categories capture functional tendencies in parliamentary discourse and are in-

[2]The ParlaMint-SI 5.1 corpus is currently only available on the CLARIN.SI internal beta installation of the NoSketch Engine.

[3]Speeches included in the subcorpora met the following criteria: Speaker_role = “Regular”, Speaker_Minister = “nonMinister”, and Speaker_MP = “MP”, i.e. speeches by non-chairpersons with confirmed MP status who were not serving as ministers at the time.

[4]Koražija, Boštjan (2021). Zapisi sej Državnega zbora Republike Slovenije, Izredna 8. mandat, 90. izredna seja (22. 12. 2021)

terpreted as indirect markers of political identity, signalling alignment, values, or group affiliation rather than explicitly stating identity. Each annotation includes a justification and an illustrative example.[5].

3.3. Keyword Analysis

Lastly, we identified keywords most strongly associated with each subcorpus by alternately treating *Our* and *Your* as the focus relative to the other. We used lowercase alphanumeric lemmas and applied a smoothing parameter of N = 20, which adjusts sensitivity: higher values emphasise more frequent items, while lower values highlight rarer ones (Kilgarriff, 2009). We applied moderate smoothing (N = 20) to reduce noise from very low-frequency lowercase lemmas while preserving mid-frequency items. The comparison captures contrasts in in-group versus out-group identity construction.

Keyword identification was based on the simple Score metric implemented in NoSketch Engine, which compares normalised frequencies in the focus and reference corpora to determine keyness (Simple maths with keywords and terms). The representative keywords were manually reviewed through close reading to assess the contexts in which they appear.

4. Results

In line with the study's main objective, the analyses produced the following findings.

4.1. Text type Analysis

To establish basic characteristics of the subcorpora, we examined several metadata fields (and their frequencies, relative frequency in text type and relative density), related to the political identity (party status and political orientation) and sentiment.

4.1.1. Sentiment

Table 2 shows the distributions across sentiment categories per individual subcorpus. The distribution of speeches in *Our* subcorpus across sentiment categories reveals a divergence between raw frequencies and normalised measures. In terms of raw frequencies, occurrences are more frequent in Negative speeches (83,080), suggesting that "our" is predominantly associated with negatively classified contexts at the speech level. However, both relative frequency (1,948.13) and relative density (133.23%) are highest in Positive speeches, indicating that "your" is more common and typical in positively classified discourse.

[5]Materials are available on the GitHub repository.

	Positive	Negative	Neutral
Our – F	17,404	**83,080**	18,956
Our – RTT	**1,948.13**	1,597.79	913.42
Our – RD(%)	**133.23**	109.27	62.47
Your – F	1,211	**34,658**	2,493
Your – RTT	135.55	**666.54**	120.13
Your – RD(%)	28.86	**141.92**	25.58
Our&Your – PF	137	**2,240**	201
Our&Your – RTT	15.34	**43.08**	9.69
Our&Your – RD(%)	48.59	**136.50**	30.69

Table 2: Distributions across sentiment categories per individual subcorpus and their respective metrics – raw frequency (F), relative frequency in text type (RTT) and relative density (RD). Additionally, the numbers in bold font highlight items and values that are most prominent for each category.

Conversely, this divergence is not observed in the *Your* and *Our&Your* subcorpora, where all three metrics consistently indicate that occurrences are more strongly associated with Negative sentiment.

4.1.2. Party status

Table 3 shows the distributions of speeches across party status (PS) categories per individual subcorpus.

	Coalition	Opposition	None
Our – F	48,216	**63,845**	7,379
Our – RTT	1,438.97	**1,785.92**	593.78
Our – RD(%)	98.41	**122.14**	40.61
Your – F	9,778	**26,663**	1,921
Your – RTT	291.82	**745.84**	154.58
Your – RD(%)	62.14	**158.81**	32.91
Our&Your – PF	716	**1,736**	126
Our&Your – RTT	21.37	**48.56**	10.14
Our&Your – RD(%)	67.71	**153.86**	32.13

Table 3: Party status text type analysis per individual subcorpus and their respective metrics – raw frequency (F), relative frequency in text type (RTT) and relative density (RD). Numbers in bold highlight prominent values in each category.

The distributions within all three corpora suggest that the occurrences are more common of Opposition speeches, both in terms of raw frequencies, as well as relative frequency in text type and relative density. This pattern might suggests that the lemmas are more systematically used in Opposition discourse, or relate to the role of Opposition in parliamentary proceedings. Additionally, relative density of Coalition category (98.41%) hovers slightly below 100%, which could suggest that occurrences are also equally typical of Coalition discourse.

4.1.3. Party Orientation

Table 4 shows distribution across party orientation (PO) categories per individual subcorpus.

	Our			Your			Our&Your		
Party orientation	F	RTT	RD(%)	F	RTT	RD	PF	RTT	RD(%)
Left	6,203	1,726.98	**118.11**	2,564	713.84	**152.00**	174	48.44	**153.49**
Centre-left	20,604	1,409.04	96.36	6,089	416.41	88.66	456	31.18	98.81
-	4,035	388.35	26.56	1,006	96.82	20.62	61	5.87	18.60
Centre to centre-left	**29,531**	1,552.23	**106.16**	8,876	466.55	99.34	600	31.54	99.93
Centre	169	2,783.86	**190.38**	12	197.67	42.09	/	/	/
Centre to centre-right	945	1,318.34	90.16	162	226.00	48.12	11	15.35	48.62
Centre-right	22,470	**2,298.11**	**157.16**	3,916	400.51	85.28	305	31.19	98.84
Right	27,501	1,380.38	94.40	**13,642**	684.75	**145.80**	**837**	42.01	**133.12**
Right to far-right	7,982	2,232.01	**152.64**	2,095	585.8	**124.74**	134	37.47	**118.72**

Table 4: Party orientation text type distributions per individual subcorpus and their respective metrics – raw frequency (F), relative frequency in text type (RTT) and relative density (RD). Numbers in bold highlight prominent values within category.

The first observation concerns relative density, which shows that in all three subcorpora, occurrences are more typical for two PO categories: *Left* and *Right to far right*, which represent the two political extremes present in the subcorpora.

Additionally, in terms of relative density, occurrences within *Our* subcorpus seem to be more typical for multiple PO categories: *Centre* (RD: 190.38%), *Centre-right* (157.16%), *Right to far-right* (152.64%), *Left* (118.11%) and *Centre to centre-left* (106.16%) speeches, suggesting typicality across (almost) all folds of political spectrum, but less so for the *Centre to centre-right* (90.16%) and not typical (or less common) for discourse of speakers without explicit orientation (i.e. independent speakers, parliamentary groups of independent speakers and MPs in transition). However, it is important to note the sizeable difference in the raw frequencies for specific cases, such as in the case of Centre (which only contain 169 speeches overall).

Conversely, the occurrences within *Your* subcorpus are most commonly present (in terms of raw frequency) in *Right* speeches, but the relative density suggest the occurrences to be more typical of the *Left* speeches (152.00%), *Right* (145.80%) and *Right to far-right* (124.74%) speeches. As with the *Our* subcorpus, we observe notable differences in the raw frequencies for specific cases, where relative density showed greater typicality (for example, the *Left* contains 2,564 speeches compared to 13,642 *Right* speeches, with relatively similar relative density, but almost 10 times the frequency of *Left* speeches).

Lastly, in the subcorpus with both lemmas present within one sentence (*Our&Your* subcorpus) the occurrences most frequently appear in *Right* speeches (F), but according to relative density, seem to be quite typical in Left speeches (153.49%). Occurrences are also more common and more typical for *Right* (133.12%) and *Right to far-right* (118.72%) speeches. It is important to note that *Centre* speeches are not present in the *Our&Your* subcorpus.

4.2. Collocation Analysis

To examine the PI building categories more closely and help us answer the question, “Who or what is ours or yours?", we analysed collocates that appear directly alongside the lemma(s). The relevant collocations and their political identity categories are shown in Table 5.

***Our* PI categories:** The top 20 collocations identified within *Our* subcorpus most frequently refer to the *Norms & Values* category, followed by *Activities/Discourse*. However, within this range, there are no collocations that identify *Aims* or *Ideology* as the dominant PI category. This does not mean that such collocations do not exist, but rather that they may be ranked lower, which is the case for the *Aims* category.

The fact that *Aims*-related collocations fall outside the top 20 may also suggest that these aims, often resulting from discursive acts such as assessments or statements, are a secondary aspect of “our" speech patterns. They are used mostly to defend actions and explain the benevolent or honourable intentions of the speakers" side. In contrast, the abundant presence of *Norms & Values*-related collocations indicates that, for “our" occurrences, norms and common virtues are extremely important and can be used to highlight the good work of our group (such as our good care for our people, homeland, etc.), while also contrasting these norms and values to criticise the opposing side (Them) and emphasise their poorly considered actions or negative traits.

Additionally, many collocations (e.g. our government, our children, our media, our politics) reveal a dual function of “our". On the one hand, it highlights positive achievements associated with the speaker’s side (e.g. “pensions were harmonised to an extraordinary degree during our government’s term"). On the other, it signals national belonging and shared values, while also emphasising the opposition’s perceived failures or their harmful consequences (e.g. “Our government still does not understand this" or “the erosion of our media").

Category	Naš	Vaš
Membership	**stranka** (party), **vlada** (government), **stališče** (position)	**stranka** (party), *poslanec* (MP), *minister*, *kandidat* (candidate)
Activities & Discourse	**mnenje** (opinion), **ocena** (evaluation), **amandma** (amendment), **predlog** (proposal)	**odgovor** (answer), **mnenje** (opinion), **izjava** (statement), **trditev** (claim), **odločitev** (decision), **predlog** (proposal), **nastop** (speech/appearance), **stališče** (position), argument
Aims	*cilj* (goal), *namen* (purpose), *interes* (interest), *zahteva* (demand), *pobuda* (initiative), *želja* (wish)	*namen* (purpose), *interes* (interest), *zahteva* (demand), *želja* (wish), *vladavina* (rule/governance), *vladanje* (governing), *pobuda* (initiative)
Norms & Values	**država** (state), **družba** (society), **državljan** (citizen), **skupen** (common), **državljanka** (citizen [female]), **gospodarstvo** (economy), **otrok** (child), **prepričanje** (belief), **človek** (human), **praven** (legal), **zakonodaja** (legislation), *nacionalen* (national), *ozemlje* (territory), *narod* (nation), *kmet* (farmer), *odgovornost* (responsibility), *rojak* (fellow citizen), *davkoplačevalec* (taxpayer), *kultura* (culture)	*otroci* (children)
Ideology	*vaš* (yours), *stran* (side)	**predsednik** (president), *vaš* (yours; *repetition*), *naš* (ours)
Group relations	**vlada**, koalicijski, sosedje	**ministrstvo**, **stran**, **predhodnik**, **koalicijski**, **vlada**, **politični**, **kolega**,
Power resources	*ustava* (constitution), *medij* (media), *moč* (power), *volivec* (voter)	*mandat* (term/mandate), *resor* (department/portfolio), *volivec* (voter)
None	/	*pozornost* (attention)

Table 5: Categorisation of the collocations by their dominant PI category. Collocations in bold appear in the top 20 list, while those not in bold represent additional collocations outside the top 20 with a logDice value of 5 or higher.

***Your* PI categories** The top 20 collocations identified within the *Your* subcorpus fall mainly under the *Activities/Discourse* category, followed closely by *Group relations*. In this range, there are no collocations indicating a relation to the *Aims* or *Norms & Values* categories. Additionally, one particular collocation, *pozornost* (attention) does not relate to any PI category; it appears only in the polite phrase *Hvala za vašo pozornost* ("Thank you for your attention"), which was the sole usage found in the manual review. The manual review also showed that collocations in the *Activities/Discourse* category are mostly used to highlight the actions of the opposing side as a point of blame, or to frame responsibility, caution, or criticism regarding decisions made by the opposing side.

Collocations in the *Group relations* category refer to various political groups – most often political opponents rather than allies, such as the government, the coalition, or the opposition. They are used to assign blame, criticise current or previous governments or parties, or imply that a subject belongs to a particular political camp. These expressions also reference shared political allies and often suggest that the opposing side benefits from the support or authority of these allies. Additionally, the collocations in the two most prominent categories (*Activities/Discourse* and *Group relations*) are often interconnected: one denotes the activity or behaviour of the opposing side, while the other identifies the explicit or implicit political group being blamed or targeted.

***Our&Your* PI categories** There are only a few stable collocations that we were able to identify within *Our&Your* subcorpus, with a long tail of low-frequency collocates or function words. (Semi-)stable collocations include *deliti* (to divide), *delitev* (division), *domovina* (homeland), *lev* (left), *desen* (right), *razlika* (difference), *skupen* (common), and *vaš* (yours). Among these, the *Ideology* category is dominant for all except *homeland* (*Norms & Values*).

4.3. Keyword Analysis

Lastly, to identify most representative words and themes within individual subcorpora, we conducted keyword analysis, comparing the two focus subcorpora (*Our* and *Your*) in comparison to one another.

4.3.1. Focal subcorpus: Our, Reference subcorpus: Your

The list of representative keywords contains relatively procedural keywords. The keywords with the highest Score include *predlagan* (adj. proposed), *novela* (amendment), *dopolnitev* (amendment), *direktiva* (directive), *naroden* (national), where all but the last keyword are related to various parliamentary activities (e.g. proposing an amendment to current legislation) and actions undertaken by MPs. This trend is also evident throughout the list of keywords. Several keywords reference speakers' parliamentary groups, both explicitly (e.g., *sab*, *desus*, *demokrat*) and implicitly (e.g., *poslanski*, *skupina*, *klub*).

Overall, there are not many words that reveal specific themes in these speeches, nor are there any strongly charged words, as opposed to keywords in the other comparison. This is understandable, as the speeches would at most emphasise their work

with words related to their activities, or use these to defend their work.

4.3.2. Focal subcorpus: Your, Reference subcorpus: Our

While the previous scenario provides relatively generic words associated with MP activities and membership, this is not the case with the keywords of the *Your* focus corpus, where the representative lemmas carry much stronger sentiment. The five keywords with the highest Score are *vaš* (yours), *ti* (you), *zanimati* (to interest), *prositi* (to ask), and *odgovor* (answer). The two lemmas with the highest scores (*vaš* and *ti*) were expected, given the composition of the *Your* sub-corpus and the interactional nature of the speeches. The lemma *ti* represents various second-person pronoun forms (e.g., *vi*, *vam*, *vas*). Contrary to the first scenario, the list includes many explicitly or contextually negative keywords, as well as politically charged keywords, which serve as attacks on the opposing side, which can be seen in Table 6.

	Keywords
Verbs (actions/processes)	*zavajati* (to mislead), *očitati* (to reproach), *odstopiti* (to resign), *lagati* (lie), *žaliti* (to insult), *oprostiti* (to forgive), *hvaliti* (to praise / boast), *sprenevedati* (to feign ignorance), *kadrovati* (to appoint personnel), *izjaviti* (to state)
Nouns (entities, concepts, institutions)	*interpelacija* (interpellation), *magnetogram* (verbatim transcript), *laž* (lie), *izjava* (statement), *obtožba* (accusation), *ovadba* (criminal complaint), *protest*, *neresnica* (falsehood / untruth), *norec* (fool / madman), *tožilstvo* (prosecution / public prosecutor's office), *neumnost* (stupidity / nonsense), *levičar* (leftist)...
Adjectives	*proceduralen* (procedural), *koalicijski* (coalition / coalitional), *ministrsko* (ministerial) ...

Table 6: Examples of keywords representative of *Your* subcorpus with lexical negative sentiment, and keywords with lexically neutral or positive sentiment but used in explicitly negative contexts.

We also identified keywords associated with specific themes, including Covid-19 terms (e.g. *maska*, *ventilator*) and references to public media (e.g. *novinar*, *rtv*), though these were less frequent. Mentions of parliamentary groups (*sd*), ministerial roles (*minister*, *ministrica*), and MPs' names (*janša*, *grims*, *gorenak*, *tanko*) were more common, often as targets of criticism.

The keyword *usta* (mouth), while less prominent, is frequently used figuratively in expressions such as "*polna usta so vas*", "*živeti iz rok v usta*", "*iz ust vaših . . .* ", and "*zapiranje ust*"..

One of the most notable characteristics of these keywords relates to the sentiment, expressed within. The list includes keywords that inherently carry (lexical) negative sentiment, such as *zavajati* (to mislead), *laž* (lie), as well as lexically neutral (or even positive) words that are used in a negative context, though such cases are rarer. An example of such case is *hvaliti* (to praise)[6]:

> *"[...] but explaining to people what is good and positive in this law is, of course, the task of the government and coalition MPs. So don't expect us in the opposition to praise **your law**."*

The keyword analysis reveals a clear discursive distinction between Ours and Yours, and thus between Us and Them. The negativity identified in the *Your* subcorpus suggests that the term is often used as a direct or figurative form of attack. In contrast, the neutral to mildly positive sentiment and the absence of keywords carrying strong sentiment in the *Our* subcorpus indicate, on the one hand, the procedural or administrative nature of the speeches and, on the other, an emphasis on the effectiveness and outcomes of "our" work.

5. Conclusion

The paper presented work on the corpus-assisted discourse analysis of political identities within Slovenian parliamentary debates. The methods used aimed to capture more general characteristics of the use of terms "ours" and "yours" through text types, collocation and keyword analysis.

Overall, the results show that occurrences of the lemma "ours" are more frequently associated with neutral or positive speeches and keywords, and are characteristic across different political orientations. Regarding PI categories, the *Norms & Values* category is more frequent, which may indicate a positive self-representation of Us and Our qualities, and what "we" value and emphasise (e.g. homeland, our people, etc.). Additionally, certain collocates reveal the dual function of "our", which can also denote (national or other) common belonging. Occurrences of the lemma " your" are predominantly associated with negative sentiment and evaluative keywords, which are more typical of contributions from ideological extremes and more frequent in the *Discourse/Activity* and *Group Relations* PI categories, reflecting negative other-representation.

Although instances in the *Our&Your* subcorpus (where both terms appear in a single sentence) are too limited for quantitative analysis, they proved valuable qualitatively: manual inspection of sentiment, collocations, and keywords revealed signs of political polarisation (i.e. *naši in vaši*). Taken together, the findings align with the ideological square: "our" foregrounds positive self-

[6]Kociper, Maša. (2021). Zapisi sej Državnega zbora Republike Slovenije, Redna 8. mandat, 91. izredna seja (27. 12. 2021).

presentation, while “your" reinforces negative other-presentation (Van Dijk, 2010).

While previous research focused on the expressions of Us versus Them dichotomy through different lenses and approaches (Kryvenko, 2024; Räikkönen, 2024; Blagus Bartolec, 2024; Al Maani et al., 2022), direct comparisons with prior work are limited. Nevertheless, this study contributes to research on discursive polarisation by showing that, beyond personal pronouns, possessive constructions also encode in-group and out-group distinctions. It further demonstrates how the *naši–vaši* pattern relates to the construction of political identities in Slovenian parliamentary debates.

These results indicate a potential direction for future research. The findings of this preliminary analysis could be extended by systematically examining shifts in terminology across successive legislative periods. Such a diachronic analysis would allow a clearer assessment of discursive change over time. Moreover, this analysis could be complemented by investigating additional dimensions, particularly the institutional roles and strategic positioning of parliamentary actors, to better contextualise patterns of term usage within broader political dynamics.

6. Acknowledgments

The work described in this paper was created within the Slovenian Research and Innovation Agency research programme P2-0103 “Knowledge Technologies", Slovenian Research and Innovation Agency infrastructure programme I0-0013 “The Research Infrastructure of Slovenian Historiography infrastructure programme" and within the DARIAH-SI and CLARIN.SI research infrastructures.

7. Literature

Bashar Al Maani, Laith Salman Hassan Hadla, Marwan Harb Alqaryouti, and Kamal Ahmad Alruzzi. 2022. The Positive-self and Negative-other Representation in Bashar Al-Assad’s First Political Speech After the Syrian Uprising. *Theory and Practice in Language Studies*, 12(10):2201–2210.

Goranka Blagus Bartolec. 2024. Mi i naši, oni i njihovi u politici: osobne deikse u govorima hrvatskih saborskih zastupnika. *Stanje in perspektive uporabe govornih virov v raziskavah govora*, pages 241–259.

Vaclav Brezina. 2018. *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.

Tomaž Erjavec, Matyáš Kopp, Taja Kuzman Pungeršek, Nikola Ljubešić, Maciej Ogrodniczuk, Petya Osenova, Rodrigo Agerri, Manex Agirrezabal, Tommaso Agnoloni, et al. 2025a. Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 5.0. Slovenian language resource repository CLARIN.SI.

Tomaž Erjavec, Matyáš Kopp, Nikola Ljubešić, Taja Kuzman, Paul Rayson, Petya Osenova, Maciej Ogrodniczuk, Çağrı Çöltekin, Danijel Koržinek, Katja Meden, et al. 2025b. Parlamint II: advancing comparable parliamentary corpora across Europe. *Language Resources and Evaluation*, 59(3):2071–2102.

Darja Fišer and Kristina Pahor de Maiti. 2021. »Prvič, sem političarka in ne politik, drugič pa . . . « Korpusni pristop k raziskovanju parlamentarnega diskurza. *Prispevki za novejšo zgodovino*, 61(1).

Jure Gašparič and Simona Kustec. 2020. Stabilna nestabilnost ali idejnopolitični (ne)značaj slovenskih strank 1992-2018. In *Narod - politika - država: Idejnopolitični značaj strank na Slovenskem od konca 19. do začetka 21. stoletja*. Inštitut za novejšo zgodovino, Ljubljana.

Cornelia Ilie. 2005. Parliamentary forms of address. *Politeness in Europe*, 127:174.

Cornelia Ilie. 2010. Identity co-construction in parliamentary discourse practices. *European parliaments under scrutiny: Discourse strategies and interaction practices*, 1:57–78.

Sylvia Jaworska and Karen Kinloch. 2018. Using multiple data sets. In *Corpus approaches to discourse*, pages 110–129. Routledge.

Adam Kilgarriff. 2009. Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference*, volume 6, pages 41–55. University of Liverpool Liverpool.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ, and Vít Suchomel. 2014. The sketch engine. *Lexicography*, 1(1):7–36.

Anna Kryvenko. 2024. Degrees of Belonging to Europe in Parliamentary Discourse: A Comparative Corpus-Assisted Study. In *Proceedings of the Conference on Language Technologies and Digital Humanities (JT-DH-2024)*. Institute of Contemporary History.

Melika Mahmutović and Marko Lovec. 2024. Exploring affective polarisation of the (digital) public sphere in slovenia: The case of marshal twito. *Javnost-The Public*, 31(3):419–439.

Nina Modrijan. 2007. Naslavljanje pri predajanju, pridobivanju in ohranjanju vloge govorca na parlamentarnih razpravah. *Jezik in slovstvo*, 52(5):3–17.

Kristina Pahor de Maiti Tekavčič and Anna Kryvenko. 2026. *From the Dispatch Box: Unlocking the Potential of ParlaMint Through noSketch Engine and TEITOK*. Inštitut za novejšo zgodovino.

Jenni Räikkönen. 2024. Pronouns separating the UK from the EU: We and us in British newspapers and parliamentary debates in 1973–2015. *Neuphilologische Mitteilungen*, 125(2):421–430.

Relative text type frequency. n.d. [link].

Jenni Riihimäki. 2019. At the heart and in the margins: Discursive construction of british national identity in relation to the eu in british parliamentary debates from 1973 to 2015. *Discourse & Society*.

Pavel Rychlỳ. 2008. A Lexicographer-Friendly Association Score. In *Raslan*, pages 6–9.

Simple maths with keywords and terms. n.d. [link].

Jure Skubic and Darja Fišer. 2022. Parliamentary discourse research in sociology: Literature review. In *Proceedings of the workshop ParlaCLARIN III within the 13th language resources and evaluation conference*, pages 81–91.

Statistic measure LogDice. n.d. [link].

Jan E Stets and Peter J Burke. 2000. Identity theory and social identity theory. *Social psychology quarterly*, pages 224–237.

Teun A Van Dijk. 2004. Text and context of parliamentary debates. In *Cross-cultural perspectives on parliamentary discourse*, pages 339–372. John Benjamins Publishing Company.

Teun A Van Dijk. 2010. Political identities in parliamentary debates. In *European parliaments under scrutiny: Discourse strategies and interaction practices*, pages 29–56. John Benjamins Publishing Company.

Teun A. Van Dijk. 2018. Discourse and migration. In *Qualitative Research in European Migration Studies.* Springer Open.

Victoria Wirth-Koliba. 2016. The Diverse and Dynamic World of 'Us' and 'Them' in Political Discourse. *Critical approaches to discourse analysis across disciplines*, 8(1).

Marko Zajc and Janez Polajnar. 2012. *Naši in vaši : iz zgodovine slovenskega časopisnega diskurza v 19. in začetku 20. stoletja.* Mirovni inštitut.

Representations of Europe and the European Union in Parliamentary Discourse from a Corpus-Assisted Perspective

Anna Kryvenko

INZ, NISS
11 Privoz, 1000 Ljubljana, Slovenia; 7-A Pyrohova str., 01054 Kyiv, Ukraine
anna.kryvenko@inz.si

Abstract

This article examines how Europe and the European Union are represented in parliamentary discourse across three contrasting European political trajectories: the United Kingdom, Slovenia, and Ukraine. Using the ParlaMint 5.0 corpora — uniformly encoded, linguistically annotated, and enriched with sentiment and topic metadata — the study applies a longitudinal, cross-linguistic corpus-assisted discourse approach. Mentions of the EU and Europe in English, Slovenian, and Ukrainian were extracted through targeted queries, supplemented by sentiment profiling, topic distribution analysis, and collocational comparison across three built-in subcorpora (Reference, COVID, COVID,War). The findings show that although the two concepts can overlap, their discursive functions diverge systematically: Europe appears as a broader cultural and geopolitical frame, while the EU attracts more policy-oriented uses. These differences intensify at moments of institutional change or crisis, with sentiment around Europe displaying sharper fluctuations than sentiment around the EU. Cross-national patterns align closely with each country's EU membership status — past, present, or aspirational — shaping how the EU is invoked, assessed, or contested. The study demonstrates the value of multilingual, longitudinal corpus analysis for tracing the evolution of political concepts and for understanding how parliaments discursively negotiate Europe's shifting institutional and geopolitical landscape.

Keywords: parliamentary discourse, divergence of overlapping concepts, British, Ukrainian and Slovenian corpora

1. Introduction

While the concept of Europe has been extensively examined in intellectual and cultural history (e.g., Delanty 1995; Pagden 2002), attitudes towards Europe in public discourse have increasingly become the focus of interdisciplinary research. It includes but is not limited to investigations of the vocabulary used to construct or deconstruct Europe at the national and supranational levels (Diez, 1999; Heinemann et al., 2022), particular 'national baggages' determining debates on European integration in Britain and Germany (Musolff et al., 2001), metaphors framing views of Europe and the EU during Brexit (Charteris-Black, 2019), academic discourse about the European Union and its antecedents (Rosamond, 2007), or the role of media in public opinion about Europe (Balks, 2016; Le, 2021). However, far less attention has been paid to how contemporary parliamentary discourse linguistically constructs and differentiates Europe and the EU.

The tendency to merge the concepts of Europe and the EU has been identified in the speeches of prominent European politicians and in major EU documents (Krzyzanowski, 2010: 92-94), as well as in the narratives of national states (Konovšek 2025, 415) and in the rhetoric of individual national leaders (Helfrich, 2022). At the same time, alongside these discursive practices of convergence between Europe and the EU, contrasting practices that reveal their divergence within European political discourse are also possible. This distinction is particularly relevant in relation to countries such as Ukraine and the UK, which are geographically, historically and culturally an integral part of Europe, yet the former remains only a candidate for EU membership while the latter has withdrawn from the EU. In fact, the three selected countries – the United Kingdom, Ukraine and Slovenia – offer contrasting political trajectories and EU membership statuses, making them particularly suitable for examining how institutional positioning shapes discursive constructions of Europe and the EU. Because Europe and the EU can overlap in political communication, a corpus-assisted approach allows us to observe where this convergence occurs and where discursive practices mark a clearer conceptual separation. Parliamentary debates are particularly suitable for this purpose, as they combine institutionalised language use with real-time political positioning.

By combining automated sentiment analysis with collocational and topical profiling, the study does not treat sentiment as a standalone measure of conceptual meaning but as one component in a triangulated analysis of evaluative framing. This mixed-method design enables a more nuanced account of how Europe and the EU are positioned in parliamentary discourse. The paper seeks answers to the following research questions:

- How consistent are similarities and differences in representations of Europe and the European Union in parliamentary discourse over time?
- To what extent does the divergence between the overlapping concepts of Europe and the

Proceedings of ParlaCLARIN V Workshop on Interoperability, Multilinguality, and Multimodality in Parliamentary Corpora @ LREC 2026, pages 22–30

European Union in parliamentary discourse relate to a country's EU membership status?

In what follows, I will briefly review quantitative and qualitative approaches to changing attitudes in parliamentary proceedings (Section 2), describe the data and method of this study (Section 3), report results of sentiment analysis, topic analysis and collocation analysis in the selected parliamentary corpora (Section 4), discuss the findings as well as limitations of the results interpretation (Section 5) and offer a brief conclusion (Section 6).

2. Quantitative, Qualitative and Mixed Approaches to Sentiment and Topic Dynamics in Parliamentary Discourse

Recent research on changing attitudes in parliamentary debates, including towards European issues, spans a wide methodological spectrum, ranging from large-scale computational models to qualitative discourse-analytic approaches, or their combination, focusing commonly on a single parliament. For instance, Pätz et al. (2025) train two machine learning models to assess topic trends based on six topic categories and sentiment dynamics based on the binary positive-negative schema in Bundestag plenary proceedings across parties between 2019 and 2024, whereas Calvo, Bäck and Carroll (2024) classify speeches according to the binary positive-negative schema based on the relative frequency of sentiment-coded words using an automated dictionary method to examine how established parties distance themselves from radical populist parties in the Swedish Riksdag between 2010 and 2022. Grijzenhout, Marx, and Jijkoun (2014) build a gold standard corpus out of transcripts of two randomly chosen plenary meetings in the Dutch House of Representatives in 2009 containing exclusively 'subjective' paragraphs communicating an opinion, judgement or emotion, which are then annotated for positive or negative orientation. Dodé and Falyuna (2024) examine the use of terminology and sentiment in speeches by the governing parties, the opposition and non-MPs government representatives in the Hungarian parliament between 2020 and 2022, adopting a corpus-assisted discourse analytic approach and drawing on data from ParlaMint-HU. Judge and Shephard (2023) provide a detailed analysis of occurrences of the phrase 'national interest' in the Commons Hansard from the UK parliament between 2016 and 2020 based on 11 variables, including party affiliations and executive positions of the speakers as well as speech sentiment and position preference. Marinova (2025) closely reads stenographic records of five 2025 debates in the newly elected European Parliament to examine how Eurosceptic parties shift agenda through thematic reframing, strategically avoid politically sensitive debates and how they alternate rational argumentation with emotional appeals.

A comparative perspective on changing attitudes in parliamentary discourse appears to be less common, especially when mixed or quality analysis is provided, arguably, *inter alia*, due to limitations in the interoperability and comparability of the datasets available for research. From a bird's-eye view, Economides, Featherstone and Hunter (2024) identify references to EU enlargement in the parliamentary proceedings of eight member states using a combination of automated and hand-coded methods and report a decline in salience, increasingly negative sentiment and a growing emphasis on identity between 1989 and 2019. Mestre-Mestre (2021) compares the lexical expression of emotions in the corpora of proceedings from the Valencian and the Scottish Parliaments during 2020 to find out "whether the lexicons used to identify emotion words in languages other than English are reliable".

3. Data and Method

The data for this study come from three interoperable and comparable corpora of parliamentary proceedings: ParlaMint-GB, ParlaMint-UA and ParlaMint-SI (Table 1), which are uniformly encoded, linguistically annotated and enriched with extensive metadata (Erjavec, Kopp, Ljubešić et al., 2025). In contrast with the older versions of the ParlaMint corpora, the present version 5.0 is additionally annotated for sentiment at the sentence level and topics at the speech level (Pahor de Maiti Tekavčič and Kryvenko, 2026).[1] Both topic and sentiment annotation were performed automatically, using a 21-topic schema supplemented by two additional categories (Others and Mix), and three sentiment scales: a six-level scale, a three-level scale, and a raw numerical output (Ljubešić, Kuzman Pungeršek and Širinić, 2025).

	ParlaMint-GB 5.0	ParlaMint-UA 5.0	ParlaMint-SI 5.0
Span	2015-2022	2002-2023	2000-2022
Speeches	670,912	429,156	311,347
Sentences	5,323,032	3,463,779	3,876,183
Tokens	139,686,402	51,376,472	81,683,385

Table 1: Basic statistics of the corpora used

[1] Unlike the other corpora at the time of writing, ParlaMint-SI is also annotated for sentiment at the speech level.

It also needs to be clarified that ParlaMint-GB includes proceedings from both Houses of the UK Parliament; ParlaMint-SI contains transcripts from Državni zbor, or the National Assembly, which is the lower house of the Slovenian Parliament; and ParlaMint-UA consists of plenary transcripts from the unicameral Verkhovna Rada of Ukraine, literally the Supreme Council of Ukraine. The British, Ukrainian, and Slovenian corpora were selected because they represent three contrasting EU-related trajectories: the UK's shift from full membership through withdrawal to the post-Brexit period, Ukraine's movement from articulated accession aspirations to candidate status, and Slovenia's progression from pre-accession discourse to full membership.

Methodologically, this paper contributes to the literature on Corpus-Assisted Discourse Studies, particularly within their temporal (Partington et al., 2013; Marchi, 2018) and cross-linguistic (Taylor and del Fante, 2020) frameworks, in their application to parliamentary corpora (Kryvenko, 2025a; Kryvenko, 2025b).

All data for this study was extracted via the noSketch Engine concordancer[2]. The terms *EU* or *European Union* in ParlaMint-GB, *ЄС* or *Європейський Союз* or *Євросоюз* 'EU / European Union / Eurounion' in ParlaMint-UA, and *EU* or *Evropska unija* 'EU / European Union' in ParlaMint-SI were queried to extract references to the EU. The hits for *ЄС* 'EU' in ParlaMint-UA were additionally filtered for quotation marks in a span one to the left and one to the right to control for the name of one parliamentary faction and party whose abbreviation is homonymous with that of the EU in Ukrainian ("*ЄС*" stands for "Європейська солідарність" 'European Solidarity'). All corpus queries in this study were restricted to contributions made by Members of Parliament, encompassing both regular speakers and those serving in chairing roles, while contributing to debates (e.g., in ParlaMint-UA, chairpersons produced 48.4% of speeches and 24.7% of tokens in the entire corpus). This delimitation reflects the analytical focus on the discursive practices of elected representatives and preserves the institutional coherence of the analysis.

To compare the sentiment dynamics and topical environments surrounding mentions of the European Union and Europe, the corresponding lexical items in English, Ukrainian and Slovenian were queried across the three built-in subcorpora: Reference (covering all data before 30 January 2020), COVID (covering data from 31 January 2020 to 23 February 2022) and COVID, War (covering data from 24 February 2024 onward, marking the start of Russia's full-scale invasion of Ukraine, as distinct from the beginning of Russia's aggression against Ukraine in 2014). This periodisation is also particularly useful for examining discursive constructions of Brexit, which was formally concluded on 31 January 2020. To obtain a more nuanced understanding of how the concepts of the European Union and Europe divergence across parliaments, top ten collocates for the corresponding lexical items in English, Ukrainian and Slovenian were extracted and compared. All collocate queries were conducted within a −5/+5 span, requiring a minimum corpus frequency of 3 and a minimum in-range frequency of 2. Selected concordance lines with scores outside the original 0–5 sentiment range were read closely to identify whether MPs expressed their attitude towards the EU or Europe, or whether these terms appeared in a broader context in which praise or criticism was directed elsewhere.

4. Results

This section quantitatively traces how references to the European Union and Europe are evaluated, thematised, and discursively patterned across the British, Ukrainian, and Slovenian parliamentary corpora, allowing cross-linguistic and cross-contextual contrasts to emerge from both sentiment and topic distributions as well as from collocational profiles. Sentiment and topic measurements are reported as relative density in text type, an indicator of whether a given value is more or less frequent in the text type than in the corpus as a whole, with values below 100% signalling underuse and those above 100% indicating overuse relative to the baseline. Collocates are ranked based on the Log Dice statistical measure.

4.1 Sentiment analysis

In the British data (Figure 1), sentences containing references to the European Union consistently show higher neutral and lower positive sentiment across all three periods, with neutrality gradually declining and negativity rising most noticeably in the COVID, War subcorpus.

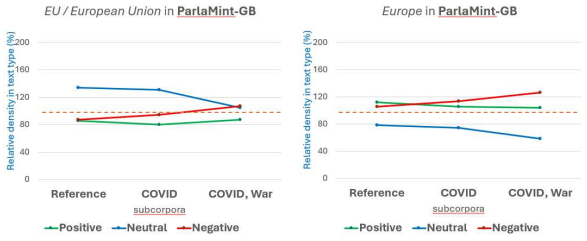


Figure 1: *EU / European Union* (left) and *Europe* (right) in sentences with the sentiment attribute in ParlaMint-GB (relative density in text type)

[2] The log-in installation of noSketch Engine concordancer at CLARIN.SI was used in this study: https://www.clarin.si/skelog

By contrast, sentences containing references to Europe display the lowest neutral values throughout, with positive sentiment remaining comparatively stable and negative sentiment increasing over time.

In ParlaMint-UA (Figure 2), contexts with references to the EU are characterised by consistently high positive sentiment, which rises further in the COVID, War subcorpus, with neutral values declining and negative underperforming consistently. By contrast, contexts with references to Europe display a more mixed evaluative profile: although neutral values remain comparatively low, negative sentiment rises in the COVID subcorpus before receding in the COVID, War subcorpus, while positive sentiment increases markedly after February 2022. The data indicate that while both terms become more positively framed over time, the EU acquires a distinctly stronger and more consistently favourable evaluative load than Europe.

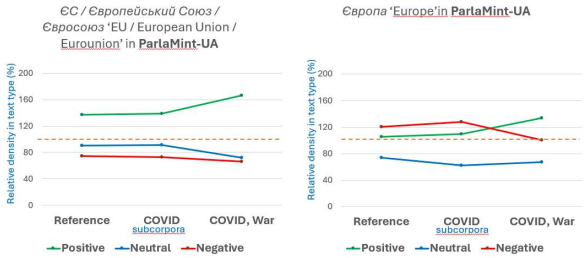
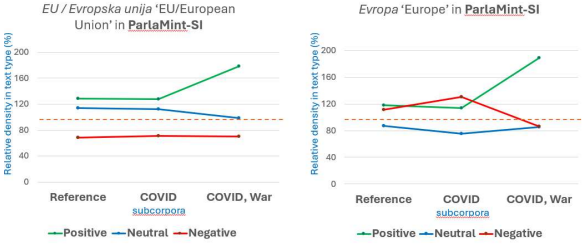


Figure 2: *ЄС / Європейський Союз / Євросоюз* 'EU / European Union / Eurounion' (left) and *Європа* 'Europe' (right) in sentences with the sentiment attribute in ParlaMint-UA (relative density in text type)

Quite similar to the Ukrainian corpus, the Slovenian data (Figure 3) show that sentences referring to the European Union are characterised by consistently high positive sentiment, rising sharply in the COVID-and-War period, while neutral sentiment remains relatively stable across the first two subcorpora before declining slightly in the third. EU contexts with negative values underperform throughout entire ParlaMint-SI.

Figure 3: *EU / Evropska unija 'EU / European Union'* (left) and *Evropa 'Europe'* (right) in sentences with the sentiment attribute in ParlaMint-SI (relative density in text type)



References to Europe, however, display a more dynamic evaluative trajectory: positive sentiment increases markedly in the COVID, War subcorpus, neutral values remain moderate and signal underuse, and negative sentiment peaks during the COVID period before falling substantially thereafter. Overall, the data indicate that while both terms become more positively framed over time, Europe exhibits greater fluctuation in evaluative load, whereas the EU retains a more consistently affirmative profile in ParlaMint-SI.

When comparing sentiment patterns across the three corpora, a common feature is the systematic underuse of neutral sentences containing references to Europe. In both the Ukrainian and Slovenian data, positive contexts for both the EU and Europe increase in the COVID, War subcorpus, reflecting a shift towards more favourable evaluative framing. By contrast, in the UK corpus, the same period is marked by a rise in negative contexts for both the EU and Europe.

4.2 Topic analysis

Regarding references to the EU and their topic associations in the UK parliamentary corpus (Figure 4), foreign trade, international affairs, immigration, agriculture and the mixed category consistently remain the top five topics across all the subcorpora, with foreign trade dominating throughout and peaking in the COVID subcorpus. For references to Europe, the topic distribution is more dispersed. Immigration and culture are the two topics that appear consistently across the whole corpus. Notably, defence and energy become more prominent in the COVID, War subcorpus, suggesting the geopolitical reorientation of parliamentary debate after February 2022. The only 'shared' topic associated with references to both the EU and Europe across all three periods is immigration.

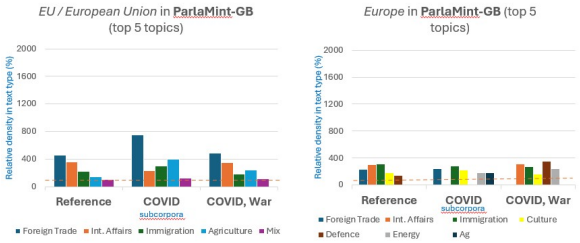


Figure 4: *EU / European Union* (left) and *Europe* (right) in sentences with the topic attribute in ParlaMint-GB (relative density in text type)

The topic profiles for both the EU and Europe in the Ukrainian parliamentary corpus (Figure 5) display a broader and more heterogeneous topical spread compared with the British data. References to the EU are concentrated primarily in the domains of international affairs and foreign trade, although the topics of immigration, environment and, to a lesser extent, technology are also salient. References to Europe are most consistently associated with international affairs, the environment, and energy. In fact, international affairs is the dominant topic linked to both the EU and Europe across all the subcorpora, followed by

environment. As expected, the salience of defence rises in the COVID, War subcorpus for references to both the EU and Europe, while health surfaces in the COVID subcorpus, although only in contexts referring to Europe.

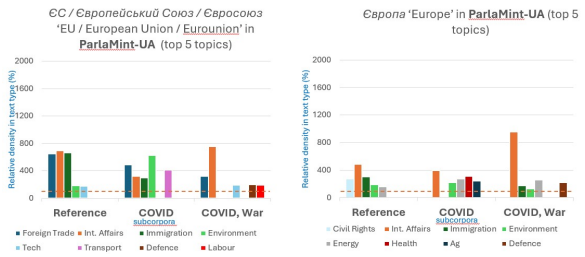


Figure 5: *ЄС / Європейський Союз / Євросоюз* 'EU / European Union / Eurounion' (left) and *Європа* 'Europe' (right) in sentences with the topic attribute in ParlaMint-UA (relative density in text type)

The Slovenian data (Figure 6) show even greater topical heterogeneity, with no single topic strongly associated with both the EU and Europe across all three periods. The most consistent is the topic of international affairs, which becomes especially prominent in the COVID, War subcorpus for both the EU and Europe. Other recurring associations, though less stable, include technology, immigration, and energy, each appearing for both concepts but not in every period. Defence emerges only for Europe in the COVID, War subcorpus, while health appears for both the EU and Europe but at different points in time and with limited overall salience.

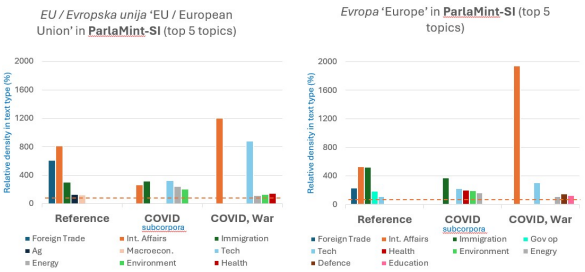


Figure 6: *EU / Evropska unija 'EU / European Union'* (left) and *Evropa 'Europe'* (right) in sentences with the topic attribute in ParlaMint-SI (relative density in text type)

4.3 Collocation analysis

In this section, references to the EU and Europe are compared across the three parliamentary corpora by examining their ten strongest collocates, providing a concise cross-linguistic overview of how MPs in each setting linguistically frame the two concepts.

In ParlaMint-GB (Table 2), the collocates of *EU / European Union* are dominated by Brexit-related vocabulary (*leave, exit, withdrawal, law, agreement, trade, from*), indicating that references to the EU occur primarily in discussions of the UK's departure and negotiations of its terms and conditions with the EU as well as Brexit's legal, political, and economic consequences. By contrast, the collocates of *Europe* point to a much broader geopolitical and cultural framing. Terms such as *eastern, western,* and *continental* situate Europe as a geographical region and political space, while *across, rest,* and *largest* reflect comparative or spatial perspectives. It is also notable that UK appears among the top collocates of EU, whereas Britain collocates with Europe, suggesting that MPs tend to frame the EU in relation to the United Kingdom as a political entity, while Europe is invoked in broader geographical and cultural terms.

Rank	Collocates of EU / European Union	Log Dice	Collocates of *Europe*	Log Dice
1	leave	11.44	eastern	9.40
2	uk	10.38	across	9.32
3	exit	10.23	council	9.11
4	citizen	10.06	rest	8.88
5	law	9.81	western	8.76
6	withdrawal	9.77	largest	8.37
7	with	9.58	continental	8.17
8	agreement	9.41	refugee	8.15
9	trade	9.40	relationship	8.13
10	from	9.39	britain	8.07

Table 2: Top ten collocates of the EU / European Union and Europe in ParlaMint-GB (MPs only)

The Ukrainian data (Table 3) show a clear divergence between the collocational profiles of *ЄС / Європейський Союз / Євросоюз* ('EU / European Union') and *Європа* ('Europe'). References to the EU are strongly institutional and integration-oriented, with top collocates such as *директива* 'directive', *асоціація* 'association', *членство* 'membership', *інтеграція* 'integration', *угода* 'agreement', *вступ* 'accession', and *адаптація* 'adaptation'. The presence of *нато* 'NATO' among the top collocates further underscores how discussions of EU integration in the Ukrainian parliament are closely intertwined with aspirations for NATO membership.

By contrast, the collocates of *Європа* point to much broader and more heterogeneous associations. Some items relate to international organisations (*асамблея* 'assembly', *парламентський* 'parliamentary'), while others evoke geographical or cultural framings (*східний* 'eastern', *світ* 'world'). A notable subset reflects sports discourse (*чемпіонат* 'championship', *футбол* 'football', *фінальний* 'final') referring to the final tournament of the 2012 UEFA European

Football Championship for men's national football teams co-hosted by Ukraine and Poland.

Rank	Collocates of *ЄС / Європейський Союз / Євросоюз* 'EU / European Union / Eurounion'	Log Dice	Collocates of *Європа* 'Europe'	Log Dice
1	директива	10.57	асамблея	10.75
2	асоціація	10.26	парламентський	10.25
3	нато	10.22	асамбль	9.95
4	членство	10.04	чемпіонат	9.53
5	інтеграція	9.68	східний	9.16
6	угода	9.62	світ	9.13
7	країна	9.24	європа	9.12
8	вступ	9.23	футбол	9.11
9	між	9.08	країна	8.99
10	адаптація	9.01	фінальний	8.96

Table 3: Top ten collocates of *ЄС / Європейський Союз / Євросоюз* 'EU / European Union / Eurounion' and *Європа* 'Europe' in ParlaMint-UA (MPs only)

In the Slovenian parliamentary corpus (Table 4), references to the EU centre on institutional, legal and membership-related language, whereas references to Europe cluster around geographical, regional and, to a lesser extent, institutional terms, revealing two rather distinct conceptual framings. In particular, the collocates for the EU such as *članica* 'member state', *vstop* 'accession', *direktiva* 'directive', *država* 'state', *predsedovanje* 'presidency', and *institucija* 'institution' suggest that the EU is framed primarily as a regulatory, membership-based political entity. The Europe collocates instead foreground a geographical and regional framing. Terms such as *vzhoden, zahoden, jugovzhoden*, and *srednji* ('eastern', 'western', 'south-eastern', 'central') show Europe being conceptualised through spatial and geopolitical subdivisions. Collocates like *svet* 'council/world', *skupščina* 'assembly', *konvencija* 'convention', and *združen* 'united' point to Europe as a broader political and organisational space, extending beyond the EU's institutional boundaries. It should be emphasised that *svet* is a polysemantic word in Slovenian. Further querying showed that in nearly 60% of its occurrences within a span of five words to the left and five words to the right of *Evropa*, it appears as part of the named entity *Svet Evrope* 'Council of Europe'.

Rank	Collocates of *EU / Evropska unija 'EU / European Union'*	Log Dice	Collocates of *Evropa 'Europe'*	Log Dice
1	članica	11.42	svet	9.89
2	vstop	10.31	evropa	9.50
3	direktiva	10.03	cel	8.83
4	država	9.80	vzhoden	8.70
5	povprečje	9.35	zahoden	8.70
6	predsedovanje	9.31	jugovzhoden	8.36
7	institucija	9.03	srednji	8.23
8	znotraj	8.97	skupščina	8.12
9	evropski	8.96	konvencija	8.05
10	slovenija	8.91	združen	8.02

Table 4: Top ten collocates of *EU / Evropska unija 'EU / European Union'* and *Evropa 'Europe'* in ParlaMint-SI (MPs only)

To sum up, across the three corpora, the collocational patterns of EU versus Europe show a consistent divide between institutional–political framings on the one hand and broader geographical–cultural framings on the other. In all three languages, references to the EU cluster around legal, regulatory, and membership-related vocabulary, while references to Europe attract more heterogeneous associations tied to geography, culture, or wider international organisations. At the same time, each corpus reflects its own political context: Brexit dominates the UK data, integration and security concerns shape the Ukrainian patterns, and the Slovenian corpus shows a clear separation between EU-institutional language and Europe as a spatial or organisational frame.

5. Discussion

As reported in the section above, the quantitative results suggest that the EU and Europe attract distinct collocational profiles in all three corpora, the cross-corpus comparison shows a consistent split between institutional-political framings of the EU and broader geographical, cultural, or organisational framings of Europe, with each national context inflecting this pattern through its own political priorities and discursive preoccupations.

The selected time periods reflect the availability of comparable parliamentary data and capture moments of heightened political salience in each country, including the post-referendum phase in the UK. Although the corpora do not cover identical chronological spans, they allow for a meaningful comparison of discursive tendencies

shaped by different membership statuses and political trajectories.

That said, it is important to note at least three potential limitations that may affect how the results should be interpreted. First, sentences containing any of the search nodes (e.g. EU, *European Union*, *Europe*) and annotated as "Negative", "Neutral", or "Positive" do not necessarily reflect the speaker's stance toward these entities. In the ParlaMint corpora, sentiment is assigned at the sentence level, whereas the present analysis focuses on patterns of use at the word or phrase level. As a result, the quantitative findings must be interpreted with caution, and close reading of selected concordance lines supported by qualitative analysis is essential for understanding how these terms are actually being used in context.

Brief close reading of concordance lines containing references to the EU, which were scored outside the original 0–5 sentiment range, showed that there is might be no direct correlation even between extremely positive or negative sentences mentioning the EU and Eurosceptical or pro-European attitudes in the UK parliamentary corpus, as illustrated in (1) and (2).

(1) *Does he further agree that the superb work done by the noble Lord, Lord Frost, and his assistant, Oliver Lewis, to try to make the EU understand that Great Britain is not a colony of the European Union but a free and sovereign state is to be applauded?* (Greville Patrick Charles Howard, Conservative Party, 2020-12-14, senti_n 5.271)

(2) *The covid crisis can no longer camouflage the deep damage that Brexit has done, and the single biggest threat to our recovery remains being dragged out of the European Union, against the wishes of those who live in Scotland*. (Ian Blackford, Scottish National Party, 2021-10-27, senti_n -0.211)

At the same time, although extreme sentiment scores may signal metaphorically rich or ideologically charged contexts, the inclusion of the neutral range is indispensable for parliamentary sentiment analysis, as it provides the necessary scope for capturing how sentiment is constructed in political discourse, thereby enabling more accurate interpretation of parliamentary debates.

Second, differences in topic distributions across the three parliaments should be interpreted with caution as well because the corpora do not cover equivalent time spans, and this is likely to affect how broad or narrow the topical range can appear. The UK data, drawn from a much tighter period, naturally produce a more concentrated set of dominant topics revolving around Brexit, while the Ukrainian and the Slovenian corpora, spanning longer and more heterogeneous periods, display wider topical dispersion that may reflect corpus design as much as genuine parliamentary divergence. This means that some contrasts in topical breadth or stability may stem from structural differences in corpus scope rather than substantive differences in how each parliament discusses the EU and Europe.

Third, the strongest collocates of EU and Europe in the entire corpus shouldn't be taken at face value as stable semantic associations without further verification. Some high-ranking collocates, such as the polysemous *svet* in the Slovenian data or sports-related terms in the Ukrainian corpus, are context-dependent or tied to specific historical events, meaning that their apparent prominence may reflect localised discourse patterns rather than broader conceptual framings. To avoid overinterpreting such items, collocates should be systematically re-queried within each subcorpus, with additional filtering, such as distinguishing capitalised forms or applying other refinements, to determine whether they reflect recurrent associations or arise from specific usages.

For instance, further analysis of collocates of *Європа* 'Europe' in ParlaMint-UA restricted to the COVID, War subcorpus returned the following strong collocates : асамблея 'assembly' (9.06), горизонт 'horizon' (8.97), восьмиразовий 'eight-time' (8.47) субконтинент 'subcontinent' (8.46), чемпіонка 'female champion' (8.39), тероризм 'terrorism' (8.03), харлан 'Kharlan (a surname)' (7.97), інновація 'innovation' (7.87), нацизм 'Nazism' (7.82), північноатлантичний 'North Atlantic' (7.78). Sports-related associations appear strongly in this list as well, now framed by the wider context of Russia's war of aggression against Ukraine. During the 2023 World Fencing Championship, Ukrainian fencer Olha Kharlan was initially disqualified after her bout with a "neutral" Russian opponent. Following the sport's no-handshake policy, Olha offered the customary blade tap instead, but officials issued her a black card, which removed her from the competition.

The top collocates of *Evropa 'Europe'* in ParlaMint-SI restricted to the COVID, War subcorpus include *prebuditi* 'to awaken' (9.25), *titanov* 'titanic' (9.01), *ruda* 'ore' (8.36), *trdnjava* 'fortress' (8.35), *rusija* 'Russia' (8.13), *ukrajina* 'Ukraine' (7.81), *ploden* 'fertile' (7.14), *napreden* 'advanced / progressive' (7.04). While the collocates *titanov* 'titanic', *ruda* 'ore' and *ploden* 'fertile' (about soil) relate to Ukraine, and *napreden* 'advanced / progressive' descibes Slovenia, the collocates *prebuditi* 'to awaken' and *trdnjava* 'fortress' point to two known conceptual metaphors EUROPE IS A HUMAN and EUROPE IS A FORTRESS (Bletsas, 2022: 35-36), as illustrated in (3) and (4).

(3) *Grob, brutalen napad Rusije na Ukrajino je prebudil Evropo in to je edino, kar je v*

tej zgodbi pozitivnega. (Branko Grims, Slovenian Democratic Party, 2022-03-16, senti_n 3.982) ['The grave, brutal attack by Russia on Ukraine has awakened Europe, and that is the only positive thing in this story.']

(4) *Žalost, ker ko gre za begunce, ki bežijo zaradi taistih vojn Zahoda, so posamezniki, države in trdnjava Evropa neizprosni in pustijo umirati ljudi v rekah, na odprtem morju, če pa pridejo do meja te Evrope, jih pa zapirajo v taborišča in pustijo zmrzovati nekje daleč stran od oči.* (Matej Tašner Vatovec, The Left, 2022-03-09, s.senti_n -0.044) ['It is tragic that when it comes to refugees fleeing the very same wars caused by the West, individuals, states, and Fortress Europe are ruthless: they let people die in rivers and on the open sea, and if they do reach Europe's borders, they lock them up in camps and leave them to freeze somewhere far from sight.']

The use of metaphor is crucial for analysing how overlapping concepts diverge and acquire distinct meanings; however, exploring this dimension in depth exceeds the scope of the present article and warrants a separate study.

Addressing the two research questions, the findings indicate that similarities and differences in how Europe and the EU are represented in parliamentary discourse are broadly consistent across the three corpora, with a stable division between institutional-political framings of the EU and broader geographical, cultural, or civilisational framings of Europe, even though each national context shapes this pattern through its own political agendas and discursive habits. At the same time, the degree to which these overlapping concepts diverge appears only partially related to EU-membership status: while the UK corpus shows a particularly sharp politicisation of the EU linked to Brexit, the Slovenian and Ukrainian corpora also display clear conceptual separation, suggesting that membership alone does not explain divergence, which is additionally shaped by corpus design, topical distributions, and context-specific discourse dynamics.

6. Conclusion

The comparative, temporal analysis of British, Ukrainian and Slovenian parliamentary discourse demonstrates that references to Europe and the European Union evolve in ways that illuminate each country's European political trajectory. While the two concepts may converge in political discourse, corpus-assisted discourse studies also make it possible to trace their divergence in parliamentary corpora with greater empirical precision.

Although not exhaustive in its approaches or corpus techniques used, this study has shown that references to Europe are typically associated with broader cultural, geographic, or normative frames, whereas references to the EU tend to appear in more explicitly political and policy-centred contexts. This divergence becomes especially visible at moments of crisis or institutional change, with sentiment and topic patterns around Europe showing sharper fluctuations than those around the EU. Cross-linguistic collocational evidence further reveals how MPs embed both concepts in nationally salient debates, from sovereignty and regulation in the UK, to security and integration in Ukraine, and, to an extent, in Slovenia.

Taken together, the findings show that EU membership status – past, present or aspirational – strongly correlates with how the EU is invoked, assessed or contested, while Europe remains a more flexible resource for articulating identity and geopolitical orientation.

The study thus underscores the value of corpus-assisted and multilingual approaches for tracing how key political concepts evolve across time, languages and parliamentary cultures.

7. Acknowledgments

The study was funded by the Slovenian Research and Innovation Agency under Programme P6-0436 "Digital humanities: resources, tools and methods", Project J6-60112 "Parliament in the Age of Europeanisation: the Czech Republic and Slovenia (ParlAgE)", Project N6–0288 "The changing discursive semantics of EU representations: identity, populism, propaganda", Horizon Europe under OSCARS Open Science cascading grant No. 101129751 "ParlaCAP - Comparing agenda settings across parliaments via the ParlaMint dataset", and DARIAH-SI.

8. Ethical Considerations

The author declares no conflict of interest.

9. Bibliographical References

Balks, A.-D. (2016). *The mirror of public opinion? Comparing the news media's perspective on European integration in Germany and the Netherlands*. Waxmann.

Bletsas, M. (2022). Founding Concepts: Metaphor and Metonymy in the (French) ECSC Treaty. In S. Heinemann, U. Helfrich and J. Visser (Eds.), *On the Discursive Deconstruction and Reconstruction of Europe* (pp. 27–45). J.B. Metzler.

Calvo, E. M. L., Bäck, H., and Carroll, R. (2024). Debating the Populist Pariah: Changing Party Dynamics and Elite Rhetoric in the Swedish Riksdag. *Political Research Quarterly*, 77(3), 950–961.

Charteris-Black, J. (2019). *Metaphors of Brexit:*

No Cherries on the Cake? Palgrave Macmillan.
Delanty, G. (1995). *Inventing Europe: Idea, identity, reality*. Macmillan Press.
Diez, T. (1999). Speaking "Europe": The politics of integration discourse. *Journal of European Public Policy*, 6(4), 598–613.
Dodé, R., and Falyuna, N. (2024). The language and motivations of expertise in political discourse. *Információs Társadalom* XXIV, 2, 48–67.
Economides, S., Featherstone, K., and Hunter, T. (2024). The changing discourses of EU enlargement: A longitudinal analysis of national parliamentary debates. *JCMS: Journal of Common Market Studies*, 62(1), 168–185.
Erjavec, T., Kopp, M., Ljubešić, N. et al. (2025). ParlaMint II: advancing comparable parliamentary corpora across Europe. *Language Resources & Evaluation* 59, 2071–2102.
Grijzenhout, S., Marx, M., and Jijkoun, V. (2014). Sentiment analysis in parliamentary proceedings. In B. Kaal, I. Maks, and A. van Elfrinkhof (Eds.), *From text to political positions: Text analysis across disciplines* (pp. 117–134). John Benjamins Publishing Company.
Heinemann, S., Helfrich, U. and Visser, J. (2022). EUROPE under Construction: Introduction and Overview. In S. Heinemann, U. Helfrich and J. Visser (Eds.), *On the Discursive Deconstruction and Reconstruction of Europe* (pp. 7–25). J.B. Metzler.
Helfrich, U. (2022). Notre Europe a besoin d'une refondation – Macron's Strategies of Political Re-Framing. In S. Heinemann, U. Helfrich and J. Visser (Eds.), *On the Discursive Deconstruction and Reconstruction of Europe* (pp. 71–100). J.B. Metzler.
Judge, D., and Shephard, M. (2023). Divining the UK's national interest: MPs' parliamentary discourse and the Brexit withdrawal process. *British Politics* 18, 579–602.
Konovšek, T. (2025). Crisis as political criticism: Slovenia, post-communism, and the conservative turn. In B. Trencsényi et al (Eds.), *East Central European crisis discourses in the twentieth century: a never-ending story?* (pp. 392–416). Routledge.
Kryvenko, A. (2025a). Is it 'the Ukraine crisis' or a war in Europe? Multiple frames of Russia's aggression in parliamentary discourse from a comparative MD-CADS perspective. *Language Discourse & Society*, 13(1), 79–106.
Kryvenko, A. (2025b). 'Maidan has become part of Ukrainian identity': the dynamics of naming and framing civil resistance in parliamentary discourse. *Corpora*, 20(3), 295–329.
Krzyzanowski, M. (2010). *The Discursive Construction of European Identities. A Multi-Level Approach to Discourse and Identity in the Transforming European Union.* Peter Lang.
Le, É. (2021). *Degrees of European Belonging: The fuzzy areas between us and them*. John Benjamins.
Ljubešić, N., Kuzman Pungeršek, T., and Širinić, D. (2025). ParlaCAP: Comparing agenda-setting across parliaments via the ParlaMint dataset [Manuscript]. GitHub.
Marchi, A. (2018). Dividing up the data: epistemological, methodological and practical impact of diachronic segmentation. In C. Taylor and A. Marchi (Eds.), *Corpus Approaches to Discourse: A Critical Review* (pp. 174–96). Routledge.
Marinova, M. (2024). Navigating change: The new European Parliament, Euroscepticism, and the global political landscape. *Papers from the International Scientific Conference of the European Studies Department, Jean Monnet Centre of Excellence*, Faculty of Philosophy at Sofia University "St. Kliment Ohridski", 12, 98–108.
Mestre-Mestre, E. M. (2021). Emotion and sentiment polarity in parliamentary debate: A pragmatic comparative study. *Corpus Pragmatics*, 5, 359–377.
Musolff, A., Good, C., Points, P., and Wittlinger, R. (Eds.). (2001). *Attitudes towards Europe: Language in the unification process*. Ashgate Publishing.
Pahor de Maiti Tekavčič, K., and Kryvenko, A. (2026). *From the dispatch box: Unlocking the potential of ParlaMint through noSketch Engine and TEITOK* (Ed. 1.0) [E-book]. Institute of Contemporary History.
Pagden, A. (Ed.). (2002). *The Idea of Europe: From Antiquity to the European Union*. Cambridge University Press.
Partington, A., Duguid, A. and Taylor, C. (2013). *Patterns and Meanings in Discourse: Theory and Practice in Corpus-assisted Discourse Studies (CADS)*. John Benjamins.
Pätz, L., Beyer, M., Späth, J., Bohlen, L., Zschech, P., Kraus, M., and Rosenberger, J. (2025). Analyzing German parliamentary speeches: A machine learning approach for topic and sentiment classification. *arXiv*:2508.03181
Rosamond, B. (2007). European integration and the social science of EU studies: The disciplinary politics of a subfield. *International Affairs*, 83(2), 231–252.
Taylor, C., and del Fante, D. (2020). Comparing across languages in corpus and discourse analysis: some issues and approaches. *Meta,* 65(1), 29–50.

10. Language Resource References

Erjavec, Tomaž; et al. (2025). *Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 5.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/2005.

Towards ParlaMint-DE: Improving the Interoperability of the GermaParl Corpus of Plenary Protocols of the German *Bundestag*

Christoph Leonhardt and Andreas Blätte
University of Duisburg-Essen
{christoph.leonhardt, andreas.blaette}@uni-due.de

Abstract

With the number of machine-readable corpora of plenary protocols continuously increasing, concerns about the potentials of harmonisation and shared encoding standards gain prominence. Interoperability of corpora can contribute to innovative research, in particular when comparative analyses are concerned. The ParlaMint encoding schema introduced by CLARIN provides comprehensive guidelines towards this goal. This contribution shows how GermaParl, a large corpus of plenary protocols of the German *Bundestag*, is transformed from a TEI-inspired XML format to the ParlaMint encoding schema. Based on previous work, this paper presents an adjusted preparation pipeline and discusses challenges of advancing an established resource into a new data format. The prospective ParlaMint-DE corpus will make the plenary debates in Germany from 1949 to 2025 available in a highly interoperable data format. Clear documentation and taxonomies increase the usefulness of the resource in comparative analyses, whereas additional metadata and linguistic annotation broaden its general applicability.

Keywords: ParlaMint, Parliamentary Data, Plenary Protocols, Corpus Creation, German Bundestag

1. Introduction

Many studies in the social sciences and beyond rely on comprehensive corpora of plenary protocols (e.g., Skubic and Fišer, 2022; Skubic and Fišer, 2024). Following a trend towards more accessible and reusable data, an increasing share of these resources is released in machine-readable formats (e.g., Agnoloni et al., 2022, p. 117; Erjavec et al., 2025a, p. 2072; Sebők et al., 2025, p. 33). However, despite the obvious merits of these efforts, the interoperability of parliamentary resources remains a challenge (e.g., Sebők et al., 2025, p. 19).

Against this backdrop, the CLARIN infrastructure has worked towards shared encoding guidelines for parliamentary corpora. Arguing that the described lack of a harmonised standard "present[s] a barrier to their interchange, re-use and comparison" (Erjavec et al., 2023, p. 418), Erjavec et al. introduced the ParlaMint encoding guidelines to facilitate comparative research and the development of new tools and methods (Agnoloni et al., 2022, p. 117; Erjavec et al., 2023). Since then, the ParlaMint project created 29 corpora of European national and regional parliaments (Erjavec et al., 2025a, p. 2072).

One parliament not yet represented in this shared format is the German *Bundestag*. The relevance of German parliamentary debates in the ParlaMint encoding standard motivates our work: The inclusion of an additional, large corpus of parliamentary debates would contribute to comparative parliamentary research at large. At the same time, the adoption of ParlaMint would benefit German parliamentary research. Aside from the immediate gain of interoperability, it would increase usability by adding many innovative features which go beyond those included in comparable available corpora of German parliamentary proceedings while also contributing to the findability of the resource by bringing it closer to the efforts of ParlaMint and CLARIN.

We base our work on GermaParl, a comprehensive and richly annotated corpus of parliamentary debates in the German *Bundestag* (Blätte and Leonhardt, 2025). GermaParl is currently provided in an XML format inspired by a standard of the Text Encoding Initiative (TEI)[1] as well as in the format of the Corpus Workbench (CWB) (Evert and Hardie, 2011). While these formats contribute to the usability of GermaParl in many aspects (Blätte and Blessing, 2018; Blätte et al., 2022), the emergence of the ParlaMint encoding guidelines presents an opportunity to further strengthen its FAIRness.[2]

Following up on an earlier description of the corpus (Blätte et al., 2022), this work presents the transformation of GermaParl to ParlaMint-DE.[3] We focus on the innovations of the new representation and updates of the corpus preparation workflow. Beyond the specific use case, we provide insights into challenges and lessons learned and discuss the updated workflow and toolset as resources which might benefit similar corpus curation projects in the future.

[1] https://tei-c.org.

[2] Referring to the FAIR data principles of Findability, Accessibility, Interoperability and Reusability (Wilkinson et al., 2016).

[3] At the time of writing, the full adoption of ParlaMint is not completed. The following description should not pre-empt the regular collaboration process of ParlaMint described by Erjavec et al. (2025a), but illustrate the workflow, challenges and potentials of transforming an existing resource into a shared data format.

Proceedings of ParlaCLARIN V Workshop on Interoperability, Multilinguality, and Multimodality in Parliamentary Corpora @ LREC 2026, pages 31–43

2. Background

The current version of GermaParl as well as the ParlaMint encoding guidelines present the starting point and the target of our efforts respectively. Discussing them briefly will inform the subsequent presentation of the corpus preparation workflow.

2.1. GermaParl

GermaParl is a comprehensive corpus of plenary protocols in the German *Bundestag*. Since the release of the first version which included all plenary protocols between 1996 and 2016 (Blätte and Blessing, 2018), it has been continuously maintained and advanced. GermaParl v2.0.0, presented in Blätte et al. (2022) and fully released in 2023 covered all protocols between 1949 and 2021. Since then, the quality of the corpus has been evaluated and improved (Leonhardt and Blätte, 2023) and new features like the addition of Named Entity Linking via `DBpedia Spotlight` (Mendes et al., 2011; Daiber et al., 2013) have been included in multiple subsequent updates. In its most recent version (Blätte and Leonhardt, 2025), GermaParl contains 290 million tokens and covers all 4559 sessions between September 1949 and March 2025.[4] The corpus includes comprehensive metadata on the level of protocols (e.g., session date, legislative period) and speeches (e.g., the name or party affiliation of a speaker) as well as linguistic mark-up in its CWB variant (Blätte et al., 2022).

Following the considerations of Blätte and Blessing (2018), GermaParl is provided in two data formats: An XML format serves as an interchange format, aimed at interoperability and especially used in more advanced workflows of users who wish to use their own toolchain. It is inspired by the TEI encoding standard for performance text (Blätte and Blessing, 2018, p. 811). Protocols are represented as nested structures of agenda items, speeches and paragraphs. To further increase usability (e.g., Blätte and Blessing, 2018, p. 813; Skubic and Fišer, 2024, p. 9), GermaParl is also provided in the format of the Corpus Workbench (Evert and Hardie, 2011) which can be analysed using a many different tools, including graphical user interfaces like `CQPweb`[5] or script-based analysis environments like `polmineR` (Blätte, 2023) which is implemented in the statistical programming language R (R Core Team, 2025). This takes into account the relevance of parliamentary corpora for social science research as well as the prevalence of R in this field (Skubic and Fišer, 2024, p. 9).[6]

GermaParl is not the only corpus of parliamentary protocols of the German *Bundestag* (see for example Abrami et al. (2024) or Richter et al. (2023)). Just like GermaParl, these resources regularly face potential trade-offs between coverage, data quality, the availability of metadata and usability. We argue that GermaParl as a high-quality resource which combines longitudinal coverage with comprehensive metadata on various levels of granularity (similarly Blätte et al., 2022) constitutes a good foundation for ParlaMint.

2.2. ParlaMint

The increased interoperability generated by a shared encoding standard provides many opportunities to lower barriers and facilitate innovative research. In this regard, the ParlaMint encoding standard markedly advanced efforts towards more interoperable data. Following the ParlaMint II project, "29 European countries and autonomous regions" have become available in the ParlaMint encoding (Erjavec et al., 2025a, p. 2072; Erjavec et al., 2025b). The temporal coverage of these corpora starts in the 1990s, with most of them beginning in the 2010s and all corpora ending between 2022 and 2024 (Erjavec et al., 2025a, p. 2080). The guidelines[7] formulate comprehensive recommendations and requirements which are sufficiently granular to allow for a truthful representation of different parliamentary proceedings in an interoperable fashion. Comprehensive metadata further contributes to the usefulness and comparative potential of the resources while the inclusion of linguistic annotation in the TEI makes the linguistic mark-up accessible for a broad audience. Furthermore, the ParlaMint project has evolved beyond corpora by making various tools and workflows easily usable for all compatible resources. This broad landscape of data, tools and workflows improves the findability of its individual resources. This yields great potentials for a German parliamentary corpus.

As summarised by Erjavec et al. (2025a, p. 2073), the current ParlaMint encoding guidelines are the result of the iterative revision of the Parla-CLARIN encoding recommendations. In ParlaMint, each corpus has the following structure:

- **Root File**: The root file of the corpus containing metadata on the corpus level including title, creators, extent, data sources and references to other files (person metadata, organisation metadata and taxonomies) (Erjavec et al., 2023, p. 435; Erjavec et al., 2025a, p. 2075).

[4] Number of tokens is based on the CWB version.

[5] `https://cwb.sourceforge.io/cqpweb.php`.

[6] Data of GermaParl is also made available interactively in `PoliCorp`, a web portal currently developed and maintained at GESIS (Smirnova et al., 2025).

[7] `https://clarin-eric.github.io/ParlaMint/`.

- **Common taxonomies**: Taxonomies shared by all corpora following the ParlaMint encoding standard (e.g., speaker types) (Erjavec et al., 2023, p. 436).
- **Local taxonomies**: Taxonomies used by a single corpus (Erjavec et al., 2025a, p. 2075).
- **List of Persons**: A list of metadata on persons, including full names, gender, affiliations to institutions (e.g., the German *Bundestag*), parties and parliamentary groups where applicable (Erjavec et al., 2023, p. 436).
- **List of Organisations**: A list of metadata on organisations (e.g., parties, parliamentary groups, cabinets), including abbreviations and full names, dates of existence as well as the political orientation of parties and parliamentary groups (Erjavec et al., 2023, p. 435; Erjavec et al., 2025a, p. 2086–2087).
- **Corpus Components**: XML files containing a single session. In the German *Bundestag*, this usually corresponds to all debates on a single day.[8] Each component contains corpus-specific and session-specific metadata as well as the speeches. These are encoded as utterances (`<u>`) which themselves contain paragraphs encoded as segments (`<seg>`) and transcriber comments (Erjavec et al., 2023, p. 438–439).

In ParlaMint, there are two versions of each corpus: While the *plain* version of ParlaMint contains the structural annotation described above, the *linguistically annotated* version of the corpus additionally includes linguistic mark-up. This comprises the segmentation of utterances into sentences and tokens, the annotation of lemmata, Part-of-Speech tags, morphological features and named entities as well as the addition of syntactic parsing (Erjavec et al., 2023, p. 439–440).[9] Generally, the encoding guidelines provide some flexibility and allow for the inclusion of additional attributes and mark-up.

In summary, the increased interoperability and resulting usefulness for comparative analyses in particular, but also the additional and more granular metadata, its comprehensive documentation within the corpus itself as well as a more accessible representation of linguistic mark-up are major arguments to move towards ParlaMint. It can be noted that GermaParl is substantively larger than the ParlaMint corpora presented in Erjavec et al., 2025a, p. 2080. However, as already indicated by the creation of other ParlaMint corpora such as ParlaMint-IL which contains over 400 million words (Goldin et al., 2025), the encoding schema itself seems well suited to accommodate large corpora. We follow up on this in the following presentation of an updated corpus preparation workflow.

[8] In the German *Bundestag*, there are a few exceptions to this, e.g., two sessions on December 16, 1949.

[9] Following the ParlaMint guidelines, Part-of-Speech, morphological features and syntactic parsing are annotated according to Universal Dependencies.

3. A Reproducible Corpus Preparation Pipeline (Revisited)

In Blätte et al. (2022), we presented a "Reproducible Corpus Preparation" pipeline for GermaParl. This workflow advanced the process described in Blätte and Blessing (2018) and still provides the foundation of the current version of GermaParl. In general, the corpus preparation pipeline follows the sequence of data *preprocessing*, its re-structuring into XML (or "*XMLification*") and the *consolidation* of metadata, in particular for speakers. In a final step, the linguistic annotation is added.[10]

Despite the differences in the encoding schemas of the current GermaParl corpus and ParlaMint, this established workflow constitutes a good foundation for the preparation of ParlaMint corpora. This benefits from the genuinely generic approach of the established workflow which relies on frameworks, scripts and R packages which can be easily adjusted for different input and output formats. In consequence, we adopt the existing workflow and resources to create a German ParlaMint corpus. To match the encoding guidelines of ParlaMint, some minor adjustments are necessary. In line with Blätte et al. (2022), we structure these updates along the dimensions of *preprocessing*, *XMLification* and *consolidation* and focus on new challenges and lessons learned during the adoption of ParlaMint.

3.1. Preprocessing

Most of the raw input is based on the data collection described in Blätte et al. (2022, p. 10–11). Mostly unstructured XML (for the period of 1949–1996), unstructured plain text (1996–2017) and structured XML (2017–2025) constitute the majority of raw data. In addition, PDF files on which Optical Character Recognition was already performed, were used if protocols were not available in sufficient quality in other data formats.[11]

[10] The documentation of the corpus is also available online: https://polmine.github.io/GermaParl2/.

[11] The input data is retrieved from the website of the German *Bundestag*, most importantly https://www.bundestag.de/services/opendata. The precise source of each protocol is documented within the resulting XML/TEI in both the current GermaParl XML/TEI and ParlaMint.

Each of these data formats poses different challenges and trade-offs. In general, we follow the considerations described in the existing workflow (Blätte et al., 2022, p. 10–11) and only change input formats of individual protocols when issues become apparent. Aside from switching input formats selectively, we fine-tune the existing preprocessing steps which among other things include the removal non-substantive contents from the raw input and the correction of OCR errors.

3.2. XMLification

When preparing GermaParl as XML/TEI, we relied on the "Framework for Parsing Plenary Protocols" (`frappp`), an R package which facilitates the identification, encoding and enrichment of agenda items and speeches in a large collection of unstructured text (Blätte et al., 2022).[12] Using `frappp`, most information was extracted from preprocessed input documents via regular expressions.

`frappp` creates one XML/TEI file per plenary session. Since the general structure of these session-specific files in the XML/TEI format is similar to those of the ParlaMint encoding, we continue to use `frappp`. Similar to the previous XML/TEI format, each ParlaMint corpus component file begins with a TEI header containing both general and session-specific metadata. This is followed by utterances which are potentially nested within agenda items.

Compared to the current XML/TEI of GermaParl, the metadata represented in each TEI header in ParlaMint is more comprehensive and potentially multilingual to further increase interoperability. The representation of utterance elements in ParlaMint and speech elements in the earlier XML/TEI format differs in some aspects. Importantly, in ParlaMint, each utterance and segment (as well as sentence and token in the linguistically annotated version) is assigned a unique identifier. In addition, each utterance element contains three attributes in the new format: A person identifier, an identifier for the utterance itself and a reference to the parliamentary role of the speaker. At this stage, we use the speaker name we extract from the protocol itself as a temporary person identifier which we replace in a later step of the workflow with a consolidated person identifier[13] as this information can be noisy (e.g., typos or errors in the protocols), incomplete (e.g., missing given names) or ambivalent (e.g., various speakers sharing the same family name). To prepare the subsequent consolidation of person metadata, the parliamentary group affiliation for Members of Parliament (MPs) is extracted from the protocols and temporarily stored in the utterance element. Similar to the temporary person identifier, we remove this information later on as it is not part of the utterance element in ParlaMint.

The classification of transcriber comments is more fine grained in ParlaMint. This is implemented via an additional mapping of regular expressions used to identify these sequences in the protocols. In addition, in the updated workflow, some regular expressions are tuned to improve the detection of speakers and transcriber comments.

In summary, turning unstructured input into XML largely follows the same workflow as before. The code base of `frappp` was modified mainly to account for running identifiers of various elements. Otherwise, the existing framework proved to be flexible enough to facilitate the creation of other data formats such as ParlaMint.

3.3. Consolidation

3.3.1. Enrichment and Metadata Structure

To consolidate the speaker information stored in each utterance element, extracted information is matched against external data sets of known parliamentary actors. Based on this disambiguation, additional metadata is then added to the ParlaMint corpus. Like above, this process is very similar to the previous workflow and only minor adjustments are necessary from a technical perspective. This mainly concerns the representation of person metadata itself: Metadata on speakers is not stored within each utterance element, but in a separate file. We adjust the consolidation mechanism of `frappp` accordingly.

3.3.2. Metadata Collection and Sources

Much of the information included about speakers in the GermaParl XML/TEI can also be found in ParlaMint, albeit often with greater granularity: Full names are provided as "*forename*" and "*surname*" and "*affiliations*" to parties and parliamentary groups are represented by references to a list of metadata on organisations. The temporal quality of metadata is made explicit: Names as well as affiliations to parties, parliamentary groups or roles can change, indicated by attributes "*from*" and "*to*" in some elements. In addition to required metadata (gender and affiliation information in particular), we provide dates of birth where possible and several external identifiers which should facilitate easier linkage with other parliamentary data sets as well as general knowledge bases like Wikidata (Vrandečić and Krötzsch, 2014).

In general, the extent of metadata required by ParlaMint greatly exceeds the information available

[12] `frappp` is available on GitHub: `https://github.com/PolMine/frappp`.

[13] We ultimately use Wikidata IDs (Vrandečić and Krötzsch, 2014) as the central identifier for persons.

in the current version of GermaParl. The limited availability of structured and date-specific information about speakers poses particular challenges when preparing ParlaMint. To address this, we rely on four major sources for metadata:

- ***Stammdaten* file**: Provided by the German *Bundestag*,[14] the *Stammdaten* file provides demographic and biographic information on every MP in the German *Bundestag*.[15] We extract date-specific full names, time-invariant gender information, date of birth, date-specific affiliations to parliamentary groups as well as the parliament itself and a parliamentary identifier.

- **Parliaments Day-by-Day Database**: Compiled by Turner-Zwinkels et al. (2022), the Parliaments Day-by-Day Database (PDBD) comprises of date-specific information on the party affiliation of MPs in Germany, the Netherlands and Switzerland. We use the date-specific party affiliation information for German MPs which is covered between 1949 and 2017.[16]

- **Wikipedia**: Wikipedia is the source for metadata about speakers other than MPs. Particularly relevant metadata includes full names, party affiliations and affiliations to cabinets and other offices. Wikipedia is also used to extend the data of PDBD for date-specific party affiliations (2017–2025).[17]

- **Wikidata**: Gender information of speakers who never have been MPs as well as Wikidata IDs have been retrieved from Wikidata (Vrandečić and Krötzsch, 2014).

[14] https://www.bundestag.de/services/opendata.

[15] The previous workflow (Blätte et al., 2022), but also other data curation projects rely on the *Stammdaten* as well, see Richter et al. (2023, p. 4) who translate it as "base data" or Turner-Zwinkels et al. (2022, p. 764) who translate it as "master data sheet".

[16] According to the online appendix of Turner-Zwinkels et al. (2022), the source of this information in PDBD is the *Stammdaten* file which contains time-invariant party affiliation information but date-specific information on parliamentary group affiliations.

[17] We manually extract information about speakers' affiliations to parties from Wikipedia. The main source are lists of MPs per legislative period in which changes in party affiliation are indicated in an unstructured fashion. Individual Wikipedia pages are used if the information on these overview pages is ambivalent. In instances in which no specific date can be found, we tried to provide the most granular information available (e.g., month-specific information).

3.3.3. Challenges

The integration of these various sources is challenging. At first glance, matching persons over these various data sets is comparatively easy: Each of the data sets contains some kind of identifier which could be related to another. However, different data sources (e.g., the *Stammdaten* file and Wikipedia) would potentially describe the same speaker (e.g., once as an MP and once as a governmental actor) using contradictory information (e.g., variations in speaker names, different party affiliations in overlapping time spans, partly due to different granularity of metadata). We consolidated varying representations of information and resolved contradictions by using both hard-coded interventions and heuristics. In particular, this should guarantee that no speaker has more than one affiliation to a parliamentary group or party or more than one name at any given time in accordance with ParlaMint.

3.3.4. Organisational Metadata

ParlaMint requires additional information about organisations, in particular parties and parliamentary groups, but also cabinets and other organisations. This information is generally gathered from Wikipedia. In contrast to their rather sparse annotation in the established version of GermaParl, full names in German and English, abbreviations, dates of existence, identifiers and information about the political orientation of parties and parliamentary groups are provided in the ParlaMint variant where available.[18] Like person metadata, this information is stored in a separate file.

3.4. Linguistic Annotation

The planned linguistic mark-up of ParlaMint-DE is presented in Table 1. The transition towards ParlaMint necessitates the addition of morphological features and syntactic parsing. Since neither of the tools used in the previous setup (Blätte et al., 2022, p. 12) extract morphological features, a modification of the processing pipeline is necessary.

For ParlaMint, we plan to use a combination of UDPipe (Straka, 2018) and Stanford CoreNLP (Manning et al., 2014).[19] This selection is motivated by the desire to integrate the tools in our established R-based corpus preparation workflow which should minimise transaction costs and facilitate easier maintenance as well as reproducibil-

[18] Political orientation is extracted manually from info boxes in the English Wikipedia.

[19] We conducted initial annotations with the model "german-hdt-ud-2.5-191206" for UDPipe which we plan to use alongside the workflow for Stanford CoreNLP presented in Blätte et al. (2022). Currently, we still evaluate various options to provide the required linguistic mark-up.

Layer	Annotation Tool
Sentence Segmentation	UDPipe (Straka, 2018)
Tokenization	
POS (UD)	
POS (STTS)	
Lemmata	
Morphological Features	
Syntactic Parsing	
Named Entities*	Stanford CoreNLP (Manning et al., 2014)
Named Entity Linking*	DBpedia Spotlight (Mendes et al., 2011; Daiber et al., 2013)

* Not fully implemented at time of writing.

Table 1: Prospective Linguistic Mark-Up of ParlaMint-DE

ity. This can be realised by using the R packages `udpipe` (Wijffels and Straka, 2026) and `bignlp`.[20] The selected set of tools should be sufficiently fast, given the size of the corpus and adequately accurate (Straka, 2018; Ortmann et al., 2019).

In addition to the required annotation layers, we strive to include Named Entity Linking to the German ParlaMint corpus at an early stage. Already a part of the current CWB variant of GermaParl, we argue that the inclusion of Named Entity Linking yields great potential for the analysis of parliamentary debates (Leonhardt and Blätte, 2024; similarly van Heusden et al., 2022 and Janssen and Kopp, 2024, p. 125). We plan to add this annotation layer via a workflow centred around the `dbpedia` R package presented in Leonhardt and Blätte (2024).[21]

3.5. Customisation

One particular feature of the ParlaMint encoding schema is its applicability to many different corpora (Erjavec et al., 2025a, p. 2073). To account for variation between parliaments, ParlaMint allows for the inclusion of "local taxonomies" (Erjavec et al., 2025a, p. 2075). Making use of this, we suggest multiple custom taxonomies to represent country-specific speaker types, agenda item types and the sources of metadata. This might require adjustments to the overall encoding schema and is thus currently experimental.

3.6. Finalisation

This updated corpus preparation workflow results in 4559 corpus component files. In a final step, we use additional scripts to create the corpus root file, format the metadata files for persons and organisations and add taxonomies. Most of the data of the corpus root file corresponds to the current headers in the individual XML files and is largely based on hard-coded information.

3.7. Lessons Learned

The described updates to our workflow provide insights for curation projects beyond the scope of GermaParl. Two aspects seem particular important: Flexible preparation pipelines which allow for a programmatic implementation of a new output format while not relying on singular hard-coded assignments as well as a dynamic and efficient organisation of required metadata.

Using the presented pipeline centred around `frappp`, it was possible to adopt the ParlaMint encoding standard comparatively quickly. This further emphasises the merits of the initial generic approach: The toolset is not limited to the preparation of a specific corpus but sufficiently flexible. For example, `frappp` allows for the definition of an XML template which provides scaffolding for plenary protocols in various output formats. Challenges emerged from data availability and one lesson learned is that the thorough organisation of metadata is important, especially when various sources are involved. To this end, we created multiple R packages which contain metadata on persons and organisations in a structured format. One particular advantage of R packages is that both the origin of the data as well as potential data processing steps can be comprehensively documented. Using semantic versioning, we can relate various versions of the same metadata to a particular corpus version, allowing for dynamic data management. This strategy pairs well with ParlaMint in which data is also documented and described in detail. This solution also underlines the lack of an integrated, authoritative data source for our use case, necessitating the combination of various country-specific (e.g., the *Stammdaten* file) and more generic (Wikipedia, Wikidata) resources.

4. ParlaMint-DE

In this section, we describe an initial version of ParlaMint-DE which is based on the workflow outlined above. The plain text variant of this version of the corpus is available on GitHub.[22] As will be

[20] https://github.com/PolMine/bignlp.
[21] https://github.com/PolMine/dbpedia.
[22] https://github.com/PolMine/ParlaMint-DE_beta.

discussed in more detail in subsection 4.3, this current state does not yet fully align with the ParlaMint encoding guidelines. We consider this to be a "beta" version of the corpus which will be further refined. In addition, since we currently still evaluate the best workflow to add linguistic annotation we focus on the plain text variant of the corpus in the following.

The protocol data used for the current version of ParlaMint-DE is the same as GermaParl v2.3.0-rc1 (Blätte and Leonhardt, 2025) and covers all plenary protocols of the German *Bundestag* between September 1949 and March 2025.[23] To move towards the ParlaMint encoding standard, the workflow underlying GermaParl v2.3.0-rc1 has been modified and metadata has been added according to the steps discussed above.

Covering the first twenty legislative periods of the German *Bundestag*, the resulting initial version of ParlaMint-DE comprises of about 260 million tokens in 1.03 million utterances in total. Figure 1 provides an overview of the number of tokens in millions by year.[24]

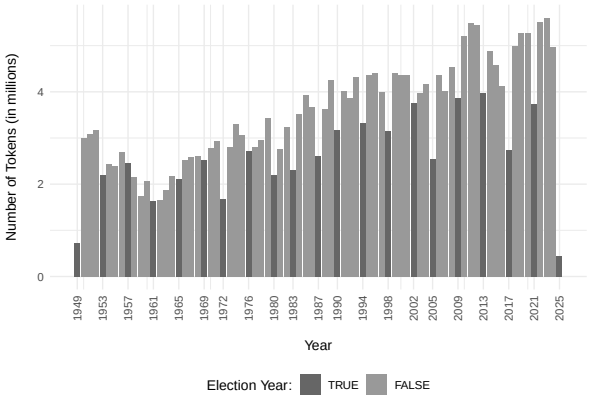


Figure 1: Number of Tokens by Year

Unsurprisingly, this is very similar to Figure 2 in Blätte et al. (2022, p. 10). As before, it is observable that election years usually contain fewer tokens than other years in the same legislative period. The trend towards more tokens per year already seen in Blätte et al. (2022, p. 10) persists.[25]

[23] GermaParl v2.3.0-rc1 is available on Zenodo as a "release candidate" due to the novel nature of the included Named Entity Linking annotations. Its general data structure is equivalent to other releases of GermaParl2. See for example GermaParl v2.1.0 which is openly available (Blätte and Leonhardt, 2024).

[24] We use `quanteda` (Benoit et al., 2018) to extract counts of tokens (including punctuation) from all segment elements in the plain text version of the corpus.

[25] Interestingly, the Open Discourse corpus exhibits the same trend (Richter et al., 2023, p. 6), underlining that this is not merely an artefact of our processing pipeline.

4.1. Metadata

Table 2 provides an overview over the metadata for persons in the German ParlaMint corpus. All in all, we extracted metadata for 4660 persons who actually take the plenary floor.

Concerning organisations, we collected metadata on 42 parties (including explicit references to unknown and missing party affiliations). For 25 out of 42 parties, we were able to extract the party's political orientation based on their English Wikipedia pages. The number of parties further underlines the extensive temporal coverage of the corpus. Information on 21 parliamentary groups (including an explicit reference for speakers without an affiliation to any parliamentary group) was gathered in a similar fashion.

4.2. Linguistic Annotation

Linguistic mark-up which will be available in the final version of ParlaMint-DE is presented in Table 1. Aside from the necessary annotations, we plan to add language-specific Part-of-Speech tags and Named Entity Linking.[26]

4.3. Data Availability

Currently, an initial version of the plain text variant of ParlaMint-DE has been prepared and is made available on GitHub.[27] Aside from all protocols between 1949 and 2025, metadata on speakers and organisations for the entire corpus as well as a draft of the root file and taxonomy files are included in this repository. We provide it as a "beta" version of an emerging ParlaMint-DE corpus to provide insights into the current state of the curation project. We identify three major aspects necessary to fully integrate the data into the ParlaMint encoding schema: Validation, the need to add further metadata and annotation layers as well as customisation.

Complete validation is a major next step towards a fully compatible ParlaMint-DE version. While we paid close attention to the ParlaMint encoding guidelines, we did not yet set up the validation pipeline in accordance with the process used within the ParlaMint project itself (Erjavec et al., 2025a, p. 2075–2076). However, given the comprehensive nature of the schema and potential country-specific characteristics, some complications cannot be ruled out before this validation has been completed. In this regard, changes to the current

[26] At the time of writing, we did not yet finalise the annotation of Named Entities and Named Entity Linking and the precise workflow used for linguistic annotation still needs to be consolidated.

[27] https://github.com/PolMine/ParlaMint-DE_beta.

Metadata	Values	Source
Speaker Name	surname, forename(s)	*Stammdaten* (MPs), Wikipedia (other speakers)
Sex	female (F) / male (M) / unknown (U)	*Stammdaten* (MPs), Wikidata (other speakers)
Date of Birth	Date in YYYY-MM-DD	*Stammdaten* (MPs)
Affiliation/Party	Party, including full labels in German and English, political orientation	PDBD (MPs), Wikipedia (MPs after 2017, other speakers, Metadata on parties themselves)
Affiliation/Parliamentary Group	Parliamentary Group, including full labels in German and English, political orientation	*Stammdaten* (MPs), Wikipedia (Metadata on parliamentary groups themselves)
Affiliation/Role	References to Organisations (e.g., cabinets, other offices)	Protocol Data, Wikipedia for duration data
Identifiers	Wikidata ID, Wikipedia URI, Parliamentary ID	Wikidata, Wikipedia, *Stammdaten*

Table 2: Metadata of ParlaMint-DE

structure of the data available on GitHub are to be expected to ensure full interoperability.

Furthermore, we are aware that some additional metadata and annotation layers still need to be added, in particular with regards to the additions of ParlaMint II (Erjavec et al., 2025a, p. 2085–2090). By making the current version available on GitHub, we aim to make these additions in a more iterative and transparent fashion in the future. At the same time, this also means that the current state of the corpus does not yet include all features available in other ParlaMint related resources.

Finally, as suggested at above, we introduced some customisations to the default ParlaMint encoding schema. In how far these adjustments should be part of the final version of ParlaMint-DE is still to be evaluated.

The release of the current version marks the start of comprehensive and strict technical validation. Going forward, this will enable us to identify remaining issues more quickly. We also aim to increase the transparency of the preparation process. Moving this process to GitHub allows for the public documentation of future adjustments and has the potential to facilitate discussions of encoding decisions with potential users and other stakeholders.

We strive to release complete versions of both the plain and the linguistically annotated variants of ParlaMint-DE in the summer of 2026. In keeping with the established dissemination strategy of GermaParl, we plan to make the corpora available on Zenodo using an open licence (CC BY). Upon release, we would encourage other repositories and platforms to make use of the resource as well.

5. Applications

As initially discussed, the potentials of ParlaMint are underlined by the broad array of workflows and tools which were developed in the context of or are compatible with ParlaMint. This section highlights but a few of these resources.

The `ParlaCAP`[28] and `ParlaSent` (Mochtak et al., 2025) classification models make topic classification and sentiment analysis, two very common tasks in substantive analyses, more accessible for parliamentary research. GermaParl as ParlaMint-DE further simplifies the adoption and comparative application of these resources for German parliamentary debates. Combined with the comprehensive metadata available in ParlaMint-DE, approaches like the sentiment analysis shown by Mochtak et al. (2025) become feasible for scholars of various technical backgrounds.

Similarly, the `dbpedia` R package was designed to lower barriers for Named Entity Linking in social science research (Leonhardt and Blätte, 2024). This can contribute to more fine-grained comparative analyses over multiple ParlaMint corpora (similarly Janssen and Kopp, 2024, p. 125; van Heusden et al., 2022). `dbpedia` makes the adoption of this approach easier by providing a wrapper and workflows for `DBpedia Spotlight` (Mendes et al., 2011; Daiber et al., 2013) for the R programming language with specific support for ParlaMint corpora (Leonhardt and Blätte, 2024).

Finally, ParlaMint corpora can be analysed and visualised with numerous tools. Platforms like `NoSketch Engine`[29] and `KonText` (Machálek,

[28] See the tutorial on GitHub: https://github.com/clarinsi/ParlaCAP-Analysis-Tutorials.

[29] `NoSketch Engine` is an open-source version of

2020) are utilized for ParlaMint corpora and particularly provide access to linguistic analyses (Janssen and Kopp, 2024, p. 121).[30] To make substantive and comparative analyses even more accessible, several additional tools like `TEITOK` (Janssen and Kopp, 2024) or the `ParlaMint NGram Viewer` (de Jong et al., 2024) are available for ParlaMint corpora (see also Erjavec et al., 2025a, p. 2079).

6. Discussion and Conclusion

This contribution presented the transition of GermaParl, a large corpus of plenary protocols of the German *Bundestag*, towards ParlaMint-DE, following a widely adopted TEI encoding standard for parliamentary proceedings. We focused on the potentials of the new standard, the challenges of the transition and possible applications. At the time of writing, ParlaMint-DE is still in development, but it should be released in the near future. A beta version of the corpus is made available on GitHub.

Once finished, the corpus will include all debates in the German *Bundestag* from September 1949 until March 2025 along with comprehensive metadata and linguistic mark-up in an interoperable format and provide access to the resources and ever growing toolset associated with ParlaMint. This immediately addresses some of the gaps presented in current-day discussions on legislative research (e.g., Sebők et al., 2025; Baden et al., 2022). By harmonising our resources, we create comparability and open up avenues for new research. ParlaMint-DE would further add to the coverage of the overall collection of ParlaMint corpora. Furthermore, shared encoding standards also contribute to the integration of tools and workflows and increase the accessibility and findability of resources, in particular for languages other than English.

We further showed how the workflow first presented in Blätte et al. (2022) could be adjusted for ParlaMint. While applied to proceedings of the German *Bundestag* in this paper, its relevance goes beyond this specific use case. By describing both the data we work with and the specific requirements of the targeted encoding guidelines, we were able to identify some generally relevant learnings: Due to its genuine flexibility, the generic approach of our corpus preparation pipeline centred around the `frappp` R package made the adoption of ParlaMint generally possible with comparatively little technical effort. The collection and representation of metadata remained more challenging. The ParlaMint encoding guidelines require comprehensive metadata while also encouraging thorough documentation. Especially when multiple sources of metadata are concerned, this can entail complex data structures. While our solution to create R data packages is specific for our R-based workflow, the need to document the provenance of data included in large corpora is important and potentially deserves more attention by data providers and curators. Ultimately, the applicability of our workflow depends on the structure and quality of both parliamentary proceedings and metadata in each use case. If both the debates themselves and metadata are not available in sufficiently structured data formats, the proposed workflow might be suitable. Lastly, aside from the file size of the resulting corpus which requires adequate infrastructure, the transition of GermaParl underlines that the volume of data is not a limiting factor of future corpus curation projects. From this perspective, parliamentary proceedings which are available in unstructured formats in large volumes might potentially be considered a next use case. The parliaments of the German regional states might be suitable candidates in this regard.

The finalisation of ParlaMint-DE should only mark a starting point for new research. Once completed, we envision a machine-translated version in English like other ParlaMint corpora (Kuzman Pungeršek et al., 2025), a closer integration of tools and workflows afforded by ParlaMint and, ultimately, more substantive analyses in the field of parliamentary research and beyond.

7. Acknowledgements

This work has been made possible by funding from the German National Research Data Infrastructure (Nationale Forschungsdateninfrastruktur/NFDI). We gratefully acknowledge funding from KonsortSWD – NFDI4Society which is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft/DFG) as part of NFDI under project number 442494171 as well as from the Text+ consortium which is funded by the German Research Foundation as part of NFDI under project number 460033370.

Furthermore, the development of GermaParl towards ParlaMint has benefited greatly from a Visiting Fellowship at the Institute of Contemporary History in Ljubljana, Slovenia. This support is appreciated.

Finally, we want to thank the anonymous reviewers for their insightful comments and suggestions.

the commercial `Sketch Engine` corpus management software by Lexical Computing (see https://www.sketchengine.eu/nosketch-engine/). See also Machálek (2020, p. 7003).

[30] The CLARIN.SI research infrastructure provides access to ParlaMint corpora in both `NoSketch Engine` (https://www.clarin.si/ske/) and `KonText` (https://www.clarin.si/kontext/).

8. Bibliographical References

Giuseppe Abrami, Mevlüt Bagci, and Alexander Mehler. 2024. German Parliamentary Corpus (GerParCor) Reloaded. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7707–7716, Turin, Italy. ELRA and ICCL.

Tommaso Agnoloni, Roberto Bartolini, Francesca Frontini, Carlo Marchetti, Simonetta Montemagni, Valeria Quochi, Manuela Ruisi, and Giulia Venturi. 2022. Making Italian Parliamentary 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*, pages 117–124, Marseille, France. European Language Resources Association.

Christian Baden, Christian Pipal, Martijn Schoonvelde, and Mariken A. C. G. van der Velden. 2022. Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures*, 16(1):1–18.

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.

Andreas Blätte. 2023. polmineR. Verbs and Nouns for Corpus Analysis. R package version 0.8.9.

Andreas Blätte and André Blessing. 2018. The GermaParl Corpus of Parliamentary Protocols. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 810–816, Miyazaki, Japan. European Language Resources Association.

Andreas Blätte, Julia Rakers, and Christoph Leonhardt. 2022. 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*, pages 7–15, Marseille, France. European Language Resources Association.

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*, pages 121–124, Graz, Austria. Association for Computing Machinery.

Asher de Jong, Taja Kuzman, Maik Larooij, and Maarten Marx. 2024. ParlaMint Ngram Viewer: Multilingual Comparative Diachronic Search Across 26 Parliaments. In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024*, pages 110–115, Turin, Italy. ELRA and ICCL.

Tomaž Erjavec, Matyáš Kopp, Nikola Ljubešić, Taja Kuzman, Paul Rayson, Petya Osenova, Maciej Ogrodniczuk, Çağrı Çöltekin, Danijel Koržinek, Katja Meden, Jure Skubic, Peter Rupnik, Tommaso Agnoloni, José Aires, Starkaður Barkarson, Roberto Bartolini, Núria Bel, María Calzada Pérez, Roberts Darģis, Sascha Diwersy, Maria Gavriilidou, Ruben van Heusden, Mikel Iruskieta, Neeme Kahusk, Anna Kryvenko, Noémi Ligeti-Nagy, Carmen Magariños, Martin Mölder, Costanza Navarretta, Kiril Simov, Lars Magne Tungland, Jouni Tuominen, John Vidler, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, and Darja Fišer. 2025a. ParlaMint II: advancing comparable parliamentary corpora across Europe. *Language Resources and Evaluation*, 59:2071–2102.

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*, 57:415–448.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*, University of Birmingham, UK.

Maarten Janssen and Matyáš Kopp. 2024. ParlaMint in TEITOK. In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024*, pages 121–126, Turin, Italy. ELRA and ICCL.

Christoph Leonhardt and Andreas Blätte. 2023. Evaluating the Quality of the GermaParl Corpus of Plenary Protocols (v2.0.0). In *Proceedings of the 3rd Workshop on Computational Linguistics*

for the Political and Social Sciences, pages 88–100, Ingolstadt, Germany. Association for Computational Linguistics.

Christoph Leonhardt and Andreas Blätte. 2024. The dbpedia R Package: An Integrated Workflow for Entity Linking (for ParlaMint Corpora). In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024*, pages 133–144, Turin, Italy. ELRA and ICCL.

Tomáš Machálek. 2020. KonText: Advanced and Flexible Corpus Query Interface. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7003–7008, Marseille, France. European Language Resources Association.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

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. Association for Computing Machinery.

Michal Mochtak, Peter Rupnik, Taja Kuzman, and Nikola Ljubešić. 2025. Parlasent: mapping sentiment in political discourse with large language models. *Political Research Exchange*, 7(1):2508377.

Katrin Ortmann, Adam Roussel, and Stefanie Dipper. 2019. Evaluating Off-the-Shelf NLP Tools for German. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers*, pages 212–222, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

R Core Team. 2025. R. A Language and Environment for Statistical Computing.

Florian Richter, Philipp Koch, Oliver Franke, Jakob Kraus, Lukas Warode, Fabrizio Kuruc, Stella Heine, and Konstantin Schöps. 2023. Open Discourse: Towards the first fully Comprehensive and Annotated Corpus of Parliamentary Protocols of the German Bundestag. SocArXiv.

Miklós Sebők, Sven-Oliver Proksch, Christian Rauh, Péter Visnovitz, Gergő Balázs, and Jan Schwalbach. 2025. Comparative European legislative research in the age of large-scale computational text analysis: A review article. *International Political Science Review*, 46(1):18–39.

Jure Skubic and Darja Fišer. 2022. Parliamentary Discourse Research in Sociology: Literature Review. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 81–91, Marseille, France. European Language Resources Association.

Jure Skubic and Darja Fišer. 2024. Parliamentary Discourse Research in Political Science: Literature Review. In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024*, pages 1–11, Turin, Italy. ELRA and ICCL.

Nina Smirnova, Muhammad Ahsan Shahid, and Philipp Mayr. 2025. Open Political Corpora: Structuring, Searching, and Analyzing Political Text Collections with PoliCorp. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 983–992, Suzhou, China. Association for Computational Linguistics.

Milan Straka. 2018. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Tomas Turner-Zwinkels, Oliver Huwyler, Elena Frech, Philip Manow, Stefanie Bailer, Niels D. Goet, and Simon Hug. 2022. Parliaments Day-by-Day: A New Open Source Database to Answer the Question of Who Was in What Parliament, Party, and Party-group, and When. *Legislative Studies Quarterly*, 47(3):761–784.

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.

Jan Wijffels and Milan Straka. 2026. udpipe: Tokenization, Parts of Speech Tagging, Lemmatiza-

tion and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit. R package version 0.8.16.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzales-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018.

9. Language Resource References

Blätte, Andreas and Leonhardt, Christoph. 2024. *GermaParl Corpus of Plenary Protocols (v2.1.0)*. Zenodo. PID https://doi.org/10.5281/zenodo.12794676.

Blätte, Andreas and Leonhardt, Christoph. 2025. *GermaParl Corpus of Plenary Protocols (v2.3.0-rc1)*. Zenodo. PID https://doi.org/10.5281/zenodo.15495748.

Erjavec, Tomaž and Kopp, Matyáš and Kuzman Pungeršek, Taja and Ljubešić, Nikola and Ogrodniczuk, Maciej and Osenova, Petya 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, Núria and Bonet Ramos, Maria del Mar and Calzada Pérez, María and Cardoso, Aida and Çöltekin, Çağrı and Coole, Matthew and Dargis, Roberts and de Libano, Ruben and Depoorter, Griet 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 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 Pol, Henk van der and Prokopidis, Prokopis and Quochi, Valeria and Rayson, Paul and Regueira, Xosé Luís and Rii, Andriana and Rudolf, Michał and Ruisi, Manuela and Rupnik, Peter and Schopper, Daniel and Simov, Kiril and Sinikallio, Laura and Skubic, Jure 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. 2025b. *Multilingual comparable corpora of parliamentary debates ParlaMint 5.0*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/2004. ISSN: 2820-4042.

Goldin, Gili and Howell, Nick and Ordan, Noam and Rabinovich, Ella and Wintner, Shuly. 2025. *Comparible corpus of parliamentary debates ParlaMint-IL 1.0*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/2032. ISSN: 2820-4042.

Kuzman Pungeršek, Taja and Ljubešić, Nikola and Erjavec, Tomaž and Kopp, Matyáš and Ogrodniczuk, Maciej and Osenova, Petya and Rayson, Paul and Vidler, John 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, Núria and Bonet Ramos, Maria del Mar and Calzada Pérez, María and Cardoso, Aida and Çöltekin, Çağrı and Coole, Matthew and Dargis, 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 Fron-

tini, 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 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 Pol, Henk van der and Prokopidis, Prokopis and Quochi, Valeria and Regueira, Xosé Luís and Rii, Andriana 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. 2025. *Linguistically annotated multilingual comparable corpora of parliamentary debates in English ParlaMint-en.ana 5.0*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/2006. ISSN: 2820-4042.

From Transcripts to Insights: A Digital Corpus and Interactive Speech Analysis Platform for Turkish Parliamentary Records

Başak Tepe, İrem Nur Yıldırım, Onur Güngör, Suzan Üsküdarlı
Department of Computer Engineering, Bogazici University
Istanbul, Turkey
basaktepe2020@gmail.com, yildirimiremnur42@gmail.com,
onurgu@pt.bogazici.edu.tr, suzan.uskudarli@bogazici.edu.tr

Abstract

Turkish parliamentary transcripts constitute a unique longitudinal record of the country's political, institutional, and linguistic evolution starting from 1920. Yet much of this archive has remained computationally inaccessible due to scanned and analog typewritten transcripts, historical orthography, and heterogeneous formats. We present a unified, machine-readable corpus of the Grand National Assembly of Türkiye (TBMM), comprising 26,648 session transcripts and 1.7 million pages encompassing ten diverse parliamentary entities spanning a century of legislative history. In addition, we introduce an open-access web platform for speech-level analysis of parliamentary debates from 1983 to 2024. The platform integrates named entity recognition, topic modeling, and diachronic semantic shift detection, enabling exploration of discourse patterns across time and parties, including the frequency and thematic focus of speech activities of specific Members of Parliament. By bridging the gap between raw archival scans and modern NLP tools, the dataset and platform support reproducible research in NLP, digital humanities, and computational social science.

Keywords: Turkish parliamentary corpus, corpus digitization, corpus exploration platform, semantic shift detection, topic modeling, speaker-topic attribution

1. Introduction

The Grand National Assembly of Türkiye (TBMM) has been in continuous session since 23 April 1920. Its parliamentary transcripts (Türkiye Büyük Millet Meclisi, 2024) constitute a comprehensive record of the country's major debates, political concerns, and institutional transformations. Over the course of a century, the TBMM has operated under four distinct constitutions (1921, 1924, 1961, and 1982), and most recently under the presidential system introduced by the 2017 constitutional amendments. Each of these transitions redefined the parliament's role and competences, contributing to significant institutional and textual diversity within the archive.

Parliamentary proceedings document these changes across multiple record types: member speeches, proposed bills, written and oral questions, memoranda, and plenary debates. Taken together, they capture the evolving positions of Members of Parliament (MPs, *Milletvekili*) and political parties, shaped by the political issues of their time. Despite their value for political science, history, and computational linguistics, substantial portions of this archive have remained difficult to use computationally, as records have primarily been distributed as scanned documents rather than machine-readable text.

The institutional evolution of the TBMM, spanning constitutional reforms, transitions between single-party and multi-party systems, and periods of military intervention, has produced significant discontinuities in legislative procedures, documentation practices, orthographic conventions, and archival standards. The result is a fragmented archive with heterogeneous formats, organizational structures, and metadata schemas, which has made systematic and longitudinal analysis difficult.

In this work, we present two openly available resources that address this gap. First, we release a unified, research-ready corpus of Turkish parliamentary session transcripts with standardized organization and metadata. Second, we provide a web platform for the analysis of parliamentary speeches, built on a separate processing pipeline. The platform applies NLP methods including named entity recognition, topic modeling, and semantic shift detection. Both the corpus and the platform source code are publicly available.[1] Together, they support both scholarly research and public exploration of Turkish parliamentary discourse.

The main contributions of this work are:

- A comprehensive, machine-readable corpus of Turkish parliamentary transcripts (1920–2024) with standardized metadata for sessions, legislative terms, and document types, addressing the previously fragmented nature of this historically significant archive. The corpus consists of 10 parliamentary bodies, 26,877 total sessions and 1,933,461 pages. The corpus and its processing code are openly available.

[1] Platform can be accessed at this URL. Source code is available at project repository. The corpus is publicly available in Parquet format on Hugging Face at turkish-parliamentary-corpus.

Proceedings of ParlaCLARIN V Workshop on Interoperability, Multilinguality, and Multimodality in Parliamentary Corpora @ LREC 2026, pages 44–55
16 May 2026.

- A web platform for the semantic analysis of Grand National Assembly of Türkiye parliamentary speeches (1983–2024), featuring MP-level topic profiling, keyword-based temporal analysis, and a speech browser. The platform uses corpus data from Term 17 onwards, featuring 1,339 total MPs, 27,662 speeches and 1,343 total sessions. The platform source code is openly available.
- Empirical analyses of Turkish parliamentary speeches, including topic modeling, semantic shift detection, and speaker-level analysis, demonstrating the utility of the released resources for computational studies of Turkish political language.

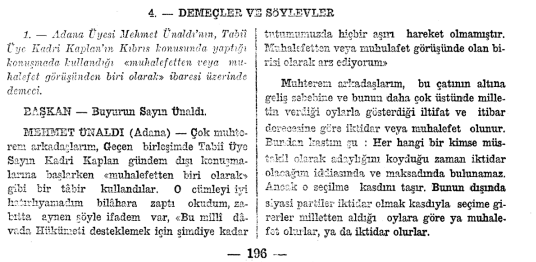
The remainder of this paper is organized as follows. Section 2 reviews related work on parliamentary corpora and computational analysis of political discourse. Section 3 describes the data collection and preprocessing steps, and the structure of the released corpus. Section 5 details the analysis platform, including MP speech extraction, the construction of a search index with enriched metadata, and the applied NLP analyses. Section 7 discusses key findings and situates them within the context of Turkish language studies, digital humanities, and NLP research, while also addressing the limitations of the presented resources.

2. Related Work

Parliamentary corpora and their computational analysis are studied internationally across many countries. Existing efforts range from large-scale comparable corpora to national archives and to tools for diachronic and semantic analysis. This section reviews these lines of work and situates the resource and platform presented in this paper.

International comparable corpora. Erjavec et al. (2025) present ParlaMint 5.0, a set of comparable corpora containing transcriptions of parliamentary debates from 29 European countries and autonomous regions. The Turkish Parliament is included from 2011 to 2022; however, the Turkish subcorpus offers neither page-level granularity nor full coverage of the TBMM archive, leaving room for a dedicated, fine-grained language resource for Turkish parliamentary research.

Turkish Parliament. For Turkish Parliament studies, the corpus of Gungor et al. (2018) is a valuable prior resource. It covers transcripts of the Grand National Assembly of Türkiye but was transcribed using the Tesseract OCR engine and was available only until 2015. It provides session-level data only, without page-level granularity, and thus



4. — DEMEÇLER VE SÖYLEVLER

1. — Adana Üyesi Mehmet Ünaldı'nın, Tabii Üye Kadri Kaplan'ın Kıbrıs konusunda yaptığı konuşmada kullandığı «muhalefetten veya muhalefet görüşünden biri olarak» ibaresi üzerinde demeci.

BAŞKAN — Buyurun Sayın Ünaldı.

MEHMET ÜNALDI (Adana) — Çok muhterem arkadaşlarım, Geçen birleşimde Tabii üye Sayın Kadri Kaplan gündem dışı konuşmalarına başlarken «muhalefetten biri olarak» gibi bir tâbir kullandılar. O cümleyi iyi hatırlıyamadım bilâhara zaptı okudum, zabıtta aynen şöyle ifadem var, «Bu millî davada Hükümeti desteklemek için şimdiye kadar tutumumuzda hiçbir aşırı hareket olmamıştır. Muhalefetten veya muhalefet görüşünde olan birisi olarak arz ediyorum»

Muhterem arkadaşlarım, bu çatının altına geliş sebebine ve bunun daha çok üstünde milletin verdiği oylarla gösterdiği iltifat ve itibar derecesine göre iktidar veya muhalefet olunur. Burdan kastım şu : Her hangi bir kimse müstakil olarak adaylığını koyduğu zaman iktidar olacağını iddiasında ve maksadında bulunamaz. Ancak o seçilme kasdını taşır. Bunun dışında siyasi partiler iktidar olmak kasdıyla seçime girerler milletten aldığı oylara göre ya muhalefet olurlar, ya da iktidar olurlar.

— 196 —

Figure 1: Example of historical TBMM scan (two-column layout, low-contrast artifacts) illustrating OCR challenges for pre-1996 material.

serves as a benchmark for subsequent work. Furthermore, data before 1996 is not digitized on the official TBMM website and requires robust OCR; historical scans often exhibit two-column layout and low-contrast artifacts (see Figure 1), which complicate traditional OCR pipelines. The dataset and platform we present address these limitations by providing a unified corpus from 1920 to 2024 with page-level granularity, an OCR pipeline suited to historical scans, and an analysis platform with speech-level and MP-level access.

Other national parliamentary resources. The Italian Parliamentary Corpus (IPSA) (Frasnelli and Palmero Aprosio, 2024) shares characteristics with the Turkish setting: a long-span dataset (over 175 years), with pre-1996 material available mainly in PDF form and thus requiring OCR. The Austrian National Council is supported by tools such as those at Somes Project Team (2024), which allow tracking MP activity over time and browsing amendments, laws, resolutions, and voting records by party. That work builds on a pre-existing, fine-grained open government data (OGD) Cooperation OGD Austria (2024) corpus. Hyvönen et al. (2023) take a data-driven approach to model parliamentary work as an ontology on the Semantic Web, using the Parliament of Finland as a case study. The MP-centric design of that approach informed the structure of our own analysis platform.

Diachronic and semantic analysis. For semantic shift analysis—one of the components of our platform—the goal is to capture how the meaning of a word changes over time. Early contextual-embedding approaches such as Giulianelli et al. (2020) operationalize semantic change by comparing word representations across *two* time periods; however, such pairwise frameworks require post-hoc cluster alignment and do not scale to fine-grained, year-by-year tracking of meaning evolution. This limitation is particularly relevant in political discourse, where terminology is subject to rapid and continuous reinterpretation across legislative peri-

ods. Periti et al. (2022) address these shortcomings with WiDiD (What is Done is Done), a memory-based clustering method that incrementally accumulates word-sense clusters along a diachronic timeline without requiring inter-period alignment. Because our corpus is partitioned by year and spans over a century of parliamentary debate, WiDiD's alignment-free, scalable design is a natural fit: we adopt it in our pipeline to analyze lexical semantic change in Turkish parliamentary discourse.

Positioning of this work. Despite these advances, no single resource has so far provided a comprehensive, page-level, machine-readable corpus of TBMM transcripts from 1920 to 2024, together with a reproducible OCR pipeline and an open platform for speech-level and MP-level analysis. Our contribution fills this gap by releasing (1) a unified dataset with standardized metadata and page- and session-level granularity, including the 1920–1928 Ottoman–Modern Turkish transition period, and (2) an interactive analysis platform that applies named entity recognition, topic modeling, and semantic shift detection to parliamentary speeches, thereby supporting reproducible research in NLP, digital humanities, and computational social science.

3. Turkish Parliamentary Data

All parliamentary transcripts are publicly available online on the Grand National Assembly of Türkiye (TBMM) website (Türkiye Büyük Millet Meclisi, 2024). This website is the origin of the data.

The organizational structure of TBMM records reflects the formal structure of the Turkish legislature. A *legislative term* (*dönem*) corresponds to a parliamentary period of typically four to five years, beginning after a general election. Within each term, parliamentary activity is divided into *session years* (*yasama yılı*). Each sitting of the assembly produces a *session transcript* (*tutanak*), which is the primary documentary unit of the corpus.

Throughout the republic's history, different types of parliamentary bodies were formed (see Figure 2). All combined, they correspond to 26,648 session transcripts between 1920-2024, consisting of 1.7 million pages. There are 10 different types of parliamentary entities that contribute to this corpus.

A format discontinuity exists in the digital archive. Proceedings published before 1996 are available only as scanned, print-layout PDFs. These documents typically follow a two-column page design and contain scan artifacts, marginal noise, degraded typography, and inconsistent print quality. As they are image-based rather than digitally typed, they are not directly machine-readable and are therefore prone to OCR errors. In contrast, post-1996 proceedings are digitally born and publicly available in structured formats (HTML and Word), in addition to PDF.

To process the material, we employed a three-step pipeline (see Figure 3). First, we obtained all historical data from the Grand National Assembly of Türkiye (TBMM) website (Türkiye Büyük Millet Meclisi, 2024), preserving the original institutional ordering of parliamentary terms and sessions. PDFs were converted to page-level PNG images at 300 DPI, with a standardized directory layout to preserve traceability to source documents. Optical character recognition was performed using the DeepSeek-OCR vision–language model (Wei et al., 2025). The conversion stage consumed approximately 500 GB of RAM and required seven days of wall time; OCR inference on a single NVIDIA H100 GPU completed in nine days.

4. Data Preparation

4.1. Information Flow

The pipeline transforms raw parliamentary documents into the released dataset and analysis platform. Raw PDFs are converted to page images (PNGs), then to text via OCR. The resulting text files are made available on Hugging Face in Apache Arrow format. The same data is then fed into speech extraction pipeline, whose output is stored in a secondary search index. The indexed speeches support three downstream analyses: topic modeling, named entity extraction, and semantic shift analysis.

4.2. Optical Character Recognition

To automate the large-scale processing of Turkish parliamentary records, we developed a pipeline that orchestrates the entire data acquisition and OCR workflow on a per-term basis.

The pipeline consists of four major stages, each designed with fault tolerance and resumability:

1. *PDF Retrieval:* We automatically scrape the official TBMM web archives to identify and download 26,648 session PDFs across all legislative terms, storing valid records in structured directories.

2. *PDF to Image Conversion:* Each PDF file is converted into page-level images at 300 DPI by default, resulting in a repository of 1.7 million PNG images. This ensures a balance between visual clarity and memory usage.

3. *OCR and Corpus Formation:* The repository of images is processed using the *DeepSeek* model. This produces 1.7 million individual page text files, which are then merged back

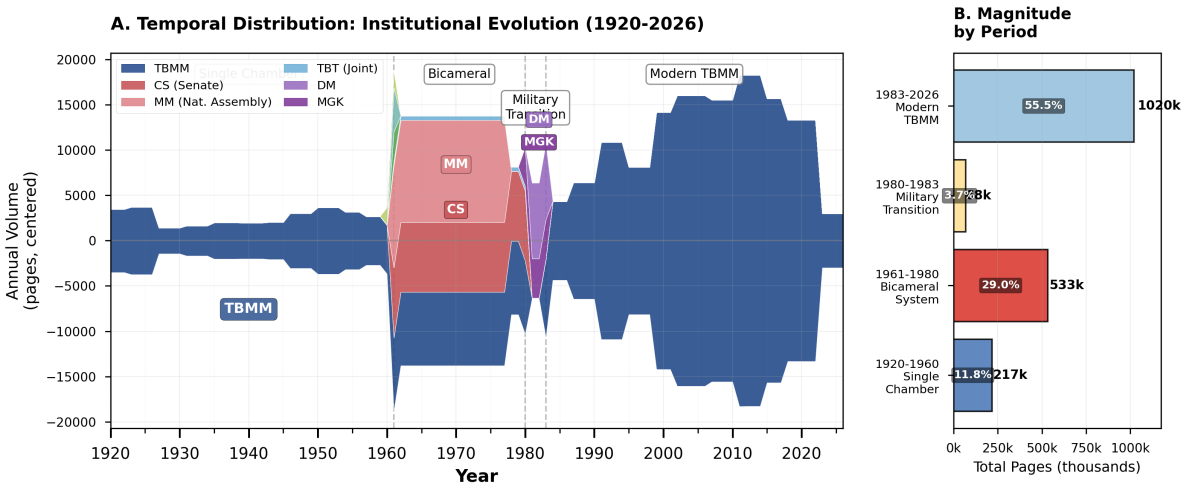


Figure 2: Corpus overview representing different parliamentary bodies' relative contribution.

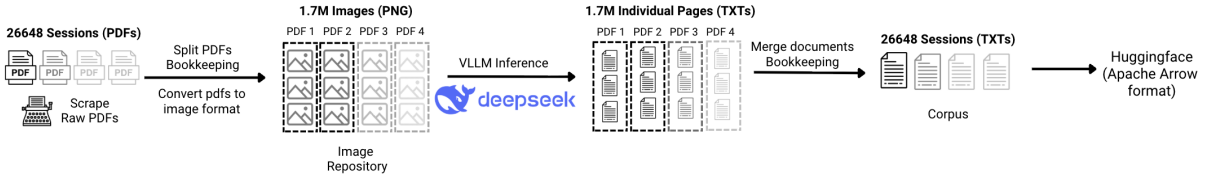


Figure 3: OCR Processing Pipeline

into session-level transcripts to form the finalized TBMM Corpus (26,648 sessions).

4. *Indexing and Storage:* The resulting transcripts and extracted metadata are indexed into *ElasticSearch*. This enables high-performance full-text search and serves as the primary data source for the analysis platform and downstream NLP tasks.

4.3. Parliamentary Speech Extraction

4.3.1. Corpus Scope and Rationale

Parliamentary speeches represent MPs' positions on legislative and societal issues. While the formed corpus spans over a century of records, we apply automated speech extraction specifically from Term 17 (1983) onward when plenary transcripts offer a consistent, comparable record of individual MP speeches. In pre-1983 structure, speeches were spontaneous and unstructured, making systematic extraction infeasible; formalized sections (e.g., "Gündem Dışı Konuşmalar" under "Başkanlığın Genel Kurula Sunuşları") emerged from Term 17, enabling rule based extraction. Many pre-1983 sittings were dominated by roll calls and failed quorums, so transcripts contain little sustained, attributable speech; much debate occurred in committees and informal settings rather than in plenary. MP-level contributions in plenary sessions were mostly motions of censure (gensoru) and formal proposals (önerge), with little direct MP–speech linkage; speeches were rare in this period. In addition, pre-1983 records are split across the National Assembly (Millet Meclisi), Joint Sessions (Birleşik Toplantı), and the Senate of the Republic (Cumhuriyet Senatosu), hindering a unified, comparable speech-level corpus.

Speeches given by MPs in plenary sessions are classified into two types: *Off-the-agenda remarks* (*gündem dışı konuşma*) are statements delivered outside the formal agenda, typically at the opening of a sitting. *Formal statements* (*açıklama*) are declarations made in response to agenda items or on behalf of a party group *(see Appendix)*. For the purposes of analysis, we define an MP's *speech* as the aggregate of their off-the-agenda remarks and formal statements within a given session.

4.4. Text Normalization Framework

4.4.1. Core Principle: Error vs. Variation

We systematically distinguish between digitization artifacts and authentic document variations. OCR-induced errors (i.e., text not present in the original documents) are corrected through normaliza-

tion. Linguistic variations present in the original transcripts are preserved to maintain fidelity to the source material. This approach ensures historical authenticity while enabling reliable text extraction.

4.4.2. OCR Error Correction

Turkish diacritic confusions: Turkish diacritic confusions frequently affect the phrase "gündem dışı" which is systematically misrecognized as gündemdeği, gündemdeş, gündemişi, or gündemiş due to character substitutions (e.g., ş→ğ, ı→e, d→i) and truncation. These variants are non-words in Turkish, do not appear in pre-digitization documents, and are not attested in manually verified original transcripts. Accordingly, all such forms are normalized to "gündem dışı" (see Table 1).

Punctuation standardization: Unicode variants such as em dash (U+2014), en dash (U+2013), and apostrophes (U+2019) are normalized to ASCII equivalents (U+002D, U+0027), as they carry no semantic distinction in the Turkish parliamentary context.

Whitespace normalization: Multiple consecutive spaces resulting from PDF extraction artifacts are collapsed into a single space.

4.4.3. Preserved Linguistic Variations

Topic prepositions: Three semantically equivalent but stylistically distinct forms introduce speech topics: *ilişkin* (formal), *hakkında* (general), and *konusunda* (common). Term 17 shows a balanced distribution (42% / 28% / 30%), while Term 22 reflects a formalization trend (58% / 30% / 12%). These variations are preserved to enable diachronic stylistic analysis.

Vowel harmony suffixes: Turkish possessive suffixes (*'nın, 'nin, 'nun, 'nün*) are determined by phonological rules. Normalization would introduce grammatical errors and is therefore avoided.

Section header formats: Historical variations such as *A)*, *A—*, and *A–)* reflect evolving style guides and are preserved for historical authenticity.

5. Analysis Platform

We developed a platform that allows researchers to browse and analyze Turkish parliamentary speeches. The platform applies a range of NLP methods, including named entity recognition, topic modeling, and semantic shift detection to the corpus. It supports browsing and searching speeches by keywords, political parties, topics, entities, speakers, and dates. It is also possible to see the thematic evolution of a given word within the parliamentary debates — how it evolved, in which different contexts it is used within each year, and when new contexts emerged or disappeared (see Table 3 and Figure 6a). Each MP has a profile page that shows a distribution of their interested topics through time. In addition, the platform profiles each member's topical emphasis relative to their political party (see Figure 6b).

Stage	Content
Input	A) GÜNDEMDEŞİ — KONUŞMALAR ... 1. - Mehmet Vedat Melik'in, Ceylanpınar Tarım İşletmeleri sınırları içerisinde yaşayan göçerlerin sorunlarına ilişkin gündemdeş konuşması...
Changes	1) GÜNDEMDEŞİ → GÜNDEM DIŞI 2) Punctuation: en/em-dash, curly quotes → ASCII (-, ') 3) Whitespace normalization (collapse doubles, fix line joins)
Preserved	Named entities and morphology kept intact: "Mehmet Vedat Melik", "Ceylanpınar", grammatical tokens like "ilişkin" and possessive "'in".

Table 1: Normalization example: preserves names and morphology while fixing OCR noise critical for section and speech detection.

5.1. Topic Modeling

The topic modeling pipeline involves:

Keyword Extraction: The Aya Expanse (Dang et al., 2024) model was used for enhanced semantic representation. Given that the model would run over more than $25{,}000$ speeches, resource and cost-efficiency was prioritized. The purpose of keyword extraction was to perform a normalization strategy before making these speeches subject to topic analysis. There is significant heterogeneity in the length and content of speeches in parliamentary texts: they range from brief speeches to multi-page discourses. There are cases where speeches are to the point, or they have rhetorical filler content. Speeches are even interrupted by automatic microphone cut-offs. By distilling these speeches into a fixed set of keywords, we achieved uniform semantic density, ensuring that the topic in each speaker's turn was fairly represented in the model regardless of the original word count or rhetorical content. The high coefficient of variation in speech length motivated the use of a fixed $k = 10$ keywords per speech.[2]

Embedding TR-MTEB turkish-embedding-model-fine-tuned Baysan and Gungor (2025) (728 dimensions) was used to embed the extracted keywords. Given the limited cardinality of the

[2]Word count statistics for statements (trimmed: bottom and top 3% excluded): $n = 25{,}611$, mean 270.9 words, standard deviation 357.5, median 140, CV 132%.

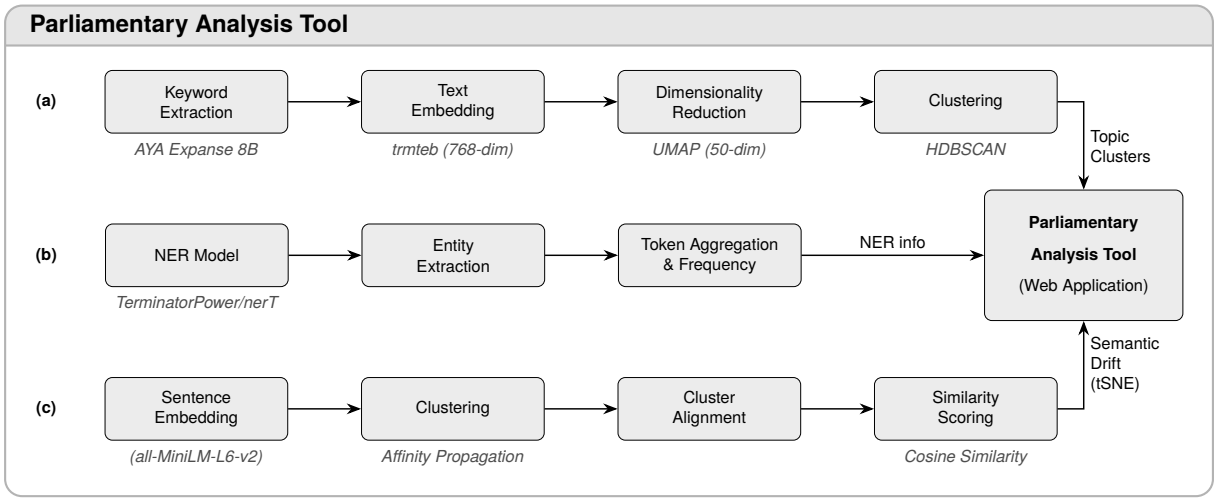


Figure 4: Analysis pipelines within the Parliamentary Analysis Tool: (a) topic modeling, (b) named entity extraction, and (c) semantic shift detection (tSNE). Process boxes list the operation name; the technology used is shown in italics below each box.

keyword set, the resulting high-dimensional representations exhibited sparsity. To mitigate this issue, Uniform Manifold Approximation and Projection (UMAP) McInnes et al. (2020) was applied to reduce the embeddings to 50 dimensions, increasing representational density while preserving the essential topological structure of the latent semantic space.

Clustering HDBSCAN was used due to its ability to perform unsupervised discovery of clusters without a pre-defined number of topics. An alternative approach would be to predefine the topics; however, this would risk missing time-specific topics related to emerging events and newly introduced terms over time (e.g., COVID-19).

Cluster refinement In this corpus, the medium is inherently political; terms such as democracy, institution names (e.g., Meclis, TBMM), ruling party abbreviations, and recurrent conflict-related vocabulary appear across a large number of speeches. Consequently, one cluster (distinct from the outlier label -1) accumulated more than one third of all speeches, with topic labels reflecting shared political rhetoric rather than substantive differences in content. For practical use of the topic assignments, a single refinement step was applied: a stop list of domain-common terms (institution and party names, frequently occurring MP names from the index) was identified and removed from the keyword strings of speeches in that cluster only. The filtered keywords were then re-embedded using the same model, and speeches were either reassigned to existing topics based on cosine similarity to topic centroids or re-clustered under the previous settings.

5.2. Named Entity Recognition

Named entity recognition (NER) was conducted using the `TerminatorPower/nerT` model (Bayraktar Ezel, 2024), targeting three entity categories: PERSON, LOCATION, and ORGANIZATION (Figure 5). As the model employs BERT-style WordPiece tokenization, subword tokens were reassembled into complete entity spans via post-processing. Resulting annotations were aligned to speech-level segments for use in downstream analyses.

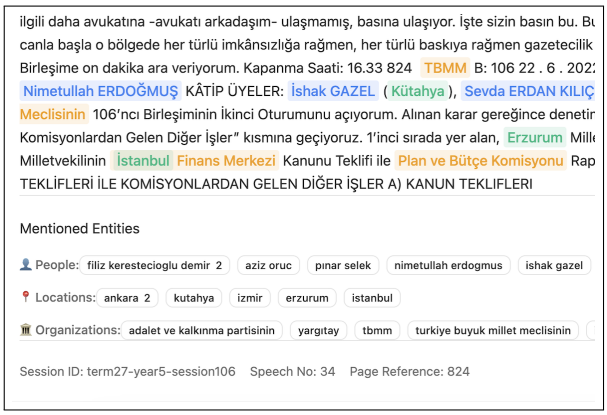


Figure 5: The persons (blue), locations (green), and organizations (yellow) in a speech.

5.3. Semantic Shift

To detect semantic shift in the parliamentary corpus, context windows of 20 tokens were extracted around each target word, capturing the local distributional context contributing to lexical meaning. Departing from the exact-match retrieval of the original WiDiD implementation (Periti et al., 2022), a

regex-based morphological expansion (root $+ \omega^*$) was employed (see Table 2), ensuring coverage of all inflected and derived forms of target keywords.

Target Word	Captured Morphological Variations
iklim *(Climate)*	iklimler (plural) iklimsel (derivational) iklimimiz (possessive) iklimin (genitive)
emekli *(Retiree)*	emeklilik (noun-forming) emeklilerimiz (plural-possessive) emekliye (dative)
döviz *(Exchange)*	dövizdeki (relational) dövizlerin (plural-genitive) dövizle (instrumental)

Table 2: Examples of Turkish morphological variations captured via root $+ \omega^*$ regex expansion, mitigating data sparsity in the embedding phase.

ID	Year	Context (Summarized)
0	1988	Migrant worker population in Council of Europe member states
0	1988	Permanent residence rights for migrant workers in Europe
0	1988	Turkish workers among 15 million migrant workers in Europe
2	1988	Discourse surrounding Palestinian refugees at the United Nations since 1940s
4	1989	Bulgarian Turks migrating to Türkiye (1949–1950)
13	2016	Refugee tragedy and migrant smuggling in Çanakkale
13	2019	Irregular migrants and Syrian refugees framed as a security issue (Aegean/Mediterranean)
13	2022	Allegations of mistreatment of irregular migrants in Greece (Lesbos)

Table 3: Topic clusters for the term *göç* across selected years, illustrating the semantic drift from labour migration discourse in the 1980s to irregular migration and border security framing in the 2010s.

The original WiDiD framework represents a target word w in corpus slice C_j via pseudo-contextual embeddings derived from Doc2Vec (Le and Mikolov, 2014), which produces static embeddings for sequences observed during training (Řehůřek and Sojka, 2010). In the present study, context windows are instead encoded using the pretrained `all-MiniLM-L6-v2` SentenceTransformer (Reimers and Gurevych, 2019), following Yin and Zhang (2024), who demonstrate its efficacy on sentence pairs that are semantically similar but structurally distinct, and vice versa. Applied without corpus-specific fine-tuning, this model yields contextualized representations that capture sense distinctions from the surrounding textual context.

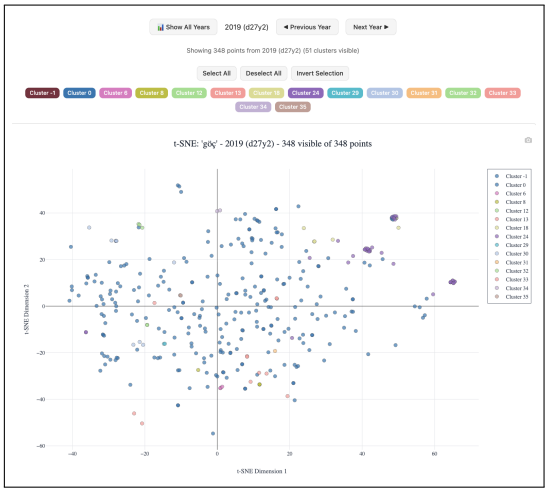
Clustering and alignment: For each term–year, affinity propagation is applied to that year's context embeddings to obtain local cluster labels, followed by a cluster-limiting step in which only the largest N clusters by size are retained and the remainder mapped to an outlier label. Centroids are computed for the retained clusters. A cluster aligner maintains a global list of centroids and assigns stable global IDs: as each new term–year is processed, its centroids are compared to the stored ones via cosine similarity; if the similarity exceeds a threshold, the cluster receives the same global ID, otherwise a new ID is created. This incremental alignment constitutes the core of WiDiD and enables consistent sense identities across time. A single t-SNE projection is then computed on the concatenated embeddings from all term–years, ensuring that spatial coordinates remain comparable across years.

5.4. Illustrative Example

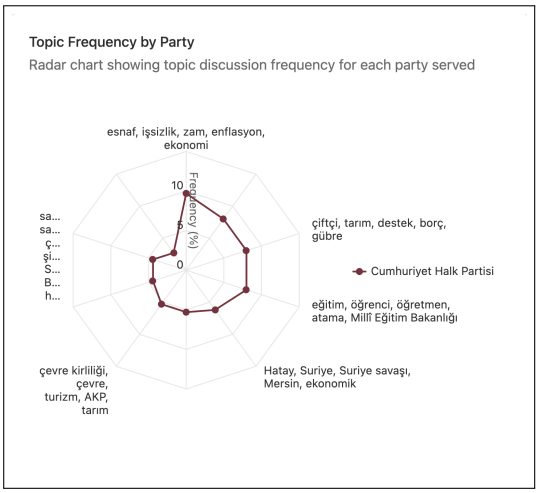
To illustrate the platform's capabilities in practice, we trace the semantic trajectory of the term *göç* (migration) across Turkish parliamentary discourse.

Figure 6(a) presents the contextual topic map associated with *göç* across time. In the late 1980s, the term is distributed across two distinct thematic clusters. Figure 6(b) presents the party-relative topical profile of a CHP member of parliament. Each axis represents a topic to which this MP makes a notable contribution within their party's discourse, with the value indicating the proportion of the party's speeches on that topic attributable to this member. The first cluster concerns labour migration, encompassing Turkish workers among the estimated 15 million migrant workers in Council of Europe member states, their rights to permanent residence, and related policy discussions (Cluster 0). The second associates *göç* with forced displacement in an international context, such as the discourse surrounding Palestinian refugees debated at the United Nations (Cluster 2). Representative contexts for each cluster and year are summarized in Table 3.

By the 2010s, the semantic environment of *göç* shifts considerably. A new cluster associated with irregular migration, refugee flows, and border security emerges (Cluster 13). Speeches in this period address the refugee tragedy in Çanakkale, migrant smuggling in the Aegean and Mediterranean, and the framing of Syrian migrants as a security concern. This transition reflects a broader reorientation of migration discourse in the Turkish parliament fol-



(a) Contextual topic map for *göç* (migration), term 27, year 2. Each point represents a context; colours denote cluster membership. The spatial layout is produced via t-SNE dimensionality reduction of the topic distributions.



(b) Party-relative topical profile of a CHP member of parliament. Each axis represents a topic to which this MP makes a notable contribution within their party's discourse.

Figure 6: Platform visualizations for member-level and semantic analysis.

lowing the onset of the Syrian migration.

The platform further enables attribution of this discourse at the member level. A pronounced contribution to the topic *Suriye* (Syria) in Figure 6(b) indicates that a given MP accounts for a disproportionate share of their party's discourse on Syrian migration, directly linking the observed semantic shift to specific parliamentary actors.

6. Conclusions

We presented a unified, machine-readable corpus of Turkish parliamentary transcripts (1920–2024) together with an interactive platform for speech-level analysis covering 1983–2024. The corpus consolidates 26,648 session transcripts and 1.7 million pages across ten parliamentary entity types into a standardized and research-ready format. The dataset will be publicly released on Hugging Face, facilitating transparent access, reuse, and integration into existing NLP workflows.

The modular structure of the data collection and indexing process makes it straightforward to incorporate newly published sessions, allowing the corpus to remain up to date as parliamentary records continue to expand. The accompanying web platform demonstrates how the resource can support topic analysis, named entity exploration, and semantic shift studies across time, parties, and individual MPs.

We see this work primarily as an enabling resource. Future directions include adding committee-level debates, proposals, bills and memoranda to broaden coverage of legislative activity and aligning the corpus with international parliamentary datasets to support comparative research. Methodologically, future developments involve building a medium-agnostic topic-modeling pipeline tuned for political discourse and refining semantic-shift clustering to automatically identify dominant discourse(s) within specific historical periods. We also aim to improve tooling for auto-updating the corpus and platform as the source data are revised. In addition, future iterations may integrate entity linking to support the construction of knowledge graphs that capture interconnections, mentions, and references among political actors, institutions, and concepts over time. We hope that this release encourages collaborative extensions and reproducible research on Turkish political language.

7. Discussion of Limitations

We acknowledge several limitations. First, the current release of the corpus extends only to 2024 and depends on the availability of data from the official parliamentary website; consequently, it is not continuously updated and requires periodic refreshes to remain current. Second, the extraction step only included MP speeches (for practical and time-related reasons); incorporating bills, committee reports, and other legislative documents would broaden coverage and analytical perspective. Third, automated speech extraction and OCR may fail on documents with exceptional formatting or substantial noise, potentially introducing gaps or transcription errors in specific sessions. Fourth, coverage for the 2011–2022 period remains imperfect and could be further improved through integration with ParlaMint Turkish corpora.

Despite these limitations, the dataset and tools presented here provide a foundation for future re-

search and have the potential to enhance public understanding and democratic accountability by making parliamentary language more accessible, searchable, and systematically analyzable.

8. Code and Data Availability

Speech Analysis Platform can be accessed at this URL. Source code is available at project repository. The corpus is publicly available in Parquet format on Hugging Face at turkish-parliamentary-corpus with different configurations for page-level and session-level transcripts as well as a tbmm-only option for eliminating other parliamentary bodies that emerges and disappears over time.

9. Acknowledgements

We are grateful to Melikşah Türker for their technical assistance and VNGRS for funding the high-performance computational resources and infrastructure required for the data processing and OCR stages of this research.

Bayraktar Ezel. 2024. Terminatorpower/nert: Turkish named entity recognition model. https://huggingface.co/TerminatorPower/nerT. Accessed: 2026-01-15.

Mehmet Selman Baysan and Tunga Gungor. 2025. TR-MTEB: A comprehensive benchmark and embedding model suite for Turkish sentence representations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8867–8887, Suzhou, China. Association for Computational Linguistics.

Cooperation OGD Austria. 2024. data.gv.at – The Austrian Open Government Data Portal. https://www.data.gv.at. Accessed: 2026-02-21.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier.

Tomaž Erjavec, Matyáš Kopp, Taja Kuzman Pungeršek, Nikola Ljubešić, 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, 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, Andriana Rii, 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. 2025. Multilingual comparable corpora of parliamentary debates ParlaMint 5.0. Slovenian language resource repository CLARIN.SI.

Valentino Frasnelli and Alessio Palmero Aprosio. 2024. There's something new about the Italian parliament: The IPSA corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16037–16046, Torino, Italia. ELRA and ICCL.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages

3960–3973, Online. Association for Computational Linguistics.

Eero Hyvönen, Petri Leskinen, and Jouni Tuominen. 2023. A data-driven approach to create an ontology of parliamentary work: Case parliament of finland on the semantic web. In *Proceedings of the International Workshop on Semantic Web and Ontology Design for Cultural Heritage (SWODCH'23)*, volume 3540 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction.

Francesco Periti, Alfio Ferrara, Stefano Montanelli, and Martin Ruskov. 2022. What is done is done: an incremental approach to semantic shift detection. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 33–43, Dublin, Ireland. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Somes Project Team. 2024. SOMES – Social Frames: Platform for Political Transparency. https://www.netidee.at/somes. Accessed: 2026-02-21.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*.

Chen Yin and Zixuan Zhang. 2024. A study of sentence similarity based on the all-minilm-l6-v2 model with "same semantics, different structure" after fine tuning. In *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, pages 677–684. Atlantis Press.

10. Appendix

- *off-the-agenda (gündem dışı):*
  - Terms 17–22 (1983–2007): Speeches appear as subsections under "BAŞKANLIĞIN GENEL KURULA SUNUŞLARI".
  - Terms 23–now (2007–): Speeches appear as main sections with the subsection "A) Milletvekillerinin Gündem Dışı Konuşmaları".
- *statement (açıklama):*
  - Terms 23–now (2007–): Introduction of the "Açıklamalar" (Statements) section.

The structural composition of the corpus is detailed in the following tables. Table 4 provides a breakdown of parliamentary bodies that existed and contributed to the corpus during the century, reflecting constitutional transitions such as the bicameral system and military interventions to current day unicameral system.

Table 5, on the other hand, presents a comprehensive profile of the constructed dataset's scale, categorized by legislative term and document type. This table quantifies the distribution of raw files and page counts formed by the institutions in Table 4: across the three primary data categories: indices, agendas, and full transcripts.

Comparative OCR Performance: A Sample Page Analysis

Figure 7 provides a comparative visualization of a representative archival transcript (Term 10, Year 1, Session 3, 24 May 1954) in three forms: the original document scan, the Tesseract-based OCR results from Gungor et al. (2018), and the output produced by the DeepSeek OCR pipeline. The historical scans frequently exhibit a two-column layout (see Figure 1). Although OCR engines can use layout-aware segmentation, the Tesseract-based transcripts from Gungor et al. (2018) still exhibit frequent character-level errors and fragmented line structure on such pages, as illustrated in Figure 7(b). The DeepSeek OCR pipeline recovers a cleaner transcript with consistent left-to-right reading order and structured formatting, as shown in Figure 7(c).

11. Language Resource References

Onur Gungor, Mert Tiftikci, and Çağıl Sönmez. 2018. A corpus of grand national assembly of turkish parliament's transcripts. In *Proceedings*

English Description	Turkish Name	Abbr.	Period
Grand National Assembly of Türkiye	Türkiye Büyük Millet Meclisi	TBMM	1920 – Curr.
GNAT Joint Session / Joint Sitting Sessions	TBMM Birleşik Toplantı	TBT	1961 – 1980
Senate of the Republic	Cumhuriyet Senatosu	CS	1961 – 1980
National Assembly Sessions	Millet Meclisi	MM	1961 – 1977
Secret Sessions	Gizli Celse	GC	1920 – 2011
Consultative Assembly	Danışma Meclisi	DM	1981 – 1983
National Security Council	Milli Güvenlik Konseyi	MGK	1980 – 1983
Constituent Assembly	Kurucu Meclis	KM	1961
Assembly of Representatives	Temsilciler Meclisi	TM	1961
National Unity Committee	Milli Birlik Komitesi	MBK	1960 – 1961

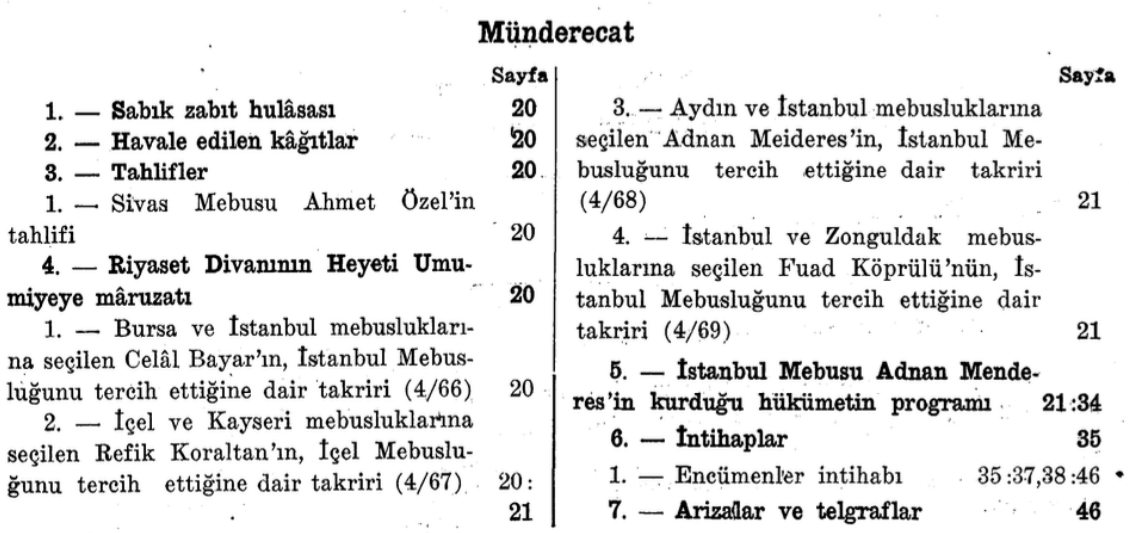
Table 4: Legislative bodies and session types with Turkish nomenclature (açılım) and corresponding date ranges. *Grand National Assembly of Türkiye.

Parliamentary Entity	Term	Index		Agenda		Transcript		Total	
		Files	Pages	Files	Pages	Files	Pages	Files	Pages
TBMM	D01	29	921	0	0	1,104	27,009	1,133	27,930
	D02	33	980	0	0	948	36,236	981	37,216
	D03	26	360	0	0	372	13,906	398	14,266
	D04	25	436	0	0	294	16,148	319	16,584
	D05	29	543	0	0	318	19,496	347	20,039
	D06	30	531	0	0	313	19,350	343	19,881
	D07	24	406	0	0	312	16,022	336	16,428
	D08	25	765	0	0	367	29,565	392	30,330
	D09	29	1,241	0	0	404	35,520	433	36,761
	D10	20	760	0	0	292	24,576	312	25,336
	D11	14	681	0	0	240	20,852	254	21,533
	D17	44	1,650	448	1,401	448	40,354	940	43,405
	D18	63	1,788	448	7,793	448	54,725	959	64,306
	D19	98	3,044	550	16,928	550	88,888	1,198	108,860
	D20	71	2,414	423	11,493	423	67,483	917	81,390
	D21	103	2,987	446	10,416	446	100,150	995	113,553
	D22	161	4,448	617	18,942	617	169,171	1,395	192,561
	D23	100	3,586	492	14,222	492	137,833	1,084	155,641
	D24	113	4,727	511	20,452	511	157,679	1,135	182,858
	D25	2	44	10	36	10	1,944	22	2,024
	D26	73	3,002	337	6,920	337	115,571	747	125,493
	D27	121	5,300	496	17,015	496	137,616	1,113	159,931
	D28	30	0	131	862	131	23,097	292	23,959
Cumhuriyet Senatosu (CS)		114	3,132	1,655	3,421	1,752	147,842	3,521	154,395
Danışma Meclisi (DM)		23	552	335	673	335	23,862	693	25,087
Gizli Celse (GC)		0	0	0	0	213	3,625	213	3,625
Kurucu Meclis (KM)		2	46	0	0	26	3,206	28	3,252
Millet Meclisi (MM)		159	6,371	2,374	18,097	2,515	167,035	5,048	191,503
Milli Birlik Komitesi (MBK)		6	222	0	0	103	3,626	109	3,848
Milli Güvenlik Konseyi (MGK)		11	372	189	220	189	16,808	389	17,400
TBMM Birleşik Toplantı (TBT)		19	370	337	593	358	8,221	714	9,184
Temsilciler Meclisi (TM)		7	170	0	0	110	4,712	117	4,882
Total		1,604	51,849	9,799	149,484	15,474	1,732,128	26,877	1,933,461

Table 5: Granular breakdown of the corpus by legislative term and document type.

of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Paris, France. European Language Resources Association (ELRA).

Türkiye Büyük Millet Meclisi. 2024. Tutanak dergisi PDFler – meclis dönemleri. Official source of parliamentary session transcripts (scanned PDFs) by legislative period. Accessed: 2026-02-23.



Münderecat

(a) Original scan

```
Münderecat
Sayfa
SayJa
1. - Sabık zabıt hulâsası
20
3. - Aydın ve istanbul mebusluklarına
2. - Havale edilen kâğıtlar
!20 seçilen Adnan Meideres'in, istanbul Me
3. - Tahlifler
20
busluğunu tercih ettiğine dair takriri
1. - Sivas Mebusu Ahmet Özel'in
(4/68)
21
tahlifi
20
4. - İstanbul ve Zonguldak mebus
4. - Riyaset Divanının Heyeti Umuluklarına seçilen Fuad Köprülü'nün, is
miyeye mâruzâtı
20
tanbul Mebusluğunu tercih ettiğine dair
```

(b) Tesseract OCR (Gungor et al., 2018)

```
Münde recat

| Sayfa | Sayfa |
|--------|--------|
| 1. - Sabık zabıt hulâsası | 20 |
| 2. - Havale edilen kâğıtlar | 20 |
| 3. - Tahlifler | 20 |
| 1. - Sivas Mebusu Ahmet Özel'in tahlifi | 20 |
| 4. - Riyaset Divanının Heyeti Umumiyeye mâruzat | 20 |
| 1. - Bursa ve İstanbul mebusluklarına seçilen Celâl Bayar'ın, İstanbul Mebusluğunu tercih ettiğine dair takriri (4/66) | 20 |
| 2. - İçel ve Kayseri mebusluklarına seçilen Refik Koraltan'ın, İçel Mebusluğunu tercih ettiğine dair takriri (4/67) | 20 |
| 1. - Enümeler intihabı | 35 |
| 2. - Arizalar ve telgraflar | 46 |

3. - Aydın ve İstanbul mebusluklarına seçilen Adnan Meideres'in, İstanbul Mebusluğunu tercih ettiğine dair takriri (4/68) | 21 |
4. - İstanbul ve Zonguldak mebusluklarına seçilen Fuad Köprülü'nün, İstanbul Mebusluğunu tercih ettiğine dair takriri (4/69) | 21 |
5. - İstanbul Mebusu Adnan Menderes'in kurduğu hükümetin programı | 21:34 |
6. - İntihaplar | 35 |
1. - Encümeler intihabı | 35:37,38:46 |
7. - Arizalar ve telgraflar | 46 |
```

(c) DeepSeek OCR pipeline

Figure 7: OCR quality comparison for a representative page (Term 10, Year 1, Session 3, 24 May 1954). Original scan (a); Tesseract-based output with recognition errors and fragmented line structure (b); DeepSeek OCR output with structured formatting (c).

Transcription and Recognition of Italian Parliamentary Speeches Using Vision-Language Models

Luigi Curini[1], **Alfio Ferrara**[2], **Giovanni Pagano**[1], **Sergio Picascia**[3*]
[1]Università degli Studi di Milano, Department of Social and Political Sciences
Via Conservatorio, 7 - 20122 Milano (Italy)
luigi.curini@unimi.it, giovanni.pagano@unimi.it

[2]Università degli Studi di Milano, Department of Literary Studies, Philology and Linguistics
Via Festa del Perdono, 7 - 20122 Milano (Italy)
alfio.ferrara@unimi.it

[3]Università degli Studi di Milano, Department of Computer Science
Via Celoria, 18 - 20133 Milano (Italy)
sergio.picascia@unimi.it

Abstract

Parliamentary proceedings represent a rich yet challenging resource for computational analysis, particularly when preserved only as scanned historical documents. Existing efforts to transcribe Italian parliamentary speeches have relied on traditional Optical Character Recognition pipelines, resulting in transcription errors and limited semantic annotation. In this paper, we propose a pipeline based on Vision-Language Models for the automatic transcription, semantic segmentation, and entity linking of Italian parliamentary speeches. The pipeline employs a specialised OCR model to extract text while preserving reading order, followed by a large-scale Vision-Language Model that performs transcription refinement, element classification, and speaker identification by jointly reasoning over visual layout and textual content. Extracted speakers are then linked to the Chamber of Deputies knowledge base through SPARQL queries and a multi-strategy fuzzy matching procedure. Evaluation against an established benchmark demonstrates substantial improvements both in transcription quality and speaker tagging.

Keywords: vision language models, document layout analysis, Italian parliamentary speeches

1. Introduction

Parliamentary proceedings constitute one of the most valuable documentary sources for the study of political, linguistic, and social change. In the Italian context, these records chronicle nearly two centuries of transformative events. The stenographic reports produced by both chambers of the Italian Parliament, the Camera dei Deputati and the Senato della Repubblica, provide a uniquely detailed account of these developments through the verbatim transcription of political discourse.

Several initiatives have sought to make these records available in machine-readable form. Cross-national projects such as ParlaMint (Erjavec et al., 2022) have assembled comparable parliamentary corpora across European countries, while Italy-specific efforts, including IPSA (Frasnelli and Palmero Aprosio, 2024) and ItaParlCorpus (Cova, 2025), have produced large-scale datasets spanning extensive historical periods. However, these resources predominantly rely on traditional Optical Character Recognition (OCR) pipelines followed by rule-based heuristics for text cleaning and speaker attribution. While effective to a degree, such approaches face well-documented limitations when applied to historical documents, contributing to transcription errors and unreliable speaker identification. The latter issue is especially acute for the earlier portion of the corpus (pre-1948), where high-quality annotated data remains scarce.

Speaker identification in Italian parliamentary documents is challenging because the typographic conventions used to mark speakers vary considerably across legislatures and historical periods. Additional variability arises from the treatment of homonymous members, for whom both surname and first name are provided, and from the occasional inclusion of the speaker's institutional role alongside the surname. Taken together, these inconsistencies make rule-based speaker attribution brittle and difficult to generalise across the full historical span of the corpus.

The recent emergence of Vision-Language Models (VLMs) offers a promising alternative to pipeline-based methods. VLMs jointly process visual and textual information through unified architectures, enabling end-to-end reasoning about document layout, content, and semantics. Specialised models such as `dots.ocr` (Li et al., 2025) have demonstrated strong performance on document layout analysis and text recognition, while general-purpose models like `Qwen2.5-VL` (Bai et al., 2025) provide complementary capabilities in semantic understanding and contextual inference. To date, however, the potential of these models for the digi-

Proceedings of ParlaCLARIN V Workshop on Interoperability, Multilinguality, and Multimodality in Parliamentary Corpora @ LREC 2026, pages 56–64

tisation and annotation of historical parliamentary documents remains largely unexplored.

In this paper, we present a pipeline for the automatic transcription, semantic segmentation, and entity linking of Italian parliamentary speeches based on Vision-Language Models. Unlike previous approaches, our method leverages the visual layout of documents alongside their textual content, enabling more accurate transcription and richer semantic annotation. We evaluate the proposed pipeline against IPSA on its released benchmark, assessing both OCR quality and speaker tagging accuracy.

The remainder of this paper is organised as follows. Section 2 reviews prior work on parliamentary corpora and vision-language models. Section 3 describes the proposed pipeline in detail. Section 4 presents the experimental evaluation and discusses the results. Finally, Section 5 summarises the contributions and outlines directions for future work.

2. Related Work

This section reviews prior work relevant to our contribution, organised into two main areas: parliamentary corpora with a focus on Italian resources, and vision-language models for document understanding and OCR.

2.1. Italian Parliamentary Resources

Parliamentary debates constitute a valuable resource for political science, linguistics, and computational social science research. The systematic collection and annotation of parliamentary proceedings has been pursued across numerous countries, resulting in large-scale corpora that enable studies of political discourse, policy preferences, and legislative behaviour. Cross-national initiatives have sought to harmonise parliamentary data across countries. The ParlaMint project (Erjavec et al., 2022) assembled comparable corpora from 29 European countries, containing over one billion words and covering at least the period 2015–2022, with linguistic annotations following the Universal Dependencies framework. Similarly, the ParlSpeech dataset (Rauh and Schwalbach, 2020) provides full-text corpora from various advanced democracies, though notably excluding Italy.

The digitisation and analysis of Italian parliamentary proceedings has received increasing attention in recent years. IPSA (Frasnelli and Palmero Aprosio, 2024) represents the most comprehensive effort to date, providing over 1.2 billion tokens of parliamentary debates from both the Camera dei Deputati and the Senato della Repubblica, spanning from 1848 to 2022. The corpus was constructed by applying Tesseract OCR to scanned documents, followed by rule-based heuristics for text cleaning and speaker tagging through fuzzy string matching against lists of parliamentarians. The ItaParlCorpus dataset (Cova, 2025) offers a machine-readable collection of Camera dei Deputati speeches from 1948 to 2022, encompassing 470 million words with metadata including speaker identification and party affiliation. ParlaMint-It (Alzetta et al., 2024) contributes a manually revised treebank of Italian parliamentary debates annotated according to the Universal Dependencies framework, addressing the underrepresentation of parliamentary language varieties in syntactic resources. The IMPAQTS corpus (Cominetti et al., 2024) takes a multimodal approach, collecting 2.65 million tokens of political discourse from 1946 to 2023 with pragmatic annotations capturing implicit content.

Despite these advances, existing Italian parliamentary corpora share common limitations: they predominantly rely on traditional OCR pipelines that struggle with historical document quality, employ rule-based approaches for speaker identification that cannot leverage visual layout cues, and lack fine-grained semantic annotations linking speakers to authoritative knowledge bases. Our work addresses these limitations by proposing a vision-language model pipeline that jointly performs transcription, semantic segmentation, and entity linking.

2.2. Vision-Language Models

Vision-Language Models (VLMs) are AI systems designed to jointly process visual and textual information (Zhang et al., 2024). These models typically employ a dual-encoder architecture, where separate encoders transform images and text into vector embeddings that are subsequently projected into a shared latent space. The aligned multimodal representations are then processed by a Transformer decoder, where visual embeddings serve as conditioning context for text generation. This architecture enables VLMs to perform a wide range of tasks requiring joint reasoning over images and text, including visual question answering, image captioning, and document understanding.

Optical Character Recognition has traditionally relied on pipeline approaches combining image preprocessing, layout analysis, character segmentation, and recognition (Islam et al., 2017). Tesseract (Smith, 2007) remains widely used due to its extensive language support and cost-effectiveness. However, traditional OCR systems face significant challenges with historical documents, including degraded print quality, non-standard typefaces, and complex multi-column layouts. The application of VLMs to document understanding and OCR represents a paradigm shift from pipeline-based approaches to end-to-end systems capable of jointly

reasoning about visual layout and textual content.

Recent benchmarks have evaluated VLMs on OCR tasks across different document types (Ouyang et al., 2025). Among end-to-end models, specialised OCR-focused VLMs such as `dots.ocr` (Li et al., 2025) have achieved strong results by combining layout analysis with text recognition in a unified framework. `dots.ocr` performs document layout analysis to identify element bounding boxes and categories, followed by text transcription respecting logical reading order. General-purpose VLMs have also demonstrated competitive OCR performance when appropriately prompted. `Qwen2.5-VL-72B` (Bai et al., 2025) achieves results comparable to specialised systems while offering additional capabilities for semantic understanding. The combination of these complementary capabilities, accurate transcription from specialised OCR models and semantic understanding from large VLMs, enables richer annotation of parliamentary documents than previously achievable with traditional approaches.

3. Methodology

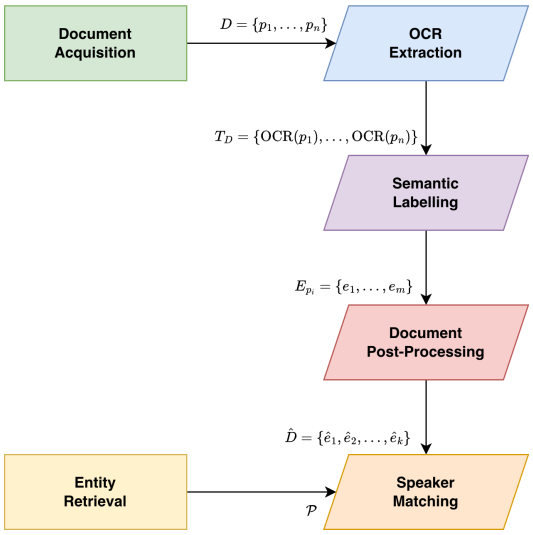


Figure 1: Pipeline diagram showing the six stages with data flow between components.

This section presents the methodology developed for the automatic transcription and semantic labelling of Italian parliamentary session reports. The proposed pipeline transforms digitised parliamentary documents into structured, semantically annotated data, enabling downstream analyses on political discourse studies. The pipeline comprises six sequential stages: (1) Document Acquisition, (2) OCR Extraction, (3) Semantic Labelling, (4) Document Post-Processing, (5) Entity Retrieval, and (6) Speaker Matching. Figure 1 provides a schematic overview of the complete processing pipeline.

To illustrate the transformations applied at each stage, we introduce a running example drawn from an actual parliamentary session. Figure 2 presents an excerpt from a document page[1] that will be traced through the pipeline, demonstrating how raw visual input is progressively transformed into structured data.

— 26 —

ATTI PARLAMENTARI — CAMERA DEI DEPUTATI — SESSIONE DEL 1874

ministrazione delle finanze, quanto in quella fatta dalla Corte dei conti, come cioè fra i risultati che io enunciai il 15 marzo 1874, quando presentai la situazione del Tesoro, vale a dire due mesi e mezzo soltanto dopo che l'esercizio 1873 era finito, fra questi risultati che io prevedeva ed annunciava alla Camera, e quelli definitivi del resoconto medesimo, vi passi una differenza minima. Dico una differenza minima, perchè, sopra un'entrata ed una spesa che cumulate insieme sorpassano i tre mila milioni, vi è un divario, tra quelle previsioni ed il resoconto definitivo approvato dalla Corte dei conti, di sole 163,000 lire. Da ciò vedrà la Camera, ed anche il paese, come l'andamento della nostra contabilità abbia fatto tale progresso che possiamo fare fermo assegnamento sopra i risultati della medesima. (*Benissimo!*)

Ho l'onore pure di ripresentare alla Camera gli stati di prima previsione dell'entrata e della spesa per l'anno 1875, che furono già presentati nella scorsa Sessione, all'epoca prescritta dalla legge. (V. *Stampati* dal n° 3 al 13 inclusivo.)

Presento similmente le note di variazioni, le quali furono da me comunicate alla Presidenza della Camera il 30 agosto e il 9 settembre scorsi, stantechè in quel tempo questa era prorogata. (V. *Stampato* n° 3 *bis*.)

In queste variazioni agli stati di prima previsione vi sono aggiunti tre quadri, secondo i desiderii che erano stati più volte espressi dalla Commissione del bilancio e dalla Camera stessa.

L'uno di essi contempla una nuova classificazione delle entrate e delle spese, specialmente compilata sulle traccie di quella che la Commissione del bilancio propose nelle ultime sue riunioni; l'altro contiene i ruoli organici delle amministrazioni dello Stato; finalmente il terzo riguarda le spese di liti sostenute nel triennio 1871-1872-1873 da tutte le amministrazioni dello Stato. Anche questo quadro fu richiesto nella Sessione passata ed io ho promesso di presentarlo.

Colla presentazione di questi documenti credo avere adempiuto agli obblighi da me assunti davanti alla Camera relativamente al bilancio di prima previsione.

PRESIDENTE. Si dà atto all'onorevole ministro per le finanze della presentazione di questi documenti, che saranno stampati e distribuiti.

MINISTRO PER LE FINANZE. Ho ancora l'onore di presentare alla Camera un piccolo disegno di legge che essa aveva già sotto gli occhi nella Sessione passata, e che riguarda il subreparto dell'imposta fondiaria nel compartimento modenese. (V. *Stampato* n° 14.)

Questo progetto non potè avere esito nella scorsa Sessione, perchè la Commissione che fu incaricata di esaminarlo non venne nominata che negli ultimi giorni della Sessione medesima.

PRESIDENTE. Do atto all'onorevole ministro per le finanze della presentazione di questo disegno di legge.

Prima di estrarre a sorte i nomi degli onorevoli deputati che dovranno comporre la Commissione incaricata di procedere allo scrutinio per la nomina della Commissione del bilancio, proporrei che, per dare tempo alla Giunta per le elezioni di preparare relazioni, il che farà, ne sono certo, poichè già diede prova di molta sollecitudine, si tenesse anche domani seduta alle tre.

In principio della medesima si potrà addivenire alla votazione per la nomina delle seguenti Commissioni: 1° delle petizioni; 2° della biblioteca; 3° per l'esame dei decreti e dei mandati registrati con riserva; 4° per la biblioteca della Camera. Si potranno pure eleggere la Commissione per l'esame dei conti consuntivi dell'amministrazione dello Stato, ed i commissari di vigilanza presso l'amministrazione della Cassa militare.

MINISTRO PER LE FINANZE. Domanderei che la Camera si occupasse prima di tutto di quella per l'esame dei resoconti amministrativi.

PRESIDENTE. Dunque sarebbero cinque le Commissioni a nominarsi. Fatte queste votazioni, si procederebbe alla verificazione delle elezioni sulle quali la Giunta avrà trasmessa la relazione. Se non vi è opposizione, s'intenderà fissata per le tre la seduta con quest'ordine del giorno.

L'onorevole Della Rocca ha presentato un disegno di legge che sarà trasmesso agli uffici.

NERVO. Chiedo di parlare per rivolgere una preghiera all'onorevole presidente.

I documenti presentati alla Camera dall'onorevole ministro per le finanze avranno, a mio avviso, una grande importanza per accertare la situazione finanziaria attuale nell'imminente discussione del bilancio del 1875; io desidererei che questi documenti così importanti fossero distribuiti alla Camera contemporaneamente al bilancio del 1875, tanto più che si tratta di consuntivi assai prossimi, le cui risultanze varranno molto per apprezzare la entità degli introiti ordinari su cui si può fare calcolo nel prossimo anno.

Quindi mi permetto di pregare l'onorevole presidente della Camera a voler disporre perchè sia accelerata la distribuzione dei documenti presentati dall'onorevole ministro.

PRESIDENTE. La Presidenza si farà dovere di tener conto della raccomandazione dell'onorevole Nervo.

Figure 2: Excerpt from the stenographic report of the session held on November 27th 1874, Legislature 12 of the Kingdom of Italy. This page excerpt serves as the running example throughout this section.

3.1. Document Acquisition

For each legislature in the history of the Italian Parliament, we retrieved the complete list of sessions (*sedute*). Each session, which can be thought of as a parliamentary meeting, is uniquely identified by a URI and associated with the specific date on which it took place. The majority of sessions are documented by a single PDF file containing the official verbatim report of the proceedings (*resoconto stenografico*). These reports tend to follow a consistent structure: they contain verbatim transcriptions of parliamentary debates, with speaker names indicated at the beginning of each intervention, typically accompanied by their institutional role

[1] https://storia.camera.it/regno/lavori/leg12/sed004.pdf#page=6

when applicable (e.g., *BIANCHI LEONARDO, MINISTER OF PUBLIC EDUCATION*).

For the session held on November 27th 1874, belonging to Legislature 12 of the Kingdom of Italy, the following metadata was retrieved from the parliamentary portal: legislature URI (http://dati.camera.it/ocd/legislatura.rdf/regno_12), session URI (http://dati.camera.it/ocd/seduta.rdf/sr12004), date (1874-11-27), and document URL (http://storia.camera.it/regno/lavori/leg12/sed004.pdf).

3.2. OCR Extraction

The digitised parliamentary reports present several challenges for automatic text extraction. The documents exhibit a two-column layout, variable print quality depending on the historical period, and occasional multilingual content. Preliminary experiments with traditional OCR engines, such as Tesseract, revealed significant limitations when facing these issues.

To address these challenges, we employed `dots.ocr` (Li et al., 2025), a specialised vision-language model designed for document understanding tasks. For each page p_i in a document $D = \langle p_1, p_2, \ldots, p_n \rangle$, the model receives the page image and a prompt requesting document layout analysis and textual transcription in reading order. The model identifies the bounding boxes of the elements in the page, labels these elements according to the corresponding category in the page layout, and transcribes their textual content respecting the logical reading sequence.

The output of this stage undergoes a further processing to extract only the textual content, discarding layout labels. This design decision was motivated by observed inconsistencies in the labelling conventions applied by the model across pages. The semantic classification of elements was therefore delegated to the subsequent processing stage, where a more capable model could leverage both visual and textual information for consistent labelling.

Formally, let $\mathrm{OCR}(p_i)$ denote the function mapping a page image to its preliminary textual transcription. The output of this stage for a document D is the sequence of page-level transcriptions $T_D = \langle \mathrm{OCR}(p_1), \mathrm{OCR}(p_2), \ldots, \mathrm{OCR}(p_n) \rangle$.

Processing the documents with a specialised model allows us to achieve high transcription quality and successfully preserves reading order, as demonstrated in the following example:

```
- 26 -

# ATTI PARLAMENTARI - CAMERA DEI DEPUTATI - SESSIONE
      DEL 1874.

ministrazione delle finanze, quanto in quella fatta
     dalla Corte dei conti, come cioe' fra i
     risultati che io enunciai il 15 marzo 1874,
     quando presentai la situazione del Tesoro, vale
      a dire due mesi e mezzo soltanto dopo che l'
     esercizio 1873 era finito, fra questi risultati
      che io prevedeva ed annunciava alla Camera, e
     quelli definitivi del resoconto medesimo, vi
     passi una differenza minima.
```

3.3. Semantic Labelling

This stage involves the employment of `Qwen-VL2.5-72B`, a large-scale vision-language model, to perform transcription refinement, textual segmentation, element type classification, and speaker identification. For each page p_i, the model receives three inputs: (i) the page image, (ii) a structured prompt that specifies the annotation schema and task requirements, and (iii) the preliminary OCR transcription $\mathrm{OCR}(p_i)$.

The prompt instructs the model to segment the page content into discrete elements, representing a segmented textual unit, each characterised by three attributes:

- *type*: a categorical label that indicates the role of the element on the page. The type is drawn from the set of labels: page-header, section-header, text, note, footnote, table. Elements labelled as text represent the main content of the report, i.e. the speeches, and are those that are typically assigned to speakers. The note type designates parenthetical remarks within the main text that record non-verbal events or reactions during the session (e.g., *Hilarity*, *The Chamber approves*);

- *content*: the verbatim transcription of the element, preserving original punctuation and orthography. Line-broken sentences within a single element are asked to be reconstructed as continuous text;

- *speaker*: the name of the person speaking, including their institutional role if explicitly mentioned. For elements that do not constitute speeches (headers, footnotes, notes, tables) or that report neutral content such as article text being read aloud, this field is set to "none". When a speech continues from a previous element on the same page without an explicit speaker indication, the model is instructed to infer the speaker from context. If inference is not possible (e.g., text continuing from a previous page), the field is marked as "unknown".

To guide the model's behaviour on challenging cases, the prompt includes a synthetic example, which aggregates multiple difficult scenarios: text continuation from previous pages, speaker role annotations, parenthetical notes interrupting speeches, footnote references, and section transitions. This one-shot example demonstrates the expected output format and handling of edge cases. The model is instructed to return its output in JSON

format, producing for each page a sequence of annotated elements, $E_{p_i} = \langle e_1, e_2, \ldots, e_m \rangle$, where each element $e_j = (\text{type}_j, \text{content}_j, \text{speaker}_j)$.

The following example shows the beginning of a structured output produced by the semantic labelling stage:


```
{
    "speaker": "none",
    "type": "page-header",
    "content": "- 26 -"
},
{
    "speaker": "none",
    "type": "page-header",
    "content": "ATTI PARLAMENTARI - CAMERA DEI
        DEPUTATI - SESSIONE DEL 1874."
},
{
    "speaker": "unknown",
    "type": "text",
    "content": "ministrazione delle finanze,
        quanto in quella fatta dalla Corte dei
        conti, come cioe' fra i risultati che io
         enunciai il 15 marzo 1874, quando
        presentai la situazione del Tesoro, vale
         a dire due mesi e mezzo soltanto dopo
        che l'esercizio 1873 era finito, fra
        questi risultati che io prevedeva ed
        annunciava alla Camera, e quelli
        definitivi del resoconto medesimo, vi
        passi una differenza minima."
}
```


In this example, the third element is a speech fragment whose speaker is marked as `"unknown"`. The text begins without any speaker heading, indicating that the speech started on a preceding page. Because the model processes each page independently, it has no access to the previous page's context and therefore cannot determine the identity of the speaker. Such cases are resolved in the subsequent phase.

3.4. Document Post-Processing

The outputs from the vision-language model undergo a post-processing phase to address artefacts arising from page-level processing and to resolve cross-page dependencies. This stage performs the following operations.

Hyphenation Resolution. Words hyphenated at line or column breaks are rejoined by detecting and removing mid-word hyphens followed by whitespace patterns indicative of line continuation.

Cross-Page Element Merging. Elements truncated at page boundaries are identified and merged with their continuations on subsequent pages. This is achieved by detecting incomplete sentences (lacking terminal punctuation) at page endings and matching them with elements marked as continuations (speaker "unknown" with type "text") at the beginning of subsequent pages.

Role Extraction. Institutional roles embedded within speaker names or element content are extracted and normalised. When a speaker is identified with an accompanying role, the role is parsed and separated by a comma from the speaker name.

Speaker Continuity Inference. For elements where the speaker could not be determined during page-level processing (marked as "unknown"), we implement a backward-looking inference mechanism. The algorithm traverses the document in reading order, propagating speaker attribution from the most recent explicitly identified speaker until a discontinuity marker is encountered. A discontinuity marker is represented by the appearance of a new section header, indicating a thematic break in proceedings. This ensures that speeches spanning multiple pages are correctly attributed to their speakers while respecting the logical structure of parliamentary proceedings.

The output of this stage is a document-level sequence of processed elements: $\hat{D} = \langle \hat{e}_1, \hat{e}_2, \ldots, \hat{e}_k \rangle$.

Figure 3 illustrates the effect of post-processing on the running example. The speech fragment on page 26, originally marked with an `"unknown"` speaker, is merged with the incomplete element at the end of page 25: the hyphenated word is rejoined, the content is concatenated, and the speaker identity is propagated from the preceding context.

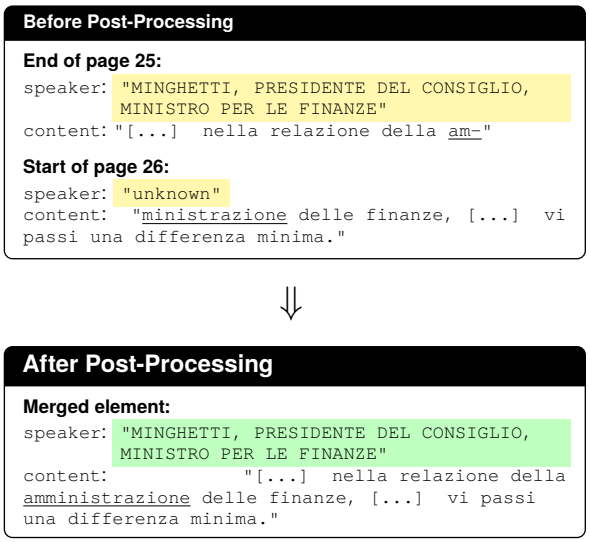


Figure 3: Effect of post-processing on a cross-page speech fragment. Highlighted regions show speaker inference (unknown → resolved); underlined text shows hyphenation resolution.

3.5. Entity Retrieval

To enable linking of extracted speakers to authoritative records, we retrieve from the Italian Chamber of Deputies knowledge base all individuals who held parliamentary positions on the date of each session. The knowledge base, accessible via a SPARQL

endpoint[2], contains structured information about deputies, government members, parliamentary officers, and members of institutional organs. For a session occurring on date d, we execute a series of SPARQL queries to retrieve the set of active entities $\mathcal{P}_d = \mathcal{P}_d^{\text{dep}} \cup \mathcal{P}_d^{\text{gov}} \cup \mathcal{P}_d^{\text{org}} \cup \mathcal{P}_d^{\text{off}}$, where $\mathcal{P}_d^{\text{dep}}$ denotes deputies, $\mathcal{P}_d^{\text{gov}}$ government members, $\mathcal{P}_d^{\text{org}}$ members of parliamentary organs, and $\mathcal{P}_d^{\text{off}}$ parliamentary officers active on date d.

For each entity $p \in \mathcal{P}_d$, we retrieve: the unique URI serving as the canonical identifier, the full name, surname, given name, and the set of institutional roles held on the relevant date. This knowledge base is linked to Wikidata, enabling subsequent enrichment with additional biographical and political information such as party affiliation and demographic attributes.

On November 27th 1874, we have 557 active parliamentarians. For each entity, we retrieve the URI, name, and roles:

```
{
    "uri": "http://dati.camera.it/ocd/persona.rdf/
         pr3028",
    "fullname": "MARCO MINGHETTI",
    "name": "MARCO",
    "surname": "MINGHETTI",
    "dep": true,
    "gov": [
        "MINISTRO: MINISTERO DELLE FINANZE",
        "PRESIDENTE: PRESIDENZA DEL CONSIGLIO"
    ],
    "org": [],
    "off": []
}
```

3.6. Speaker Matching

The final stage establishes correspondences between extracted speaker names and knowledge base entities. Given the historical nature of the corpus and the variety of conventions used to identify speakers in parliamentary proceedings, a simple string matching approach proves insufficient. Speaker names may appear as surnames only, as full names, with abbreviated given names, or solely by institutional role (e.g., *PRESIDENT*, *MINISTER OF THE INTERIOR*). Furthermore, multiple individuals may share the same surname within a single legislature. We therefore implement a multi-strategy matching pipeline that applies increasingly sophisticated matching criteria. The pipeline processes each unique speaker name extracted from a document and attempts to link it to entities in $\mathcal{P}_d$.

Generic Speaker Filtering. Certain speaker designations refer to collective or anonymous voices (e.g., *VOICES*, *A DEPUTY*) and are excluded from entity linking by pattern matching.

Role-Based Identification. When a speaker is identified solely by institutional role, the system queries the entity set for individuals holding that role on the session date. A special case handles the session president (*PRESIDENT*): the document's first page typically contains a header followed by the presiding officer's name, which is extracted and used to disambiguate among members of the *Presiding Committee*.

Name Matching with Fuzzy Strategies. For speakers identified by name, the system applies a cascade of fuzzy matching algorithms with decreasing stringency. The matching process employs multiple similarity metrics, including token-based ratios, partial matching, and token set comparisons, applied iteratively to both surnames and full names. Candidate entities are those exceeding a configurable similarity threshold.

Disambiguation. When multiple candidate entities remain after initial matching, a disambiguation cascade is applied:

1. score-based ranking: candidates are ranked by their fuzzy matching scores; if a single candidate achieves the highest score, it is selected;
2. role matching: if the speaker's role was extracted, candidates whose roles match the extracted role are prioritised;
3. full name similarity: remaining ties are broken by computing similarity between the candidate's full name and the extracted speaker name;
4. abbreviated name handling: for speaker names containing initials (e.g., "G. ROSSI"), the system generates abbreviated forms of candidate names and compares them;
5. contextual mention: candidates whose full names appear elsewhere in the document text are favoured;
6. weighted edit distance: a weighted Levenshtein distance is computed, assigning lower substitution costs to vowel-vowel substitutions to account for spelling variations common in historical documents.

If disambiguation succeeds, the speaker is linked to the entity's URI; otherwise, all high-scoring candidates are retained as potential matches. In a second pass, unresolved speakers are compared against successfully resolved speakers from the same document. If an unresolved speaker name exhibits high similarity to a resolved one, the linking from the resolved speaker is propagated.

The output of the complete pipeline is a JSON file for each session document, containing the sequence of annotated elements with speaker entity URIs where linking succeeded. This structured representation enables direct integration with the

[2] https://dati.camera.it/sparql

parliamentary knowledge base and, through its linkage to Wikidata, facilitates enrichment with political party affiliations, biographical data, and cross-references to external resources for comprehensive political discourse analysis.


```
{
    "speaker": "MINGHETTI, PRESIDENTE DEL
        CONSIGLIO, MINISTRO PER LE FINANZE",
    "type": "text",
    "content": "In questa occasione credo che la
         Camera sara' contenta di sentire quello
         che gia' vedra' distintamente tanto
        nella relazione della amministrazione
        delle finanze, quanto in quella fatta
        dalla Corte dei conti, come cioe' fra i
        risultati che io enunciai il 15 marzo
        1874, quando presentai la situazione del
         Tesoro, vale a dire due mesi e mezzo
        soltanto dopo che l'esercizio 1873 era
        finito, fra questi risultati che io
        prevedeva ed annunciava alla Camera, e
        quelli definitivi del resoconto medesimo
        , vi passi una differenza minima.",
    "speaker_uri": "http://dati.camera.it/ocd/
        persona.rdf/pr3028",
    "wikidata_uri": "Q597155"
}
```


4. Evaluation

To assess the effectiveness of the proposed pipeline, we conduct a comparative evaluation against IPSA (Frasnelli and Palmero Aprosio, 2024), a previously published system for Italian parliamentary corpus construction. We evaluate both the OCR transcription quality and the speaker tagging accuracy using the benchmark dataset released by the authors.

4.1. Evaluation Setup

Benchmark Dataset. The evaluation relies on the benchmark dataset released alongside IPSA. The dataset consists of 60 scanned parliamentary pages paired with manual transcriptions, which serve as the ground truth for OCR evaluation. These pages span the period from 1848 to 1996 and cover documents from both the Camera dei Deputati and the Senato della Repubblica across multiple legislatures. For the speaker tagging task, annotations are available for 58 of these 60 pages. We retrieved the page images and reference annotations from the authors' repository and processed the same page images through our pipeline.

Metrics. We adopt Word Error Rate (WER) and Character Error Rate (CER) for OCR quality assessment. Both metrics quantify the edit distance between the predicted transcription and the ground truth:

$$\text{WER} = \frac{S + D + I}{N}, \quad \text{CER} = \frac{S + D + I}{N} \qquad (1)$$

where S denotes the number of substitutions, D the number of deletions, I the number of insertions, and N the total number of words (for WER) or characters (for CER) in the ground truth. Lower values indicate higher transcription quality.

For the tagging task, we report Precision, Recall, and F1 score. A true positive corresponds to a correctly identified speaker-entity link, a false positive to an incorrect identification, and a false negative to a missed identification.

Experimental Configuration. We employed `dots.ocr`[3] for the initial OCR processing, and then `Qwen-VL2.5-72B`[4] with 8-bit quantisation to enable semantic inference. The experiments were run on a Nvidia H100 NVL 94 GB graphics card.

4.2. OCR Evaluation

Table 1 presents the OCR evaluation results on the 60 benchmark pages, comparing the Tesseract baseline used in IPSA against the two stages of our pipeline: (i) `dots.ocr` alone, and (ii) `dots.ocr` followed by `Qwen-VL` transcription refinement.

Method	CER	WER
Tesseract (IPSA)	0.030	0.071
dots.ocr	0.031	0.050
dots.ocr + Qwen-VL	**0.009**	**0.024**

Table 1: OCR evaluation results on the 60-page benchmark. Lower values indicate better performance.

The results indicate that when applied in isolation, `dots.ocr` achieves a comparable CER to Tesseract (0.031 vs. 0.030) while yielding a notable improvement at the word level, reducing WER by approximately 30% (from 0.071 to 0.050). The integration of `Qwen-VL` for transcription refinement delivers substantial gains: the complete pipeline achieves an error reductions of approximately 70% in CER and 66% in WER compared to the Tesseract baseline.

4.3. Tagging Evaluation

Table 2 presents the speaker tagging evaluation results, comparing the rule-based approach of IPSA against our pipeline.

The released benchmark provides only page images without the relative session metadata. Since our pipeline requires this information, in particular the session date for Senate document, in order to retrieve the set of active parliamentarians, we restrict our evaluation to the 29 *Chamber of Deputies*

[3] https://github.com/rednote-hilab/dots.ocr

[4] https://huggingface.co/RedHatAI/Qwen2.5-VL-72B-Instruct-FP8-dynamic

pages for which the date could be reliably extracted from the filename.

Method	Precision	Recall	F1
IPSA (Global)	**0.939**	0.880	0.909
Ours (Global)	0.885	**0.970**	**0.925**
IPSA (Pre-WW2)	**0.953**	0.850	0.898
Ours (Pre-WW2)	0.883	**0.970**	**0.924**
IPSA (Post-WW2)	**0.916**	0.942	0.929
Ours (Post-WW2)	0.892	**0.971**	**0.930**

Table 2: Speaker tagging evaluation results.

Our pipeline achieves a higher global F1 score, driven by substantially higher Recall, while the IPSA baseline retains an advantage in Precision. The lower precision of our method is partly explained by a systematic difference in segmentation granularity: when a speech is interrupted by a parenthetical note (e.g., *The Chamber approves*), our pipeline splits the surrounding text into two distinct speech elements, whereas the ground truth treats it as a single continuous speech. This produces a higher number of predicted speech segments and, consequently, additional false positives.

A notable finding is that our pipeline exhibits no significant performance gap between the Pre-WW2 and Post-WW2 subsets, whereas the IPSA baseline shows a marked sensitivity to document quality across historical periods. This degradation is consistent with the lower typographic quality of Pre-WW2 documents, which exhibit greater variability in print clarity, adversely affecting both OCR accuracy and the reliability of the downstream tagging task. The stability in performance of our VLM-based approach suggests that the superior transcription quality does not merely benefit the exploitation of the text itself, but also carries over to subsequent tasks, keeping them uneffected by the degradation inherent in older source documents.

4.4. Limitations

Several factors limit the direct comparability of tagging performance between our pipeline and the IPSA baseline and may partially account for observed differences. As noted above, the absence of session date metadata in *Senate* document filenames restricted the tagging evaluation to the *Chamber of Deputies* subset. For these pages, the date was extracted from the filename and used to query the set of active parliamentarians. In a full-corpus processing scenario, session dates would be readily available from the document metadata, removing this constraint entirely.

Furthermore, the fragmented nature of the benchmark, consisting of isolated pages rather than complete documents, prevents our pipeline from demonstrating its full capabilities in cross-page speaker inference. In a standard workflow, a speech that begins on a previous page and continues onto the current one would be attributed to the correct speaker through the Speaker Continuity Inference mechanism. In the benchmark setting, however, no previous page is available, and the ground truth accordingly leaves the first speech unlabelled when no explicit speaker heading appears on the page. Our system is therefore evaluated without exercising one of its core design strengths.

A related limitation concerns the identification of the presiding officer. In a complete document, the President of the session is explicitly named on the first page, allowing our system to propagate this identity throughout subsequent pages. Because benchmark pages are processed in isolation without access to this introductory context, our method must approximate the presiding officer based solely on the legislature, a heuristic that is prone to error when mid-term changes in presidency occur or when a Vice President substitutes for the main President.

Finally, while our approach requires significantly higher computational resources, necessitating GPUs for Vision-Language Model inference, we argue that this cost is justified by the superior quality of the final output, which is crucial for enabling accurate downstream NLP tasks such as topic modelling or sentiment analysis.

5. Conclusion

In this paper, we presented a pipeline for automatic transcription, semantic segmentation, and entity linking of Italian parliamentary speeches based on Vision-Language Models. The proposed approach combines a specialised OCR model (`dots.ocr`) with a large-scale VLM (`Qwen2.5-VL-72B`) to jointly perform text extraction, element classification, and speaker identification, followed by entity retrieval from the Chamber of Deputies knowledge base and a multi-strategy fuzzy matching procedure for linking speakers to records.

Evaluation against the IPSA benchmark demonstrated that the VLM-based pipeline achieves substantial improvements in transcription quality, with relative error reductions of approximately 70% in Character Error Rate and 66% in Word Error Rate compared to the Tesseract baseline. The results of the speaker tagging showed competitive performance, with our method achieving higher F1 scores than the baseline across all temporal subsets. Notably, the pipeline exhibited consistent performance in both pre- and post-WW2 documents, suggesting robustness to the considerable variation in source document quality across historical peri-

ods. The evaluation on isolated benchmark pages represents a conservative estimate of performance, as the pipeline's cross-page inference mechanisms could not be fully exploited in this setting.

We are currently preparing the public release of a dataset spanning from 1848 to 1948 extracted with the proposed approach. The released data will include OCR transcriptions, structural annotations, and speaker-entity links in a structured, machine-readable format. Processing complete documents rather than isolated pages improves speaker tagging accuracy, as the system can fully exploit cross-page inference mechanisms such as session president identification and speaker continuity propagation. The link to external knowledge bases, such as Wikidata, opens the possibility of enriching the corpus with structured metadata, including party affiliations and demographic attributes, enabling richer downstream analyses in political science and computational social science.

Acknowledgements

Computational resources provided by INDACO Core facility, which is a project of High Performance Computing at the University of MILAN http://www.unimi.it

Bibliographical References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report.

Noman Islam, Zeeshan Islam, and Nazia Noor. 2017. A survey on optical character recognition system. *arXiv preprint arXiv:1710.05703*.

Yumeng Li, Guang Yang, Hao Liu, Bowen Wang, and Colin Zhang. 2025. dots.ocr: Multilingual document layout parsing in a single vision-language model.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. 2025. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24838–24848.

R. Smith. 2007. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644.

Language Resource References

Alzetta, Chiara and Montemagni, Simonetta and Sartor, Marta and Venturi, Giulia. 2024. *Parlamint-it: an 18-karat UD treebank of Italian parliamentary speeches*. Springer Science and Business Media LLC.

Cominetti, Federica and Gregori, Lorenzo and Lombardi Vallauri, Edoardo and Panunzi, Alessandro. 2024. *IMPAQTS: a multimodal corpus of parliamentary and other political speeches in Italy (1946-2023), annotated with implicit strategies*. ELRA and ICCL.

Cova, Joshua. 2025. *A new database for Italian parliamentary speeches: introducing the ItaParlCorpus dataset*.

Erjavec, Tomaž and Ogrodniczuk, Maciej and Osenova, Petya and Ljubešić, Nikola and Simov, Kiril and Pančur, Andrej and Rudolf, Michał and Kopp, Matyáš and Barkarson, Starkaður and Steingrímsson, Steinþór and undefinedöltekin, undefinedağrı and de Does, Jesse and Depuydt, Katrien and Agnoloni, Tommaso and Venturi, Giulia and Pérez, María Calzada and de Macedo, Luciana D. and Navarretta, Costanza and Luxardo, Giancarlo and Coole, Matthew and Rayson, Paul and Morkevičius, Vaidas and Krilavičius, Tomas and Darǵis, Roberts and Ring, Orsolya and van Heusden, Ruben and Marx, Maarten and Fišer, Darja. 2022. *The ParlaMint corpora of parliamentary proceedings*. Springer Science and Business Media LLC.

Frasnelli, Valentino and Palmero Aprosio, Alessio. 2024. *There's Something New about the Italian Parliament: The IPSA Corpus*. ELRA and ICCL.

Rauh, Christian and Schwalbach, Jan. 2020. *The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies*. Harvard Dataverse.

Beyond OCR: Structural Segmentation and Speaker Attribution in Historical Italian Parliamentary Debates

Claudia Corbetta, Samuele Mazzei, Alessio Palmero Aprosio
University of Trento, Corso Bettini 84, Rovereto (Italy)
claudia.corbetta@unitn.it, samuele.mazzei@studenti.unitn.it, a.palmeroaprosio@unitn.it

Abstract

Historical parliamentary debates are essential for longitudinal political and linguistic research, yet much early material remains available only as scanned images. In the Italian context, proceedings from 1848–1996 lack large-scale, structurally annotated, machine-readable representations. This paper addresses the challenge of transforming historical Italian parliamentary debates into structured corpora by moving beyond plain Optical Character Recognition (OCR) toward functional block segmentation and speaker attribution. We present detailed annotation guidelines and a manually annotated dataset of 300 randomly sampled pages. Two approaches are compared: (i) direct multimodal Large Language Model (LLM) annotation and (ii) a modular pipeline combining OCR with LLM-based structural reconstruction under zero-shot and few-shot prompting. Evaluation on a held-out test set shows that separating transcription from structural reasoning improves performance, with few-shot prompting yielding the most reliable results. The study demonstrates the feasibility of integrating LLM-based reasoning into historical parliamentary digitisation workflows.

Keywords: Historical Parliamentary Corpora, Optical Character Recognition (OCR), Speaker Attribution

1. Introduction

Parliamentary debates represent the most comprehensive and continuous record of a nation's political, social, and linguistic evolution, capturing the direct confrontation between ideologies and the formal legislative process as it unfolds.

The study of parliamentary proceedings has been revolutionized by the emergence of large-scale computational text analysis. Systematic infrastructures like the ParlaMint project (Erjavec et al., 2023, 2024) have paved the way for comparative European legislative research by providing harmonized, multilingual datasets. However, there is a significant disparity in the availability of these resources.

Most existing studies rely on recent transcriptions, as modern data is collected digitally and already annotated semantically, making it easily accessible for researchers. In contrast, research on older historical data remains largely "vertical", limited to small time intervals or specific subsets of debates, primarily due to the lack of structured, machine-readable data.

In the Italian context, although recent efforts have produced machine-actionable corpora for contemporary parliamentary data, most historical proceedings from the Kingdom and early Republic remain available only as scanned images. Exploratory initiatives such as IPSA (Frasnelli and Palmero Aprosio, 2024) and more recent structured resources like ItaParlCorpus (Cova, 2025) address parts of this landscape, yet large-scale, structurally annotated representations of long-span historical debates are still missing.

Although established OCR and layout analysis pipelines exist (Breuel, 2008; Wick et al., 2018), they do not integrate higher-level reasoning mechanisms capable of jointly addressing block segmentation and speaker attribution in complex historical parliamentary layouts.

Accurate block segmentation and speaker attribution are essential for downstream analyses of political discourse and representation. Longitudinal studies require distinguishing between communicative units, separating oral interventions from legislative articles, procedural notes, and other structural elements.

This paper addresses the challenges posed by Italian historical parliamentary documents through a multi-step framework centered on the following research question: which computational approach most effectively integrates accurate OCR with reliable structural reconstruction of document layout and speaker segmentation in long-span historical data? To answer this question, we first introduce comprehensive guidelines for the annotation of textual blocks in Italian parliamentary debates, designed to support downstream tasks such as discourse analysis and speaker tracking. These guidelines make explicit the structural complexity and information density of the documents, highlighting how multiple textual layers (discursive, procedural, and editorial) coexist within the same pages. We then present a dataset of 300 pages, randomly sampled from the data and manually annotated accord-

Proceedings of ParlaCLARIN V Workshop on Interoperability, Multilinguality, and Multimodality in Parliamentary Corpora @ LREC 2026, pages 65–76

ing to this framework. Building on this resource, we survey the main methodological approaches to the task, encompassing both optical character recognition and block/speaker identification strategies. Finally, we conduct a systematic comparative evaluation, measuring the extent to which different systems succeed not only in accurately transcribing the textual content, but also in reconstructing the structural organization of the documents. This evaluation ultimately allows us to identify the approach that achieves the best balance between textual fidelity and structural comprehension across nearly 150 years of Italian parliamentary proceedings.

2. Related Work

Parliamentary transcripts constitute a fundamental resource for longitudinal research in political science and linguistics, enabling the analysis of ideological trends, agenda-setting, political representation, and the evolution of political discourse over time, as illustrated, for example, by studies focusing on the Italian parliamentary context (Curini et al., 2024; Cominetti et al., 2022). The large-scale digitisation of parliamentary archives across Europe has led to the creation of computational corpora such as Hansard (Wattam et al., 2014; Nanni et al., 2019; Coole et al., 2020), GERPARCOR (Abrami et al., 2022, 2024), and GePaDe (Rehbein et al., 2024), among others, which integrate OCR, structural reconstruction, metadata enrichment, and linguistic annotation to support systematic analysis of legislative debates. Similarly, infrastructures such as Parla-CLARIN[1] and workflows like OCR4all[2] demonstrate the feasibility of transforming historical parliamentary scans into structured, machine-readable corpora through iterative OCR and annotation pipelines (Kavčič et al., 2024). These initiatives highlight the importance of integrating OCR with downstream linguistic processing and structural annotation to enable large-scale computational analysis.

However, the **digitisation of historical parliamentary** records presents significant technical challenges due to the limitations of OCR when applied to degraded materials, non-standard typography, and complex layouts typical of historical printings, which increase recognition errors and complicate text reconstruction (Reul et al., 2019; Greif et al., 2025). In particular, Document Layout Analysis (DLA) remains a critical bottleneck, as errors in segmenting multi-column formats, speaker markers, and procedural elements can disrupt reading order and affect downstream tasks such as speaker attribution and discourse segmentation. To mitigate these issues, recent workflows combine layout segmentation, deep learning–based OCR models, and iterative training with manual correction to improve accuracy and robustness.

Within this context, **annotation workflows** are essential for transforming OCR-derived text into reliable research corpora, as OCR output often contains structural inconsistencies and recognition errors that must be resolved during annotation. For instance, the IsraParlTweet corpus required iterative preprocessing combining rule-based extraction and human validation to accurately identify speakers and segment debates (Mor-Lan et al., 2024), highlighting that annotation is an integral part of the digitisation process rather than a separate downstream task. Clear annotation guidelines and structured schemes are crucial to ensure consistency and resolve ambiguities caused by OCR noise (Reinig et al., 2024), while multilayer workflows incorporating double annotation and expert validation, such as those used in the CitiLink-Minutes dataset, further improve annotation reliability and correct digitisation-induced inconsistencies (Guimarães et al., 2025).

3. Annotation Guidelines

As stated in Section 2, the annotation process represents a critical step in transforming OCR-derived parliamentary transcripts into reliable research corpora, as raw OCR output often contains structural ambiguities, segmentation errors, and incomplete speaker attribution.

The annotation of historical parliamentary debates was therefore carried out following a structured set of guidelines[3] designed to segment transcripts into coherent functional units and assign consistent metadata to each segment. By systematically identifying and labeling functional textual blocks, annotation enables the reconstruction of the logical and communicative structure of parliamentary proceedings, ensuring that speeches, legislative content, and procedural elements can be accurately distinguished and analyzed.

While these guidelines provide a robust framework for the Italian parliamentary context, they have been specifically designed to accommodate the idiosyncratic layout conventions, document structure, and procedural norms of the Italian Parliament. Consequently, although the underlying principles of functional segmentation and metadata annotation are broadly applicable, the schema may require adaptation or extension to address the diplomatic, legal, and linguistic conventions of parliamentary debates in other national contexts.

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

[2] https://github.com/ocr4all

[3] All the material related to this article is available on the main Github of the IPSA project: https://github.com/dhfbk/ipsa

3.1. Annotation workflow

Annotation was performed at the level of textual blocks, defined as continuous stretches of text with a single communicative function, such as a speech turn, legislative article, procedural description, or quoted document.

Each block was assigned a unique instance identifier within the page and annotated with a type label, speaker information, and optional notes to clarify ambiguous cases. Specifically, for each instance, a structured set of annotation fields was defined, including:

- **Page ID**: the corresponding PDF page number;
- **Instance/Round ID**: a progressive integer starting at 1 for each new page;
- **Type Label**: a functional category identifying the communicative nature of the block (refer to Table 1);
- **Speaker**: the individual or institutional role responsible for the speech or text (refer to 3.2);
- **Notes**: optional clarifications, particularly to specify the nature of blocks labeled as "other".

Moreover, the annotation scheme relied on a predefined taxonomy of block types (i.e., type label), including speech, article, text, description, title, and other, to capture the structural and functional diversity of parliamentary records. Table 1 provides a detailed description of the communicative function associated with each label, clarifying the criteria used to distinguish between oral interventions, legislative content, procedural annotations, and other structural elements of the parliamentary proceedings.

Segmentation of block type is based on functional and graphical criteria. A new unit is annotated whenever the text changes its communicative function and this shift is clearly signaled by layout features such as line breaks, indentation, or other formatting cues. Typical cases include the beginning of a new speaker's turn, the introduction of a legislative article, the presence of a procedural note, the appearance of a title, or the insertion of attachments or tables. Consecutive non-speech elements are annotated as separate units whenever they represent distinct functional blocks.[4] However, when multiple components form a single integrated unit, they are annotated as one segment. The proposed annotation schema is partially aligned with existing standards for parliamentary corpora, such as the ParlaMint schema.[5] In particular, core elements such as speaker attribution and speech segmentation can be directly mapped to ParlaMint components (e.g., <u> elements with speaker metadata). However, our annotation framework introduces a finer-grained distinction of functional block types (e.g., description, speechnotext, presidencydeclaration). These categories are motivated by the need to capture the structural and procedural complexity of historical parliamentary documents, where non-speech elements (e.g., procedural notes, legislative articles, titles, and editorial insertions) play a central role in structuring the document. More generally, this design choice reflects a different modelling perspective: while standard parliamentary encoding schemes such as ParlaMint are primarily designed for TEI-compliant textual representation, our framework aims to support structuring and linking of information in a Linked Open Data (LOD) setting. As a result, certain distinctions are made explicit at the annotation level in order to facilitate downstream semantic integration.

3.2. Speaker Attribution

Given the parliamentary nature of the texts, accurate identification of the speaker is a fundamental requirement for enabling reliable analysis of political discourse, speaker behavior, and institutional dynamics. For this reason, particular attention was devoted to the annotation of speaker attribution.

Speaker attribution was designed to make speaker identification as explicit and interpretable as possible. The annotation captures speaker mentions at the textual level, without enforcing cross-document identity resolution, which is deferred to a later linking stage (e.g., via LOD). To this end, a set of labels (explicit, inferred, hidden, unknown) was introduced to represent different degrees of speaker identifiability. More specifically, the identification of the **speaker**, recorded in the *speaker* column, includes both the personal name and any associated institutional role when explicitly provided in the source text:

"Baccarini, ministro dei lavori pubblici";

"'Rosadi".

In cases of generic or collective interjections where no individual speaker can be identified, the prefix `[unknown]` is used:

"`[unknown]` Una voce".

When the speaker is not explicitly named in the current segment but can be reliably inferred from

[4] Similarly, consecutive elements classified as "other" (e.g., attachments, tables, or related materials) are annotated separately if they are visually and structurally distinct.

[5] See https://clarin-eric.github.io/ParlaMint/.

the context (for instance, as a continuation from a previous speech), the prefix `[inferred]` is applied:

"[inferred] Presidente".

Conversely, if the speaker cannot be identified, as in cases where the speech is ongoing and the speaker was mentioned on a previous page, the label `[hidden]` is assigned:

"[hidden]".

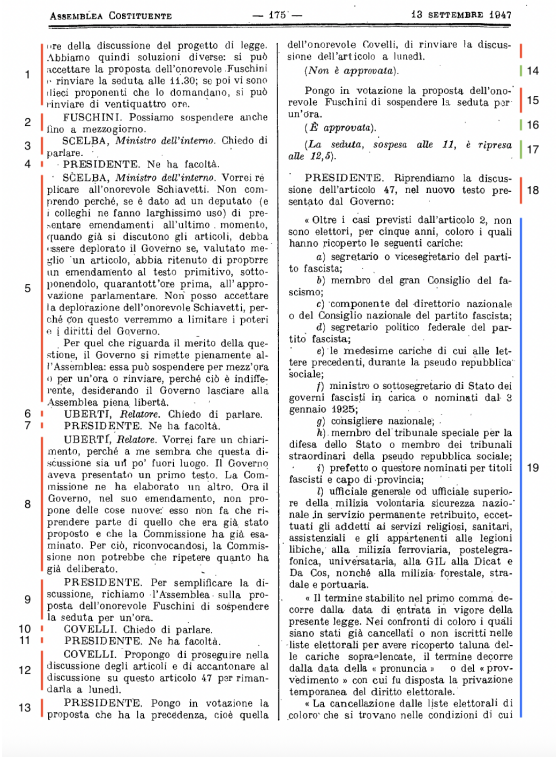
4. Annotated Dataset

Given the complexity of the annotated documents, which necessarily require detailed and articulated guidelines, an initial inter-annotator agreement phase was conducted on a randomly selected sample of 20 pages prior to proceeding with the annotation of the gold-standard data. Inter-annotator agreement (Table 2) was measured using Krippendorff's α for nominal data. Speaker attribution yielded substantial agreement ($\alpha = 0.76$), whereas agreement on type labels was considerably lower ($\alpha = 0.44$).

The discrepancy between the two scores reflects the different nature of the tasks. Speaker attribution is largely referential, whereas type label assignment requires interpretative distinctions between structurally adjacent categories (e.g., speech vs. description). The analysis of disagreements led to a refinement of the annotation guidelines before proceeding to the full gold-standard annotation, thereby improving overall consistency.

Following the agreement phase and the subsequent reconciliation of annotation discrepancies, we carried out a manual annotation on a randomly selected subset of the corpus comprising 300 pages, which were independently annotated by three annotators. Agreement was not recomputed on the full dataset, as the annotation process adopted an adjudication-based workflow aimed at producing a consistent gold-standard resource. More specifically, all disagreements concerning either type labels or speaker attribution were systematically identified and resolved through a joint adjudication process. Annotators collectively reviewed conflicting cases with reference to the annotation guidelines, refining their interpretation when necessary. This iterative reconciliation ensured consistency across the dataset and resulted in a fully harmonized gold standard.

Table 3 presents an example of the manual annotation of page 274 from the annotated subset, while Figure 1 provides a visual representation of the same page with color-coded and numbered boxes corresponding to the annotated segments (red for speech, green for description, and blue for text). Each box identifies a distinct functional block and is associated with its instance identifier, shown alongside the segment.


ASSEMBLEA COSTITUENTE — 175 — 13 SETTEMBRE 1947

re della discussione del progetto di legge. Abbiamo quindi soluzioni diverse: si può accettare la proposta dell'onorevole Fuschini o rinviare la seduta alle 11.30; se poi vi sono oltri proponenti che lo domandano, si può rinviare di ventiquattro ore.

FUSCHINI. Possiamo sospendere anche fino a mezzogiorno.

SCELBA, *Ministro dell'interno*. Chiedo di parlare.

PRESIDENTE. Ne ha facoltà.

SCELBA, *Ministro dell'interno*. Vorrei replicare all'onorevole Schiavetti. Non comprendo perché, se è dato ad un deputato (e i colleghi ne fanno larghissimo uso) di presentare emendamenti all'ultimo momento, quando già si discutono gli articoli, debba essere deplorato il Governo se, valutato meglio un articolo, abbia ritenuto di proporre un emendamento al testo primitivo, sottoponendolo, quarantott'ore prima, all'approvazione parlamentare. Non posso accettare la deplorazione dell'onorevole Schiavetti, perché con questo verremmo a limitare i poteri e i diritti del Governo.

Per quel che riguarda il merito della questione, il Governo si rimette pienamente all'Assemblea: essa può sospendere per mezz'ora o per un'ora o rinviare, perché ciò è indifferente, desiderando il Governo lasciare alla Assemblea piena libertà.

UBERTI, *Relatore*. Chiedo di parlare.

PRESIDENTE. Ne ha facoltà.

UBERTI, *Relatore*. Vorrei fare un chiarimento, perché a me sembra che questa discussione sia un po' fuori luogo. Il Governo aveva presentato un primo testo. La Commissione ne ha elaborato un altro. Ora il Governo, nel suo emendamento, non propone delle cose nuove: esso non fa che riprendere parte di quello che era già stato proposto e che la Commissione ha già esaminato. Per ciò, riconvocandosi, la Commissione non potrebbe che ripetere quanto ha già deliberato.

PRESIDENTE. Per semplificare la discussione, richiamo l'Assemblea sulla proposta dell'onorevole Fuschini di sospendere la seduta per un'ora.

COVELLI. Chiedo di parlare.

PRESIDENTE. Ne ha facoltà.

COVELLI. Propongo di proseguire nella discussione degli articoli e di accantonare al discussione su questo articolo 47 per rimandarla a lunedì.

PRESIDENTE. Pongo in votazione la proposta che ha la precedenza, cioè quella dell'onorevole Covelli, di rinviare la discussione dell'articolo a lunedì.

(*Non è approvata*).

Pongo in votazione la proposta dell'onorevole Fuschini di sospendere la seduta per un'ora.

(*È approvata*).

(*La seduta, sospesa alle 11, è ripresa alle 12,5*).

PRESIDENTE. Riprendiamo la discussione dell'articolo 47, nel nuovo testo presentato dal Governo:

« Oltre i casi previsti dall'articolo 2, non sono elettori, per cinque anni, coloro i quali hanno ricoperto le seguenti cariche:

a) segretario o vicesegretario del partito fascista;

b) membro del gran Consiglio del fascismo;

c) componente del direttorio nazionale o del Consiglio nazionale del partito fascista;

d) segretario politico federale del partito fascista;

e) le medesime cariche di cui alle lettere precedenti, durante la pseudo repubblica sociale;

f) ministro o sottosegretario di Stato dei governi fascisti in carica o nominati dal 3 gennaio 1925;

g) consigliere nazionale;

h) membro del tribunale speciale per la difesa dello Stato o membro dei tribunali straordinari della pseudo repubblica sociale;

i) prefetto o questore nominati per titoli fascisti e capo di provincia;

l) ufficiale generale od ufficiale superiore della milizia volontaria sicurezza nazionale in servizio permanente retribuito, eccettuati gli addetti ai servizi religiosi, sanitari, assistenziali e gli appartenenti alle legioni libiche, alla milizia ferroviaria, postelegrafonica, universitaria, alla GIL alla Dicat e Da Cos, nonché alla milizia forestale, stradale e portuaria.

« Il termine stabilito nel primo comma decorre dalla data di entrata in vigore della presente legge. Nei confronti di coloro i quali siano stati già cancellati o non iscritti nelle liste elettorali per avere ricoperto taluna delle cariche sopraelencate, il termine decorre dalla data della « pronuncia » o del « provvedimento » con cui fu disposta la privazione temporanea del diritto elettorale.

« La cancellazione dalle liste elettorali di coloro che si trovano nelle condizioni di cui

Figure 1: Visual representation of the annotation for page 274. Color-coded and numbered boxes indicate individual functional blocks, each labelled with its instance identifier, illustrating the segmentation process and its alignment with the document layout.

As can also be observed from the annotation example reported in 3, the annotation scheme is characterized by multiple layers of variation and complexity.

For instance, the same page includes different structural units, such as speech, description, and text, reflecting the heterogeneous nature of parliamentary proceedings. The majority of segments are labeled as speech, each associated with an explicit speaker attribution (e.g., Presidente, Scelba, Ministro dell'interno, Uberti, relatore), while other segments correspond to procedural descriptions (e.g., voting outcomes or session interruptions), which do not involve an identifiable speaker. The example also illustrates additional annotation decisions, such as the use of special markers for partially unavailable speaker information (e.g., [hidden]) or inferred attributions ([inferred] Presidente). This layered structure requires distinguishing between structural segmentation (ID-level units), discourse function (type labels), and speaker meta-

Label	Description
`speech`	Oral discourse delivered by MPs or institutional figures, typically introduced by a name. This includes interruptions such as *"Voci"* or *"Molte voci"*.
`speechnotext`	Instances where a speech act is referenced (e.g., a secretary reading a formal record) but the verbatim content is not provided in the transcript.
`article`	Legislative articles, usually introduced by the abbreviation "Art." followed by a numerical identifier.
`text`	Verbatim quoted written texts, official statements, or legal provisions read during a session. These are distinguished from oral speech via formatting such as indentation or angle quotes («. . . »).
`description`	Procedural notes regarding session management (e.g., opening/closing times) or standalone bracketed comments.
`presidencydeclaration`	Explicit indicators of the session's presidency, typically appearing in small caps below the title.
`title`	Structural headings for sessions, topics, or legislative proposals.
`other`	Non-discursive elements including signatures, tables, attachments, indexes, and content-related footnotes.

Table 1: Taxonomy of annotation labels for block types (label type)

Annotation task	Krippendorff's α
Speaker attribution	0.76
Type label classification	0.44

Table 2: Inter-annotator agreement measured using Krippendorff's α for nominal data on the initial 20-page sample.

Page	ID	Type	Speaker
274	1	speech	[hidden]
274	2	speech	Fuschini
274	3	speech	Scelba, Ministro dell'interno
274	4	speech	Presidente
274	5	speech	Scelba, Ministro dell'interno
274	6	speech	Uberti, relatore
274	7	speech	Presidente
274	8	speech	Uberti, relatore
274	9	speech	Presidente
274	10	speech	Covelli
274	11	speech	Presidente
274	12	speech	Covelli
274	13	speech	Presidente
274	14	description	
274	15	speech	[inferred] Presidente
274	16	description	
274	17	description	
274	18	speech	Presidente
274	19	text	

Table 3: Structured representation of page 274, including Page ID (Page), Instance ID (ID), type label (type), and speaker attribution (Speaker).

data, thereby increasing the overall annotation complexity.

Out of the 300 annotated pages, the complete set of type labels identified in the dataset is reported in Table 4.

Type label	Frequency	Percentage (%)
article	127	5.7
description	213	9.6
other	60	2.7
pres. declaration	11	0.5
speech	1585	71.2
speechnotext	22	1.0
text	89	4.0
title	121	5.4
Total	**2228**	**100.0**

Table 4: Distribution of type labels across the 300 annotated pages, with absolute and percentage frequencies.

Table 4 shows that the distribution of type labels is heavily skewed towards the speech category, which accounts for 71.2% of all annotated segments. This is expected given the dialogic and interactional nature of parliamentary proceedings, where the primary structural unit is the individual speech turn. At the same time, the presence of several additional categories, such as description (9.6%), article (5.7%), title (5.4%), and smaller classes like text, speechnotext, and presidency-declaration, highlights the structural heterogeneity of the documents. These categories capture procedural notes, editorial insertions, structural markers,

and non-speech material, which coexist with spoken interventions and may interrupt or frame them. Therefore, although speech segments clearly dominate the corpus quantitatively, the interaction and alternation between speech and non-speech units contribute significantly to the overall structural complexity of the annotation.

5. Experiments

The task of identifying and segmenting functional blocks in historical parliamentary debates is highly challenging. Unlike plain OCR transcription, this problem requires the reconstruction of the logical and communicative structure of the document.

Errors in reading order, missing typographical cues, or imperfect speaker markers can easily propagate into incorrect block segmentation and speaker attribution. Moreover, communicative boundaries are often signaled by subtle graphical features (e.g., indentation, capitalization, spacing) that may be partially lost during digitisation. For these reasons, block identification cannot be reduced to a simple text classification problem, but rather requires joint reasoning over layout, discourse function, and institutional conventions.

To evaluate the feasibility of automating this task, we experimented with two different approaches:

- Single-step direct block identification via LLMs
- Two-step pipeline: OCR followed by LLM-based block extraction

The experiments were conducted on the manually annotated corpus of 300 pages described in Section 4.

The prompt used for the LLM queries can be found in the Appendix.

5.1. Direct Block Identification

As a first baseline, we tested whether modern multimodal Large Language Models could directly perform block segmentation and speaker attribution in a single step.

In this configuration, the model was provided with:

- the annotation guidelines (in PDF format);
- the PDF page to be annotated;
- explicit instructions to segment the page into blocks following the guidelines.

This approach represents an upper-bound scenario in which the model is asked to jointly solve OCR, layout interpretation, functional classification, and speaker attribution within a single reasoning process.

However, preliminary observations show that this task is particularly demanding: performance degrades in pages containing dense legislative articles, embedded attachments, or unclear typographical separation between speech turns and procedural notes.

5.2. OCR + LLM-Based Block Extraction

Given the complexity of the single-step configuration, we designed a modular two-step pipeline that separates transcription from structural reconstruction.

5.2.1. Step 1: OCR

The first step consists of extracting raw textual content from scanned pages using different OCR systems. We evaluated the following engines: Mistral (multimodal OCR capabilities), Azure Document Intelligence, AWS Textract, Google Vision (Visual API), Tesseract.

The output of each OCR system consists of plain text (or structured text when available), which is then passed to the second stage. While these systems differ substrantially in terms of layout sensitivity and robustness to degraded scans and historical typography, Tesseract is the only one that is open source and available for free.

5.2.2. Step 2: LLM-Based Block Extraction

In the second step, the OCR output is provided to a Large Language Model tasked with:

- segmenting the text into functional blocks;
- assigning a type label from the predefined taxonomy;
- performing speaker attribution according to the annotation guidelines.

We evaluated the output of all the OCR systems with the following LLMs: Mistral Large, Gemini 3, GPT mini 5, Llama 4 Scout 17B 16e.

This modular configuration allows us to isolate the impact of transcription errors on structural reconstruction and the reasoning capabilities of different LLMs when operating on noisy OCR text.

5.3. Zero-Shot vs Few-Shot Configurations

The second approach (OCR + LLM extraction) was evaluated under two prompting strategies:

5.3.1. Zero-Shot Setting

In the zero-shot configuration, the model receives:

- the annotation guidelines;
- the OCR-extracted page text;
- instructions to produce the structured annotation.

No annotated examples are provided. This setting measures the model's ability to generalize directly from the schema description.

5.3.2. Few-Shot Setting

In the few-shot configuration, we provide the model with manually annotated examples before presenting the target page.

To ensure methodological rigor, the dataset of 300 annotated pages was split as follows:

- Development set (100 pages)
- Test set (200 pages)

Few-shot examples were sampled exclusively from the development set, while evaluation was conducted strictly on the held-out test set. No page in the test set was used as demonstration material.

The few-shot examples were selected to maximize structural diversity (e.g., pages dominated by speeches, pages with multiple legislative articles, presence of attachments, complex procedural notes).

5.4. Evaluation

The evaluation was designed to assess system performance along two complementary dimensions: (i) the correct identification and ordering of functional blocks, and (ii) the accurate transcription and attribution of speakers within speech segments.

The evaluation is performed on the 200 pages included in the test set (see Section 5.3.2).

5.4.1. Block Identification

For the first dimension, we evaluated the system's ability to reconstruct the correct sequence of annotated blocks on each page. For every test page, we generated two ordered sequences of labels:

- a gold sequence derived from the manually annotated corpus;
- a predicted sequence produced by the system.

Each element in the sequence corresponds to the type label assigned to a block (e.g., speech, description, article, text, etc.). Evaluation consists of measuring the distance between the gold and predicted sequences, taking into account insertions, deletions, substitutions, and ordering errors. This formulation allows us to capture both segmentation mistakes (e.g., merged or split blocks) and misclassification errors.

ChatGPT	Blocks	1338
	Speakers	639
Gemini	Blocks	715
	Speakers	373
Mistral	Blocks	1952
	Speakers	504

Table 5: Results of the single-step baseline (measured using Levenshtein distance).

5.4.2. Speaker Transcription and Attribution

The second dimension focuses specifically on speech segments. In this case, we constructed sequences composed only of blocks labeled as "speech" (see Section 3.1). For each speech block, the label corresponds to the extracted speaker string. Gold and predicted sequences were then compared after filtering out all non-speech elements. This setup evaluates both the correctness of speaker recognition and the ability to maintain the proper discourse order.

In both evaluation dimensions, correctness is measured using an edit-distance–based similarity metric grounded in Levenshtein distance. The metric is conceptually analogous to Word Error Rate (WER), which is commonly used to compare token sequences in speech recognition, and to its generalizations to higher-level semantic units, such as Slot Error Rate (SER). By operating on structured label sequences rather than individual tokens, the metric captures structural reconstruction quality rather than raw textual overlap (Ákos Tündik et al., 2020).

5.5. Results

The experimental results highlight clear performance differences across system configurations. Table 6 shows the edit-distance–based similarity metric grounded in Levenshtein distance. Table 5 shows the results for single-step approach (see next Section).

5.5.1. Baseline (Single-Step LLM)

The single-step direct block identification approach, in which the model jointly performs OCR, layout interpretation, classification, and speaker attribution, consistently yields the lowest performance across both evaluation dimensions. This confirms

			Mistral	Google	Tesseract	AWS	Azure
Llama 4	Zero-shot	Blocks	1418	1616	1566	1592	1368
		Speakers	468	666	696	752	604
	Few-shots	Blocks	1156	1060	636	970	652
		Speakers	464	636	570	630	418
GPT 5 mini	Zero-shot	Blocks	476	710	634	672	420
		Speakers	110	236	292	244	116
	Few-shots	Blocks	436	692	538	658	414
		Speakers	90	224	278	230	94
Gemini 3	Zero-shot	Blocks	546	756	562	546	498
		Speakers	96	202	76	98	56
	Few-shots	Blocks	580	560	388	400	338
		Speakers	88	210	92	66	56
Mistral large	Zero-shot	Blocks	1394	2410	1534	2038	2136
		Speakers	288	418	472	492	422
	Few-shots	Blocks	460	848	712	700	630
		Speakers	180	414	308	400	190

Table 6: Results of the accuracy (measured using Levenshtein distance) of the OCR + LLM approach (the lower, the better).

that solving OCR and structural reasoning simultaneously represents a particularly demanding task, especially in the presence of complex layouts and degraded historical scans.

5.5.2. Zero-Shot OCR + LLM

The modular two-step configuration (OCR followed by LLM-based block extraction) significantly improves results. In the zero-shot setting, performance varies depending on the specific combination of OCR engine and language model. The separation of transcription from structural reasoning allows LLMs to operate on textual input, reducing multimodal complexity. However, transcription noise from OCR systems still affects downstream segmentation and speaker attribution.

5.5.3. Few-Shot Configuration

The few-shot configuration achieves the best results for the block identification task, while achieving comparable results for speaker transcription. Providing annotated examples from the development set substantially improves structural alignment accuracy.

Interestingly, using OCR and LLM tools from the same provider (e.g., Mistral OCR + Mistral LLM, or Google Vision + Gemini) does not lead to systematic improvements. This suggests that performance is not determined by ecosystem coherence, but rather by the intrinsic quality of the transcription and the reasoning capabilities of the language model independently.

5.5.4. Language Model Comparison

Among the evaluated LLMs, Gemini consistently achieves the strongest performance across configurations and OCR inputs. It demonstrates robust handling of noisy OCR output and greater stability in maintaining correct block order and speaker attribution. Other models show higher variance depending on input quality and page complexity.

5.5.5. OCR System Comparison

Regarding transcription quality, Azure Document Intelligence emerges as the most reliable OCR system overall. Its outputs yield the best downstream structural reconstruction results, indicating superior robustness to historical typography and layout variability.

Notably, Tesseract, despite being open source and freely available, performs remarkably well. While it does not consistently surpass commercial systems, its results remain competitive, especially considering cost-effectiveness and reproducibility constraints.

6. Release

To foster transparency and reproducibility, we plan to publicly release both the annotated dataset and the annotation guidelines upon completion of the anonymized review process. An anonymized Google Drive folder has been shared with the re-

viewers.[6]

The release will include:

- the manually annotated corpus of 300 randomly sampled pages in structured format;
- the full annotation guidelines used during the project;
- the evaluation scripts required to reproduce the experimental results reported in this paper.

The dataset will be distributed in a machine-readable format designed to facilitate reuse. All resources will be made available under an open license compatible with research and academic use. By releasing both data and guidelines, we aim to provide a replicable benchmark for future work on historical parliamentary digitisation and to support the development of robust OCR and structural segmentation pipelines for long-span political corpora.

7. Conclusion and Future Work

This paper addressed the digitisation of historical Italian parliamentary proceedings (1848–1996), moving beyond plain OCR to tackle functional block segmentation and speaker attribution. We introduced tailored annotation guidelines and a manually annotated dataset of 300 randomly sampled pages.

We compared a single-step multimodal LLM approach with a modular OCR+LLM pipeline under zero-shot and few-shot settings. Results show that separating transcription from structural reasoning yields more reliable outputs, and that few-shot prompting significantly improves performance. Overall results are strongly influenced by OCR quality: Azure Document Intelligence provided the most robust inputs, Tesseract proved competitive as an open-source alternative, and Gemini achieved the best balance between textual accuracy and structural reconstruction across configurations.

Future work will focus on conducting additional experiments to identify the most robust and scalable configuration before applying the pipeline to the tens of thousands of pages of historical Italian parliamentary debates still awaiting structured digitisation. We also plan to evaluate the portability of both the annotation guidelines and the proposed pipeline to other languages and parliamentary traditions, assessing the degree of adaptation required across different institutional and layout conventions.

[6] https://drive.google.com/drive/folders/191ixt3e31EeTpa-ct36nzmiDxRmmY6wl

8. Limitations

The modular pipeline clearly demonstrates that structural reconstruction performance is strongly conditioned by transcription quality. OCR errors affecting capitalization, punctuation, indentation cues, or speaker markers propagate to downstream segmentation and attribution tasks.

Although Azure Document Intelligence performed best overall, commercial OCR systems introduce reproducibility and cost constraints. Conversely, while Tesseract remains competitive and reproducible, its sensitivity to degraded typography may limit scalability to lower-quality scans. Therefore, the overall framework remains critically dependent on the availability of high-quality OCR systems capable of handling historical Italian typography.

The annotation guidelines were designed specifically for the Italian parliamentary tradition and its layout conventions. Although the underlying principles of functional segmentation and speaker attribution are broadly applicable, direct transfer to other parliamentary corpora (e.g., Hansard, GERPARCOR, ParlaMint extensions) would likely require schema adaptation.

While the few-shot OCR+LLM pipeline achieves promising results, large-scale application to the entire parliamentary archive would entail:

- substantial computational costs for LLM inference;
- potential latency issues;
- and careful monitoring of error propagation across millions of pages.

Further optimization, model distillation, or hybrid rule-based post-processing may be required to ensure sustainable large-scale deployment.

9. Bibliographical References

Thomas M. Breuel. 2008. The OCRopus open source OCR system. In *Document Recognition and Retrieval XV*, volume 6815, page 68150F. International Society for Optics and Photonics, SPIE.

Gavin Greif, Niclas Griesshaber, and Robin Greif. 2025. Multimodal llms for ocr, ocr post-correction, and named entity recognition in historical documents.

Nuno Guimarães, Purificação Silvano, Ricardo Campos, Alipio Jorge, Ana Filipa Pacheco, Dimitar Iliyanov Dimitrov, Nikolaos Nikolaidis, Roman Yangarber, Elisa Sartori, Nicolas Stefanovitch,

Preslav Nakov, Jakub Piskorski, and Giovanni Da San Martino. 2025. NarratEX dataset: Explaining the dominant narratives in news texts. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20408–20434, Suzhou, China. Association for Computational Linguistics.

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 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8287–8300, Torino, Italia. ELRA and ICCL.

Christian Reul, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner, and Frank Puppe. 2019. Ocr4all—an open-source tool providing a (semi-)automatic ocr workflow for historical printings. *Applied Sciences*, 9(22):4853.

Christoph Wick, Christian Reul, and Frank Puppe. 2018. Calamari - a high-performance tensorflow-based deep learning package for optical character recognition.

Máté Ákos Tündik, Balázs Tarján, and György Szaszák. 2020. A low latency sequential model and its user-focused evaluation for automatic punctuation of asr closed captions. *Computer Speech & Language*, 63:101076.

10. Language Resource References

Giuseppe Abrami, Mevlüt Bagci, Leon Hammerla, and Alexander Mehler. 2022. German parliamentary corpus (gerparcor). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1900–1906.

Giuseppe Abrami, Mevlüt Bagci, and Alexander Mehler. 2024. German parliamentary corpus (gerparcor) reloaded. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7707–7716.

Francesca Cominetti, Lorenzo Gregori, Edoardo Lombardi Vallauri, and Alessandro Panunzi. 2022. Impaqts: un corpus di discorsi politici italiani annotato per gli impliciti linguistici. In *Corpora e studi linguistici = Corpora and linguistic studies: Atti del 54. Congresso internazionale di studi della Società di linguistica italiana (online, 8–10 settembre 2021)*, pages 151–164. Officinaventuno.

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.

Jacopo Cova. 2025. A new database for italian parliamentary speeches: Introducing the itaparl-corpus dataset. *Italian Political Science Review / Rivista Italiana di Scienza Politica*, 55(1):77–86.

Luigi Curini, Silvia Decadri, Alfio Ferrara, Stefano Montanelli, Fedra Negri, and Francesco Periti. 2024. The gender gap in issue attention and language use within a legislative setting: An application to the italian parliament (1948–2020). *Politics & Gender*, 20(1):182–211.

Tomaž Erjavec, Matyáš Kopp, Nikola Ljubešić, Taja Kuzman, Paul Rayson, Petya Osenova, Maciej Ogrodniczuk, Çağrı Çöltekin, Danijel Koržinek, Katja Meden, Jure Skubic, Peter Rupnik, Tommaso Agnoloni, José Aires, Starkaður Barkarson, Roberto Bartolini, Núria Bel, María Calzada Pérez, Roberts Darģis, Sascha Diwersy, Maria Gavriilidou, Ruben van Heusden, Mikel Iruskieta, Neeme Kahusk, Anna Kryvenko, Noémi Ligeti-Nagy, Carmen Magariños, Martin Mölder, Costanza Navarretta, Kiril Simov, Lars Magne Tungland, Jouni Tuominen, John Vidler, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, and Darja Fišer. 2024. Parlamint ii: Advancing comparable parliamentary corpora across europe. *Language Resources and Evaluation*.

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*, 57:415–448.

Valentino Frasnelli and Alessio Palmero Aprosio. 2024. There's something new about the Italian parliament: The IPSA corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources*

and Evaluation (LREC-COLING 2024), pages 16037–16046, Torino, Italia. ELRA and ICCL.

Alenka Kavčič, Martin Stojanoski, and Matija Marolt. 2024. Historical parliamentary corpora viewer. In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024*, pages 127–132, Torino, Italia. ELRA and ICCL.

Guy Mor-Lan, Effi Levi, Tamir Sheafer, and Shaul R. Shenhav. 2024. IsraParlTweet: The israeli parliamentary and Twitter resource. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9372–9381, Torino, Italia. ELRA and ICCL.

Federico Nanni, Stefano Menini, Sara Tonelli, and Simone Paolo Ponzetto. 2019. Semantifying the uk hansard (1918-2018). In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 412–413.

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 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12553–12563, Torino, Italia. ELRA and ICCL.

Stephen Wattam, Paul Rayson, Marc Alexander, and Jean Anderson. 2014. Experiences with parallelisation of an existing NLP pipeline: Tagging Hansard. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4093–4096, Reykjavik, Iceland. European Language Resources Association (ELRA).

A. Prompts

We utilized differentiated prompting strategies for text extraction across the two experimental setups: a direct LLM-based extraction and a combined OCR-LLM pipeline. This ensured that the prompts were adapted to the unique input requirements of each method. The placeholders *{GUIDELINES}*, *{TEXT}*, *{FILE}* represent the task-specific instructions, the target input string and the source file, respectively.

A.1. LLM ONLY

Role: You are an expert archival researcher specializing in the digitization and annotation of Italian Parliamentary Debates. Your task is to segment a provided page of text into coherent "Textual Blocks" and label them according to strict linguistic and functional guidelines.

Core Task:

Analyze the provided page. Identify every distinct textual unit (Instance). For each unit, extract the required metadata into a CSV format.

Annotation Fields (CSV Columns):

1. **Page ID**: The integer page number.
2. **Instance/Round ID**: Progressive integer (1, 2, 3...) restarting at 1 for every new page.
3. **Type Label**: Must be exactly one of: `speech`, `speechnotext`, `article`, `text`, `description`, `presidencydeclaration`, `title`, `other`.
4. **Speaker**: The name/role of the speaker. Use `[unknown] Name` for generic voices, `[hidden]` if not printed, or `[inferred] Name` if clearly understood from context. Leave blank for `description` or `title`.
5. **Notes**: Brief description if Type is `other` (e.g., "signature", "table").

Classification Rules:

{GUIDELINES}

Constraint:

Return **ONLY** the CSV data. Do not provide introductory text or conversational fillers. Use a comma as a delimiter. Wrap text fields in double quotes if they contain commas.

Example Format:

```
Page ID,Instance/Round ID,Type label,
Speaker,Notes
12,1,presidencydeclaration,CASATI,
12,2,title,,Verificazione di poteri
12,3,speech,PRESIDENTE,
12,4,description,,(Approvato)
12,5,other,,signature
```

More instructions:

- Do not hallucinate extra rows. Only record distinct, visible structural blocks.
- A single 'speech' block usually contains multiple paragraphs. Do NOT create a new row for every paragraph or line; group them by the speaker.
- For this type of document, there are typically between 5 and 20 blocks per page. If you are exceeding 30 blocks, you are likely being too granular.
- If the speaker is unknown, do not guess; use '[unknown]'. However, do not use this to create repetitive filler rows.

{FILE}

A.2. OCR + LLM

Role: You are an expert archival researcher specializing in the digitization and annotation of Italian Parliamentary Debates. Your task is to segment a provided page of text into coherent "Textual Blocks" and label them according to strict linguistic and functional guidelines. The page is provided as an OCR transcription, therefore it can contain headers or footer that should be ignored.

Core Task:

Analyze the provided page. Identify every distinct textual unit (Instance). For each unit, extract the required metadata into a CSV format.

Annotation Fields (CSV Columns):

1. **Page ID**: The integer page number (it can be "X" if the number is not found/identified).
2. **Instance/Round ID**: Progressive integer (1, 2, 3...) restarting at 1 for every new page.
3. **Type Label**: Must be exactly one of: ‘speech‘, ‘speechnotext‘, ‘article‘, ‘text‘, ‘description‘, ‘presidencydeclaration‘, ‘title‘, ‘other‘.
4. **Speaker**: The name/role of the speaker. Use ‘[unknown] Name‘ for generic voices, ‘[hidden]‘ if not printed, or ‘[inferred] Name‘ if clearly understood from context. Leave blank for ‘description‘ or ‘title‘.
5. **Notes**: Brief description, not empty only if Type Label is ‘other‘ (e.g., "signature", "table").

Constraint:

Return **ONLY** the CSV data. Do not provide introductory text or conversational fillers. Use a comma as a delimiter. Wrap text fields in double quotes if they contain commas.

Example Format:

```
Page ID,Instance/Round ID,Type label,
Speaker,Notes
12,1,presidencydeclaration,CASATI,
12,2,title,,
12,3,speech,PRESIDENTE,
12,4,description,,
12,5,other,,signature
```

More instructions

- Do not hallucinate extra rows. Only record distinct, visible structural blocks.
- A single ’speech’ block usually contains multiple paragraphs. Do NOT create a new row for every paragraph or line; group them by the speaker.
- For this type of document, there are typically between 5 and 20 blocks per page. If you are exceeding 30 blocks, you are likely being too granular.
- If the speaker is unknown, do not guess; use ’[unknown]’. However, do not use this to create repetitive filler rows.

{GUIDELINES}

Could you extract the textual units from the text below, obtained through OCR?

{TEXT}

Computational Political Landscape of the Netherlands and Prime Minister Schoof's Position

Wessel Ledder[1] and Iris Hendrickx[2]
[2]Centre for language and Speech Technology, Centre for Language Studies
[1,2]Radboud University, Nijmegen, The Netherlands
{wessel.ledder, iris.hendrickx}@ru.nl

Abstract

This study presents a computational model of the Dutch political landscape during the Schoof government period, constructed using debate speeches from the House of Representatives. We construct a two-dimensional representation of the Dutch political landscape by fine-tuning a BERT model on parliamentary debate speeches and applying dimensionality reduction techniques to the resulting embeddings. We evaluate the validity of this model by comparing it to an independently developed model from an external research institute, finding that both models reveal similar patterns along the socio-economic left–right dimension. We also examine content patterns and word frequency distributions in targeted samples located at distinct regions of the landscape to interpret the model. We further evaluate the stability of the landscape to ensure that the observed patterns are not driven by random variation. Finally, we position Prime Minister Schoof within this computational landscape. Schoof was intended to be a neutral Prime Minister without any party affiliation that would represent the coalition parties of the government equally. Our analysis will show whether Schoof was indeed neutral in his statements or not.

Keywords: political ideological landscape, parliamentary debate speeches, BERT embeddings

1. Introduction

Over the last years in the Netherlands, fragmentation of political parties is becoming more common in the Dutch political landscape (Sipma et al., 2021). The national election results of the last decade show a broad distribution of votes across multiple parties, with more parties getting a seat in the House of Representatives (HR)[1].

The latest government (July 2024 – June 2025) consisted of a coalition of four political parties (BBB, NSC, PVV and VVD) and the formation of this government required a considerable amount of time. As these parties could not agree on which party should deliver the prime minister, they opted for an external person that was not affiliated with any political party as to have a neutral prime minister to represent all coalition parties equally. The coalition appointed Dick Schoof, who had a background in civil service in the areas of Justice and Security, as Prime Minister (Van Holsteyn and Irwin, 2025).

In this study we aim to investigate the Dutch political landscape with computational methods and the role of Prime Minister Schoof in particular. Was Schoof really as neutral in his statements as was intended? We used Bidirectional Encoder Representations from Transformers (BERT) embeddings (Devlin et al., 2018), which carry semantic information of debate transcripts from the House of Representatives, as the basis for a computational political landscape of the political parties in the Netherlands. We aim to capture the underlying ideological political stances.

A political ideology is a set of ideas about how the society and economy should be organized. This definition is rather vague; it can be applied to a broad interpretation such as social-economic views of left versus right, or to much more narrow scope (e.g. in favor or against one United Europe). The ways in which political ideologies surface have been studied extensively, especially for the political system in the United States, e.g. (Poole, 2005; Diermeier et al., 2012). In this study we focus on the Dutch parliament. We investigated the following research questions:

- **RQ1:** How can we make a computational political landscape using debate speeches?
- **RQ2:** Where does Prime Minister Schoof fall on this landscape and is he truly neutral in his statements?

In this paper, we introduce a method to computationally create a model of the Dutch political landscape, based on speeches in the House of Representatives during the Schoof government. We do this by first fine-tuning a BERT model, after which we reduce dimensionality of the resulting BERT embeddings to create a two-dimensional landscape. We evaluate whether our political land-

[1]Recap of Dutch Parliament formation: during national elections, people vote for the political parties in House of Representatives (HR). Based on the number of votes, each party will get a number of seats in HR. As we do not have one party with the overall majority in the HR, a group of political parties agree on a coalition to form the active government consisting of the prime minister, the other ministers and the state secretaries.

Proceedings of ParlaCLARIN V Workshop on Interoperability, Multilinguality, and Multimodality in Parliamentary Corpora @ LREC 2026, pages 77–86

scape is interpretable and meaningful by comparing our landscape against a manually designed model that was created by an independent research institute (KiesKompas BV., 2023) and by inspecting the content and word frequencies in small samples in different corners of the landscape. We additionally test the stability of our semantic space to rule out random effects. Finally, we will place Prime Minister Schoof on our political landscape.

2. Related Work

Political debate speeches are a rich source to study the relation between language use and political ideologies. Schoonvelde et al. (2019) revealed structural differences in complexity of language use in political speeches from liberal versus conservative politicians. The claim that populist leaders use simple language in their speeches has also been investigated and this claim has been disputed (McDonnell and Ondelli, 2022; Zanotto et al., 2024). Neiman et al. (2016) studied language use of political speeches of Democrats and Republicans and their expressions of moral values. They did not observe systematic differences between the two parties in their study. Jordan et al. (2019) examined the language use of US Presidents and political leaders in other countries over a 20-year period to determine whether linguistic changes had occurred, particularly in analytical thinking and confidence. Using a method called LIWC (Tausczik and Pennebaker, 2010) to analyse patterns in word frequencies, they found an overall decline in analytical thinking and an increase in confidence during this period.

Furthermore, previous work has shown earlier attempts to create computational models of political landscapes. For example, Poole (2005) showed that voting decisions in the US congress can be modelled with a one-dimensional statistical model that expressed the dimensions of liberal versus conservative values. Diermeier et al. (2012) have shown credible evidence that political debate speeches from the US congress are expressing these ideologies in such a consistent way that one can train a machine learning classifier on speeches from conservative and liberal politicians to predict the position on this dimension for new unseen speeches. Their study also revealed that politicians tend to use certain terms that are characteristic for their ideology and that such terms are mostly related to culture rather than the economy.

Previous studies showed that word embeddings can be used effectively to capture underlying constructs such as political ideologies, to offer a unified framework for analysing political language (Rheault and Cochrane, 2020) or to model stances in political debates (Konjengbam et al., 2018).

These earlier findings form the underpinning for our approach where we aim to use parliamentary speeches as the source for modelling the underlying ideologies of Dutch political parties.

We did not use the obvious alternative source for getting insights in the ideologies from Dutch political parties, the political party program reports. This source has been used to construct a two-dimensional representation of the Dutch political landscape by the research institute Kieskompas (KiesKompas BV., 2023) in the run-up for the 2023 Dutch national elections. This landscape is created manually and consists of a social-economic left–right dimension and a cultural conservative–progressive dimension based on a list of stances that pertains relevant topics for the 2023 elections and can help voters to get insights in their own stances in light of the stances of the political parties on these topics (Wall et al., 2014). We will use the Kieskompas 2023 political landscape as an external source to validate the political computational landscape that we are creating in this study.

3. Method

We created a digital representation of the Dutch political landscape by fine-tuning a BERT embedding model on political debate speeches. We describe the model and fine-tuning in more detail in the next section. We evaluate whether this model indeed captures the political ideologies from the Dutch political parties with the following steps. We apply factor analysis (Spearman, 1904) to reduce the high-dimensional embedding vectors to two latent semantic dimensions that organize the embedding space. We compare this reduced space to an existing political model manually created by the company Kieskompas (KiesKompas BV., 2023). We validate the stability of the landscape as detailed in section 3.2. To answer our second research question on where Prime Minister Schoof falls on this computational political landscape, we collected the speeches of Schoof separately from the other debate speeches and project them into the embedding space to reveal which political parties are closest to his statements.

We first present the dataset that we use in our studies in the next section, followed by the details of the experimental setup to create and validate the political landscape.

3.1. Data

The data used in this study are the plenary reports of the House of Representatives (Tweede Kamer der Staten Generaal, 2024), in the period of the Schoof government (July 2nd, 2024 to June 3rd, 2025). These reports are the transcripts of the

#	Text	Translation	Party	Speaker
1	(...) Het is belangrijk dat we bij zorg rekening houden met de diversiteit in onze zorg. Daarom de volgende motie. De Kamer, gehoord de beraadslaging, constaterende dat mantelzorgers vaak zorg krijgen van hulpverleners met verschillende culturele achtergronden; overwegende dat deze diversiteit helpt om beter in te spelen op de behoeften in de ouderenzorg; verzoekt de regering om inclusiviteit en diversiteit in de zorg en mantelzorg actief te ondersteunen en mee te nemen in beleid, en gaat over tot de orde van de dag.	(...) It is important we keep in mind diversity in healthcare. Hence the next motion. The House, having heard the deliberation, noting that informal caregivers often work with healthcare professionals from different cultural backgrounds; considering that this diversity helps to better respond to the needs in elderly care; calls on the government to include inclusivity and diversity in their policy for healthcare and informal caregiving, and moves on to the order of the day.	DENK	De heer El Abassi
2	(...) De PVV vindt dat ons belastinggeld in de eerste plaats moet worden uitgegeven aan Nederland en aan de Nederlanders. We zijn er trots op dat er vandaag eindelijk gehoor wordt gegeven aan deze oproep. Dit kabinet, deze coalitie, zet nu echt forse stappen: een forse bezuiniging van 300 miljoen in 2025, die oploopt tot maar liefst 2,4 miljard in 2027. Daar komt in 2027 ook nog eens de bezuiniging van 1,6 miljard op de eerstejaarsopvang voor asielzoekers bij. Die wordt ook betaald uit ontwikkelingsgeld. Van het strengste immigratiebeleid ooit naar de grootste bezuiniging ooit op ontwikkelingshulp: dat is bij elkaar een prachtig mooie bezuiniging van 4 miljard op ontwikkelingshulp. Dat klinkt alle rechtse, hardwerkende Nederlanders, onze Henk en Ingrid, die een sterke overheid willen die hun belangen wél behartigt, als muziek in de oren.	(...) The PVV believes that our tax money should primarily be spent on the Netherlands and the Dutch people. We are proud that today, this call is finally being heeded. This cabinet, this coalition, is now taking major steps: a major cut of 300 million in 2025, rising to no less than 2.4 billion in 2027. In 2027, there will also be an additional cut of 1.6 billion for first-year asylum seeker reception, which is also funded from development aid. From the strictest immigration policy ever to the biggest cut on development aid: that together is a nice cut of 4 billion on development aid. That sounds like music to the ears to all right-wing, Dutch citizens, our Henk and Ingrid, who want a strong government that truly looks after their interests.	PVV	De heer Ram
3	(...) Afgelopen dinsdag sprak ik Jan uit Rotterdam, die zei: "Ik werk in de gehandicaptenzorg en die 310 miljoen euro bezuinigingen daarop ... Het kan gewoon niet!" En u gooit er deze kabinetsperiode nog een keer in totaal 4,6 miljard aan bezuinigingen op de zorg voor mensen bovenop. U doet dat, meneer Wilders. (...)	(...) Last Tuesday, I spoke with Jan from Rotterdam, who said: "I work in disability care, and these 310 million euros of cuts ... It is just not possible!" And during this cabinet's term, you add another 4.6 billion of cuts to people's healthcare. You are doing that, mister Wilders. (...)	SP	De heer Dijk
4	Ik heb vertrouwen in het kabinet-Schoof, de introductie die minister-president Schoof hier net hield, in zijn ethos, zijn wil en ook zijn trackrecord om de rechtsstaat te respecteren. En ja, ik denk dat hij een goede minister-president is en dat deze ploeg het in zijn geheel in zich heeft om die rechtsstaat te respecteren.	I have confidence in the Schoof cabinet, the introduction that Prime Minister Schoof just delivered here, in his ethos, his determination and also his track record in respecting the rule of law. And yes, I believe he is a good Prime Minister, and that this group as a whole is capable to respect the rule of law.	NSC	De heer Omtzigt

Table 1: Exemplar data. For each speech in the debate, the speech, speaker and respective party is stored. Note that the English translations were not part of the political landscape, and only used in our automatic analysis of the landscape.

political debates in the HR. The spokespersons from the political groups in the HR discuss with the ministers or state secretaries or with each other in these debates. We assume that the collection of debate speeches of all people from the same party together represent the stances of their political party and reflect their ideological stances.

The transcripts of the debates were manually typed out by transcribers from the government. This has the effect that even though these are transcribed speeches, these transcripts do adhere to the writing style conventions of using sentence boundaries and punctuation. Typical speech elements such as fillers or laughter occur sparsely. After scraping the data, we first filtered out any metadata so that only the transcripts of the report are kept. We removed any speeches from the chair of the HR, Martin Bosma, as the chair leads the debate, and never debates for his own opinion in the debate. Next, we matched each transcript to their speaker and their respective party. Any speeches containing less than 15 words were omitted to ensure that each speech sample has at least some semantic content. This resulted in a total of 42882 speech samples with an average length of 152.8 words.

Some speech examples can be found in Table 1 and demonstrate that some words can strongly show the ideological tendency. For example, in speech 1, words such as 'diversity' and 'inclusivity' tend to be uttered by left-wing parties, while in speech 2, words such as 'the Dutch people' and 'cut' are more associated with right-wing parties. However, these words alone do not show ideology, as the context is important: in speech 3, the word 'cuts' is mentioned twice, but the speaker is against the cuts. Finally, speech 4 shows an example where the ideology of the speaker is less clear from this speech utterance.

Due to the non-uniform representation of parties in the House of Representatives, the distribution of the number of speeches per party is skewed, as is visible in Figure 1. The political parties with the lowest amount of speech samples in Figure 1 have only a few speakers in the debates: JA21 (1 seat), Volt (2 seats), CU, SGP, FVD, PvdD and

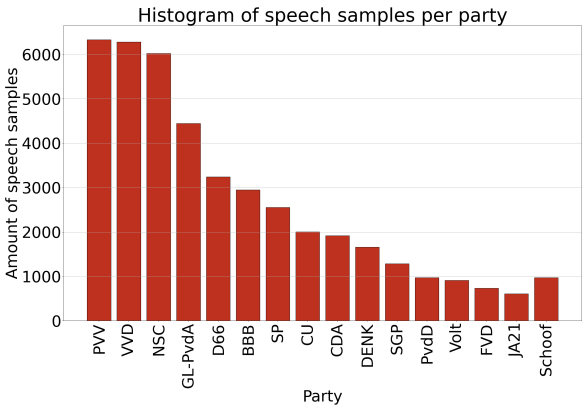


Figure 1: Distribution of total speeches per party.

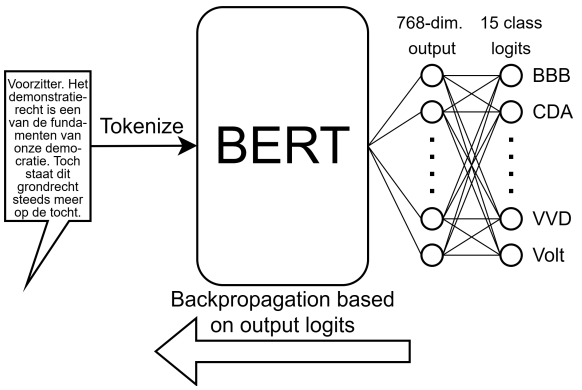


Figure 2: Graphical representation of the fine-tuning process. During training, the model must predict the party with each speech sample in its classification layer, allowing all weights to be fine-tuned accordingly.

DENK (3 seats) (Van Holsteyn and Irwin, 2025). We refer to Appendix A for the full name of each party, as well as the European Parliament group each party is affiliated to. The distribution of speakers within each party is also uneven, as party leaders, ministers and state secretaries generally have a larger role in the political debates.

3.2. Experimental Setup

For answering both research questions and to make a computational political landscape, we decided to use the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2018), which is a transformer neural network model, capable of transforming an input text sample into a numeric vector representation while capturing the semantics of the input text as a higher level embedding. These can in turn be used to create the landscape, as will be discussed later. In particular, we used a Dutch version of BERT, called BERTje (De Vries et al., 2019), which has been pre-trained to embed Dutch text.

For our task we fine-tuned the BERTje model on the stances of the Dutch political parties. The default embedding space that resulted from pre-training BERTje on a large and diverse training set tends to cluster words with similar meaning close together in the semantic space. However, for our digital political landscape we do not aim to cluster together on topics (like climate change or migration) but focus on ideological perspectives.

To achieve this, we fine-tuned the BERTje model to learn the stances and ideological perspectives of the political parties. First, we randomly split the dataset in a training set (80%), a validation set (10%) and a test set (10%), while keeping apart any speeches from Prime Minister Schoof in a separate dataset. To fine-tune the BERTje model weights, we first append a classification layer to the model, with one node for each party. Then, using the training set, we let the model predict the political party of the speaker of each speech. Weights are updated based on a cross-entropy loss function. The classification performance is evaluated with the validation set during fine-tuning, where after early stopping, we achieve an accuracy of 0.59 on the training set, and an accuracy of 0.42 on the validation set[2]. Note that for a complex linguistic 15-label classification task, the accuracy is expected to be low. The fine-tuning process is shown graphically in Figure 2.

Each sample from the training set is fed through the fine-tuned BERTje model (without classification layer), resulting in a 768-dimensional output vector embedding, which gives an aggregate representation of the full input, following the approach of Devlin et al. (2018). Factor analysis is used to project the high-dimensional vector embeddings of the training samples into two latent semantic dimensions that structure the embedding space. We then created the political landscape by projecting the samples from the held-out test set into the reduced space using the projection fitted on the training set. We computed the centroids for each political party based on the mean embeddings of speech samples from political party members in the reduced space.

3.3. Evaluation

We evaluate the resulting political landscape in various ways. First, we verify whether the represen-

[2]For any specifics on training or hyperparameters, please refer to code and data in our Github repository at github.com/wledderw/ComputationalPoliticalLandscape.

tations in the political landscape are stable. We did not only create the political landscape for the held-out test set, but created another version of the space using the training set data. We compare the resulting landscapes and the centroids per party of both the training and test set to see whether the centroid positions in relation to each other stay similar when using different data samples to create the landscape. Furthermore, we validate whether factor analysis provides an interpretable lower-dimensional representation of the parties, and hence a political landscape that indicates the underlying ideologies. Moreover, we validate our computational political landscape by comparing it to a manually-made political landscape by the company Kieskompas (KiesKompas BV., 2023).

We also inspected a random sample of the speech samples in different corners of the semantic space to interpret what that the underlying dimensions represent. While the first latent factor gave us a clear interpretation, the second latent factor was not so straightforward. We tried TF*IDF word ranking (Sparck Jones, 1972) to see if we could find words that might be important to distinguish this dimension direction, but that did not lead to any insights. Our manual inspection of the random sample did indicate that we saw more substantiated argumentations in the bottom speeches in the second dimension and more unfounded assertions or irrational oppositions at the top of the dimension. Linguistic Inquiry and Word Count (LIWC) is an automatic text analysis tool developed by psychologists to count words and group these in psychologically meaningful categories. Analysing someone's writing or speech style can give insights in emotions, social relationship and thinking styles (Tausczik and Pennebaker, 2010). Therefore, we applied LIWC-22 (Boyd et al., 2022) as this can be used to measure the amount of 'analytical thinking' in texts which seemed to fit with our intuition on how to interpret this dimension. We applied LIWC in the following way: we took the top 200 and bottom 200 speech samples from the second dimension of the test set, we translated speech transcriptions to English[3] and applied LIWC.

For answering research question 2, finding Prime Minister Schoof's place on this landscape, we use the same procedure as stated above to find Schoof's average embedding and hence his location on the political landscape. Note that we did not fine-tune the BERTje model to Schoof's speeches, allowing us to compare Schoof with different parties on the embeddings space.

[3]Translations adapted from Google Translate, accessed February 17, 2026, https://translate.google.com/.

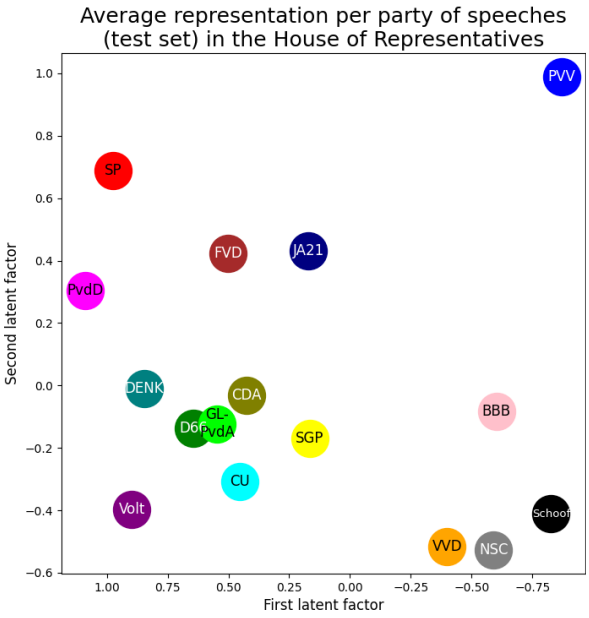


Figure 3: Computational political landscape based on samples of the test set.

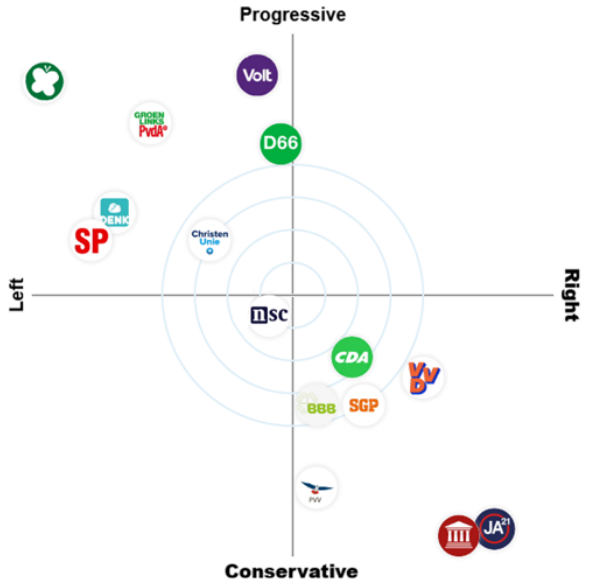


Figure 4: Political landscape created by Kieskompas for the 2023 Dutch elections (KiesKompas BV., 2023). For most parties, the acronym of the party is clear from the logo, except for the butterfly (PvdD) and the temple (FVD). Note that we removed parties from this political landscape that are not in the HR.

4. Results

We present the computational political landscape that is the result from the factor analysis on the fine-tuned BERTje model applied to the test set reduced to its two main dimensions, in Figure 3.

Let us start with the position of Prime Minister Schoof in this political landscape. Schoof is located closest to the coalition party NSC in the bottom right corner of the landscape. We can observe that Schoof is also close to coalition parties VVD and BBB.

First landscape dimension We observe in the first dimension (x-axis) that there is a clear separation between the coalition parties (BBB, NSC, PVV and VVD) on the right side and the other opposition parties on the left side. Since the coalition is right-leaning, this dimension also reflects a left–right spectrum.

When we compare our first dimension to the left–right dimension in the political landscape in Figure 4 that has been designed manually by KiesKompas BV. (2023), we can also see a large overlap in orientation. In both plots, parties with a social-economic left ideology (PvdD, SP, DENK, Volt) are on the left-most side while the right oriented parties PVV, VVD and BBB are on the right side of the first latent dimension[4]. We also observe differences between the two plots. The parties JA21 and FVD are much further to the right in the Kieskompas landscape then in our landscape, while for NSC we see the opposite as it is placed to the right in our landscape while the Kieskompas placed it in the middle of the left–right dimension.

Second landscape dimension The second latent factor is more difficult to interpret. It does not match with progressive–conservative axis in the Kieskompas political landscape shown in Figure 4. On the second latent factor (y-axis), we see a cluster of parties BBB, CDA, D66, DENK, GL-PvdA and SGP in the middle to bottom region of the axis. Parties FVD, JA21, PvdD, PVV and SP are located at the top of the second latent dimension, while CU, NSC, Volt and VVD are found at the bottom of the y-axis.

After manual inspection of a random sample on the y-axis, we noticed that the top speech samples contained more emotional arguments while speeches at the bottom seemed to consist of supported and critical arguments. To study this in a more systematic way, we computed the LIWC-22 (Boyd et al., 2022) 'analytical thinking' summary variable for the top 200 (higher latent factor values) and bottom 200 samples (lower latent factor values) from the y-axis. The spread of the scores for each of the two sets of samples can be found in Figure 5. For the speeches of lower latent factor values, 'analytical thinking' had an average score of 59.8, while the speeches for higher latent factor values had an average score of 44.0, on a scale running from 1 (personal and intuitive) to 99 (formal and logical). Hence, we hypothesize that our second latent factor indicates the degree of analytical thinking, where a lower second latent factor means a higher degree of analytical thinking.

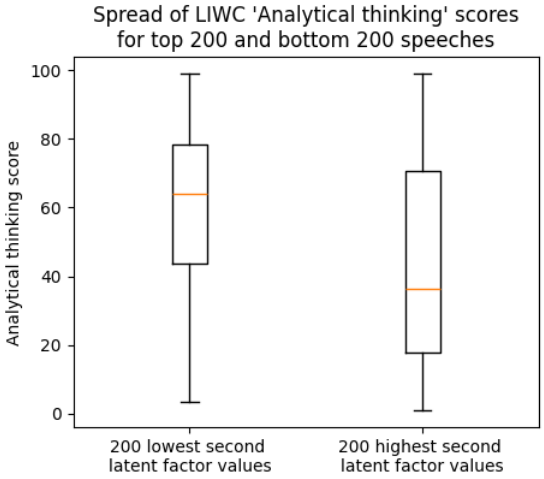


Figure 5: Spread of LIWC analytical thinking scores for the two extremes of the second dimension of the political landscape.

4.1. Stability of the political landscape

We investigated the stability of the landscape in two ways: we reviewed the stability of the semantic space based on different sets of speech samples, and we looked at the dimensionality reduction technique.

Stability over different samples First, we verified whether the semantic space created on the basis of the held-out test set samples keeps the relative positions between the political party centroids in similar positions when compared to the landscape created on the basis of the training set.

The semantic space has been trained using the training set, and the factor analysis dimensionality reduction also has been fitted on the embedded vectors of the training set. To analyse whether our model is stable, we plot the centroids for each party of the training set and the test set in the same plot, as is visible in Figure 6. The centroids from the training set are more transparent than the centroids from

[4]Note that we have reversed the axis of the first latent factor in our landscape to match left and right parties to the left and the right of the plot.

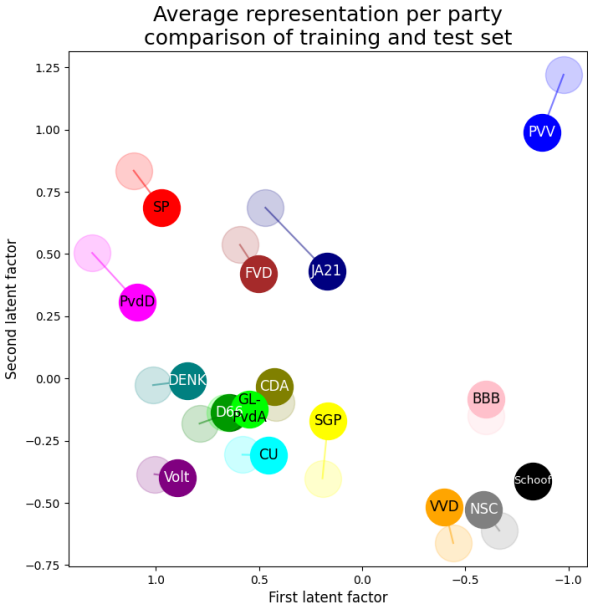


Figure 6: Comparison of computational political landscape for the training and test set. The transparent colours show the locations on the political landscape of parties by using speeches from the training set, while the opaque colours show the locations on the political landscape of the parties by using speeches from the test set.

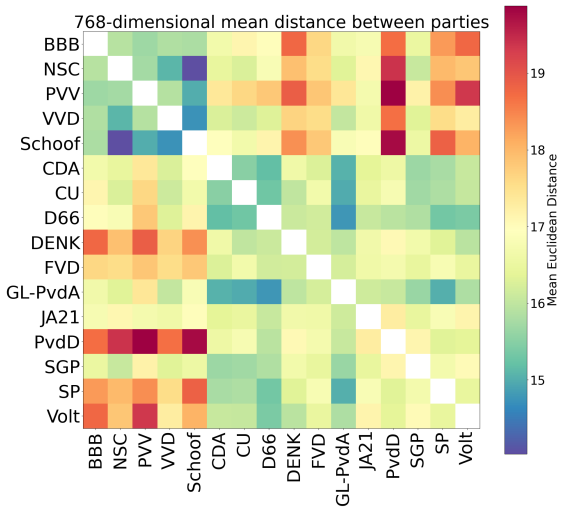


Figure 7: A Euclidean distance matrix between parties in the 768-dimensional semantic space. Note: coalition parties (BBB, NSC, PVV, VVD) and Prime Minister Schoof have been put together, the rest of the parties are sorted alphabetically.

the test set. In the plot, we see that the centroids for most parties move between the training set and the test set towards a central point. Due to this, the relative order of the parties does not change much. Hence, we declare our computational landscape to be stable from this perspective.

Stability of dimensionality reduction We want to know whether the two-dimensional landscape resulting from factor analysis keeps the relative distance between parties intact. Hence, we compare the distance between the parties in the two-dimensional landscape of Figure 3 to the Euclidean distances between the party centroids for the full 768-dimensional semantic space, as is displayed in the distance matrix in Figure 7.

The distance matrix shows that the coalition parties are located close to each other, and further away from any of the opposition parties. This is also visible in our political landscape, where there is a big gap over the first latent factor between the coalition and opposition parties. The parties that are located furthest from the coalition on the political landscape (DENK, PvdD, SP, Volt) also have the largest distances to the coalition in high-dimensional space. Moreover, we see a cluster of opposition parties CDA, CU, D66 and GL-PvdA. In the mean distance matrix, we also see that these parties are close together.

5. Discussion

Our results showed that we were able to create a computational political landscape on the basis of the parliamentary debate speeches for the Schoof government period.

5.1. RQ1 Creation of the Political Landscape

We made a computational political landscape by fine-tuning BERT and reducing dimensionality using factor analysis. The resulting landscape is stable over an unseen test dataset, and the dimensionality reduction keeps relative distances between parties intact.

While analysing the first latent factor in Figure 3, a clear divide between opposition parties and the four coalition parties is visible. When looking at the order of the opposition parties from left to right, we see it closely resembles the left-wing to right-wing axis of Kieskompas's political landscape in Figure 4. It also is logical that the right-wing and centrist parties are closer to the coalition than the left-wing parties, as the coalition is also right-winged (Van Holsteyn and Irwin, 2025).

NSC was closer to the middle in the Kieskompas landscape while in our landscape it was located

on the right of the first dimension. This might be explained by the difference in source material and timeline on which the two political landscapes were created. The Kieskompas used the party programs written before the elections in 2023 while we base our landscape on speeches from the period after the coalition formation where NSC clearly chose the right-oriented direction (van den Berg, 29 November 2024).

We would have expected that JA21 and FVD (the temple icon right under in Figure 4) were placed much further to the right in our political landscape than they actually were, as these are known as strongly right-oriented parties. A possible explanation could be that these parties have the least amount of speech samples as was shown in Figure 1, which limited the speakers' overall contribution to the semantic space.

Analysis of the second latent factor was more difficult. Manual examination of the individual speeches revealed a clear pattern of less grounded speeches that sometimes attack other politicians, to grounded speeches based on arguments. After running LIWC over the speeches, we found that the 'analytical thinking' score corresponds to our findings. Hence, we give the second axis an analytical thinking scale, where a lower second latent value means a higher degree of analytical thinking. Many of the parties that are on the less analytical side (or more intuitive/personal side) of the landscape are populist parties (FVD, JA21, PVV and SP (Rooduijn, 2021). As Zanotto et al. (2024) claimed, populists tend to use rhetorical strategies in their speeches. These rhetorical claims are usually not supported by arguments, and hence these parties end up at the side of the analytical thinking axis representing lower levels of analytical thinking. This is in line with the findings of the study of Jordan et al. (2019) who showed that populist US president Trump is on the lower end of the analytic thinking scale of the LIWC tool.

5.2. RQ2 Schoof in the Political Landscape

Our results showed that both in the high dimensional representation and in the reduced semantic space, Schoof was located most closely to NSC. The Euclidean distance matrix (Figure 7) also indicated that Schoof was positioned close to all coalition parties, while those parties were farthest from DENK, PvdD, SP and Volt, which are all left-oriented. When we soleley look at the first latent factor, the left–right opposition–coalition dimension, we can safely assume that the statements of Schoof were well in line with the statements from the right-oriented coalition parties. However, when we examine the full distance matrix in detail, we find that Schoof is closest to NSC among the coalition parties. The landscape also shows a large distance on the second analytical-thinking dimension between Schoof and PVV. From this perspective, we can conclude that although Schoof was ideologically neutral within the coalition, his analytical-thinking levels aligned closely only with NSC and VVD.

6. Conclusion

We made the following contributions: we created a stable computational model of the Dutch political landscape during the Schoof government period based on debate speeches from the House of Representatives. We validated this model against a model created by an independent research institute and confirmed that both models show similar trends on the socio-economic left–right axis. Finally, our analysis indicates that, based on his parliamentary statements, Prime Minister Schoof can be described as ideologically neutral within the coalition, yet his analytical-thinking style does not align with both BBB and PVV.

Our study has several limitations. The data distribution for the different parties in our model was skewed and not all parties were represented equally. When reducing the multi-dimensional space to two dimensions, many aspects are lost and were not taken into account in our analysis. In the comparison against the Kieskompas landscape, we use both a different data source — political party programs and debate speeches — and a different time frame as the Kieskompas was constructed before the elections and the debates cover the period after the formation process. As an additional validation we could use the political party programs to create a BERT model and compare that against the manually constructed Kieskompas and our current landscape.

We envision several paths for future work. We would like to study party dynamics by zooming in on prominent party members and study how close or distant they are from their average party representation: do we see closely clustered members or are they perhaps closer to other parties? Furthermore, we would be very interested to expand our analysis over a longer timeline and investigate whether we could predict some of the dynamics of changes in the political landscape such as the fusion between GL and PvDA (Van Holsteyn and Irwin, 2025) or the separation by Pieter Omzigt from CDA to start the new party NSC (NOS Nieuws, 20 August 2023).

This study focused on the Dutch parliamentary debates of the Schoof government but the proposed method is certainly easily applicable to datasets from other countries and parliaments as the approach is language independent.

7. Acknowledgments

We would like to thank Prof. James Pennebaker for his advice and suggestions for our study. This work contributed to the HAICu project with file number NWA.1518.22.105 which is financed by the Dutch Research Council (NWO).

8. Bibliographical References

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of LIWC-22. Technical report, University of Texas at Austin, Austin, TX. https://www.liwc.app.

Wietse De Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. *arXiv preprint arXiv:1912.09582*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Daniel Diermeier, Jean-François Godbout, Bei Yu, and Stefan Kaufmann. 2012. Language and ideology in congress. *British Journal of Political Science*, 42(1):31–55.

Kayla N. Jordan, Joanna Sterling, James W. Pennebaker, and Ryan L. Boyd. 2019. Examining long-term trends in politics and culture through language of political leaders and cultural institutions. *Proceedings of the National Academy of Sciences*, 116(9):3476–3481.

KiesKompas BV. 2023. Kieskompas Tweede Kamer verkiezingen in 2023. https://tweedekamer2023.kieskompas.nl/.

Anand Konjengbam, Subrata Ghosh, Nagendra Kumar, and Manish Singh. 2018. Debate Stance Classification Using Word Embeddings. In *Big Data Analytics and Knowledge Discovery*, pages 382–395, Cham. Springer International Publishing.

Duncan McDonnell and Stefano Ondelli. 2022. The Language of Right-Wing Populist Leaders: Not So Simple. *Perspectives on Politics*, 20(3):828–841.

Jayme L. Neiman, Frank J. Gonzalez, Kevin Wilkinson, Kevin B. Smith, and John R. Hibbing. 2016. Speaking different languages or reading from the same script? word usage of democratic and republican politicians. *Political Communication*, 33(2):212–240.

NOS Nieuws. 20 August 2023. Omtzigt doet mee aan verkiezingen met eigen partij: Nieuw Sociaal Contract.

NU.nl. 2 June 2024. Welke partijen zitten voor welke fractie in het Europees Parlement?

Keith T. Poole. 2005. *Spatial Models of Parliamentary Voting*. Analytical Methods for Social Research. Cambridge University Press.

Ludovic Rheault and Christopher Cochrane. 2020. Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora. *Political Analysis*, 28(1):112–133.

Matthijs Rooduijn. 2021. Populisme, Nederland en verkiezingen. https://www.uva.nl/shared-content/faculteiten/nl/faculteit-der-maatschappij-en-gedragswetenschappen/nieuws/2021/02/verkiezingen-populisme-nederland-en-verkiezingen.html.

Martijn Schoonvelde, Anna Brosius, Gijs Schumacher, and Bert N Bakker. 2019. Liberals lecture, conservatives communicate: Analyzing complexity and ideology in 381,609 political speeches. *PloS one*, 14(2):e0208450.

Take Sipma, Marcel Lubbers, Tom Van der Meer, Niels Spierings, Kristof Jacobs, et al. 2021. Versplinterde vertegenwoordiging. Nationaal Kiezersonderzoek 2021.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Charles Spearman. 1904. "General Intelligence" Objectively Determined and Measured.

Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54.

J.Th.J. van den Berg. 29 November 2024. Het treurig lot van NSC. https://www.parlement.com/column/202411/het-treurig-lot-van-nsc.

Joop JM Van Holsteyn and Galen A Irwin. 2025. The Dutch parliamentary elections of November 2023. *West European Politics*, 48(2):464–477.

Matthew Wall, André Krouwel, and Thomas Vitiello. 2014. Do voters follow the recommendations of voter advice application websites? A study of the effects of kieskompas.nl on its users' vote choices in the 2010 Dutch legislative elections. *Party Politics*, 20(3):416–428.

Sergio E. Zanotto, Diego Frassinelli, and Miriam Butt. 2024. Language Complexity in Populist Rhetoric. In *Proceedings of the 4th Workshop on Computational Linguistics for the Political and Social Sciences: Long and short papers*, pages 61–80, Vienna, Austria. Association for Computational Linguistics.

9. Language Resource References

Tweede Kamer der Staten-Generaal. 2024. *Plenaire verslagen*. Retrieved December 22, 2024 from https://www.tweedekamer.nl/kamerstukken/plenaire_verslagen.

A. Appendix A. Acronyms of political parties

Acronym	Party name	English translation	EP group
BBB	BoerBurgerBeweging	Farmer-Citizen Movement	EPP
CDA	Christen-Democratisch Appèl	Christian Democratic Appeal	EPP
CU	ChristenUnie	Christian Union	EPP
D66	Democraten 66	Democrats 66	RE
DENK	DENK	DENK	-
FVD	Forum voor Democratie	Forum for Democracy	NI
GL-PvdA	GroenLinks - Partij van de Arbeid	GreenLeft - Labour Party	G/EFA - S&D
JA21	Het Juiste Antwoord '21	The Correct Answer '21	ECR
NSC	Nieuw Sociaal Contract	New Social Contract	EPP
PvdD	Partij voor de Dieren	Party for the Animals	Left
PVV	Partij voor de Vrijheid	Party for Freedom	PfE
SGP	Staatkundig Gereformeerde Partij	Reformed Political Party	ECR
SP	Socialistische Partij	Socialist Party	Left
Volt	Volt	Volt	G/EFA
VVD	Volkspartij voor Vrijheid en Democratie	People's Party for Freedom and Democracy	RE

Table 2: The full version of the acronym for each party name, the translation of the party name to English, and their European Parliament (EP) group as of the start of the Schoof government in 2024 (NU.nl, 2 June 2024).

Author Index