Italian Political Science Review/Rivista Italiana di Scienza Politica (2025), **55**, 77–86
doi:10.1017/ipo.2025.6

IPSR RISP

RESEARCH NOTE: DATASET

A new database for Italian parliamentary speeches: introducing the *ItaParlCorpus* dataset

Joshua Cova

Max Planck Institute for the Study of Societies, Cologne, Germany
Email: joshua.cova@mpifg.de

(Received 3 September 2024; revised 16 February 2025; accepted 18 February 2025)

Abstract

A common challenge in studying Italian parliamentary discourse is the lack of accessible, machine-readable, and systematized parliamentary data. To address this, this article introduces the *ItaParlCorpus* dataset, a new, annotated, machine-readable collection of Italian parliamentary plenary speeches for the Camera dei Deputati, the lower house of Parliament, spanning from 1948 to 2022. This dataset encompasses 470 million words and 2.4 million speeches delivered by 5830 unique speakers representing 77 different political parties. The files are designed for easy processing and analysis using widely-used programming languages, and they include metadata such as speaker identification and party affiliation. This opens up opportunities for in-depth analyses on a variety of topics related to parliamentary behavior, elite rhetoric, and the salience of political themes, exploring how these vary across party families and over time.

Keywords: Italy; parliament; political parties; research methods; text analysis

Introduction

As big data and quantitative text analysis techniques have advanced, political scientists have identified in parliamentary speeches, rich data collections that explicitly illustrate policymakers' preferences, a valuable tool for analyzing political parties' stances on a wide range of issues (Rauh and Schwalbach, 2020; Sebők *et al.*, 2025). Among many different applications, parliamentary speeches have been found to provide an excellent source of data for examining policymakers' policy preferences, tracking shifts in political discourse over time, analyzing how elite rhetoric varies across party families, studying the salience of different policies, and conducting sentiment analyses (Proksch and Slapin, 2015). While parliamentary data are readily available for several countries, researchers focusing on Italian politics have often faced challenges due to the limited availability of machine-readable parliamentary texts.

This article introduces the *ItaParlCorpus* dataset, a new comprehensive, annotated, and machine-readable dataset of parliamentary speeches for Italy's lower house of parliament, the *Camera dei Deputati*, for the period 1948–2022. This new database, which covers the parliamentary plenary debates of 18 legislatures includes over 470 million words, 2.4 million interventions, from 5830 unique speakers representing 77 different political parties and parliamentary groups. Scholars of Italian politics, who have previously faced challenges when analyzing parliamentary debates due to the prevalence of non-machine-readable scans and poor data quality, can now benefit from this resource. The *ItaParlCorpus* provides structured .csv files, with each row containing information on the name of the speaker, party affiliation, and the date of each

parliamentary intervention. Additionally, this corpus of parliamentary debates includes unique identifiers for parliamentarians, which can be linked to the frequently employed Comparative Legislators Database (Göbel and Munzert, 2022); a dataset, which provides extensive socio-demographic information on parliamentarians. The large .csv files which make up the *ItaParlCorpus* dataset can easily be processed and analyzed with common programming languages such as R and Python, allowing researchers to understand how the salience of different political themes has changed in time and between parties.

This article is structured as follows. First, I review various efforts to digitize and annotate parliamentary debates in Italy and other democracies and the range of research questions that have recently been addressed by using corpora of parliamentary debates. Second, I detail the data collection process for the *ItaParlCorpus* database and present the structure of the dataset. The third section illustrates a few concrete applications of what one can do with this new database. It does so by analyzing how the topics of abortion and the mafia have been discussed in Italian parliamentary debates, both over time and across different party groups. Finally, the last section concludes by summarizing the added value this database could provide to researchers studying Italian politics.

Parliamentary speeches in big data and natural language processing research

With the growing availability of digitized textual data, political science researchers seeking to assess policymakers' ideological positions can now draw on a wide range of empirical sources, from social media posts to electoral manifestos. Consequently, text data have also been widely used to analyze Italian politics. For instance, Ceron (2024) employs a text analysis of Italian presidents' investiture speeches and television addresses to explore how Italian presidents' ideological leanings manifest themselves in these settings. Quantitative text analysis has also been employed to investigate Italian social media texts and have helped address a variety of different research questions. These include exploring levels of intra-party conflicts (Ceron, 2017), analyzing immigration-related discussions (De Rosa *et al.*, 2021), and forecasting electoral outcomes (Caldarelli *et al.*, 2014). Additionally, scholars of Italian politics have employed newspaper corpora as a resource to analyze how policymakers adopted different policy narratives, for example during the Euro crisis (Bobba and Seddone, 2018) or the COVID-19 pandemic (Crabu *et al.*, 2021).

In this context, researchers have shown growing interest in leveraging collections of parliamentary speeches to study political actors' policy preferences. Using parliamentary speeches makes good sense as parliaments are crucial venues for the formulation of policy agendas, as they constitute the formal institution where parties compete to fulfill their representative mandates in between elections and thus a *forum* in which policymakers can not only signal policy positions toward other parties, but also to the electorate as a whole (Vliegenthart *et al.*, 2013; Proksch and Slapin, 2015). However, researchers' ability to draw inferences on policymakers' positions on a range of political questions is predicated upon the ease with which it is possible to access and analyze these texts. Parliamentary speeches have proven to be a rather challenging source of textual data as parliamentary records are often not adequately digitized and data quality frequently deteriorates the further back in time one goes. As noted by Sebők *et al.* (2025) for several European democracies, there exist significant hurdles in obtaining data that can then be readily parsed and analyzed using commonly used programming languages. This means that researchers must invest significant efforts in scraping parliamentary debates from national parliaments' websites and even so, scraped files can often not be readily processed and analyzed because of poor data quality or inconsistent data infrastructures and naming conventions.

However, when accessible, legislative data from parliamentary corpora have become an increasingly valuable resource, widely applied in diverse studies utilizing various methodological designs. Thus, parliamentary corpora have been used to assess the salience of policy issues

(Greene and Cross, 2017; Cova and Schmitz, 2024), conduct sentiment analyses (Proksch *et al.*, 2019), and investigate speech complexity in elite rhetoric (Osnabrügge *et al.*, 2021), for example. Additionally, quantitative text analyses of legislative debates have explored how different parliamentary rules shape legislative behavior, both in comparative contexts and specifically for Italy (Giannetti and Pedrazzani, 2021). Researchers have also applied text analysis to corpora of parliamentary debates to investigate how the socio-demographic characteristics of parliamentarians influence the likelihood of certain policy themes being discussed. Numerous studies have thus illustrated the extent to which personal and socio-demographic factors, such as gender and socio-economic background, shape legislative behavior, including variations in speaking time and topic selection (Bäck and Debus, 2019; O'Grady, 2019).

Considering the numerous research questions that can be addressed using parliamentary speeches as well as the different automated text analysis methods that can be applied, recent years have seen increased efforts to publish machine-readable collections of parliamentary speeches, which have taken the form of country-specific as well as cross-national projects. For instance, Remschel and Kroeber (2022) released a comprehensive dataset on German parliamentary proceedings that includes not only parliamentary speeches but also other types of parliamentary data, such as bills, written responses, communications, requests, and replies. Beelen *et al.* (2017) have harmonized and standardized Canadian parliamentary archives, dating back to the 19th century, into machine-readable formats. Recent years have also marked the emergence of comparative projects, which have sought to map out and harmonize digital infrastructures, including parliamentary corpora (Erjavec *et al.*, 2023). Most notably, the widely used *ParlSpeech* dataset (Rauh and Schwalbach, 2020) contains full-text corpora of parliamentary speeches from various advanced democracies over the past two to three decades and has been extensively employed in comparative analyses of legislative activities and parliamentary behavior. However, Italy is not represented in the dataset.

As far as the Italian case is concerned, as part of the broader CLARIN project, which aims to collect machine-readable and annotated corpora of European countries' parliamentary proceedings, Agnoloni *et al.* (2022) have collected 79,000 speeches containing ca. 31 million words for the *Senato* for 2013–2020. While this constitutes an important achievement, the limited timespan offered by the database limits the possibility of conducting historical analyses. Within this context, it is also important to mention the Italian Legislative Speech Dataset: a historical collection (1946–2022) of investiture speeches delivered in Parliament during votes of confidence.[1] The texts are segmented into quasi-sentences to facilitate classification of parliamentary interventions based on the salience of different policy themes. While this dataset is a valuable resource, its focus on investiture speeches restricts its scope, thus limiting opportunities to analyze a broader range of parliamentary speeches that are reflective of ordinary legislative interactions. Due to the challenges of obtaining machine-readable plenary speeches, researchers studying Italian parliamentary behavior and political practices have shifted their focus to other types of discourse recorded in parliament. To name a few examples, Cavalieri and Froio (2022) examine the behavior of populist parties through parliamentary questions, while Salvati (2021) analyzes prime ministers' speeches during votes of confidence from 1994 onward.

The dataset

The *ItaParlCorpus* dataset is a comprehensive, annotated, and machine-readable database of Italy's parliamentary speeches spanning from 1948 to 2022. It provides data on parliamentary plenary transcripts, which allows researchers to investigate a wide range of topics, including party positioning on various policy areas, elite rhetoric, and the salience of different issues within the Italian parliamentary context. The dataset as well as the codebook are open access and are

[1]See, "ILSD: Italian Legislative Speech Dataset" (https://andreaceron.com/projects/ilsd/).

precisazione che ho sentito il dovere di fare (Interruzione del deputato Villa). Si è arrivati a questo dibattito con una parola d'ordine; « Niente trattative ». VILLA. Quello che hai detto era atteso da tutto il paese! (Proteste del deputato Pinto). PRESIDENTE. Onorevole Pinto, ella ha a disposizione un tempo limitato: se ne serva per dire quello che ha da dire. PINTO. La parola d'ordine era: « Niente trattative ». Non sono d'accordo, in primo luogo per il modo cinico, che non è inconsapevole, ma che è consapevole, con cui si rischia di determinare la morte dell'onorevole Moro. Può sembrare strano che io, che vedevo e vedo ancora in Moro (perché spero che non sia morto) uno dei leaders autorevoli e indiscussi di quel partito che voglio continuare a combattere fermamente e al quale voglio contribuire a togliere il potere, mi debba fare paladino e difensore della vita dell'onorevole Aldo Moro. Un giornale scrive che forse si è scelta la strada più facile: quella di mostrare al paese e al popolo che lo Stato democratico non viene sconfitto. Ma io penso che non in questo modo si eviti la sconfitta dello Stato democratico, che è stato più volte sconfitto, anche se non abbattuto. Non intendo, infatti, spazzar via le conquiste che la classe operaia ha saputo ottenere in questi trent'anni. Lo Stato democratico, tuttavia, è stato sconfitto più volte: è stato sconfitto quando le libertà costituzionali sono state violate. Non è demagogia, non è populismo dire che esso è stato sconfitto ogni volta che un emigrante lasciava il sud e andava all'estero, quando i compagni venivano ammazzati sulle strade, quando le leggi non venivano applicate, quando i vostri sostenitori, colleghi della democrazia cristiana, portavano i soldi all'estero, quando morivano nelle fabbriche operai di cui non si ricorda il nome, quando avvenivano il dramma di Seveso e le alluvioni o il terremoto del Friuli, e si gestivano in un parti. colare modo anche le disgrazie naturali e umane. Non penso che con questa risposta diamo un'immagine riabilitata e forte dello Stato. Ho sentito parlare di ritorno alle libertà, di ritorno alla pace. Non vi può essere pace se non abbiamo la capacità di vedere il perché delle azioni delle Brigate rosse, il perché esse oggi esistano. È strano che si affermi che con esse non si tratta perché in tal modo si darebbe loro un riconoscimento. Credo di aver capito, collega Bozzi, che questo ella intendesse dire. È inutile nasconderci la realtà. Non è vero che, se non trattiamo, le Brigate rosse non esistono o non sono riconosciute. Il loro riconoscimento sono i loro colpi d'arma da fuoco e i morti che hanno seminato, Esse esistono, e non dobbiamo pensare di dimostrare che non esistono nel momento PAJETTA. Il loro riconoscimento è anche il processo! PINTO. Scusa, Pajetta, con te sono più educato. Il modo con cui si è risposto all'azione delle Brigate rosse, la campagna del tipo «caccia alle streghe » nei confronti dei « sostenitori » e dei « simpatizzanti » dimostrano che, anche se nessuno parla di leggi speciali, le leggi speciali esistono. Si sono fatti i proclami, si sono avuti gli interventi dell'onorevole Ugo La Malfa, il quale riteneva che con la pena

Figure 1. Original .txt files of parliamentary discussions on April 4, 1978.

freely available on the Harvard dataverse.[2] In what follows, I briefly describe the construction of the database and the operationalization of the variables.

As discussed above, a significant hurdle for researchers studying historical parliamentary speeches is the poor quality of available material, which largely depends on existing digitization efforts and the availability of archival resources. In the Italian case, the *Camera dei Deputati* only offers images and scans for records prior to 1996. The further back one goes, the more unstructured do these files tend to be. The first necessary step is thus converting these scans and images into text files using optical character recognition (OCR) technology. This prior invaluable work was carried out by Frasnelli and Aprosio (2024), who made the .txt files available on a GitHub repository. However, these text files are unstructured as illustrated by the example (Figure 1) and therefore are not readily utilizable by researchers interested in substantive political science questions. Moreover, a common challenge when using OCR technology for text files is ensuring the accurate conversion of scans into readable text. In the Appendix, I assess data quality by employing the Italian Hunspell spellchecker to identify improperly converted words. As illustrated in Figure A2, the proportion of misspelled words in parliamentary interventions remains relatively low, typically ranging between 1 and 3% across most years.

The unstructured text files derived from the OCR scans present a challenge because they do not clearly separate the content of parliamentary speeches from the identities of the speakers. As illustrated in Figure 1, the speech content and the names of parliamentarians appear on the same line. Once the repository of these unstructured text files is downloaded, the task is thus to accurately differentiate between speakers' ID and the speech content that is associated to them. This distinction is essential for political science research, as correctly linking party affiliation to specific speeches is crucial for analyzing substantive policy discussions. As shown in Figure 1, parliamentarians are indicated in the text files using words which only contain capitalized letters. However, uncritically relying solely on words with capitalized letters to identify speakers is problematic because not all such words can be tied back to parliamentarians (e.g. acronyms or Roman numerals). Furthermore, inconsistent naming conventions add complexity: in some legislatures, only surnames are used, while in others, names may appear as "first name-surname" or "surname-first name". Another complication is that not all parliamentary interventions are from members of the *Camera dei Deputati*. Invited speakers, such as ministers, technocrats without parliamentary seats, and members of the *Senato*, also intervene in plenary debates.

To determine the party affiliations of speakers in the text files, I utilized a list of Italian legislators from the Comparative Legislators Database (Göbel and Munzert, 2022). This multinational dataset includes detailed information on national-level policymakers, such as their names, party affiliation as well as socio-demographic data. To ensure that I have accurately identified government members who do not serve in parliament, I supplemented this data by scraping

[2]See, *ItaParlCorpus* (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/KUARWD).

date	doc_id	row_id	legislati	speaker	pageid_wiki	party_nam	party_family	party_id_parlgov	party_id_itaparl	chair	cabinet	text
4/4/1978	19780404_	19780404_	7	BIAGIO PINTC	5059706	Partito Repubbl	Liberal	93	53	FALSE	FALSE	PINTO. Signor Presidente del Consiglio, anch'io non mi reputo sodisfatto dell
4/4/1978	19780404_	19780404_	7	RUGGERO VIL	428-miss	Democrazia Cr	Christian democ	1633	13	FALSE	FALSE	VILLA. Quello che hai detto era atteso da tutto il paese! (Proteste del deputa
4/4/1978	19780404_	19780404_	7	PRESIDENTE		Chair	PRESIDENTE		11	TRUE	FALSE	PRESIDENTE. Onorevole Pinto, ella ha a disposizione un tempo limitato: se n
4/4/1978	19780404_	19780404_	7	BIAGIO PINTC	5059706	Partito Repubbl	Liberal	93	53	FALSE	FALSE	PINTO. La parola d'ordine era: << Niente trattative >>. Non sono d'accordo,
4/4/1978	19780404_	19780404_	7	GIAN CARLO I	128533	Partito Comun	Communist/Far-	1088	39	FALSE	FALSE	PAJETTA. Il loro riconoscimento e anche il processo!
4/4/1978	19780404_	19780404_	7	BIAGIO PINTC	5059706	Partito Repubbl	Liberal	93	53	FALSE	FALSE	PINTO. Scusa, Pajetta, con te sono piu educato. Il modo con cui si e risposto

Figure 2. Converted .csv files of parliamentary discussions on April 4, 1978.

the Italian government's website to collect the names of cabinet members across different governments.[3] Additionally, I incorporated data from prior data collections on Italian technocrats (Improta, 2021). This comprehensive approach allowed me to compile a complete list of legislators, encompassing both parliamentarians and non-elected government officials, enabling precise separation of speakers from speech content. In order to address discrepancies between the text files and the compiled list of speakers, I conducted manual checks across legislative periods. Mismatches typically arose from differences in naming conventions, such as the use of middle names or maiden vs. married names in one source but not the other. The Appendix provides further information on how these manual checks were conducted.

The final output is a series of .csv files, in which every row corresponds to a parliamentary intervention/speech, recording the following information: the day in which the speech took place (*date*), the year (*year*), a document identifier (*doc_id*), a unique row identifier (row_id), the legislature (*legislature*), the speaker's name (*speaker*), the speaker's unique numerical identifier (*pageid_wiki*), which coincides with that used by the Comparative Legislators Database (Göbel and Munzert, 2022), the party name (*party_name*), the party family (*party_family*) to which the party belongs to as recorded in the ParlGov Database (Döring and Manow, 2024), the ParlGov unique numerical party identifier (*party_id_parlgov*), the *ItaParlCorpus* unique numerical party identifier (*party_id_itaparl*), a Boolean variable denoting whether the speaker is the chair (*chair*), another variable, which denotes whether the speaker is a cabinet member (*cabinet*) without being recorded as a member of the *Camera dei Deputati* (i.e. a member of the Senate or a technocrat), and finally the raw text (*text*). An example of the final output is shown in Figure 2.

Potential applications

The *ItaParlCorpus* database enables researchers to perform quantitative text analyses of Italy's parliamentary discourse, offering valuable insights for various subfields of political science and related disciplines. Beyond the raw text data, the database also provides contextual political information, such as party affiliation, which supports the development of more sophisticated research designs. Here, I present some brief examples to demonstrate potential applications of this new corpus of Italian parliamentary speeches. The analysis focuses on two particularly prominent and contentious topics in Italy's post-war republican history: abortion and the mafia. Specifically, it examines the salience of these topics in the parliamentary context, as well as the manner in which these issues have been debated by policymakers. To be clear, the purpose of this article is not to provide a substantive analysis of these well-studied topics, but rather to illustrate how this corpus can be used in ways that may be of interest to scholars of Italian politics.

In many European democracies, the struggle for women's reproductive rights, particularly access to safe and legal abortion, emerged as a pivotal civil rights issue in the latter half of the 20th century. In Italy, the introduction of abortion rights in 1978 through the *Legge* 194 marked a watershed moment, granting women the legal right to terminate a pregnancy within the first 12 weeks. At the time, this was a deeply contentious issue, influenced also significantly by the

[3]See, *Presidenza del Consiglio dei Ministri*, "I Governi nelle Legislature" (https://www.governo.it/it/i-governi-dal-1943-ad-oggi/i-governi-nelle-legislature/192).

Catholic Church's prominent role in Italian political life. While left-wing parties and lay, liberal-centrist groups such as the Partito Liberale Italiano and the Partito Repubblicano Italiano supported the legislation, the Christian Democrats (DC) faced internal divisions, and the far-right Movimento Sociale Italiano opposed it.

A very different, but very politically salient issue which significantly shaped Italy's post-war history is that of the mafia. For much of this period, the mafia was a topic rarely addressed publicly by political elites, who often downplayed the infiltration of organized crime in the upper echelons of power. However, escalating public displays of violence, culminating in the high-profile assassinations of judges and politicians during the 1980s and 1990s, forced a shift in political discourse and policy action.

To explore the changing salience of these two key terms using the *ItaParlCorpus* database, I examine the share of parliamentary interventions discussing the terms "abortion" and "mafia" in the period 1948–1992 (Figure 3). I operationalize salience as the share of parliamentary interventions discussing these topics as a share of the total number of parliamentary interventions made by party groups. Abortion emerges as a highly salient topic, with nearly 10% of parliamentary interventions addressing this topic at its peak. Notably, the center-left and the PCI (Italian Communist Party) demonstrated significantly higher salience on this issue compared to the DC. In contrast, discussions of the mafia exhibited low salience during the early decades, gradually increasing in prominence throughout the 1980s and 1990s. While this analysis is indicative of the parliamentary salience of these terms and which political party emphasized these issues more than others, the next step is to examine the way in which parliamentary discourse has changed. In this article, I showcase two different analytical approaches.

To analyze how discussions of the mafia evolved in parliamentary discourse, I conduct a descriptive analysis of the words most frequently employed in these debates. To do this, I extract *sentences* from parliamentary interventions that explicitly mention the mafia and conduct my analysis across four distinct historical periods: 1948–1959, 1960–1979, 1980–1999, and 2000–2022. As highlighted in Figures 4 and 5, one can observe notable temporal shifts in the most frequently used nouns and adjectives in sentences in which politicians discuss the mafia

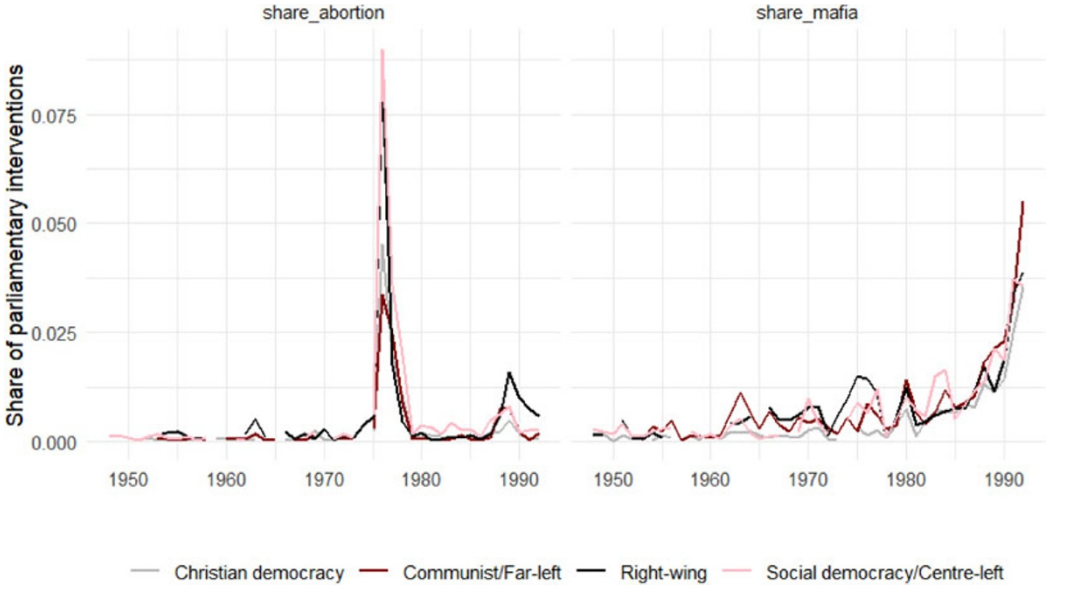

Figure 3. Share of parliamentary interventions discussing abortion (left) or the mafia (right) as a share of all parliamentary interventions by party (1948–1992).

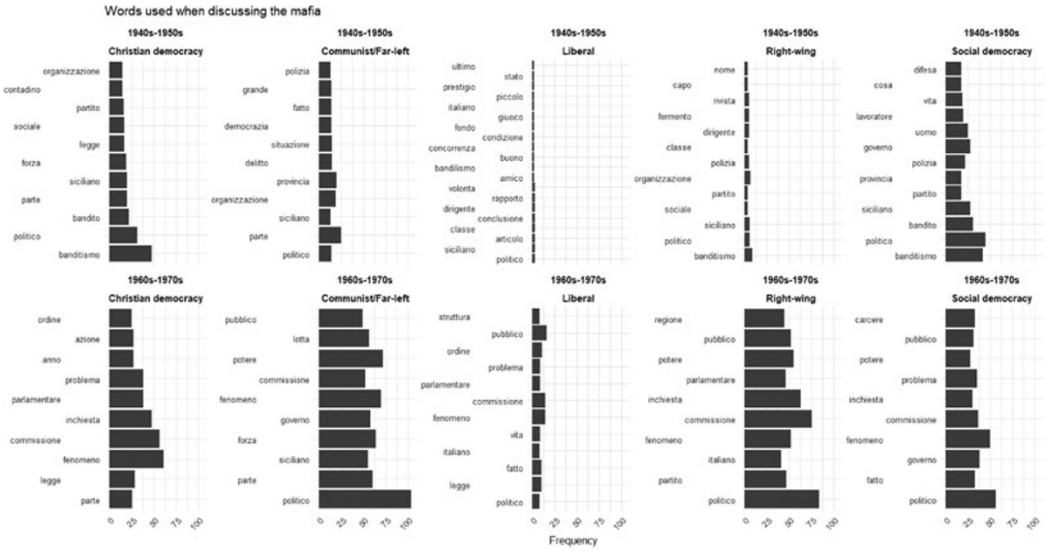

Figure 4. Most common nouns and adjectives used by political party families when discussing the mafia (1948–1979).

in parliament. During the initial period under analysis (1948–1959), parliamentary discussions frequently associated the mafia with banditry (*banditismo*) while also highlighting possible political connections. As the 1960s and 1970s unfolded, possibly mirroring a heightened awareness of the mafia, discussions across party lines increasingly incorporated terms such as "phenomenon," "enquiry," and "problem." In the 1980s and 1990s, as the mafia emerged as a more prominent and widely discussed topic, there seems to have been greater alignment in the terminology

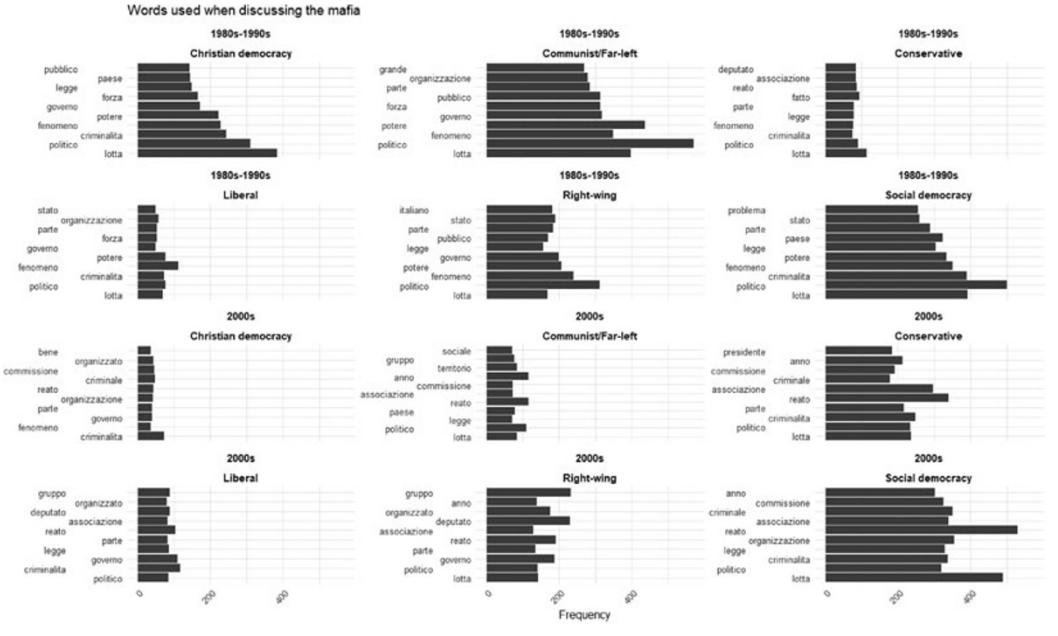

Figure 5. Most common nouns and adjectives used by political party families when discussing the mafia (1980–2022).

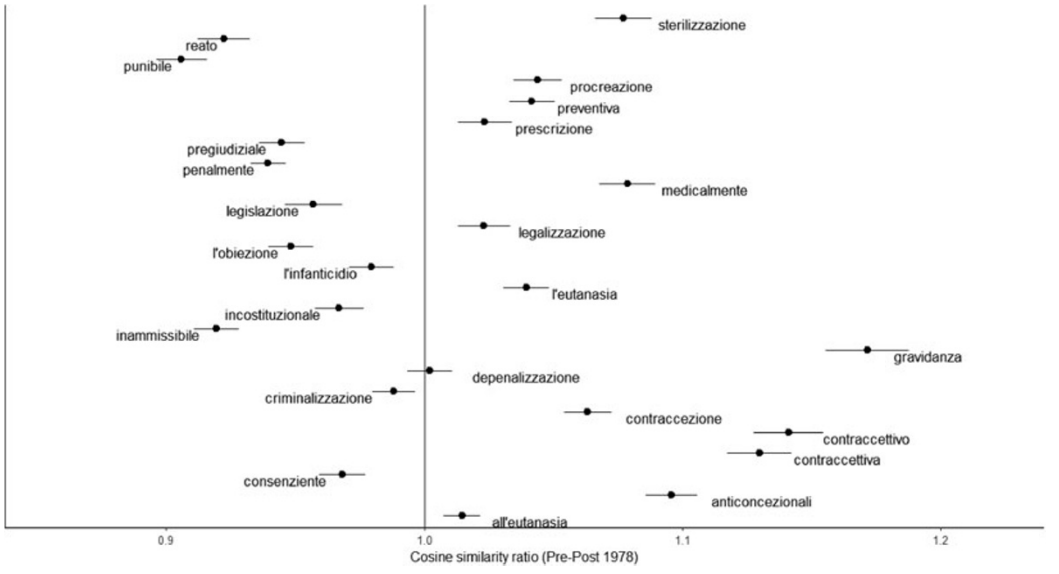

Figure 6. Cosine similarity for speeches discussing abortion before and after the introduction of the *Legge* 194 (1978).

employed by various political groups. This period also saw a growing emphasis on the term "public," possibly reflecting increased awareness about the mafia's impact on society. Finally, in the 2000s, parliamentary language shifted toward a more legalistic framework, with terms like "criminality," "laws," and "criminal offense" becoming more prominent, signaling a possible focus on institutional and judicial responses to organized crime.

While the analysis presented above has relied on frequency-based methods, in recent years, political scientists have increasingly adopted more sophisticated text analysis techniques. Leveraging advances in natural language processing (NLP), such as word embeddings, researchers can in fact examine semantic relationships within large text corpora, capturing subtler connections and meanings that simple word counts or co-occurrence metrics might miss. As highlighted in the recent work of Rodriguez *et al.* (2023), one specific application is the use of cosine similarity, which leverages word embeddings to identify the most distinctive words across different groups or categories.

In this context, a cosine similarity analysis of parliamentary interventions on abortion before and after the introduction of Law 194 in 1978 highlights the most distinctive terms of each period. Interestingly, as illustrated in Figure 6, parliamentary discourse prior to 1978 often emphasized topics and frames related to crime and morality. In contrast, post-1978 discussions shifted toward a medicalized vocabulary, focusing on terms such as contraceptives and sterilization, as well as broader bioethical issues like euthanasia.

Conclusion

This article introduces the *ItaParlCorpus* database, a new resource for studying Italian politics. The dataset covers Italy's post-war republican period (1948–2022) and consists of a machine-readable corpus of parliamentary plenary speeches, which can be easily processed and analyzed using popular programming languages, such as R and Python. The *ItaParlCorpus* is an extensive dataset, featuring over 2.4 million parliamentary speeches across 18 different legislatures, along with metadata that includes speaker identification and party affiliation.

As large text corpora gain prominence in empirical political science, digitized records of parliamentary debates have become essential for exploring key questions about party positioning on policy issues. Beyond assessing the salience of political topics and tracking the evolution of discourse, this corpus can be integrated with other datasets, such as the Comparative Legislators Database, which provides detailed socio-demographic and electoral data on parliamentarians. Such integration would for example enable investigations into how parliamentarians' socio-demographic profiles influence their legislative activities.

The *ItaParlCorpus* dataset also facilitates research into how party positioning (e.g. on the left-right or the GAL-TAN dimension of political competition) affect parties' emphasis on particular policy issues. Moreover, advanced NLP tools can be employed to analyze levels of party conflict over time, both within and between parties, thus offering insights into the changing dynamics of parliamentary discourse (see e.g. Rheault and Cochrane, 2020). Beyond political science, the large-scale digitization of texts and the computational, quantitative study of texts has become an emerging trend in cultural studies and linguistics (Michel *et al.*, 2011). In recent years, comparative political scientists have increasingly turned to parliamentary debates to explore questions of political representation and examine how issue salience evolves over time. For these analyses, researchers have relied on comparative datasets of parliamentary corpora, such as *ParlSpeech* (Rauh and Schwalbach, 2020) and *ParlaMint* (Erjavec *et al.*, 2023). However, Italy is either absent or has a limited time series in these datasets. As such, the *ItaParlCorpus* dataset emerges as a valuable resource, holding relevance across multiple disciplines and offering rich potential for scholarly investigations into Italy's post-war political language.

Funding. This research received no specific grant from any public or private funding agency.

Data. The *ItaParlCorpus* dataset is available at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/KUARWD and the replication files are available at: https://dataverse.harvard.edu/dataverse/ipsr-risp.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/ipo.2025.6.

Acknowledgments. I would like to thank the editor and the two anonymous reviewers for their helpful feedback and their constructive comments, which have significantly improved this article and the associated dataset. I would also like to thank Gabriele Beretta for his support and encouragement.

Competing interests. The author declares none.

References

Agnoloni T, Bartolini R, Frontini F, Montemagni S, Marchetti C, Quochi V, Ruisi M and Venturi G (2022) Making Italian parliamentary records machine-actionable: the construction of the ParlaMint-IT corpus. *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference* (pp. 117–124).

Bäck H and Debus M (2019) When do women speak? A comparative analysis of the role of gender in legislative debates. *Political Studies* **67**, 576–596.

Beelen K, Thijm TA, Cochrane C, Halvemaan K, Hirst G, Kimmins M, Lijbrink S, Marx M, Naderi N, Rheault L and Polyanovsky R (2017) Digitization of the Canadian parliamentary debates. *Canadian Journal of Political Science/Revue canadienne de science politique* **50**, 849–864.

Bobba G and Seddone A (2018) How do Eurosceptic parties and economic crisis affect news coverage of the European Union? Evidence from the 2014 European elections in Italy. *European Politics and Society* **19**, 147–165.

Caldarelli G, Chessa A, Pammolli F, Pompa G, Puliga M, Riccaboni M and Riotta G (2014) A multi-level geographical study of Italian political elections from Twitter data. *PLoS One* **9**, e95809.

Cavalieri A and Froio C (2022) The behaviour of populist parties in parliament. The policy agendas of populist and other political parties in the Italian question time. *Italian Political Science Review/Rivista Italiana di Scienza Politica* **52**, 283–296.

Ceron A (2017) Intra-party politics in 140 characters. *Party Politics* **23**, 7–17.

Ceron A (2024) Ideological disagreement and the rejection of laws by Italian heads of state. *Swiss Political Science Review* **30**, 66–79.

Cova J and Schmitz L (2024) *A Primer for the Use of Classifier and Generative Large Language Models in Social Science Research (No. r3qng_v1)*. Center for Open Science.

Crabu S, Giardullo P, Sciandra A and Neresini F (2021) Politics overwhelms science in the COVID-19 pandemic: evidence from the whole coverage of the Italian quality newspapers. *PLoS One* **16**, e0252034.

De Rosa AS, Bocci E, Bonito M and Salvati M (2021) Twitter as social media arena for polarised social representations about the (im)migration: the controversial discourse in the Italian and international political frame. *Migration Studies* **9**, 1167–1194.

Döring H and Manow P (2024) ParlGov 2024 Release V1, Harvard Dataverse, https://doi.org/10.7910/DVN/2VZ5ZC

Erjavec T, Ogrodniczuk M, Osenova P, Ljubešić N, Simov K, Pančur A, Rudolf M, Kopp M, Barkarson S, Steingrímsson S, Çöltekin Ç, de Does J, Depuydt K, Agnoloni T, Venturi G, Pérez MC, de Macedo LD, Navarretta C, Luxardo G, Coole M, Rayson P, Morkevičius V, Krilavičius T, Darǵis R, Ring O, van Heusden R, Marx M and Fišer D (2023) The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation* **57**, 415–448.

Frasnelli V and Aprosio AP (2024) There's something new about the Italian Parliament: the IPSA corpus. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 16037–16046).

Giannetti D and Pedrazzani A (2021) Italy: legislative speeches under changing electoral rules. In Back H, Debus M and Fernandes JM (eds), *The Politics of Legislative Debates*. Oxford: Oxford University Press, pp. 505–527.

Göbel S and Munzert S (2022) The comparative legislators database. *British Journal of Political Science* **52**, 1398–1408.

Greene D and Cross JP (2017) Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Analysis* **25**, 77–94.

Improta M (2021) Inside technocracy: features and trajectories of technocratic ministers in Italy (1948–2021). *Italian Political Science* **16**, 220–240.

Michel J-B, Shen YK, Aiden AP, Veres A, Gray MK, Google Books Team, Pickett JP, Hoiberg D, Clancy D, Norvig P, Orwant J, Pinker S, Nowak MA and Aiden EL (2011) Quantitative analysis of culture using millions of digitized books. *Science* **331**, 176–182.

O'Grady T (2019) Careerists versus coal-miners: welfare reforms and the substantive representation of social groups in the British Labour Party. *Comparative Political Studies* **52**, 544–578.

Osnabrügge M, Hobolt SB and Rodon T (2021) Playing to the gallery: emotive rhetoric in parliaments. *American Political Science Review* **115**, 885–899.

Proksch S-O and Slapin JB (2015) *The Politics of Parliamentary Debate*. Cambridge: Cambridge University Press.

Proksch S-O, Lowe W, Wäckerle J and Soroka S (2019) Multilingual sentiment analysis: a new approach to measuring conflict in legislative speeches. *Legislative Studies Quarterly* **44**, 97–131.

Rauh C and Schwalbach J (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* **1**, 1.

Remschel T and Kroeber C (2022) Every single word: a new data set including all parliamentary materials published in Germany. *Government and Opposition* **57**, 276–295.

Rheault L and Cochrane C (2020) Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis* **28**, 112–133.

Rodriguez PL, Spirling A and Stewart BM (2023) Embedding regression: models for context-specific description and inference. *American Political Science Review* **117**, 1255–1274.

Salvati E (2021) Politicization and conflict in the relationship with the European Union: an analysis of Italian Prime Ministers' parliamentary speeches. *Italian Political Science Review/Rivista Italiana di Scienza Politica* **51**, 1–24.

Sebők M, Proksch S-O, Rauh C, Visnovitz P, Balázs G and Schwalbach J (2025) Comparative European legislative research in the age of large-scale computational text analysis: a review article. *International Political Science Review* **46**, 18–39.

Vliegenthart R, Walgrave S and Zicha B (2013) How preferences, information and institutions interactively drive agenda-setting: questions in the Belgian parliament, 1993–2000. *European Journal of Political Research* **52**, 390–418.

Cite this article: Cova J (2025). A new database for Italian parliamentary speeches: introducing the *ItaParlCorpus* dataset. *Italian Political Science Review/Rivista Italiana di Scienza Politica* **55**, 77–86. https://doi.org/10.1017/ipo.2025.6