Codebook

Introducing ItaParlCorpus: A new annotated corpus for the Italian Parliament (1948-2022)

Joshua Cova*

2025-02-01

Introduction

As more advanced tools for quantitative text analysis and natural language processing (NLP) continue to emerge, political science researchers are increasingly utilizing techniques such as classifier models and large language models to address a broad spectrum of research questions. These include estimating the ideological positions of political parties, identifying and classifying policy topics and themes, and analyzing the linguistic complexity of elite rhetoric (Carammia, Iacus, and Porro 2024; Cova and Schmitz 2024). However, the ability to accurately assess and infer policymakers' ideas and positions depends on the prior acquisition of extensive text corpora in which policymakers' ideas are made apparent. This necessity has driven significant efforts in recent years to collect, standardize, and harmonize parliamentary corpora — vast collections of texts that can capture how political contestation changes in time. There have been country-specific efforts (Abrami et al. 2022; Richter et al. 2023; Marx, Schuth, et al. 2010) as well as comparative projects, which have harmonized parliamentary corpora from different countries (Rauh and Schwalbach 2020; Erjavec et al. 2023; Sylvester et al. 2023). These undertakings do, however, hinge upon researchers' ability to draw from digitized archival material from national parliamentary debates.

Not only does the availability of digitized parliamentary texts vary significantly between countries, but the quality of this data is also likely to diminish the further back in time one goes. This is certainly the case for Italy's parliament; indeed, to the best of the author's knowledge, the only annotated data collection available for the Italian Parliament is as part of the *Parlamint* project, which collects data for Italy's upper chamber of parliament, the *Senato*, for the years 2013-2020 (Erjavec et al. 2023; Agnoloni et al. 2022).

To address the paucity of available Italian parliamentary data, I introduce the newly developed *ItaParlCorpus;* a *machine-readable* and *annotated* corpus of parliamentary protocols of plenary

*Max Planck Institute for the Study of Societies

Table 1: Descriptive statistics

legislature	dates	# of sessions	# of speeches	avg. words per speech
1	08.05.1948 - 24.06.1953	812	151,472	245.20
2	25.06.1953 - 11.06.1958	586	106,803	288.82
3	12.06.1958 - 15.05.1963	609	104,747	298.95
4	16.05.1963 - 04.06.1968	700	143,838	317.45
5	05.06.1968 - 24.05.1972	466	108,210	307.96
6	25.05.1972 - 04.07.1976	461	871,15	347.44
7	05.07.1976 - 19.06.1979	404	101,875	185.56
8	20.06.1979 - 11.07.1983	642	229,243	160.56
9	12.07.1983 - 01.07.1987	607	177,744	182.21
10	02.07.1987 - 22.04.1992	747	179,086	259.38
11	23.04.1992 - 14.04.1994	305	60,284	244.21
12	15.04.1994 - 08.06.1996	313	58,310	202.92
13	09.06.1996 - 29.05.2001	838	241,837	196.50
14	30.05.2001 - 27.04.2006	756	188,673	265.56
15	28.04.2006 - 28.04.2008	275	59,689	298.04
16	29.04.2006 - 14.03.2013	736	145,689	312.98
17	15.03.2013 - 22.03.2018	863	188,241	257.29
18	23.03.2018 - 12.10. 2022	739	121,862	368.95

debates of Italy's lower chamber of Parliament, the *Camera dei Deputati*.[1] This comprehensive corpus contains all parliamentary speeches from the Italian post-war republican period (1948-2022) and is the first fully annotated corpus that links parliamentary interventions to specific parliamentary groups across the country's entire republican history. Covering 18 legislatures, 68 different governments, the corpus includes over 470 million words, 2.4 million speeches, from 5,830 unique speakers representing 77 different political parties.

This work rests, and would not have been possible, without the use of the following resources. Frasnelli and Aprosio (2024) made available the .txt files of the OCR scans of parliamentary speeches/debates from the beginnings of the Kingdom of Italy to today. The resulting files are large, unstructured text documents from which it is necessary to separate the speaker from the content (i.e. the text of the speech). This task proved challenging due to significant variations in how speakers were distinguished from the text across different legislative periods. To identify unique speakers and link them to their party affiliations, I utilized a database of Italian parliamentarians provided by the Comparative Legislators Database (Göbel and Munzert 2022). Additionally, I collected data from the Italian Government's website to include ministers and cabinet members who were not part of parliament, such as technocrats.[2]

The parliamentary debates have been saved into csv files which can easily be processed and

[1] Available here: Harvard Dataverse, ItaParlCorpus

[2] *Governo Italiano, Presidenza del Consiglio dei Ministri: I Governi dal 1943 ad oggi*

analyzed by commonly used programming languages such as R and Python. Due to the size of these files as well as limitations on the number of rows that can be saved on a Microsoft Excel workbook, the debates have been grouped into four time periods and are available in csv and RData file formats:

- **Legislatures I-V**; Years: 1948-1972
- **Legislatures VI-X**; Years: 1972-1992
- **Legislatures XI-XIV**; Years: 1992-2006
- **Legislatures XV-XVIII**; Years: 2006-2022

Codebook

The variables which are contained in the dataset are as follows:

- **date**: Stores the date of the parliamentary session in this format YYYY-MM-DD.
- **year**: A numerical variable representing the year.
- **doc_id**: A character variable, containing the date of the parliamentary session as well a letter to denote whether there have been multiple sessions of parliamentary debates on the same day, for example one in the morning and one in the afternoon (e.g. a, b).
- **row_id**: A unique row identifier.
- **legislature**: A numerical variable containing the number of the legislative term (1-18) in which the speech was given.
- **speaker**: A character vector containing the name of the person giving the speech.
- **pageid_wiki:** A numerical variable containing the speaker's unique Wikipedia page ID as derived from the Comparative Legislators Database (Göbel and Munzert 2022).
- **party_name:** A character variable denoting the party affiliation of the speaker.
- **party_family:** A character variable denoting the party family the party belongs to, following the categorization in the ParlGov Database (Döring and Manow 2024).
- **party_id_itaparl:** A unique party ID for the *ItaParlCorpus* dataset.
- **party_id_parlgov:** A unique party ID for the *ParlGov* dataset.
- **chair**: a Boolean variable denoting whether the speech is given by the presidents of the Chamber of Deputies.
- **cabinet**: a Boolean variable denoting whether the speech was given by a member of the cabinet (minister, prime minister, secretary).

- **text:** raw text of the speech given in parliament.

Table 2: Party coding (ItaParlCorpus and ParlGov)

Name of party	Party ID (*ItaParlCorpus*)	Party ID (*ParlGov*)
+Europa	1	2737
Alleanza dei Progressisti	2	NA
Alleanza Nazionale	3	373
Associativo Italiani all'Estero	4	927
Blocco Nazionale	5	NA
Casa delle Libertà	6	NA
Centro Cristiano Democratico / Cristiani Democratici Uniti	7	627
Centro Democratico	8	2153
Centrodestra	9	67
Centrosinistra	10	NA
Chair	11	NA
Democratici di Sinistra	12	809
Democrazia Cristiana	13	1633
Democrazia Cristiana per le Autonomie	14	52
Democrazia è Libertà – La Margherita	15	1005
Democrazia Proletaria	16	1424
Federazione dei Verdi	17	910
Forza Italia – Il Popolo della Libertà	18	596
Fratelli d'Italia	19	2154
Fronte Democratico Popolare	20	NA
Government	21	NA
Italia dei Valori	22	693
L'Ulivo	23	NA
L'Unione-Prodi	24	1246
Lega	25	1436
Lega d'Azione Meridionale	26	436
Lega Nord	27	1436
Liga Lombarda	28	1560
Liga Veneta	29	1080
Lista Dini – Rinnovamento Italiano	30	630
Lista Pannella	31	NA
MAIE	32	NA
Movimento 5 Stelle	33	2155
Movimento Comunità	34	NA
Movimento per l'Autonomia	35	1040

Name of party	Party ID (*ItaParlCorpus*)	Party ID (*ParlGov*)
Movimento per la Democrazia – La Rete	36	366
Movimento Sociale Italiano	37	831
Others	38	NA
Partito Comunista Italiano	39	1088
Partito Comunista Italiano-Indipendenti	40	NA
Partito dei Comunisti Italiani	41	1661
Partito dei Contadini d'Italia	42	618
Partito della Rifondazione Comunista	43	1321
Partito Democratico	44	382
Partito Democratico della Sinistra	45	NA
Partito Democratico Italiano di Unità Monarchica	46	1702
Partito di Unità Proletaria per il comunismo	47	NA
Partito Liberale Italiano	48	487
Partito Monarchico Popolare	49	421
Partito Nazionale Monarchico	50	1050
Partito Popolare Italiano	51	142
Partito Popolare Sudtirolese	52	NA
Partito Repubblicano Italiano	53	93
Partito Sardo d'Azione	54	335
Partito Socialista Democratico Italiano	55	242
Partito Socialista di Unità Proletaria	56	1593
Partito Socialista Italiano	57	1475
Patto per l'Italia	58	NA
Patto Segni	59	827
PCI-PSIUP	60	NA
Polo del Buon Governo	61	NA
Polo delle Libertà	62	NA
Polo per le Libertà	63	NA
Popolari per Prodi	64	840
PSI-PSDI Unificati	65	NA
Radicali	66	1296
Rosa nel Pugno	67	NA
Scelta Civica	68	2156
Sinistra Ecologia Libertà	69	NA
Südtiroler Volkspartei	70	1030
SVP-PATT	71	1030
Union Valdôtaine	72	974
Unione / Centro	73	226
Unione Democratici per l'Europa	74	399

Name of party	Party ID (*ItaParlCorpus*)	Party ID (*ParlGov*)
Unità Socialista	75	NA
USEI	76	2671
Verdi Arcobaleno	77	325

References

Abrami, Giuseppe, Mevlüt Bagci, Leon Hammerla, and Alexander Mehler. 2022. "German Parliamentary Corpus (Gerparcor)." *arXiv Preprint arXiv:2204.10422*.

Agnoloni, Tommaso, Roberto Bartolini, Francesca Frontini, Simonetta Montemagni, Carlo Marchetti, 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*, 117–24.

Carammia, Marcello, Stefano Maria Iacus, and Giuseppe Porro. 2024. "Rethinking Scale: The Efficacy of Fine-Tuned Open-Source LLMs in Large-Scale Reproducible Social Science Research." *arXiv Preprint arXiv:2411.00890*.

Cova, Joshua, and Luuk Schmitz. 2024. "A Primer for the Use of Classifier and Generative Large Language Models in Social Science Research." Center for Open Science.

Döring, Holger, and Philip Manow. 2024. "Parliament and Government Composition Database (ParlGov):information on Parties, Elections and Cabinets in Established Democracies."

Erjavec, Tomaž, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, et al. 2023. "The ParlaMint Corpora of Parliamentary Proceedings." *Language Resources and Evaluation* 57 (1): 415–48.

Frasnelli, Valentino, 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)*, 16037–46.

Göbel, Sascha, and Simon Munzert. 2022. "The Comparative Legislators Database." *British Journal of Political Science* 52 (3): 1398–408.

Marx, Maarten, Anne Schuth, et al. 2010. "DutchParl. A Corpus of Parliamentary Documents in Dutch." *Proceedings Language Resources and Evaluation (LREC)*, 3670–77.

Rauh, Christian, and Jan Schwalbach. 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.

Richter, Florian, 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 the Parliamentary Protocols of the German Bundestag."

Sylvester, C, A Ershova, A Khokhlova, N Yordanova, and Z Greene. 2023. "ParlEE Plenary Speeches V2 Data Set: Annotated Full-Text of 15.1 Million Sentence-Level Plenary Speeches of Six EU Legislative Chambers." *Harvard Dataverse*.