There's Something New about the Italian Parliament: the IPSA Corpus

Valentino Frasnelli, Alessio Palmero Aprosio
Università di Trento, Fondazione Bruno Kessler
Trento, Italy
valentino.frasnelli@studenti.unitn.it, aprosio@fbk.eu

Abstract

Parliamentary debates constitute a substantial and somewhat underutilized reservoir of publicly available written content. Despite their potential, the Italian parliamentary documents remain largely unexplored and most importantly inaccessible in their original paper-based form. In this paper we attempt to transform these valuable historical documents into IPSA, a digitally readable structured corpus containing speeches, reports of the Standing Committees, and law proposals spanning 175 years of Italian history, from the issuing of the Statuto Albertino in 1848, up to the present day. At first, the PDF documents, available on the official websites of *Senato della Repubblica* and *Camera dei Deputati*, the two chambers that form the Italian Parliament, are digitized using Optical Character Recognition (OCR) techniques. Then, the speeches are tagged with the corresponding speakers. The final dataset is released both in textual and structured format.

Keywords: Parliamentary Debates, Linked Data, Optical Character Recognition

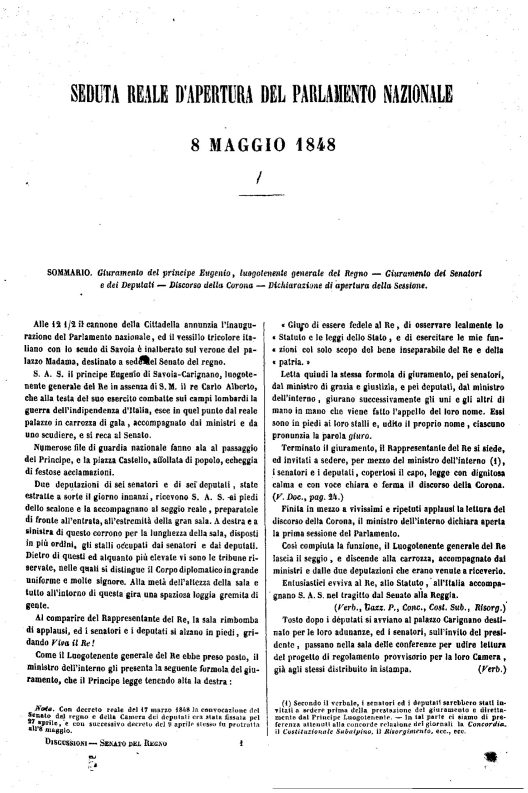
1. Introduction

Analyzing parliamentary debates holds significant importance across various research domains. Beyond its relevance in political science, this kind of datasets offers valuable insights into how a language and its associated culture have evolved throughout history. In particular, over the past two centuries, Italian society has undergone profound transformations.

Beginning with the shift from an absolute monarchy to a parliamentary monarchy in 1848, Italy has witnessed a series of pivotal historical events, including both world wars, the era of fascist dictatorship, the exile of the royal family, the establishment of universal suffrage, Italy's accession to the European Union, and a multitude of other significant developments. These crucial milestones, as well as the broader spectrum of Italian political and social life, are chronicled within the parliamentary records.

Many research groups worldwide have developed and shared datasets of political debates in different languages, spanning diverse areas of study such as religion (Cheng, 2015), gender (Paoletti, 1991), multilinguality (Bayley, 2004), and more.

One notable dataset, GerParCor (Abrami et al., 2022), comprises German-language parliamentary records spanning three centuries and four nations. Similarly, siParl (Pancur and Erjavec, 2020), DutchParl (Marx and Schuth, 2010), and the Polish Parliamentary Corpus (Ogrodniczuk and Nitoń, 2020) represent collections of political debates in Slovenian, Dutch, and Polish languages, respectively.



SEDUTA REALE D'APERTURA DEL PARLAMENTO NAZIONALE

8 MAGGIO 1848

SOMMARIO. *Giuramento del principe Eugenio, luogotenente generale del Regno — Giuramento dei Senatori e dei Deputati — Discorso della Corona — Dichiarazione di apertura della Sessione.*

Alle 12 1/2 il cannone della Cittadella annunzia l'inaugurazione del Parlamento nazionale, ed il vessillo tricolore italiano con lo scudo di Savoia è inalberato sul verone del palazzo Madama, destinato a sede del Senato del regno.

S. A. S. il principe Eugenio di Savoia-Carignano, luogotenente generale del Re in assenza di S. M. il re Carlo Alberto, che alla testa del suo esercito combatte sui campi lombardi la guerra dell'indipendenza d'Italia, esce in quel punto dal reale palazzo in carrozza di gala, accompagnato dai ministri e da uno scudiere, e si reca al Senato.

Numerose file di guardia nazionale fanno ala al passaggio del Principe, e la piazza Castello, affollata di popolo, echeggia di festose acclamazioni.

Due deputazioni di sei senatori e di sei deputati, state estratte a sorte il giorno innanzi, ricevono S. A. S. ai piedi dello scalone e la accompagnano al seggio reale, preparatole di fronte all'entrata, all'estremità della gran sala. A destra e a sinistra di questo corrono per la lunghezza della sala, disposti in più ordini, gli stalli occupati dai senatori e dai deputati. Dietro di questi ed alquanto più elevate vi sono le tribune riservate, nelle quali si distingue il Corpo diplomatico in grande uniforme e molte signore. Alla metà dell'altezza della sala e tutto all'intorno di questa gira una spaziosa loggia gremita di gente.

Al comparire del Rappresentante del Re, la sala rimbomba di applausi, ed i senatori e i deputati si alzano in piedi, gridando *Viva il Re!*

Come il Luogotenente generale del Re ebbe preso posto, il ministro dell'interno gli presenta la seguente formola del giuramento, che il Principe legge tenendo alta la destra:

« Giuro di essere fedele al Re, di osservare lealmente lo « Statuto e le leggi dello Stato, e di esercitare le mie funzioni col solo scopo del bene inseparabile del Re e della « patria. »

Letta quindi la stessa formola di giuramento, pei senatori, dal ministro di grazia e giustizia, e pei deputati, dal ministro dell'interno, giurano successivamente gli uni e gli altri di mano in mano che viene fatto l'appello del loro nome. Essi sono in piedi ai loro stalli e, udito il proprio nome, ciascuno pronunzia la parola *giuro*.

Terminato il giuramento, il Rappresentante del Re si siede, ed invitati a sedere, per mezzo del ministro dell'interno (1), i senatori e i deputati, copertosi il capo, legge con dignitosa calma e con voce chiara e ferma il discorso della Corona. (*V. Doc., pag.* 2.)

Finita in mezzo a vivissimi e ripetuti applausi la lettura del discorso della Corona, il ministro dell'interno dichiara aperta la prima sessione del Parlamento.

Così compiuta la funzione, il Luogotenente generale del Re lascia il seggio, e discende alla carrozza, accompagnato dai ministri e dalle due deputazioni che erano venute a ricevorlo.

Entusiastici evviva al Re, allo Statuto, all'Italia accompagnano S. A. S. nel tragitto dal Senato alla Reggia.

(*Verb.*, Ediz. P., Conc., *Cost. Sub.*, *Risorg.*)

Tosto dopo i deputati si avviano al palazzo Carignano destinato per le loro adunanze, ed i senatori, sull'invito del presidente, passano nella sala delle conferenze per udire lettura del progetto di regolamento provvisorio per la loro Camera, già agli stessi distribuito in istampa. (*Verb.*)

(1) Secondo il verbale, i senatori ed i deputati sarebbero stati invitati a sedere prima della prestazione del giuramento e direttamente dal Principe Luogotenente. — In tal parte ci siamo di preferenza attenuti alla concorde relazione dei giornali la *Concordia*, il *Costituzionale Subalpino*, il *Risorgimento*, ecc., ecc.

Figure 1: The very first session of the Italian Parliament, on 8th May 1848.

Since the establishment of the European Union, the political debates of the European Parliament have been accessible in multiple languages, offering a valuable resource for machine translation

LREC-COLING 2024, pages 16037–16046

(Koehn, 2005).

Furthermore, as the prevalence of large language models (LLMs) continues to grow, having large collections of texts has become of great importance for their constructions, especially the ones belonging to a particular domain (such as legal domain texts in the context of parliamentary speeches). Past research has shown that the utilization of domain-specific LLMs leads to enhanced performance across various tasks, including document classification and information extraction (Chalkidis et al., 2022).

In this paper, we present IPSA (Italian Parliamentary Speeches and other Acts), the first version of the Italian Parliamentary Corpus, a collection of documents covering 175 years of Italian history and containing all the documents redacted by the two houses of the bicameral Italian Parliament (*Camera dei Deputati*, the lower house, and *Senato della Repubblica*, previously *Senato del Regno*, the upper house).

Besides the parliamentary speeches, we also collect two more sets of documents: report and speeches from the parliamentary Standing Committees, and law proposals along with the corresponding amendments. Each speech related to the parliamentary debates is automatically tagged with the politician who delivered it.

Documents before 1996 are only available as scanned PDF, therefore an Optical Character Recognition (OCR) software has been used to extract their textual content. For this reason, data dating back to that period has to be considered 'silver', since it may contain errors. A comprehensive analysis of the errors and the description of some approaches used to fix them are described in Section 4.

More recent texts, on the contrary, are available in electronic format, therefore both the texts and the politician tags are supposedly correct. In total, IPSA contains 1.2 billion tokens of structured documents belonging to the Italian Parliament debates, along with more than 5 million tagged speeches, and additional 1.2 billion tokens taken from Standing Committees and law proposals.

The rest of the article is structured as follows. In Section 2 we examine similar projects worldwide. Section 3 describes how the raw data has been collected. Section 4 we show some analysis about the texts obtained through OCR. Section 5 gives information about how the data is structured. Section 6 contains some statistics of the corpus. Finally, both the source code and the dataset are available for download, as described in Section 7.

2. Related work

Parliamentary debates in various languages have been collected as a testament to the importance of preserving and analyzing the discourse that shapes the policies and governance of nations.

DutchParl (Marx and Schuth, 2010) was among the first parliamentary corpora available, being published in 2010. DutchParl had the aim of bringing parliamentary documents in the Dutch language together under one uniform schema, incorporating documents from Belgium, the European Union, the Netherlands and Suriname under one metadata schema.

The German Parliament Corpus (GerParCor) (Abrami et al., 2022) aims to fill the gap in German-language parliamentary corpora. It is a genre-specific corpus consisting mainly of historical German-language parliamentary proceedings spanning three centuries and four countries, including state and federal level data, and both transcribed and OCR-processed versions of scanned protocols, even those in Fraktur typeface.[1] The corpus gathers parliamentary discussion documents of Germany since 1867 (both at national and regional level), of Austria since 1918, of Liechtenstein since 1997 and of Switzerland since 1999, for a total of more than one billion tokens.

The Polish Parliamentary Corpus (Ogrodniczuk and Nitoń, 2020), first created in 2018, is a comprehensive corpus of transcripts of Polish parliament proceedings dating back to 1919, including Sejm (lower house) sittings, Senate sittings, committee sittings, interpellations, and questions. It contains more than 750M tokens.

The dataset "Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts" (Gentzkow et al., 2018) provides processed text from 1.6 million documents belonging to the United States Congressional Records. These records include all speeches occurred on the floors of both the House of Representatives and the Senate spanning from the 43rd Congress (1873) to the 114th (2017).

The CzechParl corpus (Jakubíček and Kovář, 2010) consists of stenographic protocols of the Czech parliament, for a total of more than 80 million tokens. The corpus primarily focuses on debates that are stenographically recorded, in particular, since 1993.

The siParl corpus (Pančur et al., 2022) is a collection of legislative documents from Slovenia, including minutes from various legislative periods and bodies. It covers sessions from 1990 to 2022, with

[1] Fraktur is a style of blackletter typeface that was historically used for printed material in the German-speaking world. It is characterized by its distinctive and ornate script with elaborate, angular letterforms.

over 11,000 sessions, one million speeches, and 200 million words.

The Aalto Finnish Parliament ASR Corpus (Virkkunen et al., 2023) consists of both audio recordings and transcriptions extracted from the Finnish parliamentary plenary sessions from 2008 to 2020. It contains 19M tokens and 3000 hours of audio files.

Finally, ParlaMint (Erjavec et al., 2023) is a big collection of comparable corpora of parliamentary debates of 29 European countries and autonomous regions, covering at least the period from 2015 to 2022, containing over 1 billion words, and including also Italian data.

The above-described datasets and collections are summarized in Table 1.

3. Data collection

We obtained all the documents that were accessible online from the official websites of the two chambers of the Italian Parliament.

The *Camera dei Deputati* website offers the complete catalog of digital data and documents dating from the first Legislature of the Kingdom of Sardinia to the present Republic. Conversely, for data related to the *Senato della Repubblica*, we could directly download only documents issued after 1948. The debates that took place between 1848 and 1940 have already been digitized but were not yet published at the time of our research. We were able to obtain them with assistance from the *Servizio dei Resoconti e della Comunicazione istituzionale del Senato della Repubblica*.

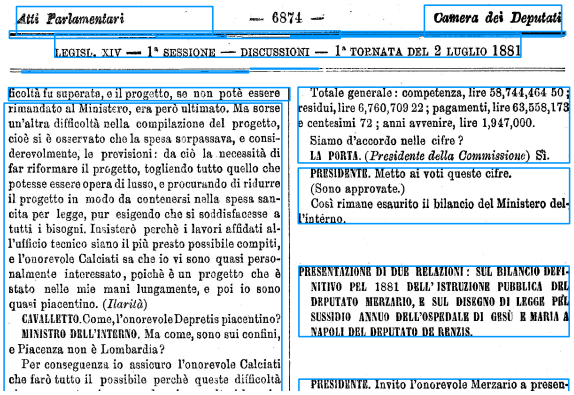
In both cases, documents created prior to 1996 were not originally produced in digital formats, so they are only available in scanned PDF format. Starting from 1996 (Republic Legislature number XIII), debates have been published in text format on the web.

Section 5.1 contains more information about how data is structured in the two chambers.

4. OCR processing

To transform scanned PDF documents into editable text, we employed Optical Character Recognition (OCR) technology, specifically Tesseract (Kay, 2007), originally developed by Hewlett-Packard and subsequently released as an open-source solution. It is both cost-free and provides extensive language support, encompassing more than 100 languages right out of the box, including Italian.

After downloading all the Parliament documents of both Chambers, the OCR phase was straightforward: each page of each PDF was converted into a 400 PPI [2] image, and fed to the Tesseract engine. Since Tesseract is a computationally expensive engine, it was necessary to perform the OCR of the documents on 8 parallel processes using five Xeon machines, which allowed the whole operation to be performed in about one month.


Atti Parlamentari — 6874 — Camera dei Deputati

LEGISL. XIV — 1ª SESSIONE — DISCUSSIONI — 1ª TORNATA DEL 2 LUGLIO 1881

ficoltà fu superata, e il progetto, se non potè essere rimandato al Ministero, era però ultimato. Ma sorse un'altra difficoltà nella compilazione del progetto, cioè si è osservato che la spesa sorpassava, e considerevolmente, le previsioni: da ciò la necessità di far riformare il progetto, togliendo tutto quello che potesse essere opera di lusso, e procurando di ridurre il progetto in modo da contenersi nella spesa sancita per legge, pur eseguendo che si soddisfacesse a tutti i bisogni. Insisterò perchè i lavori affidati all'ufficio tecnico siano il più presto possibile compiti, e l'onorevole Calciati sa che io vi sono quasi personalmente interessato, poichè è un progetto che è stato nelle mie mani lungamente, e poi io sono quasi piacentino. (*Ilarità*)

CAVALLETTO. Come, l'onorevole Depretis piacentino?

MINISTRO DELL'INTERNO. Ma come, sono sui confini, e Piacenza non è Lombardia?

Per conseguenza io assicuro l'onorevole Calciati che farò tutto il possibile perchè queste difficoltà

Totale generale: competenza, lire 58,744,464 50; residui, lire 6,760,709 22; pagamenti, lire 63,558,173 e centesimi 72; anni avvenire, lire 1,947,000.

Siamo d'accordo nelle cifre?

LA PORTA. (*Presidente della Commissione*) Sì.

PRESIDENTE. Metto ai voti queste cifre.

(Sono approvate.)

Così rimane esaurito il bilancio del Ministero dell'interno.

PRESENTAZIONE DI DUE RELAZIONI: SUL BILANCIO DEFINITIVO PEL 1881 DELL'ISTRUZIONE PUBBLICA DEL DEPUTATO MERZARIO, E SUL DISEGNO DI LEGGE PEL SUSSIDIO ANNUO DELL'OSPEDALE DI GESÙ E MARIA A NAPOLI DEL DEPUTATO DE RENZIS.

PRESIDENTE. Invito l'onorevole Merzario a presen-

Figure 2: Division of text into blocks by Tesseract.

Tesseract attempts to reconstruct the text in the images it scans, and in doing so optically detects the structure of the original document by subdividing portions of text in a hierarchical manner: pages, blocks, paragraphs, lines and words. Figure 2 shows an example of block extraction.

Following the conversion process, we applied rule-based heuristics to clean the data.

In particular, the operation consisted in the following steps:

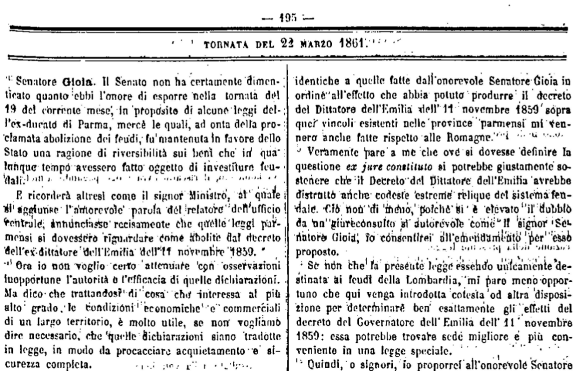
- rid the texts of all unwanted elements from the original documents, such as the heading of each page, the division into two columns in almost all the documents, blank space between pages, white pages;
- reconstruct the text by handling the division into paragraphs, combine truncated words, connect the text of two different pages as seamlessly as possible, as well as concatenating the words in the correct order.

In the final step, we sought to assess the quality of the OCR output. To accomplish this, we assembled a benchmark dataset, comprising 58 pages that had been manually transcribed. These pages were drawn from various legislative documents spanning the period from 1848 to 1996.

[2] PPI (Pixels Per Inch) is a measurement that quantifies the pixel density of a digital image or display, indicating how many individual pixels are present in one linear inch. Three different values for image quality were tested: 300, 400 and 600 PPI. The choice fell on using 400 PPI images, since 300 PPI scans were visibly blurrier and at the same time 600 PPI images did not have significant visual changes from the 400 PPI version, allowing to save disc space and processing power.

Corpus Name	Size	Time span
DutchParl	Over 800 million tokens	1971-2009
GerParCor	Over 1 billion tokens	1867-present
Polish Parliamentary Corpus	Over 750 million tokens	1919-present
US Congressional Record	1.6 million documents	1873-2017
CzechParl	Over 80 million tokens	1993-2020
siParl	Over 200 million tokens	1990-2020
Aalto Finnish Parliamentary ASR Corpus	Over 19 million tokens	2008-2020
IPSA (Italian Parliament)	Over 1.2 billion tokens	1848-2022
ParlaMint (29 languages)	Over 1 billion words	2015-2022

Table 1: Information about existing parliamentary corpora.



— 495 —

TORNATA DEL 22 MARZO 1861

Senatore Gioia. Il Senato non ha certamente dimenticato quanto ebbi l'onore di esporre nella tornata del 19 del corrente mese, in proposito di alcune leggi dell'ex-ducato di Parma, mercè le quali, ad onta della proclamata abolizione dei feudi, fu mantenuta in favore dello Stato una ragione di riversibilità sui beni che in qualunque tempo avessero fatto oggetto di investiture feudali.

E ricorderà altresì come il signor Ministro, al quale mi aggiunse l'onorevole parola del relatore dell'ufficio centrale, annunciasse recisamente che quelle leggi parmensi si dovessero riguardare come abolite dal decreto dell'ex-dittatore dell'Emilia dell'11 novembre 1859.

Ora io non voglio certo attenuare con osservazioni inopportune l'autorità e l'efficacia di quelle dichiarazioni. Ma dico che trattandosi di cosa che interessa al più alto grado le condizioni economiche e commerciali di un largo territorio, è molto utile, se non vogliamo dire necessario, che quelle dichiarazioni siano tradotte in legge, in modo da procacciare acquietamento e sicurezza completa.

identiche a quelle fatte dall'onorevole Senatore Gioia in ordine all'effetto che abbia potuto produrre il decreto del Dittatore dell'Emilia dell' 11 novembre 1859 sopra quei vincoli esistenti nelle province parmensi mi vennero anche fatte rispetto alle Romagne.

Veramente pare a me che ove si dovesse definire la questione *ex jure constituto* si potrebbe giustamente sostenere che il Decreto del Dittatore dell'Emilia avrebbe distrutto anche codeste estreme reliquie del sistema feudale. Ciò non di meno, poichè si è elevato il dubbio da un giureconsulto di autorevole come il signor Senatore Gioia, io consentirei all'emendamento per esso proposto.

Se non che la presente legge essendo unicamente destinata ai feudi della Lombardia, mi pare meno opportuno che qui venga introdotta codesta od altra disposizione per determinare ben esattamente gli effetti del decreto del Governatore dell'Emilia dell' 11 novembre 1859: essa potrebbe trovare sede migliore e più conveniente in una legge speciale.

Quindi, o signori, io proporrei all'onorevole Senatore

Figure 3: Example of noisy text: OCR algorithms perform poorly on documents with these types of printing errors

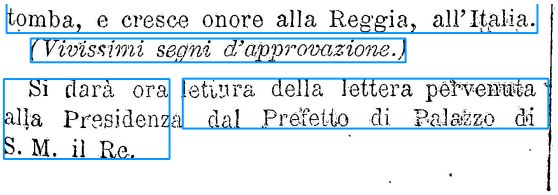
Apart from just transcribing the debates, each speech featured on the page is also associated with the respective politician. This allows this collection of pages to be utilized for assessing the accuracy of tag extraction (see next Section 5).

4.1. OCR errors

OCR systems are very much prone to error also considering the age of many of the documents present in the data, many pages of which are unreadable even for a human because of printing errors. Figure 3 displays a particularly noisy page. In such cases, OCR algorithms perform poorly, since discerning printing errors from the original content is a difficult task to perform even for state-of-the-art techniques.

In order to evaluate the error rate of the OCR process, and also to evaluate the possible error correction phase, two metrics of comparison were chosen based on the state-of-the-art OCR evaluation literature (Neudecker et al., 2021): Word Error Rate (WER) and Character Error Rate (CER).

Word Error Rate measures the percentage of words that are missing, incorrectly placed or spelled in the predicted document when compared to its



tomba, e cresce onore alla Reggia, all'Italia.

(*Vivissimi segni d'approvazione.*)

Si darà ora lettura della lettera pervenuta alla Presidenza dal Prefetto di Palazzo di S. M. il Re.

Figure 4: Example of Tesseract wrong block identification.

correct counterpart. Similarly, Character Error Rate conducts the same calculation as the WER, but character-wise, meaning it measures the percentage of characters that are missing, incorrectly placed or different in the predicted document when compared to its correct counterpart. Their formulas are as follows:

$$WER = \frac{S + D + I}{N} \qquad CER = \frac{S + D + I}{N}$$

where: S is the number of word/character substitutions; D is the number of words/chars deletions; I is the number of words/chars insertions; and N is the total number of words/chars in the test set counterpart.

Figure 5 shows the plot of text quality over intervals of five legislatures: overall, a downward trend is visible in the WER metric, with a seemingly unjustified peak in the XI-XV interval. This indeed is justifiable by the fact that three documents in that time interval caused the Tesseract OCR to erroneously separate lines and paragraphs, leading to the reconstructed files having segments of text in the wrong placement (see Figure 4). Nonetheless, Figure 5 allows to visualize the downward trend in CER and WER (corresponding to an upward trend in the quality and readability of the original documents) since the older legislatures.

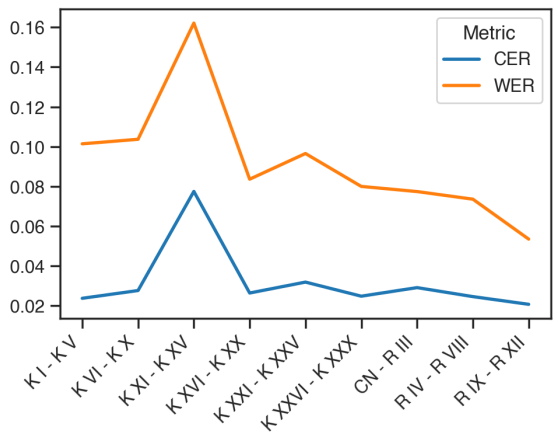


Figure 5: Trend of Word Error Rate (WER) and Character Error Rate (CER) in intervals of five legislatures. The X axis represents the time intervals, expressed with the kind of legislature (K = Kingdom, CN = Consulta Nazionale, R = Republic) plus the roman numerals describing its sequential number.

Correction method	CER	WER
Original	0.030	0.071
SymSpell	0.036	0.121

Table 2: Mean CER and WER against the test set (the lower, the better).

4.2. Data cleaning preliminary tests

Following the approaches described in previous work (Abrami et al., 2022), we try to fix OCR errors using a spell-checker, specifically SymSpell,[3] which is freely available and has high processing speed.

Unfortunately, the WER and CER metrics applied to the test set before and after the corrections led to worse results (see Table 2), therefore this first version of IPSA is released without any intervention on OCR output.

5. Documents tagging

After collecting the transcriptions of the debates, we attempt to map each speech made by a Parliament member or a Government member in a parliamentary sitting to the precise reference of the involved politician. To this purpose, we first provide a brief introduction to the Italian Parliament data structures.

5.1. Italian Parliament data structures

Both *Camera dei Deputati* and *Senato della Repubblica* release their recent data using the paradigm of Linked Open Data,[4] following Tim Berners-Lee modern vision of the Web where data and information is not exclusively available in human readable form, but it also accessed by machines using a common standard, usually RDF.[5]

Following the directives from the *Agenzia per l'Italia Digitale*,[6] both chambers of the Italian Government have designed two ontology descriptions in XML/RDF and two specific namespaces, OCD (Ontologia Camera dei Deputati)[7] and OSR (Ontologia Senato della Repubblica).[8] Unfortunately, the two structures have been developed by different working groups, therefore they are not completely compatible and interchangeable, although they share some similar elements, such as having a unique ID for each politician, act, session and speech.

The data management is very different among the two administrations. For instance, data from *Camera dei Deputati* is available for download in bulk directly from its open data website,[9] in RDF format, or can be queried through a SPARQL endpoint.[10] On the contrary, data from *Senato della Repubblica* from the corresponding website[11] is available only by applying some filters and can be obtained exclusively in XML, JSON, or CSV, although a SPARQL endpoint[12] does exist for making queries.

Apart from technical differences, both ontologies are written in OWL and are modelled on other Open Government Data produced in other states, following the guidelines and spreading best practices.

Notably, the Chamber of Deputies not only maintains up-to-date RDF datasets specific to its own chamber but also manages highly comprehensive cross-sectional data, encompassing information about both Chambers and Governments. For this reason, we concentrate our efforts on producing the additional data using the OCD namespace.

5.2. Tagging procedure

In the OCD ontology, each politician is identified by a unique HTTP URI. For example, the Member of Parliament Ivanoe Bonomi is defined by the URI `https://dati.camera.it/ocd/persona.rdf/p27370`.

Overall, the mapping of the speeches to their speaker included the following steps:

[3] https://github.com/wolfgarbe/SymSpell

[4] https://www.w3.org/standards/

[5] https://www.w3.org/RDF/

[6] https://www.agid.gov.it/it

[7] http://dati.camera.it/ocd

[8] https://dati.senato.it/sito/21

[9] https://dati.camera.it/

[10] https://dati.senato.it/sparql

[11] https://dati.senato.it/

[12] https://dati.senato.it/sparql

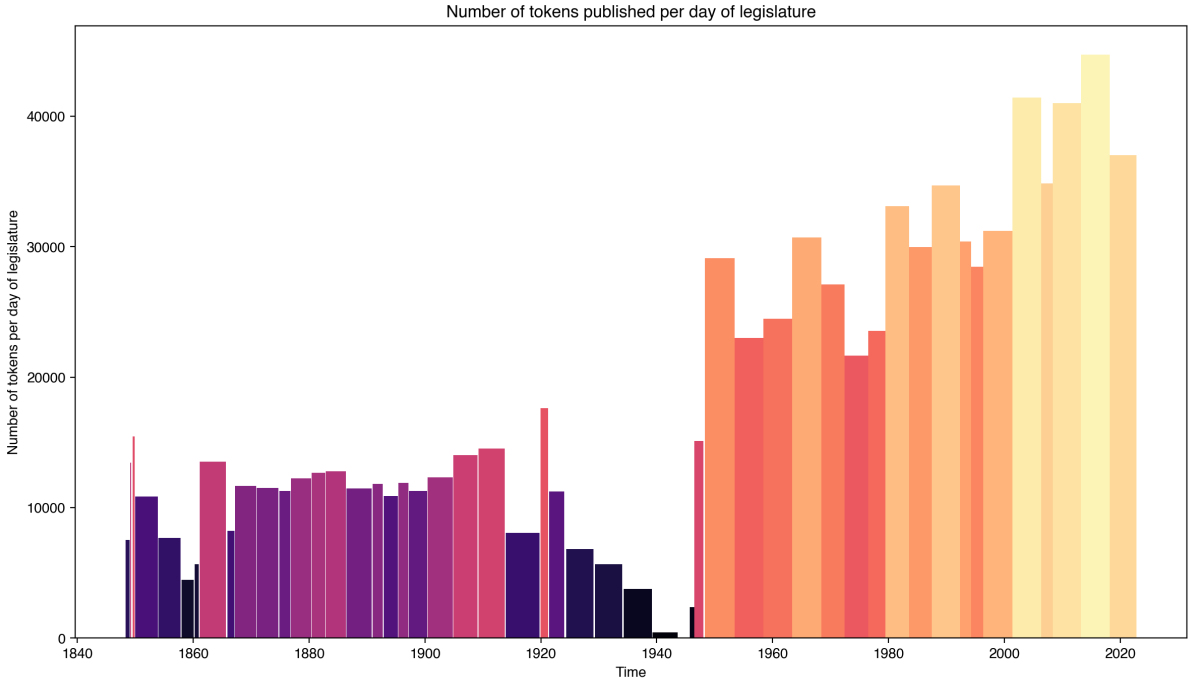


Figure 6: Bar plot of the average number of words per day contained in the parliamentary sittings by legislature with thickness proportional to time span of the legislature.

Senatore PRIMERANO. Domando di parlare.

PRESIDENTE. Ne ha facoltà.

Senatore PRIMERANO. Mi pare che con questa dizione si ritorni all' antico corpo di stato maggiore e si escluda la dicitura che è tutto il concetto dell' articolo, cioè: « Capo di stato maggiore dell' esercito ».

Senatore TAVERNA, *relatore*. Non è che una questione di forma: in luogo di mettere l' aggiunta votata in fine dell' articolo, per maggiore chiarezza la si pone a metà.

Figure 7: Example of a series of speeches where speakers are recognizable by being uppercase, although in come cases their are preceded by their role in the Parliament

Presidente. Senta questo non ha a che fare col processo verbale.

Pellegrini. La prego di credere e le dichiaro che può darsi, ciò che a nessuno cale e neppure a me, che questa sia l'ultima volta che io ho l'onore di chiedere la facoltà di parlare... (*Rumori*)

Presidente. Onorevole Pellegrini, io auguro che Ella chieda mille volte ancora la facoltà di parlare...

Figure 8: Example of a series of speeches where speakers are in lowercase notation, in this case they are discernible from the speech by being in boldface

MIRABELLI ROBERTO. Prima Spaventa.

MURATORI. Prima Spaventa con la giustizia nell'amministrazione. Ed ebbe in Napoli un largo consenso di applausi. Ma la critica fatta allora dei costumi parlamentari, rileggendo quel discorso, è sempre di attualità. E prima ancora di lui un altro uomo insigne, onore del Mezzogiorno, Francesco De Santis, scrisse sulla degenerazione dei nostri costumi parlamentari nelle celebri lettere pubblicate sul giornale *Il Diritto*.

Figure 9: Example of a series of speeches where speakers are recognizable by being exclusively uppercase

- gathering of all possible speakers and their URI in a parliamentary sitting for each legislature;
- identifying when a speech is occurring inside of a document;
- pairing the speech to its corresponding speaker.

Speakers in Italian Parliament summary reports are usually identified by their surname, however they are not marked in the same manner in all legislatures, and by simply observing the examples in figures 7, 8 and 9, one can make several important observations:

- some legislatures exclusively mark speakers in uppercase (9), some exclusively in lowercase with boldface (7), others a mix of the two (8);
- presidents in charge of the Chamber in that specific instance are always marked as “presidente";
- in case of Parliament members being homonyms both surname and name of the speaker are specified (see 9, MIRABELLI ROBERTO);
- occasionally, the surname of the speakers is accompanied by the position of the speaker in the parliament (“Senatore", “Relatore", “Presidente del Consiglio").

Depending on the legislature and the format of the text, the first part of each paragraph is compared to the list of politicians belonging to that legislature. This list includes not only the people assigned to that particular chamber, but all the people involved in the entire legislature, such as politician from both Camera and Senato, and the Government representative, that often do not sit in the Parliament. As far as presidencies go, a new presidency is detected by finding the word “Presidenza" and checking inside that line whether a surname is present, while when the speaker is identified as “Presidente", the current president of the sitting is assigned as the speaker.

Since text could contain spelling errors caused by OCR, the comparison was not straightforward, and a fuzzy algorithm was used to match as closely as possible the right surname. This was done with `fuzzywuzzy`,[13] an easy-to-use Python library which performs efficient fuzzy string matching.

The tagging task has been evaluated against the benchmark test set already described in Section 4, and is composed of 58 pages. Table 3 shows the results of the evaluation in terms of precision, recall, and F1. In our calculations, we consider true positive a correctly identified politician, a false positive a wrong identification, and a false negative a missing identification. As expected from this kind of task, precision is higher than recall, meaning that when a speech is identified, the involved politician is correctly marked. The lower recall shows that the main difficulty of the task is the identification that a new speech has started.

To underline how a bad quality of the OCR process affects the results, we add two rows in Table 3, showing the difference in performances when we consider documents before and after 1945. Similarly, Figure 10 shows the F1 trend over the years. Apart for a negative peak in legislatures IV-VIII, probably due to the small size of the test set, the chart shows a constant increase in the performance of the tagger.

[13] https://github.com/seatgeek/fuzzywuzzy

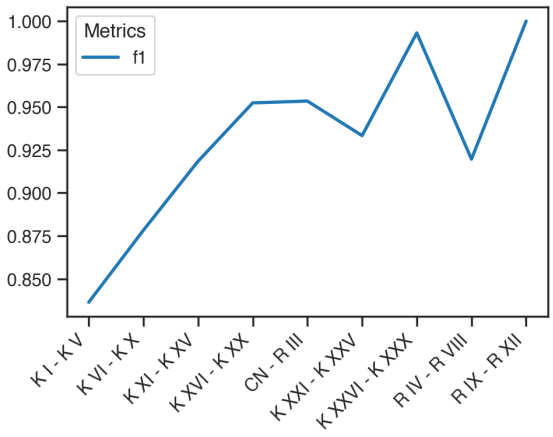


Figure 10: Trend of tagging task F1 over time (see caption of Figure 5 for the description of the X axis).

Metric	**Precision**	**Recall**	**F1 Score**
Global	0.939	0.880	0.909
pre-WW2	0.953	0.850	0.898
post-WW2	0.916	0.942	0.929

Table 3: Performance on the tagging task.

5.3. Data format

Law proposals and reports from Standing Committees are released in text-only format, since for this kind of documents no post-processing has been done. Parliamentary speeches, on the contrary, are released in two formats: text-only and XML. The XML version contains the tags that link each speech to the politician who gave it.

Beside these two formats, we also release the RDF triples describing the same information included in the XML files, i.e. the association between the politicians and the debates. As described in Section 5.1, this information follows the guidelines of the OCD ontology.

The *Camera dei Deputati* LOD infrastructure is already tracking every speech of the Chamber of Deputies sittings, but only started from the XVII legislature (2013), therefore it was necessary to expand the RDF from just two legislatures to a total of 50 legislatures, with each speech being tracked by the dataset built in the process. In addition to this, we had also to add all the speeches from the *Senato della Repubblica*.

In adding the reference to the textual data (since before 1996 no structured data related to the

speeches is available), we use the `StartEndPointer` class from the W3C RDF pointers methods,[14] which was perfectly fitting for the use in this context.

6. Dataset statistics

The IPSA dataset is released in different formats, depending on the type of documents. In general, documents from Kingdom I (1848) to Republic XII (1996) are available only as scanned PDF, therefore the corresponding text may be incorrect. Starting from Republic XIII, data has been uploaded in electronic format, therefore the quality should be very high.

Table 4 contains all the statistics (number of documents, pages, tokens, and tags) from the parliamentary debates, along with quantitative data about the side datasets containing the reports from the Standing Committees and the law proposals.

Finally, Figure 6 gives a visual representation of the size of the parliamentary speeches. The thickness of each bar is proportional to the time, while the height represents the average words per day in that legislature. At a first glance, one can see how the quantity of data has grown after WW2 (that can be also related to the increment of the number of components for each chamber).

7. Release

All the data described in this paper is available for download under the CC-BY 3.0 legal code,[15] similarly to what both Italian chambers did for the original data.

In addition to text-only, XML and RDF files, the Github page of the project[16] contains all the source code used to download, parse and tag the data, along with the test documents used for the evaluation.

8. Conclusions and Future Work

In this paper, we presented the first version of IPSA, a collection of parliamentary debates, along with reports from Standing Committees and law proposals.

The dataset contains around 2.4 billion tokens, half of which belonging to the Italian Parliament, where each speech is tagged with the corresponding politician. The second half contains texts extracted from committees and law proposals, and is not post-processed nor tagged.

Texts before 1996 are available in scanned PDF, therefore they are error-prone, while some processing has been done to remove headings and clean noisy contents such as indexes and tables.

In the future, we plan to implement state-of-the-art methods and try to get better results in terms of both OCR correction and tagging. Some recent works take advantage of seq2seq models to clean OCR results (Hämäläinen and Hengchen, 2019), but need training data, usually difficult to produce, and often created artificially by adding noise to clean texts taken from the same domain (Schaefer and Neudecker, 2020). Since OCR errors are often related to single characters, also approaches that involve byte-level approaches could be explored (Stankevičius et al., 2022).

Regarding the tagging task, our goal is to increase recall by splitting the task into speech identification and politician attribution, so to avoid misattribution of spans of texts.

9. Bibliographical References

Paul Bayley. 2004. Cross-cultural perspectives on parliamentary discourse. *Cross-Cultural Perspectives on Parliamentary Discourse*, pages 1–390.

Jennifer E Cheng. 2015. Islamophobia, muslimophobia or racism? parliamentary discourses on islam and muslims in debates on the minaret ban in switzerland. *Discourse & Society*, 26(5):562–586.

Mika Hämäläinen and Simon Hengchen. 2019. From the paft to the fiiture: a fully automatic NMT and word embeddings method for OCR post-correction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 431–436, Varna, Bulgaria. INCOMA Ltd.

Anthony Kay. 2007. Tesseract: An open-source optical character recognition engine. *Linux J.*, 2007(159):2.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Clemens Neudecker, Konstantin Baierer, Mike Gerber, Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. 2021. A survey of ocr evaluation tools and metrics. In *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing*, HIP '21,

[14] https://bit.ly/pointers-rdf
[15] https://bit.ly/cc-30-legal
[16] https://github.com/dhfbk/ipsa

		Speeches				Committees	Law proposals
Legislature	**Time span**	**Docs**	**Pages**	**Tokens**	**Tags**	**Tokens**	**Tokens**
Kingdom I	8 May 1848 - 30 Dec 1848	172	1,742	1,884,530	3,659		
Kingdom II	1 Feb 1849 - 30 Mar 1849	81	850	1,111,824	3,942		
Kingdom III	30 Jul 1849 - 20 Nov 1849	123	1,756	1,937,255	6,997		
Kingdom IV	20 Dec 1849 - 20 Nov 1853	1,010	14,740	15,336,297	41,056		
Kingdom V	19 Dec 1853 - 25 Oct 1857	692	10,484	10,549,093	19,657		
Kingdom VI	14 Dec 1857 - 21 Jan 1860	243	3,987	3,632,438	4,742		
Kingdom VII	2 Apr 1860 - 17 Dec 1860	112	1,475	1,448,936	3,222		
Kingdom VIII	18 Feb 1861 - 7 Sep 1865	1,260	25,403	22,526,607	59,282		
Kingdom IX	18 Nov 1865 - 13 Feb 1867	215	4,431	3,652,425	15,974		
Kingdom X	22 Mar 1867 - 2 Nov 1870	832	18,593	15,431,672	73,099		
Kingdom XI	5 Dec 1870 - 20 Sep 1874	839	20,234	15,753,401	57,796		
Kingdom XII	23 Nov 1874 - 3 Oct 1876	380	10,864	7,709,064	25,246		
Kingdom XIII	20 Nov 1876 - 2 May 1880	802	21,688	15,003,399	64,088		
Kingdom XIV	26 May 1880 - 2 Oct 1882	546	15,556	11,218,642	53,795		
Kingdom XV	22 Nov 1882 - 27 Apr 1886	798	23,202	15,743,332	69,560		
Kingdom XVI	10 Jun 1886 - 22 Oct 1890	947	27,208	17,801,332	90,345		
Kingdom XVII	10 Dec 1890 - 27 Sep 1892	370	11,843	7,676,947	42,949		
Kingdom XVIII	23 Nov 1892 - 8 May 1895	470	15,571	9,517,141	48,417		
Kingdom XIX	10 Jun 1895 - 2 Mar 1897	358	12,048	7,552,127	42,344		
Kingdom XX	5 Apr 1897 - 17 May 1900	679	20,444	12,682,897	70,444		
Kingdom XXI	16 Jun 1900 - 18 Oct 1904	931	30,238	19,337,926	99,856		
Kingdom XXII	30 Nov 1904 - 8 Feb 1909	861	34,824	20,959,150	111,053		
Kingdom XXIII	24 Mar 1909 - 29 Sep 1913	925	38,787	23,615,958	107,512		
Kingdom XXIV	27 Nov 1913 - 29 Sep 1919	595	26,598	16,866,634	65,329		
Kingdom XXV	1 Dec 1919 - 7 Apr 1921	317	13,353	8,537,827	34,246		
Kingdom XXVI	11 Jun 1921 - 25 Jan 1924	416	16,965	10,729,054	53,071		
Kingdom XXVII	24 May 1924 - 21 Jan 1929	462	21,240	11,416,742	41,701		
Kingdom XXVIII	20 Apr 1929 - 19 Jan 1934	447	16,821	9,893,523	28,199		
Kingdom XXIX	28 Apr 1934 - 2 Mar 1939	288	10,204	6,596,115	23,621		
Kingdom XXX	23 Mar 1939 - 5 Aug 1943	51	1,278	801,341	2,144		
Consulta Nazionale	25 Sep 1945 - 1 Jun 1946	44	1,032	810,899	1,124		
Assemblea Costituente	25 Jun 1946 - 31 Jan 1948	621	13,049	10,060,817	47,479		
Republic I	8 May 1948 - 24 Jun 1953	2,098	81,878	53,870,019	249,543	22,896,326	23,767,034
Republic II	25 Jun 1953 - 11 Jun 1958	1,391	63,353	41,405,718	168,004	32,345,009	29,219,090
Republic III	12 Jun 1958 - 15 May 1963	1,486	69,046	43,170,257	162,984	30,277,524	38,314,873
Republic IV	16 May 1963 - 4 Jun 1968	1,648	90,809	56,172,083	238,068	30,674,666	51,745,102
Republic V	5 Jun 1968 - 24 May 1972	1,146	63,660	38,765,442	156,511	20,572,607	36,146,586
Republic VI	25 May 1972 - 4 Jul 1976	1,055	54,004	32,156,179	130,752	26,490,431	41,647,883
Republic VII	5 Jul 1976 - 19 Jun 1979	813	45,438	24,953,487	100,593	18,112,002	29,664,196
Republic VIII	20 Jun 1979 - 11 Jul 1983	1,291	94,024	48,376,002	196,305	32,168,053	46,563,783
Republic IX	12 Jul 1983 - 1 Jul 1987	1,236	85,906	42,999,226	147,954	111,266,296	56,041,838
Republic X	2 Jul 1987 - 22 Apr 1992	1,434	149,160	59,809,776	199,471	72,460,247	73,710,470
Republic XI	23 Apr 1992 - 14 Apr 1994	599	51,233	20,904,850	109,644	21,830,412	39,068,058
Republic XII	15 Apr 1994 - 8 May 1996	636	50,660	20,381,344	124,807	22,417,350	43,391,506
Republic XIII	9 May 1996 - 29 May 2001	1,937		63,022,325	472,397	39,910,122	25,167,396
Republic XIV	30 May 2001 - 27 Apr 2006	1,721		70,992,327	479,756	36,071,576	34,080,638
Republic XV	28 Apr 2006 - 28 Apr 2008	561		27,309,928	137,743	15,347,848	19,323,212
Republic XVI	29 Apr 2008 - 14 Mar 2013	1,598		75,138,607	350,453	39,826,686	37,251,923
Republic XVII	15 Mar 2013 - 22 Mar 2018	1,786		85,207,397	366,955	31,704,661	58,053,568
Republic XVIII	23 Mar 2018 - 12 Oct 2022	1,204		67,004,681	204,486	28,142,682	73,298,526
Total		**40,527**		**1,209,434,993**	**5,408,375**	**632,514,498**	**756,455,682**

Table 4: Statistics of the dataset.

page 13–18, New York, NY, USA. Association for Computing Machinery.

Andrej Pancur and Tomaž Erjavec. 2020. The siParl corpus of Slovene parliamentary proceedings. In *Proceedings of the Second ParlaCLARIN Workshop*, pages 28–34, Marseille, France. European Language Resources Association.

Aglaia Paoletti. 1991. La presenza femminile nelle assemblee parlamentari: Per un'analisi comparata. *Il Politico*, 56(1 (157)):77–96.

Robin Schaefer and Clemens Neudecker. 2020. A two-step approach for automatic ocr post-correction. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 52–57.

Lukas Stankevičius, Mantas Lukoševičius, Jurgita Kapočiūtė-Dzikienė, Monika Briedienė, and Tomas Krilavičius. 2022. Correcting diacritics and typos with a byt5 transformer model. *Applied Sciences*, 12(5).

10. Language Resource References

Giuseppe Abrami, Mevlüt Bagci, Leon Hammerla, and Alexander Mehler. 2022. German parliamentary corpus (gerparcor). In *Proceedings of the Language Resources and Evaluation Conference*, pages 1900–1906, Marseille, France. European Language Resources Association.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

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

Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy. 2018. Congressional record for the 43rd-114th congresses: Parsed speeches and phrase counts. Palo Alto, CA: Stanford Libraries [distributor].

Miloš Jakubíček and Vojtěch Kovář. 2010. Czechparl: Corpus of stenographic protocols from czech parliament. In *Fourth Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2010)*, pages 41–46, Karlova Studánka, Czech Republic.

Maarten Marx and Anne Schuth. 2010. DutchParl. the parliamentary documents in Dutch. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Maciej Ogrodniczuk and Bartłomiej Nitoń. 2020. New developments in the Polish parliamentary corpus. In *Proceedings of the Second ParlaCLARIN Workshop*, pages 1–4, Marseille, France. European Language Resources Association.

Andrej Pančur, Tomaž Erjavec, Katja Meden, Mihael Ojsteršek, Mojca Šorn, and Neja Blaj Hribar. 2022. Slovenian parliamentary corpus (1990-2022) siParl 3.0. Slovenian language resource repository CLARIN.SI.

Anja Virkkunen, Aku Rouhe, Nhan Phan, and Mikko Kurimo. 2023. Finnish parliament asr corpus: Analysis, benchmarks and statistics. *Language Resources and Evaluation*, 57(4):1645–1670.