A preliminary release of the Italian Parliamentary Corpus

Valentino Frasnelli[1,†], Alessio Palmero Aprosio[2,*,†]

[1]Università di Trento, Via Giuseppe Verdi 26, I-38122 Trento, Italy
[2]Fondazione Bruno Kessler, Via Sommarive 18, I-38121 Trento, Italy

Abstract

English. Political debates have been used for years in political and social studies on languages and their cultures. In this paper, we release a preliminary version of the Italian Parliamentary Corpus, a dataset containing 1.2 billion words that includes the political debates in the Italian Parliament from 1848 to 2018. The data has been collected applying an Optical Character Recognition (OCR) software to the original documents, available in PDF format on the websites of *Camera dei Deputati* and *Senato della Repubblica*.

Italian. I dibattiti politici vengono usati da anni in studi sociali e politici sulle lingue e le loro culture. In questo articolo, rilasciamo una versione preliminare dell'Italian Parliamentary Corpus, un dataset contenente 1.2 miliardi di parole che include i dibattiti politici del Parlamento Italiano dal 1848 al 2018. I dati sono stati collezionati applicando un software di Optical Character Recognition (OCR) ai documenti originali, disponibili in formato PDF sui siti web della *Camera dei Deputati* e del *Senato della Repubblica*.

Keywords

Parliamentary Corpus, Political debates, OCR post-correction, Italian Parliament

1. Introduction

The analysis of parliamentary debates is very important from many research perspectives. Apart from political science, this kind of data can be used to understand how a language and its culture evolves in history. In particular, in the last two centuries the Italian society has changed under a lot of points of view. Since the transition from the absolute monarchy to the parliamentary monarchy, that took place in 1848, Italy went through historical events such as two world wars, the fascist dictatorship, the exile of the royal family, the universal suffrage, the accession to the European Union, and much more. Such important milestones, along with all the rest of the Italian political and social life, are traced in the parliamentary reports.

Most research groups around the world have already collected and released corpora of political debates in various languages, used in diversified fields, such as religion [1] and gender [2] studies, multilinguality [3], and so on. GerParCor [4] is a dataset containing the German-language parliamentary protocols from three centuries and four countries. Similarly, siParl [5], DutchParl [6], and the Polish Parliamentary Corpus [7] are collection of political debates, in Slovenian, Dutch, and Polish languages respectively. In addition, since the creation of the European Union, political debates of the European Parliament have been made available in multiple languages, becoming a precious resource for machine translation [8].

In this paper, we present the preliminary version of the Italian Parliamentary Corpus, a collection of documents covering 200 years 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).

The rest of this article is structured as follows. In Section 2 we describe how the raw data has been collected. Section 3 we show the steps performed to get the clean texts. Section 4 contains some statistics of the dataset. Finally, both the source code and the dataset are available for download, as described in Section 5.

2. Data collection

We downloaded all the available documents available online on the websites of the two houses of the Italian Parliament.

While each website is managed by a different administration, both of them released the data in structured format (RDF for the *Camera dei Deputati*, and CSV/JSON/XML for the *Senato della Repubblica*). The *Camera dei Deputati* website contains complete catalogue of digital data and documents from the Legislature of the Kingdom of Sardinia to all the data of the Republic. Differently, for the same data belonging to the Senato della Repubblica we could directly download only documents produced after 1948. Since the debates in the 1848-1940

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
*Corresponding author.
†These authors contributed equally.
valentino.frasnelli@studenti.unitn.it (V. Frasnelli);
aprosio@fbk.eu (A. Palmero Aprosio)
0000-0002-1484-0882 (A. Palmero Aprosio)

time interval have already been digitalized, but not yet published at the time of writing, we could obtain them thanks to the precious help from the *Servizio dei Resoconti e della Comunicazione istituzionale del Senato della Repubblica*.

In both cases, documents dated before 1996 were not produced natively in a digital format, therefore are available only in PDF scanned format. Starting from 1996 (Republic Legislature number XIII), debates have been published also in text format on the web.

3. Processing

To convert PDF scanned documents to text, we used Optical Character Recognition (OCR), in particular Tesseract [9], a software originally developed by Hewlett-Packard, and subsequently released as open source. Tesseract is free to use and can support more than 100 languages out-of-the-box (among them, Italian).

After the conversion, the data is cleaned using some rule-based heuristics: headers, footers and indexes are removed, hyphenated words are joined, and pages are merged.

Finally, we needed to test the OCR output quality. To do this, we compiled a gold standard consisting of 30 pages manually transcribed, taken from different legislatures spanning from 1848 to 1996.

To evaluate the accuracy of the extraction, we use two metrics: word error rate (WER), and character error rate (CER). The error rates are derived from Levhenstein distance [10] and quantify the number of operation – insertions, deletions and substitutions – needed to transform one string in the other. They are common metrics for evaluating the performance of speech recognition and machine translation systems, but are often used also for OCR [11].

They are computed as follows:

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

where I, S, and D represent the number of insertions, substitutions, and deletions respectively. N is the total number of instances (words or character, depending on which metric is considered). The lower the value, the higher the accuracy.

As a baseline, we first evaluated the accuracy of the extraction on the output of Tesseract. Then, we applied the spell-checker software SymSpell.[1] Since SymSpell only works on words (or word-like strings), we removed all the punctuation marks from the text. We also ignore case and consider every word as lowercase.

SymSpell makes use of dictionaries for the correction of documents in the format `<word> <frequency>` for all words one wants to insert in the dictionary. Since SymSpell Italian default dictionary is build on top of recent and general purpose texts, we attempted to create dictionaries using the lexicon present in the documents themselves, trying to filter out those words containing errors. The idea is to create custom dictionaries for each legislature, containing only words coming from the time period of that legislature, in order to better capture the historical nuances for each legislature. To avoid as much as possible inserting words with spelling errors into the dictionaries, only words with a Tesseract confidence score over a user-set threshold (meaning that their recognition is likely accurate) were inserted in the dictionary. Furthermore, in order to make its creation more robust, the dictionary for a specific legislature is merged with those chronologically adjacent, meaning that dictionaries contained words from both its legislature of origin and a user-selected window of adjacent legislatures (for instance, a span of 7 legislatures mean the dictionary having on average a span of around 35 years). Figure 1 shows how the windowed dictionary system works. In theory, this allowed SymSpell to have access to both more domain specific and historically realistic lexicon in the dictionaries, instead of the Italian dictionary that comes out-of-the-box with the software.

By looking at the error made by SymSpell, it seems that most of the problems belong to proper names (such

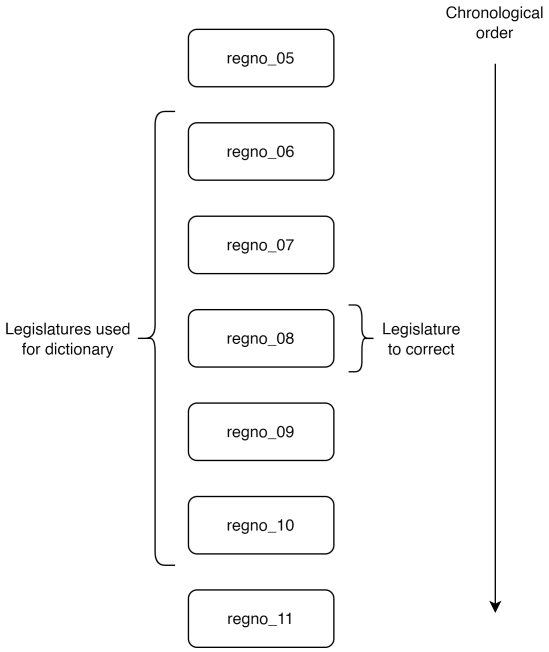

Figure 1: Example of how the dictionary used to correct the VIII legislature of the Kingdom of Italy would be constructed, with the parameter window set to 5.

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

Correction method	CER	WER
Original	0.030	0.071
SymSpell	0.036	0.121
Windowed	0.033	0.102
Windowed lower cased	0.031	0.087

Table 1
Mean CER and WER against the test set (the lower, the better).

as persons and geographical entities), that often are not included into the dictionary and are replaced by existing words very close to the apparently-misspelled term.

We then compare four different approaches: OCR plain output from Tesseract, SymSpell with the original dictionary, Symspell with the windowed dictionary, Symspell with the windowed dictionary applied only to lowercased words.

Table 1 shows the results of the four configurations. The CER and WER value calculated without applying SymSpell are lower than the other ones, resulting in a more accurate extraction. However, the use of the custom frequency list and the removal of proper nouns seems promising when compared to SymSpell applied with the original model.

By looking at the data, we can infer some useful insights. First of all, the raw text returned by Tesseract is already very precise: the Italian documents are written in a very clear font, and the digitalization has been done at a good level. The errors show that SymSpell replaced right words with wrong ones in case of proper names and very technical words, as expected.

In this first release, then, we will not use any spelling correction software, and provide the raw text extracted by Tesseract.

4. Dataset statistics

Table 2 shows some statistics of the dataset. In particular, for each legislature, one can see the number of words, pages and documents. In recent legislatures (since 1996) data is published in HTML format on the web, therefore the number of pages is not available.

5. Release

Both the data and the scripts (written in Python) are free to use and released on Github.[2]

The data contained in the *Camera dei Deputati* and *Senato della Repubblica* websites is released under the Creative Commons Attribution 3.0.[3] We use the same policy and distribute the text data under the same license.

[2]https://github.com/valefras/Italian_Parliament_Symspell
[3]https://creativecommons.org/licenses/by/3.0/

6. Conclusion and Future Work

In this paper we describe a preliminary version of the Italian Parliamentary Corpus, containing the Italian Parliament debates since 1848. In total, around 1.2 billion words have been collected.

In the future, we will further investigate OCR post-correction solutions to get cleaner data. We will also complete the data collection, by downloading and processing attachments to the parliamentary sessions, bulletins, law proposals, and reports of the Standing Committees, already available on the Italian Parliament houses websites.

We are also planning to assign each speech to the corresponding politician, and release the dataset so that anyone can use the tagging to make comparative and social studies.

References

[1] J. E. Cheng, Islamophobia, muslimophobia or racism? parliamentary discourses on islam and muslims in debates on the minaret ban in switzerland, Discourse & Society 26 (2015) 562–586.

[2] A. Paoletti, La presenza femminile nelle assemblee parlamentari: Per un'analisi comparata, Il Politico 56 (1991) 77–96.

[3] P. Bayley, Cross-cultural perspectives on parliamentary discourse, Cross-Cultural Perspectives on Parliamentary Discourse (2004) 1–390.

[4] G. Abrami, M. Bagci, L. Hammerla, A. Mehler, German parliamentary corpus (gerparcor), in: Proceedings of the Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 1900–1906.

[5] A. Pancur, T. Erjavec, The siParl corpus of Slovene parliamentary proceedings, in: Proceedings of the Second ParlaCLARIN Workshop, European Language Resources Association, Marseille, France, 2020, pp. 28–34.

[6] M. Marx, A. Schuth, DutchParl. the parliamentary documents in Dutch, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010.

[7] M. Ogrodniczuk, Polish Parliamentary Corpus, in: D. Fišer, M. Eskevich, F. de Jong (Eds.), Proceedings of the LREC 2018 Workshop *ParlaCLARIN: Creating and Using Parliamentary Corpora*, European Language Resources Association (ELRA), Paris, France, 2018, pp. 15–19.

[8] P. Koehn, Europarl: A parallel corpus for statistical machine translation, in: Proceedings of Machine Translation Summit X: Papers, Phuket, Thailand, 2005, pp. 79–86.

Table 2
Statistics of the dataset.

Legislature		Camera dei Deputati			Senato della Repubblica (del Regno)		
		Words	**Pages**	**Docs**	**Words**	**Pages**	**Docs**
Regno 01	8 May 1848 - 30 Dec 1848	1,479,030	1,365	125	345,124	329	47
Regno 02	1 Feb 1849 - 30 Mar 1849	661,745	628	59	123,546	123	22
Regno 03	30 Jul 1849 - 20 Nov 1849	1,444,180	1,344	87	345,144	331	36
Regno 04	20 Dec 1849 - 20 Nov 1853	13,028,979	11,841	691	2,891,139	2,821	319
Regno 05	19 Dec 1853 - 25 Oct 1857	9,357,294	8,702	496	1,700,528	1,758	196
Regno 06	14 Dec 1857 - 21 Jan 1860	3,059,049	3,324	180	473,000	555	63
Regno 07	2 Apr 1860 - 17 Dec 1860	1,244,804	1,159	76	261,535	304	36
Regno 08	18 Feb 1861 - 7 Sep 1865	18,131,690	19,286	809	5,026,679	5,919	451
Regno 09	18 Nov 1865 - 13 Feb 1867	3,320,700	3,801	161	532,399	612	54
Regno 10	22 Mar 1867 - 2 Nov 1870	13,281,102	15,291	603	2,656,412	3,131	229
Regno 11	5 Dec 1870 - 20 Sep 1874	12,673,461	14,827	566	3,932,011	5,305	273
Regno 12	23 Nov 1874 - 3 Oct 1876	5,895,113	7,518	245	2,247,260	3,235	135
Regno 13	20 Nov 1876 - 2 May 1880	12,637,227	16,246	530	3,646,076	5,427	272
Regno 14	26 May 1880 - 2 Oct 1882	9,465,573	12,102	396	2,030,825	3,286	150
Regno 15	22 Nov 1882 - 27 Apr 1886	13,980,213	18,326	586	2,973,795	4,737	212
Regno 16	10 Jun 1886 - 22 Oct 1890	14,748,962	19,784	633	4,672,145	7,358	314
Regno 17	10 Dec 1890 - 27 Sep 1892	6,292,409	8,633	246	1,998,816	3,116	124
Regno 18	23 Nov 1892 - 8 May 1895	8,159,069	11,820	321	2,339,573	3,751	149
Regno 19	10 Jun 1895 - 2 Mar 1897	6,092,924	8,794	233	1,979,404	3,131	125
Regno 20	5 Apr 1897 - 17 May 1900	10,369,144	14,942	432	3,460,008	5,427	247
Regno 21	16 Jun 1900 - 18 Oct 1904	16,167,079	22,135	594	4,833,899	7,842	337
Regno 22	30 Nov 1904 - 8 Feb 1909	17,480,096	25,020	574	5,774,328	9,804	287
Regno 23	24 Mar 1909 - 29 Sep 1913	19,179,900	26,890	588	6,706,423	11,711	337
Regno 24	27 Nov 1913 - 29 Sep 1919	15,438,966	21,444	394	3,184,990	5,154	201
Regno 25	1 Dec 1919 - 7 Apr 1921	6,964,613	9,728	194	2,477,546	3,609	123
Regno 26	11 Jun 1921 - 25 Jan 1924	8,050,058	11,150	243	3,599,413	5,695	173
Regno 27	24 May 1924 - 21 Jan 1929	6,770,726	9,778	246	5,909,658	11,330	216
Regno 28	20 Apr 1929 - 19 Jan 1934	6,562,034	9,616	239	4,084,703	7,001	208
Regno 29	28 Apr 1934 - 2 Mar 1939	4,017,010	5,628	150	3,174,981	4,460	138
Regno 30	23 Mar 1939 - 5 Aug 1943	424,944	626	28	350,342	562	23
Consulta Nazionale	25 Sep 1945 - 1 Jun 1946	732,609	1,012	44			
Assemblea Costituente	25 Jun 1946 - 31 Jan 1948	9,377,227	12,866	621			
Repubblica 01	8 May 1948 - 24 Jun 1953	32,044,382	42,385	1,114	26,357,526	39,282	984
Repubblica 02	25 Jun 1953 - 11 Jun 1958	27,444,116	36,615	738	17,231,858	26,559	653
Repubblica 03	12 Jun 1958 - 15 May 1963	27,924,486	37,418	789	19,417,065	31,667	697
Repubblica 04	16 May 1963 - 4 Jun 1968	33,501,680	45,096	844	27,874,392	45,368	804
Repubblica 05	5 Jun 1968 - 24 May 1972	24,326,333	33,912	549	18,257,542	29,670	597
Repubblica 06	25 May 1972 - 4 Jul 1976	19,351,600	27,820	483	15,978,505	25,972	572
Repubblica 07	5 Jul 1976 - 19 Jun 1979	17,601,367	28,463	418	10,195,047	16,975	395
Repubblica 08	20 Jun 1979 - 11 Jul 1983	35,750,707	63,837	674	17,913,360	30,014	617
Repubblica 09	12 Jul 1983 - 1 Jul 1987	29,672,360	55,945	639	17,320,038	29,961	597
Repubblica 10	2 Jul 1987 - 22 Apr 1992	42,453,808	96,131	769	22,650,712	52,571	665
Repubblica 11	23 Apr 1992 - 14 Apr 1994	11,642,821	22,920	312	11,328,792	28,313	287
Repubblica 12	15 Apr 1994 - 8 May 1996	9,778,180	20,029	326	12,897,363	30,619	310
Repubblica 13	9 May 1996 - 29 May 2001	34,825,006		878	27,613,322		1,059
Repubblica 14	30 May 2001 - 27 Apr 2006	33,929,341		757	39,427,915		964
Repubblica 15	28 Apr 2006 - 28 Apr 2008	11,492,227		278	13,066,129		283
Repubblica 16	29 Apr 2008 - 14 Mar 2013	28,685,376		739	41,579,478		859
Repubblica 17	15 Mar 2013 - 22 Mar 2018	29,946,098		863	48,645,737		923
Total		726,857,818		22,560	471,486,483		16,763

[9] A. Kay, Tesseract: An open-source optical character recognition engine, Linux J. 2007 (2007) 2.

[10] V. I. Levenshtein, Binary codes capable of correcting deletions, insertions and reversals., Soviet Physics Doklady 10 (1966) 707–710. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

[11] S. Schulz, J. Kuhn, Multi-modular domain-tailored OCR post-correction, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2716–2726. doi:10.18653/v1/D17-1288.