Compiling and Exploring a Portuguese Parliamentary Corpus: ParlaMint-PT

José Aires, Aida Cardoso, Rui Pereira, Amália Mendes
University of Lisbon - School of Arts and Humanities / Center of Linguistics
Alameda da Universidade, 1600-214 Lisboa, Portugal
joseaires74@gmail.com, aidacard@gmail.com
ruifilipedebarrospereira@gmail.com, amaliamendes@letras.ulisboa.pt

Abstract

As part of the project ParlaMint II, a new corpus of the sessions of the Portuguese Parliament from 2015 to 2022 has been compiled, encoded and annotated following the ParlaMint guidelines. We report on the contents of the corpus and on the specific nature of the political settings in Portugal during the time period covered. Two subcorpora were designed to enable comparisons of the political speeches between pre- and post-COVID-19 pandemic. We discuss the pipeline applied to download the original texts, ensure their preprocessing and encoding in XML, and the final step of annotation. This new resource covers a period of changes in the political system in Portugal and will be an important source of data for political and social studies. Finally, we have explored the political stance on immigration in the ParlaMint-PT corpus.

Keywords: parliamentary corpus, Portuguese, political views

1. Introduction

Providing access to the sessions of the national parliaments is an important tool for monitoring the democratic system. Session transcripts are frequently available and can be consulted by the population, ensuring greater transparency in a representation system carried out through elections. To ensure access to this data, several initiatives have sought to apply well-established NLP techniques, which use standards in metadata encoding and linguistic annotation, to be carried out on the data. The availability of sessions' transcripts from different national parliaments on a single website adds a new level of transparency and contributes to citizen empowerment. With this objective, the ParlaMint I project (Erjavec et al., 2023) established a first set of corpora with transcriptions of the sessions of 17 European national parliaments, uniformly encoded and with a rich set of metadata and annotated with Universal Dependencies (Erjavec et al., 2021). The first phase of the project was expanded to other European languages (ParlaMint II), including Portuguese. We report on the compilation, preprocessing and annotation of the Portuguese corpus ParlaMint-PT. This new resource allows, on the one hand, to explore the political views of the parties in relation to different topics and, on the other hand, to make contrastive analyses of the policies followed in several European countries, taking advantage of the English translation of the corpora. As a result, it is possible to compare, for example, how parties in southern European countries positioned themselves in relation to vaccination during the COVID-19 pandemic or the position of the different right-wing European parties in relation to immigration.

We present the new ParlaMint-PT corpus and provide information about its content and the levels of annotation added, as well as details about the political situation in Portugal during the period covered by the corpus, that is, from 2015 to 2022. The last legislature of the Portuguese parliament included in the corpus already points to an ongoing change in the Portuguese party system. Until then, and unlike many European countries, the political parties traditionally represented in parliament maintained great influence. The tendency for new parties to be represented in the Parliament starts in 2019 (especially parties on the right of the political spectrum) and increases in the next legislature, in 2022. The corpus thus keeps track of the first moments of an ongoing change in Portuguese politics. The ParlaMint corpora are openly available via the CLARIN.SI repository for download, as well as through the NoSketch Engine and KonText concordancers and the Parlameter interface for online exploration and analysis. The detailed metadata included in the corpora associated with the advanced searches enabled by the query programs allow for intuitive and user-friendly data analysis. We intend to show how the corpus can be useful by presenting preliminary results of work on political stances regarding immigration issues in the parliament sessions.

We discuss in section 2 other initiatives to collect and explore the transcriptions of the sessions of the Portuguese Parliament, and in section 3 we provide information on the political parties represented in the Parliament during the time frame of our corpus.

ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora, pages 12–20
May 20, 2024.

Section 4 addresses the raw data of the ParlaMint-PT corpus: we first provide some quantitative information, then details on the set of metadata in 4.1, and finally information on the structure of the sessions in 4.2. The pipeline for preprocessing, encoding texts in XML and annotation are described in 5. We provide preliminary results of a case study on the parliamentary views about immigration in section 6, and we conclude in 7.

2. Related Work on Portuguese Parliamentary Speeches

A growing number of initiatives have targeted the collection of Parliamentary data for a large set of languages. In this section, we will be concerned specifically with previous resources and projects that have collected and processed data from the Portuguese Parliament.

The first initiative to collect and explore the Portuguese parliamentary speech sessions was undertaken in the framework of a general corpus of Portuguese, the Reference Corpus of Contemporary Portuguese – CRPC (Généreux et al., 2012). The sessions of the Portuguese Parliament, as well as some legislation, are included in a large section called "politics" that contains 163M words. The CRPC corpus can be queried online in the CQP-web platform, and users can restrict their query to the subcorpus "politics"[1]. Although the data cover a large time frame of the XIX century, this corpus lacks detailed metadata that would enable queries regarding time period, political parties and speakers, and the internal sections of the speeches are not structured and properly encoded. A 1M words subset of the section "politics", the PTPARL corpus, is freely available for academic purposes on the PORTULAN CLARIN infrastructure[2], with POS annotation. This subset also lacks detailed metadata and the encoding of the internal structure of the speeches.

The website Demo.cratica was an effort to make accessible the transcripts of the Parliament sessions (A. I Carvalho and R. Lafuente (2010)). Another initiative is the Portuguese Observatory on Parliamentary Dynamics (POPaD)(Giorgi and Dias, 2019). The data has been explored in several analyses of the political system in Portugal and in a contrastive perspective with the patterns found in Europe (De Georgi and Moury, 2015).

More recently, a new compilation of Portuguese Parliamentary speeches, from 2000 to 2015, was used to explore how saliency, government dynamics, and party size affect the use of members of the parliament who specialize in specific areas of expertise in debates (Fernandes et al., 2019). The authors gathered 50,000 speeches and 6,000 bills from the Portuguese Parliament Official Website. A second initiative is the compilation of the corpus PTPARL-D, an annotated corpus of debates in the Portuguese Parliament, covering the years 1976 to 2019 (Almeida et al., 2021).

In spite of these initiatives, there was still no fully accessible set of parliamentary speeches, making use of widely known query tools, providing structured data, following standards established in the community working on parliamentary data, and providing a comparable corpus for Portuguese in line with the growing efforts congregated in the ParlaMint project. We believe that the ParlaMint-PT corpus, by using comparable structure, encoding and annotation to the other corpora of the project, will provide a crucial resource for studies on the Portuguese Parliament and for contrastive studies of the European political system.

3. The Constitution of the Portuguese Parliament – 2015-2022

A single chamber of Members constitutes the Portuguese Parliament (Assembleia da República – AR). The Members of Parliament (MPs) are elected by universal, direct and secret suffrage in legislative elections that take place every four years. The Portuguese Parliament is constituted by the Plenary (corresponding to the elected MPs' seats) and the Bureau. At the start of the legislature, the Assembly elects its President and the remaining members of the Bureau (four vice presidents, four secretaries, and four vice secretaries). The Parliament has a total of 230 seats. Parliamentary proceedings include periods for plenary sittings, parliamentary committee and parliamentary group meetings, and for MPs to spend time on constituency business. The ParlaMint-PT corpus focuses on the plenary sittings and the transcripts of these sessions, which occur, typically, three times a week. Still, it also includes solemn sessions (e.g., commemorative sessions of the *25 de Abril*, or the inauguration session of the President of Portugal).

The ParlaMint-PT corpus was created to cover the temporal period before, during and immediately after the COVID-19 pandemic, which had such an impact on the health and lives of European and global citizens. The data was intended to observe how national parliaments had addressed public health issues and the relationship between political orientation and type of proposals (for example, on vaccination).

In the Portuguese case, in addition to the pandemic period, the years covered by the corpus are also a time of major changes in the configuration of the party system. The corpus covers the last

[1] http://gamma.clul.ul.pt/CQPweb/crpc

[2] https://portulanclarin.net

10 months (January to October 2015) of the XII Legislature, and the full XIII (2015-2019) and XIV (2019-2022) Legislatures. A Legislature (Term of Office) covers the period between legislative elections.

In the 2015 legislative elections, despite the austerity policy imposed by the government of the PSD party (center-right) in the XII Legislature, PSD was surprisingly the most voted party. Nevertheless, it did not succeed in establishing a stable majority in the Parliament, and the PS (socialist party) took office, supported by governance agreements signed with the Left (PCP, PEV, BE). In the 2019 elections, the PS was the party with the most votes, although without an absolute majority. No coalition agreements were signed with parties to the left of the PS, but there were specific agreements in the Parliament for passing bills.

Between 2015 and 2022, the Parliament's configuration underwent major changes. Table 1 presents the number of speakers per party in each Legislature. Parties are identified by their acronym and are listed from Far-Left (FL), Left (L), Center-Left (CL), Center (C), Center-Right (CR), Right (R) and Far-Right (FR). No numerical information is provided in the Table when the party did not exist at the time. The ParlaMint-PT corpus covers the first three Legislatures, from 2015 to 2022. From 2019 to 2022, some of the parties saw their number of speakers decrease. This is the case of the PCP, perhaps reflecting a negative reaction from their electorate to the support given to the PS (De Giorgi and Russo, 2018), its traditional opponent since the revolution of April 25, 1974. Also, the CDS-PP significantly reduces its electorate from 24 speakers in the XII Legislature to 5 speakers in the XIV Legislature. The BE has 19 speakers in this period, surpassing the communist party PCP. The Livre party finally managed to elect a speaker during this period. And several new parties were created and quickly succeeded in electing Parliament members. The PAN party, with environmental concerns, elected 1 speaker in 2015 and increased its representation to 4 speakers in 2019. On the right wing of the political spectrum, two new parties emerged, Iniciativa Liberal (IL) and Chega, which elected 1 speaker each in 2019. With the election of a speaker from the populist party Chega, Portugal ceased to be the only country in Europe that did not have a populist far-right party with parliamentary representation.

Party	XII Leg.	XIII Leg.	XIV Leg.
PCP (FL)	14	15	10
PEV (FL)	2	2	2
BE (FL)	8	19	19
Livre (L)	0	1	1
PS (CL)	74	86	108
PAN (C)	0	1	4
PSD (CR)	108	79	77
IL (CR)	-	-	1
CDS (R)	24	18	5
Chega (FR)	-	-	1

Table 1: Number of speakers per party in each Legislature
XII=01.01.2015-22.10.2015; XIII=23.10.2015-24.10.2019; XIV=01.11.2019-01.02.2022

When the Left parties refused to approve the budget proposed by the PS in 2022, the President of the Republic dissolved the AR and called elections, resulting in the XV Legislature. This Legislature is not included in the corpus (nor in Table 1). Still, it is worthwhile to provide some information about its composition, as it shows how the 2019 vote was not an isolated moment but rather pointed to trends in the reconfiguration of the political party system. In 2022, the PS has an absolute majority; PCP and BE suffer a drastic reduction to 6 and 5 speakers, respectively; the CDS party no longer has parliamentary representation; on the contrary, the two new parties on the right increase the number of speakers from 1 to 8, in the case of Iniciativa Liberal, and from 1 to 12 in the case of CHEGA. Recently, a corruption investigation in which the Prime Minister's name was mentioned led him to resign, and the President of the Republic dissolved the Parliament. In the elections of March 2024, the parliamentary group of the party CHEGA increased to 50 speakers, a process that is reminiscent of the growth of Marine Le Pen's party in France, and of the political situation in other countries in Europe.

It would naturally be interesting to enlarge the corpus in the future to include the XV and XVI Legislatures, to study the evolution of the activities in the Parliament, the topics discussed, and also the type and register of the interventions in the sessions.

4. Parliamentary Raw Data

The Portuguese Parliamentary Corpus' raw data consists of transcripts of Portuguese Parliament sessions. These transcripts were gathered from the official Portuguese Parliament website. On the website, each transcript of the parliamentary sessions is available via the publication of the official journal of the Parliament, the Journal of the *Assembleia da República* (*Diário da Assembleia da República*). The transcripts are available for download in two file formats: TXT and PDF.

The Portuguese Parliamentary Corpus comprehends transcripts of sessions in the time period from 1 January 2015 until 22 March 2022. The cor-

Reference corpus
XII (01.01.2015-22.10.2015)
XIV (01.11.2019-22.03.2022)
XIII (23.10.2015-24.10.2019)
COVID Corpus
XIV (25.10.2019-31.10.2019)

Table 2: Time period of the Reference subcorpus and the COVID subcorpus

	Reference	COVID
Session days	499	205
Number of utterances	121,317	49,620
Number of words	11,570,662	5,882,413

Table 3: Contents of the Reference and of the COVID subcorpora

pus was divided into two subcorpora, according to the period each one covers: (i) the reference subcorpus covers sessions from 1st January 2015 until 31st October 2019; (ii) the COVID subcorpus comprehends sessions between 1st November 2019 and 22nd March 2022. The time periods considered, as well as the division into two subcorpora taking into account the start of media coverage about COVID, follow Parla-CLARIN general guidelines and proceedings for parliamentary corpora (Erjavec and Pančur, 2019). The time period of each subcorpus is provided in Table 2. Quantitative information about the number of session days, utterances and words in each subcorpus is given in Table 3.

4.1. Metadata Collection

Regarding metadata, the Portuguese Parliamentary Corpus makes available information concerning the corpus data, the speakers, the political parties, and the session files. More general information is also included, such as the type of parliament (unicameral) and the structure of the proceedings (taxonomy with types of meetings, types of speakers, legislative periods).

The Portuguese corpus provides information regarding the speaker's ID, name and surname(s), birth date, death date, gender, political affiliation (only for MPs, not for occasional speakers), and the status of the speaker (role and role description). The information regarding political parties consists of the abbreviation of the party, the full name of the party (in Portuguese), and the party ID (which is the same as the abbreviation). Finally, the metadata concerning the session files encompasses date-stamped mandates, sessions and speeches. Each session contains the transcripts of the speeches divided into utterances and paragraphs. However, the transcripts also contain the transcribers' commentary, which was retained and encoded. Each speech turn (i.e. utterance) is accompanied by the date, speaker ID, and role of the speaker (chair, regular or guest).

As for the roles attributed to speakers, the *chair* corresponds to the President of the Parliament, designated in Portuguese as *Presidente da Assembleia da República*; *regular* encompasses different situations: the prime minister, ministers and state secretaries (members of the Government), regular MPs from each party elected in legislative elections for the Portuguese Parliament, and MPs that were elected by the Parliament members as vice presidents and secretaries of the Parliament and aid the chair; the term *guest* identifies any visitor, often a member of a foreign country's Government, invited to speak in a Parliament session. The metadata files contain a description of the different roles fulfilled in public office by each member of the Parliament in different time periods and Legislatures.

The information compiled in the metadata was gathered from the official Portuguese Parliament website. This website provides webpages with political and biographic information for each politician who is or was an elected MP, secretary, vice president, or President of the Portuguese Parliament. In a few cases, the information available on the Parliament website was complemented by further research on newspaper articles or on Wikipedia pages of Portuguese politicians.

4.2. Structure of the Portuguese Parliament Plenary Sittings

In building the corpus, we must consider the structure of each plenary session and the particularities of the transcripts published in the Journal of the *Assembleia da República* (*Diário da Assembleia da República*). As it will be made clear, identifying different and regular parts of the political debates and transcriptions was crucial to the production and processing of the XML corpus.

The Portuguese Parliament plenary sittings transcripts are structured in distinctive moments, each providing various types of information that need to be encoded accordingly. The first one is the *Preamble*, which includes the identification of key features and figures in the session: (i) the date, series and number of the Journal of the *Assembleia da República*; (ii) the Legislature and session number; (iii) the date; (iv) the chair, and (v) the secretaries. After the *Preamble*, we find the *Summary*, a brief description of the interventions and votings that took place during the session. Then, we have the *Beginning of the Session*, which overlaps with the chair's first intervention and includes a time stamp. Next is the *Debate*, which corresponds to the core of the session, where we find the different speeches and interventions of the MPs. After the *Debate*, the session usually proceeds to vote on bills, and, thus, we have a section that corresponds to *Voting*. The *Closing* section follows the chair's last intervention, including a time stamp. Finally, some transcriptions end with *Written Voting Declarations*, an appendix to the session. They are not part of the debate itself but correspond to written declarations that the MPs may deliver to the Bureau in order to further justify or explain their voting during the session. We used the linguistic markers that we consistently found associated with each of these moments of the debate sessions to automatically identify the moments in the transcription files, as shown below.

As mentioned, the transcriptions include commentary by the transcribers, which were annotated by type in the XML files. These comments can occur at any moment of the debate. They pertain to pieces of information such as time, date, indication of sections such as summary, or of moments in the debate such as voting, indications of pauses in the debate, and events (e.g., an MP shows a visual aid during their intervention; the chair is replaced by the vice-chair). They may also indicate non-vocalized communicative phenomena (e.g., clapping) or vocalized, but not necessarily lexical, communicative phenomena (e.g., shouting, laughing, protests).

5. Production of the XML Corpus

In this section, we will describe how the Portuguese Parliamentary corpus was prepared for the XML generation, which required information about the actual sessions, as well as all the entities involved in those sessions.

The information about the several entities involved (people, governments, legislatures and so on) required some research so a few TSV files could be compiled and then used as a source for the needed elements. On the other hand, the information about the sessions was only available in text format, which meant they had to be processed in order to produce the corresponding XML files. However, the texts appeared to have been obtained from PDF files, which, in turn, seemed to have been obtained from OCR of the physical paper documents, considering the many issues found in them. Fortunately, after a brief inspection, we realized the texts had a fairly regular structure, with several text sections that could be used as anchors, contributing to simplifying the automation process.

We divided the text processing into several stages, which had the advantage of allowing us to focus on specific issues and keeping them localized. All the stages were carried out iteratively since a failure on a given stage might result from an error on an earlier stage. The several stages are described in the subsections below.

5.1. Preprocessing of Texts

There was a significant number of issues found in the texts, and since we planned on using text markers to identify and extract relevant information, we introduced a first stage in which we focused on fixing those issues.

This way, we could rely more confidently on such markers by ensuring a more uniform structure of the texts and avoiding the introduction of exceptions when looking for such markers. Many of the issues found consisted of cases like the following:

- missing (or extra) spaces, parentheses or dashes;
- mistaking the letter 'o' with the digit '0', and vice-versa;
- Unicode characters which looked similar and required uniformization.

These corrections were accomplished using simple regular expression replacement. Then, we proceeded to discard page headers and footers like numbers, dates, or series, which sometimes ended up between paragraphs spanning more than one page, trying to reestablish text paragraphs. A given number of empty lines, some specific separator symbols, and an initial letter casing were also considered.

Once the paragraphs were identified, we moved on to removing line breaks that did not correspond to new sentences. This was accomplished by checking the end of a line and the start of the next for composed words separated by a dash, letter casing, specific symbols and exception cases.

5.2. Main Sections Identification

At this point, we opted to identify section limits like summaries, interventions, and interruptions, which was done by looking at expressions that would indicate such cases. In our case, we could identify the following main sections:

- head: which in turn had date, session, permanent, title, president and secretary sections;
- summary: which implicated the identification of the time in which the session started;
- main: which implicated the identification of the time in which the session ended;
- final: used for any voting information.

Once these main sections were identified, we were able to improve the paragraphs further by eliminating additional line breaks that did not correspond to new sentences. At this stage, the XML document creation can be carried out in a much simpler way.

5.3. Automatic XML Generation and Checking

The preprocessing of the session files facilitated the implementation of the procedures to produce the XML files.

This time, the information about the entities was also considered to produce the final version. Even though the text files complied with a fairly regular structure, as mentioned above, we had to account for the possibility of errors, which raised the need to check if things were fine. This is why, after creating the XML document, we carried out an additional checking stage that allowed us to identify several situations in which there was missing or unexpected information, which in turn enabled us to look further into the problematic files and fix them.

During this checking stage, we found recurring errors throughout the documents, which affected the identification of utterances, paragraphs, and different types of transcribers' commentaries and events. A close reading of the texts allowed us to identify specific linguistic elements that were consistently used to introduce those sections in the transcriptions (e.g. specific adverbial expressions are used to indicate events or votings, such as *Entretanto* 'In the meanwhile', *Neste momento* 'At this moment' or *De seguida* 'Then') and what specific textual elements were associated with processing errors (e.g. punctuation marks were often associated with errors: every utterance was identified by a colon followed by a dash in the transcription, but these punctuation marks were not always correctly identified as the beginning of an utterance; periods after abbreviations were, in some cases, misidentified as an indicator of the end of paragraph). A set of expressions was compiled from these errors in order to allow an automatic search throughout the XML files. To do so, we automatized the search task by recording a macro using Notepad++, which allowed us to perform searches simultaneously in multiple files. The search results enabled us to focus our attention on a reduced set of possible problematic areas to correct any identified errors manually.

5.4. Syntactic Annotation and Main XML Files

Additional information about the session files needed to be included, namely the POS tagging and Universal Dependency Relations (UDR) identification for the session interventions, which could only be carried out after the basic XML files were produced.

The POS tagging was established using the MBT tagger (Daelemans et al., 1996) trained over the CINTIL corpus (Barreto et al., 2006). We adapted the tagset to be conformant to the UD POS tags used in ParlaMint. The CINTIL corpus includes NER annotation. We lemmatized the corpus with MBLEM (van den Bosch and Daelemans, 1999), which combines a dictionary lookup with a machine learning algorithm to produce lemmas. As a basis for the dictionary, we used a list of wordform – POS-tag combinations mapped to lemmas. This list was produced in-house. The dictionary used in MBLEM contains 102,196 word forms combined with 27,860 lemmas, leading to 120,768 wordform-lemma combinations. The adaptation of the MBT tagger and MBLEM lemmatizer are described in (Généreux et al., 2012).

The UD Relations were established using the LX-UD dependency parser[3], adapted to the set of POS and relation types used in ParlaMint. The UDR tool took a very long time to run, particularly considering the great number of session files, so it became really important to run tasks in parallel. Such parallel processing was implemented within a single file, in which we were able to carry out more than one process per sentence, as well as within a set of files, in which we were able to process several files at once. This approach allowed us to obtain results seven times faster.

6. Using ParlaMint-PT to Explore Political Views on Immigration

The topic of immigration is controversial and is frequently addressed in the programs of the political parties. As such, we expect the discussion of immigration issues and legislation proposals to be identified in the transcriptions of the Parliament sessions and to shed some light on the position of the government and of the opposition regarding the topic. The ParlaMint corpora enable us to test whether some variables are relevant to the political position of the MPs, for instance, political orientation or gender. In Europe, migration routes in the Mediterranean have put pressure on South-East countries, such as Greece and Italy, but they also affect countries in the North. Portugal has not been on the route of this migration, but, according to official numbers in the PORDATA portal[4], the foreign population officially residing in Portugal has been increasing, especially

[3] https://portulanclarin.net/workbench/lx-udparser

[4] https://www.pordata.pt/subtema/portugal/migracoes-34

since 2016, and, in 2022, reached around 800,000 people (with the total population of Portugal being around 10 million). Of the nationalities that immigrate to Portugal, the most notable are immigrants who originate from Portuguese-speaking countries, especially Brazil with 240,000 residents, and more recently, immigrants from the South-East, such as India. The latter work in large agricultural productions and, in some cases, they outnumber the local population, creating some concerns and the need for the local authorities to prepare lines of action for better integration (see, for instance, the town hall program for the integration of immigrants in Odemira (AAVV, 2015-2017)), in the South.

The press and social media have been a frequent source of data related to the perception of migration (Taylor, 2014), but parliamentary speeches are also an interesting source of data, as shown in the project "Who is the enemy now?" based on the UK and Italian ParlaMint corpora (Del Fante and Zorzi, 2023). The project reports similarities between the discourse used in both countries in spite of differences in their political backgrounds, such as the fact that the UK government was of the Conservative Party, while ministers in Italy were mostly from the left wing.

To query the corpus, we establish a list of keywords (and inflected variants) related to the foreign population living in Portugal, such as *imigração* (immigration), *imigrante* (immigrant), *migrante* (migrant), and *refugiado* (refugee). We use the version of the corpus available on Sketch Engine (Kilgarriff et al.) and extract concordances and frequencies. Here, we discuss the word *migrante(s)* that occurs 409 times in the corpus. The list of sessions where the word was most used is presented in Table 4 with the frequency of the word and the relative density (above 100% shows that the word is more frequent in this text type (session) than in the corpus). It shows a significant increase in occurrences from 2016 and 2021. This is in line with the rise of the foreign population in Portugal reported in PORDATA. The results are also aligned with the frequencies found in (Del Fante and Zorzi, 2023) for UK and Italian corpora: the word *migrant* in English and its equivalent in Italian show a strong increase in frequency in both corpora, independently of the political orientation of the government of both countries. This increase is also found in the Portuguese data, where a Center-Left party was in government during the XIII and XIV Legislature, with support from the Left parties.

Two other variables seem to be related to the use of the word *migrante* 'migrant'. One of them is the gender of the speaker, as reported in Table 5. Speakers of the feminine gender use the word more frequently than speakers of the masculine gender (243 vs. 166). Although feminine members of the Parliament are in the minority, they account for a higher number of occurrences: the relative density shows that the term is not typical of masculine Parliamentary discourse (under 100%), while it is typical of the feminine Parliamentary discourse (above 100%). Another variable is political orientation, as shown in Table 6. As the number of speakers from each party differs considerably (see Table 1), Relative density is a better indicator than raw frequency. The values in Table 6 points to a higher use of the word *migrante* by Left to Far-Left and Center-Left parties. The Right to Far-Right orientation party "Chega" has a single speaker in the Parliament and shows the highest relative density, with 179.43%).

date	**freq.**	**rel.** (%)
16-03-2016	12	1,675.86%
02-03-2017	11	1,669.85%
22-06-2018	11	2,284.92%
16-12-2020	25	2,841.95%
27-05-2021	37	6,653.24%
09-07-2021	44	8,152.84%

Table 4: Sessions with the higher frequencies of the word *migrante(s)* in ParlaMint-PT

gender	**freq.**	**rel.** (%)
F	243	180.57%
M	166	60.49%

Table 5: Distribution of the word *migrante(s) per gender of the speaker*

The concordances of the “Chega” party refer to the need to control an unbelievable flux of migrants and connect the reference to migrants to the traffic of human beings, as in example 1.

(1) precisamos de controlar o fluxo inacreditável quer de **migrantes**, quer de *tráfico de seres humanos* ‘we need to control the unbelievable flow of migrants and of the human being

party orientation	**frequency**	**rel** (%)
Left to Far-Left	147	142.08%
Left	10	48.74%
Center-Left	189	131.10%
Center-Right	38	49.03%
Center-Right to Right	18	33.58%
Right to Far-Right	7	179.43%

Table 6: Distribution of the word *migrante(s)* per the political orientation of the speaker

traffic'

The reference to a flow uses a metaphorical representation of migration as a liquid, also present in the UK and IT corpora (Del Fante and Zorzi, 2023). Three other contexts of the party "Chega" refer to the concern of the government and of the Left parties with the life/work conditions of the migrants, in contrast with those that were "born in our land" (*quem nasceu na nossa terra*). It would be interesting to analyse the XV Legislature when the party "Chega" increased its number of speakers from 1 to 12.

While political orientation is certainly important, one also needs to take into consideration the political parties. For instance, the two parties with a Left to Far-Left orientation, the Communist Party and the Bloco de Esquerda, differ in the frequency of use of the word 1. The relative density of 240.27% of the Bloco de Esquerda contrasts with the 37.86% in the case of the Communist party.

7. Final Remarks

The new open-access corpus ParlaMint-PT provides an opportunity to explore the interventions of the speakers of the Portuguese Parliament, by giving information on the topics that are addressed in the Parliament and on the views of individual speakers or political parties, or general patterns of use related to genre, time period and political orientation.

We reported on the contents of the corpus, the metadata, and the syntactic annotation. As a case study using this resource, we provided some data on the political views on immigration in the Portuguese Parliament. The automatic translation to English of the national corpora also enables comparative studies on national views over topics that are of relevance to the social and political situation of Europe today.

8. Acknowledgements

This work was partially supported by CLARIN ERIC, *ParlaMint: Towards Comparable Parliamentary Corpora*, and by Fundação para a Ciência e a Tecnologia, under the grant UIDP/00214/2020. We wish to thank the anonymous reviewers for their helpful comments.

9. Bibliographical References

AAVV. 2015-2017. Odemira integra - plano municipal para a integração dos imigrantes.

P. Almeida, M. Marques-Pita, and J. Gonçalves-Sá. 2021. PTPARL-D: an annotated corpus of forty-four years of Portuguese parliamentary debates. *Corpora*, 16(3):337–348.

F. Barreto, A. Branco, E. Ferreira, A. Mendes, M.F.P. Bacelar do Nascimento, F. Nunes, and J. Silva. 2006. Open resources and tools for the shallow processing of Portuguese. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006). Genoa, Italy*.

W. Daelemans, J. Zavrel, P. Berck, and S. Gillis. 1996. MBT - a memory-based part of speech tagger generator. *Proceedings of the 4th ACL/SIGDAT Workshop on Very Large Corpora. pp. 14–27.*

E. De Georgi and C. Moury. 2015. Government-opposition dynamics in Southern European countries during the economic crisis. great recession, great cooperation? *Journal of Legislative Studies*, 21.

E. De Giorgi and F. Russo. 2018. Portugal: The unexpected path of far left parties, from permanent opposition to government support. In E. De Giorgi and G. Ilonszky, editors, *Opposition parties in European legislatures*. Routledge.

D. Del Fante and V. Zorzi. 2023. ParlaMint - a resource for democracy. Https://www.clarin.eu/impact-stories/parlamint-resource-democracy.

T. Erjavec and A. Pančur. 2019. Parla-clarin: TEI guidelines for corpora of parliamentary proceedings. Technical report.

T. Erjavec et al. 2023. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 57(1):415–448.

J. M. Fernandes, M. Goplerud, and M. Won. 2019. Legislative bellwethers: The role of committee membership in parliamentary debate. *Legislative Studies Quarterly*, 44(2):307–343.

M. Généreux, I. Hendrickx, and A. Mendes. 2012. Introducing the Reference Corpus of Contemporary Portuguese On-Line. In *LREC'2012 – Eighth International Conference on Language Resources and Evaluation*, pages 2237–2244, Istanbul, Turkey. European Language Resources Association (ELRA).

A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlỳ, and journal=Lexicography volume=1 number=1 pages=7–36 year=2014 publisher=Springer Suchomel, V. The Sketch Engine: ten years on.

C. Taylor. 2014. Investigating the representation of migrants in the UK and Italian press: A cross- linguistic corpus-assisted discourse analysis. *International Journal of Corpus Linguistics*, 19(3):368–400.

A. van den Bosch and W. Daelemans. 1999. Memory-based morphological analysis. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, ACL'99. pp. 285–292.*

10. Language Resource References

A. I Carvalho and R. Lafuente. 2010. *Demo.cratica*. PID HTTP://demo.cratica.org/.

T. Erjavec and others. 2021. *Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1*. PID HTTP://hdl.handle.net/11356/1431.

E. D. Giorgi and A. Dias. 2019. *Portuguese Observatory on Parliamentary Dynamics Database (POPAD): information on legislative process, scrutiny activity and speeches in the Portuguese Parliament*. PID HTTPS://popad.org/.