Procesamiento del Lenguaje Natural, Revista nº 71, septiembre de 2023, pp. 125-136 recibido 31-03-2023 revisado 13-05-2023 aceptado 14-05-2023

Catalan Parliamentary Plenary Session Transcriptions from 2015 to 2022. The ParlaMintCAT Corpus

Las transcripciones de las sesiones plenarias del Parlamento de Cataluña desde 2015 a 2022, el corpus ParlaMintCAT

Marilina Pisani, Rodolfo Zevallos, Núria Bel
Universitat Pompeu Fabra
marilinapisani@gmail.com, rodolfojoel.zevallos@upf.edu, nuria.bel@upf.edu

Abstract: Parliamentary speeches are considered to be of interest for different research areas because they are publicly available transcriptions, produced under controlled and regulated procedures that add totally reliable sociodemographic data like gender, age, and other details of the speakers. Moreover, speeches are rich in topics and domains, and they are actually public domain data, not subject to copyright restrictions. The ParlaMint project: Towards Comparable Parliamentary Corpora is developing a comparable and uniformly annotated multilingual corpus with the data from 33 different parliaments in Europe. This paper describes the details of building the ParlaMintCAT corpus, for which the transcriptions of the Catalan Parliament General Assembly sessions from 2015 to 2022 have been compiled, processed and annotated.
Keywords: parliamentary corpora, ParlaMint, linguistic annotation, metadata, Catalan.

Resumen: Los discursos parlamentarios pueden resultar de interés para distintos ámbitos de investigación ya que son textos públicos, elaborados con arreglo a procedimientos regulados, a los que se han añadido datos sociodemográficos totalmente fiables como el género, la edad y otros detalles de los oradores. Además, los discursos son ricos en temas y dominios y son realmente datos de dominio público, es decir, que no están sujetos a restricciones de copyright. El proyecto ParlaMint: Towards Comparable Parliamentary Corpora está desarrollando un corpus multilingüe comparable y uniformemente anotado con datos de 33 parlamentos diferentes de Europa. Este artículo describe los detalles de la construcción del corpus ParlaMintCAT, para el que se han recopilado, procesado y anotado las transcripciones de las sesiones plenarias del Parlamento de Cataluña desde 2015 hasta 2022.
Palabras clave: corpus parlamentario, ParlaMint, anotación lingüística, metadatos, catalán.

1 Introduction

The ParlaMint project: Towards Comparable Parliamentary Corpora (Erjavec et al., 2022), financially supported by the European Infrastructure CLARIN[1], is developing a comparable and uniformly annotated multilingual corpora with texts of parliamentary sessions from 33 different Parliaments in Europe. The ParlaMint project aim is based on the interest of parliamentary speeches as there are publicly available transcriptions, produced under controlled and regulated circumstances, which are rich in topics and domains as well as related to valuable sociodemographic data like gender, age, origins of the speakers. In addition, they are publicly available and are not subject to copyright restrictions.

[1] http://www.clarin.eu

ISSN 1135-5948 DOI 10.26342/2023-71-10

The ParlaMint project has been conducted in two stages: ParlaMint I, which created and made available corpora from 17 parliaments and 16 languages of the European Union with texts from 2015 to 2020. These corpora are now available at CLARIN.si web page, and in corpus applications like NoSketch Engine[2] and SketchEngine[3]; and ParlaMint II, which has upgraded some technical aspects like the XML schema and its validation, has extended the existing corpora to cover data at least to July 2022, has added corpora for new languages including non-EU official languages like Basque, Catalan and Galician, which are co-official languages in Spain, and other languages like Turkish and Ukranian, and has proposed to further enrich the corpora with additional metadata.

In this paper, we describe the development of ParlaMintCAT corpus, which is one of the results of ParlaMint II. The ParlaMintCAT corpus is made of the transcriptions of the Plenary Assembly sessions of the Parlament de Catalunya (Parliament of Catalonia). The Parlament de Catalunya is the unicameral legislature of the autonomous region of Catalonia, in Spain. The Parlament is currently made up of 135 members, known as deputies (diputats/deputats/diputados), who are elected for four-year terms chosen by universal suffrage in lists of four constituencies, corresponding to the Catalan provinces. The Plenary Assembly is the meeting of all the deputies in which the debates referring to the following list of functions take place: the Parlament elects the President of the Generalitat de Catalunya; it passes the Catalan legislation in the business of its competence; it passes the budget and controls the action of the Government of Catalonia and the autonomous agencies, public companies and all other bodies answerable to it.

The ParlaMintCAT corpus comprises the transcriptions of the plenary sessions from January 2015 to August 2022, and includes a number of relevant historical situations like the discussions about Catalan independence process and the emergency state created with the COVID-19 pandemics, just as two examples. Plenary assembly sessions[4] of the Parlament are recorded and spoken interventions are transcribed by the Department of Editions, which oversees the production of the official publications of the Parliament and manages and coordinates the transcription of the parliamentary sessions. These are linguistically corrected and made publicly available at the web pages of the institution[5] as pdf documents, ready for paper printing as the official texts, i.e. Diari de Sessions del Parlament de Catalunya (journal of the sessions) with ISBN, and they reproduce only the spoken interventions made during the sessions.

One of the most interesting features of this corpus is its multilinguality. The speeches are in Catalan or Spanish, the languages most used by the deputies, or Aranese, a standardized variant of the Pyrenean Gascon variety of Occitan[6].

In what follows we describe the details of the building of the ParlaMintCAtT corpus. Section 2 gives information on other ParlaMint corpora, as well as other parliamentary corpora collected in different countries for different research purposes. Section 3 describes the schema proposed by ParlaMint project to rule a collection of comparable and uniformly annotated corpora from 31 different parliaments. Special emphasis is given to the implementation of the guidelines in the ParlaMintCAT corpus. Section 4 describes how the source data was processed to become ParlaMint TEI-compliant .xml files. Section 5 describes the decisions taken to linguistically annotate the data as well as the used tools. Section 6 summarizes the contributions and conclusions.

2 Related work

The ParlaMint initiative (Erjavec et al., 2022) has largely contributed to the existing catalogue of parliamentary corpora with the joint preparation of 31 parliamentary corpora from different European countries (Ogrodniczuk et al., 2022). Some of the preliminary works to ParlaMint have been described in specific publications like the Polish corpus by Ogrodniczuk and Nitoń (2020) or the Danish

[2] `https://clarin.si/noske/parlamint21.cgi/`

[3] `www.sketchengine.eu`

[4] A session is meant to be the working time to cover a particular agenda, and a meeting is meant to cover the session held in a single day.

[5] `https://www.parlament.cat/web/documentacio/publicacions/diari-ple/index.html`

[6] Aranese is spoken by about 5000 people in the Val d'Aran, in northwestern Catalonia close to the Spanish border with France. Source Wikipedia.

corpus by Jongejan, Hansen, and Navarreta (2022).

Also parliamentary corpora are the Corpus of Grand National Assembly of Turkish Parliament's Transcripts (Onur Gungor and Çağıl Sönmez, 2018) of 208 million tokens ranging from 1920 to 2015, and the Corpus of Quebec's Parliamentary Debates (Ménard and Aleksandrova, 2022) of 33.3 billion words ranging from 1908 to 2021 and with texts in French and English. Other corpora containing similar data were compiled for social sciences studies. The debate transcripts from the Hansard UK Parliament with 1.6 billion words from 1803 to 2005 were compiled by Abercrombie and Batista-Navarro (2020) for analysing sentiment and position taking; Osnabrügge, Ash, and Morelli (2023) studied how politicians use emotional resources and emotive rethoric to attract voters with two million word speeches delivered in the House of Commons and in the Dáil Éireann, the lower houses of parliament of Great Britain and Ireland, respectively; (Naderi and Hirst, 2018) created a corpus of 14,000 questions and answers from the Oral Question period of the Canadian parliamentary proceedings to study how politicians implement reputation defence strategies in their speeches.

Specifically for Catalan, Külebi et al. (2022) in the corpus ParlamentParla have used the registrations and transcripts of the Parlament de Catalunya for increasing the training data of an automatic speech recognition system. ParlamentParla contains more than 600 hours of speech that have also been gathered from the recordings of the Catalan Parliament plenary sessions from July 2007 to July 2018 that are available at the institutional website. Speech files have been aligned with the transcripts provided as pdf files. This alignment implied both the matching of the metadata coming from two different sources, i.e. speech and text, and the process to create the ASR training ready corpus. Külebi et al. (2022) reported that the most time consuming tasks to set up their corpus were the processing of the pdf files and the identification of speeches in other official languages than Catalan. As we explain in the next sections, the ParlaMintCAT corpus source files were .docx files what saved us from the problems that usually arise when using converted pdf files. Despite of this, the exploitation of .doc and .docx files has been minoritary in the ParlaMint project with only the Greek and our corpus having declared to have used this source file format. Most of the ParlaMint corpora have been obtained from .txt, .html or .pdf, in most cases requiring quite an important effort for adding the structure and metadata required by the project.

3 ParlaMintCAT corpus structure and ParlaMint schema

ParlaMintCAT Corpus is made of the transcriptions of the Plenary Assembly sessions of the Parlament de Catalunya. The source files were provided to us by the Departament d'Edicions del Parlament de Catalunya to whose director we are indebted. Table 1 provides a quantitative description of the corpus.

Item	**Quantity**
Speakers	365
Documents/Sessions	286
Tokens Total	15,667,673
Tokens Catalan	13,115,625
Tokens Spanish	2,505,953
Tokens other languages	46,095
Temporal span	01-2015/08-2022

Table 1: Quantitative description of ParlaMintCAT corpus.

As just mentioned, the ParlaMint project's aim is the creation of comparable and uniformly annotated multilingual corpus of parliamentary sessions out of 31 different corpora, in different languages and following different formats. ParlaMint has focused on maximizing the interest of these corpora by developing a framework of interoperability and homogeneous encoding. To that purpose, ParlaMint has defined a specialized schema based on the Text Encoding Initiative (TEI) Guidelines (Truan and Romary, 2021). The ParlaMint schema rules the corpus structure, the formal specifications of the documents that compose the corpora, the corpus metadata, the description of speakers and political parties, the encoding of transcriber's notes, and the structure of the texts and speeches transcripts. ParlaMint also requires to process the corpus to get linguistically annotated versions of the

transcripts including annotation of part of speech, morphosyntactic description, named entities, and grammatical dependencies.

Therefore, the source files were processed according to the ParlaMint schema and requirements. The final corpus is composed of two master files (ParlaMint-ES-CT.xml and ParlaMint-ES-CT.ana.xml), that contain the top level element `teiCorpus`, the `teiHeader` with corpus-wide metadata and the `Xinclude` elements gathering the 286 files that are the components of the corpus.

Following ParlaMint convention, the TEI-compliant xml version of the source files were named with the prefix `ParlaMint-ES-CT` to identify the source country and language of the data with the international standard ISO3166-2. After the prefix, each file was identified with the official index of the session it contains. This official index is made of the date and the number and part of the session.

Each corpus component file has its own TEI header followed by the transcription text. The component header contains specific information: the corpus-wide unique title and ID, the type of meeting according to the ParlaMint taxonomy, the number of the term, session, etc. The main content of each component file is, obviously, the transcription of the speeches, which are encoded as utterance elements (`<u>`), with a reference to the speaker unique ID and the type of speaker, ie., chair, regular member or guest. Each utterance is described as a number of segments (`<seg>`), which are enriched with the notes of the transcribers like 'applause', 'noise', etc., if available in the source data. In Figure 1, a sample of the xml shows the encoding of utterances, segments, and notes.

For every such a Parlamint-ES-CT TEI .xml compliant file, there is a file named with the suffix 'ana.xml' with the linguistically annotated texts: part of speech, morphosyntactic description, named entities and grammatical dependencies. Other files identified as Parlamint-ES-CT are: the speakers and the organizations files, that contain the schema of the ParlaMintCAT metadata: names of speakers, organizations as parliamentary groups, political parties, etc. All the encoding is explained in the ParlaMint Guidelines: The structure and encoding of ParlaMint Corpora, that it is publicly available at the github of the ParlaMint project[7].

In order to check that the component files, both TEI .xml and ana.xml were correct in terms of the ParlaMint schema, a number of validation tools[8] were provided by the project. If the files pass the validation, 5 additional files with the following suffixes are created: text(.txt), characters(.tbl), speakers(.meta), vertical format(.vert) and CoNLL format morphosyntactic analysis(.conllu).

In the next sections we describe the different steps to convert source documents from .docx to TEI-xml format, and to further process the resulting .xml files to build the ana.xml with the required linguistic information.

4 Processing the source files, from .docx to TEI compliant .xml

The source files were provided to us in .docx format, allowing us to retrieve the .xml document containing relevant information encoded in different styles' metadata (Pisani, 2022). In the source files of the ParlaMint-CAT corpus, various Microsoft Word tools were used by transcribers to encode:

- Information of the nature of the texts. For instance: Agenda, Summary, Introduction, Speeches, Notes.
- The name, position, or both, of the speaker whose speech follows.
- The language of the speech, as identified by the spell checker.

We extracted the structure and metadata from the .xml file within the .docx document with xml.etree[9] and stored them in a Pandas' DataFrame. Figure 2 shows in a DataFrame the relation of text and styles as collected.

Table 4 describes the columns of the DataFrame where information gathered from .docx is stored. Utterance (`<u>`) and segments (`<seg>`) are numbered after being identified from the style 'D3Textnormal'. The language of the segments provided by source files was only referring to other languages than Catalan, and some processing

[7] `https://clarin-eric.github.io/ParlaMint/`

[8] `https://github.com/clarin-eric/ParlaMint/blob/main/Makefile`

[9] `https://docs.python.org/3/library/xml.etree.elementtree.html`

```
   <seg xml:id="ParlaMint-ES-CT_2018-01-17-0101.16.0.27" xml:lang="ca">Molt bé. Completada,
   <note>(Es procedeix a l'escrutini i, després, al recompte.)</note>
   <seg xml:id="ParlaMint-ES-CT_2018-01-17-0101.16.0.28" xml:lang="ca">Molt bé. Procedit, c
   <vocal type="murmuring">
      <desc>(Veus de fons.)</desc>
   </vocal>
   <seg xml:id="ParlaMint-ES-CT_2018-01-17-0101.16.0.29" xml:lang="ca">Em sembla que els re
   <seg xml:id="ParlaMint-ES-CT_2018-01-17-0101.16.0.30" xml:lang="ca">Tenim, doncs, per or
   <seg xml:id="ParlaMint-ES-CT_2018-01-17-0101.16.0.31" xml:lang="ca">Queden, doncs, procl
   <seg xml:id="ParlaMint-ES-CT_2018-01-17-0101.16.0.32" xml:lang="ca">Fins aquí les votaci
   <kinesic type="applause">
      <desc>(Aplaudiments perllongats.)</desc>
   </kinesic>
   <seg xml:id="ParlaMint-ES-CT_2018-01-17-0101.16.0.33" xml:lang="ca">Aquí acaba la nostra
   <seg xml:id="ParlaMint-ES-CT_2018-01-17-0101.16.0.34" xml:lang="ca">Moltes gràcies a tot
   <note>(Elisenda Alamany Gutiérrez demana la paraula.)</note>
   <seg xml:id="ParlaMint-ES-CT_2018-01-17-0101.16.0.35" xml:lang="ca">Perdó, hi ha una der
</u>
<note type="speaker">Elisenda Alamany Gutiérrez</note>
<u xml:id="ParlaMint-ES-CT_2018-01-17-0101.17.0" who="#AlamanyElisenda" ana="#regular" xml:
   <gap reason="inaudible">
      <desc>(L'oradora comença a parlar sense fer ús del micròfon, motiu pel qual no n'han
   </gap>
   <seg xml:id="ParlaMint-ES-CT_2018-01-17-0101.17.0.0" xml:lang="ca">Nosaltres volíem mani
   <seg xml:id="ParlaMint-ES-CT_2018-01-17-0101.17.0.1" xml:lang="ca">La nostra decepció, p
   <seg xml:id="ParlaMint-ES-CT_2018-01-17-0101.17.0.2" xml:lang="ca">En primer lloc, aques
   <seg xml:id="ParlaMint-ES-CT_2018-01-17-0101.17.0.3" xml:lang="ca">Per tant, crec que qu
   <seg xml:id="ParlaMint-ES-CT_2018-01-17-0101.17.0.4" xml:lang="ca">Gràcies.</seg>
   <kinesic type="applause">
      <desc>(Aplaudiments.)</desc>
   </kinesic>
</u>
```

Figure 1: Example of the TEI-xml encoding of utterances, segments, speaker, and of different transcribers' notes: murmuring, applause and inaudible.

Columns	Description
Text	Text per Paragraph
Lang	Language of the text
Utterance_id	Speech text ID
Segment_id	Speech text paragraph ID
Tag	Type of Paragraph
Style	Word style formatting
Curs	cursive or not
Speaker_id	Speaker ID
Speaker_role	Speaker Role
Speaker_name	Speaker Name

Table 2: All the metadata and features used to convert .docx to ParlaMint .xml were collected in a DataFrame with the columns shown in this table.

was required as explained below in Section 4.3. The speaker was identified because of the source 'D3Intervinent' style encoding.

Moreover, the analysis of the resulting DataFrame was very useful to detect source inconsistencies that needed to be corrected before further processing as described below in sections 4.1 and 4.2 (Pisani, 2022). For instance, a considerable number of times, the name of the speaker was encoded as normal text; transcriber's notes describing non-linguistic events in the room were embedded in the text and were not identified with the corresponding style; differences in the identification of speakers needed to be harmonized, for instance, when speakers are named by the position, not by the person name, or because there were spelling variations in the name of the speakers that had to be identified uniquely.

Once the DataFrame has all the data obtained from source files, ParlaMint TEI-compliant .xml files[10] are built.

[10] https://github.com/IULATERM-TRL-UPF/ParlaMint_ES-CT/blob/main/src/to_xml.py

index	text	style
0	XI legislatura - primer període - sèrie P - nú...	D3Textnorma
1	Sessió 3, tercera i darrera reunió, dijous 12 ...	D3Textnorma
2		D3Textnorma
3	Ple del Parlament	Crgar
4	Presidència de la M. H. Sra. Carme Forcadell i...	CPresidncia
5		D3Textnorma
6	SESSIÓ 3.3	D2Davantal-Sessic
7	La sessió, suspesa el dia 10 de novembre, es r...	D2Davanta
8	Al banc del Govern seu el president de la Gene...	D2Davanta
9		D3Textnorma
10	La presidenta	D3IntervinentObertura
11	Es reprèn la sessió.	D3Textnorma
12	D'acord amb l'article 4.3 de la Llei de la pre...	D3Textnorma
13	(Veus de fons.) Senyor García Albiol, per què ...	D3Textnorma
14	Xavier García Albiol	D3Intervinen

Figure 2: Sample of the information encoded in .docx sytles.

4.1 Notes and comments of transcribers as metadata

The most successful example of reusing source .docx metadata was the identification of the different comments that transcribers' used in the source files and its conversion into ParlaMint metadata (Pisani, 2022). After the analysis of the source documents, three different types of comments were identified: title, note and interruption.

Interruptions were the most frequent case of transcriber's comments in the ParlaMint-CAT corpus, however there were cases where the interruptions were added by transcribers' in parenthesis in the middle of the speech, with no particular style format. In these cases, resulting TEI-xml encoding takes the note as a discourse marker that divides the original paragraph into two segments.

The notes, which contained textual information about interruptions of the speech and their nature (applause, noise, etc.), were also extracted after a thorough analysis of the data. An example of the resulting encoding is shown in Figure 1. The actual xml encoding followed the recommendations of TEI (Truan and Romary, 2021) that suggest the following types:

- `<pause>` to mark short voluntary interruptions of the speech
- `<vocal>` to identify vocal noises like laughs or disagreement signals
- `<kinesic>` to identify gestures or applause
- `<incident>` for other events that are not vocalizations or non vocalized communicative expressions.

In addition to TEI recommendations, ParlaMint schema introduces the element `<gap>` to indicate omissions, that is, parts of a speech that were lost because of technical reasons, for instance, like shown in Figure 1, where the transcriber describes the reasons of the gap, that the speaker has started to speak without microphone, and therefore there is no recording, as (`<gap reason="inaudible">`).

4.2 Identification of speakers and their metadata

ParlaMint has made an important effort to include a rich number of metadata about the speakers that participated in the parliamentary sessions. This rich information is meant to promote political, sociological, sociolinguistic and linguistic studies for which speaker-related information is necessary[11]. For ParlaMintCAT corpus, the information about the 365 speakers was mainly found in the website of the Parliament, the website of the Catalan Autonomous Government and in Viquipèdia, the Catalan version of Wikipedia.

All speakers are described at the file `ParlaMint-ES-CT-listPerson.xml`. Each speaker, identified as `<person>`, gets there a unique identifier that is used to mark all his or her speeches. In Figure 2 we can see an example of such encoding, that, according to the project guidelines, includes full name, birth year, political party membership, parliamentary role or position (i.e. regular member, secretary, chair person), or if applicable, role or position in other organizations, mostly applicable to members of the government. In turn, the references to parties, governments, etc. are found in the file `ParlaMint-ES-CT-listOrg.xml` with the information recommended by the ParlaMint project Guidelines. In Figure 3, we show a sample of the encoding of a speaker, also with his roles at the autonomous government.

As mentioned before, it was not unusual a certain variation in the spelling of speakers

[11] For instance, an example of the use of these metadata is a recent result of a ParlaMint Hackaton about citation networks in different parliaments to discover a bias about citing females. Consulted in `https://www.clarin.eu/impact-stories`.

```
<person xml:id="AragonèsPere">
   <persName>
      <surname>Aragonès</surname>
      <nameLink>i</nameLink>
      <surname>Garcia</surname>
      <forename>Pere</forename>
   </persName>
   <sex value="M"/>
   <birth when="1982"/>
   <affiliation ref="#party.ERC" role="member"/>
   <affiliation role="member" ref="#PG.JxSi-XI"/>
   <affiliation role="member" ref="#PG.ERC-XIV"/>
   <affiliation ref="#PC" role="member" from="2015-10-26" to="2016-01-21"/>
   <affiliation ref="#PC" role="member" from="2018-01-17" to="2020-12-18"/>
   <affiliation ref="#PC" role="member" from="2021-03-12" to="2021-05-25"/>
   <affiliation ref="#GOV" role="head" from="2021-05-26" to="2022-11-06"/>
       <affiliation ref="#GOV" role="member" from="2021-05-26" to="2022-11-06"/>
   <affiliation ref="#GOV" role="minister" from="2018-05-29" to="2021-05-25"/>
       <affiliation ref="#GOV" role="member" from="2018-05-29" to="2021-05-25"/>
</person>
```

Figure 3: Example of the encoding of one speaker, with biographic information (birth date and gender) and with information about his belonging to a political party, the terms he was deputy and the terms and roles as member of the Catalan Autonomous Government.

names along the different sessions. This is mainly due to the use of Catalan and Spanish spellings that might differ in accentuation rules. For instance, 'María', with an accented 'i' is the Spanish correct spelling, while in Catalan the correct spelling is 'Maria', with no accent. Note also that both in Catalan and Spanish, two family names are the norm for identifying a person. In Catalan, these two names can be written with a coordination or without it. It depends on the transcriber that a person is identified, for instance, as 'Pere Aragonès i Garcia' or 'Pere Aragonès Garcia', but also 'Pere Aragonès García' with the Spanish spelling of the second name, a very typical Spanish family name. In order to make that in all these cases the speaker gets the same unique person identifier, we compiled a list of possible variations. The name of the person in the person's file is the most frequent spelling found in the files. Another list for identifying the speaker when only the name of the position was mentioned in the source files was also used, for instance: 'La presidenta' (the chairwoman) in Figure 2, index 10, referring to 'Carme Forcadell'. Note that the name of the position also showed some variation.

4.3 A multilingual corpus with language identified at <seg> level

As already mentioned, the Catalan Parliament has three official languages: Catalan, Spanish and Aranese. Aranese use is occasional, Catalan is the most frequent language, as shown in Table 1. While the normal case would be that each speaker consistently uses only one particular language, there are cases of code switching: a speaker can switch from one language to another in the same speech and therefore the language has to be encoded at the level of <seg>. The correct identification of the language is not only important because of the metadata, but also because it is unavoidable for linguistic processing: the correct Natural Language Processing (NLP) tools must be selected.

Although the .docx files already contained a tag identifying the language of paragraphs written in a language different to Catalan, we additionally used a language identifier[12] that could guess the language with less than 200 characters to verify that indeed, unmarked paragraphs were written in Catalan.

[12]Google language detector cld3

5 *Linguistic processing: creating the .ana.xml files*

ParlaMint project guidelines require the linguistic annotation of the speeches, more technically, of the `<seg>` elements. Therefore, ParlaMintCAT `<seg>` elements have been segmented into sentences, or `<s>`, and each sentence has been tokenized into wordforms, `<w>`, punctuation symbols, `<pc>`, and inter-token spaces.

A special tokenization case for Catalan was because this language, as other Romance languages, writes contractions of prepositions and articles: 'pel' instead of 'per el' (by the), and personal pronouns attached to verb forms with an hyphen like in example 1. ParlaMint guidelines propose to split these wordforms into their components. In example 1, we see how these wordforms are analysed as words associated with the attribute norm and the linguistic annotation described above.

```
<w xml:id="59"> governar-se
  <w xml:id="57"
  msd="UPosTag=VERB|Mood=Inf"
  norm="governar"
  lemma="governar" />
  <w xml:id="58"
  msd="UPosTag=PRON"
  norm="-se"
  lemma="es" />
</w>
```

Example 1: xml construction of the word 'governar-se', with detailed information about the parts that make up the construction.

For each wordform, the lemma, Universal Dependencies[13] (UD) part-of-speech, and morphological description are encoded in respective attributes (lemma, pos and msd) of the `<w>` element. Named entities are identified, classified into the standard four classes: location (LOC), person (PER), organization (ORG) and miscellaneous (MISC) and marked like `<name>` elements and type attribute.

Additionally, the UD dependency parse is added as the encoding of the grammatical relations among sentence's wordforms (Nivre et al., 2020) and stored in the `<linkGrp>` element. The `<linkGrp>` is composed of `<link>` elements, which encode the syntactic relation (`ana` attribute) between two wordforms, which are represented by its unique ID. The tags are defined in the linguistic taxonomies that are common to all ParlaMint corpora.

For linguistic processing we tested a number of libraries for getting the required annotation of Catalan texts: Spacy (Honnibal and Montani, 2017), Stanza (Qi et al., 2020), UDPipe 2 Models (Straka, 2022) and FreeLing (Padró and Stanilovsky, 2012). After a comparison of the performance of these tools, for Catalan we noted that:

- Stanza had no NER tool for Catalan.
- At the time of testing, Spacy modules for lemmatization performed with difficulties with the parliamentary texts [14].
- UDpipe also had some difficulties for segmenting and lemmatizing our texts. For instance 'segregui' lemmatizes to 'segreure' instead of 'segregar'.
- FreeLing delivered NER with the standard notation, but a special script was required to transform Freeling format into BIO format compliant with ParlaMint xml encoding.
- FreeLing splits contractions and verb-clitic forms, but we had to recreate the full wordform to follow ParlaMint requirements.
- FreeLing syntactic dependencies use a tagset that is not compatible with the UD tagset.

Eventually, named entities, lemmatization and PoS tagging both for Catalan and Spanish, were done with FreeLing PoS tagger v4.2[15], and tags and morphosyntactic descriptions were mapped to UD PoS tags and msd descriptions to follow ParlaMint guidelines. For dependency relations annotation, the Catalan and Spanish modules (version 220711) of the Universal Dependencies 2-10 models were used UDpipe2[16]. In order to make it compatible with the FreeLing segmentation and tokenization, the input to the parser was the vertical format as produced by FreeLing.

For the merging of the information coming from FreeLing and from UDPipes, for both

[13] `https://universaldependencies.org/`

[14] The lemmatizer was updated and delivered a standard performance while we were already processing the texts with FreeLing

[15] `https://nlp.lsi.upc.edu/freeling/node/1`

[16] `https://ufal.mff.cuni.cz/udpipe/2`

```
<s xml:id="ParlaMint-ES-CT_2018-01-17-0101.16.0.13.3" xml:lang="ca">
   <w xml:id="ParlaMint-ES-CT_2018-01-17-0101.16.0.13.3.1" msd="UPosTag=VERB|Mood=Ind|Tense=Pres|Person=3|Number=Sing" lemma="començar">Comença</w>
   <w xml:id="ParlaMint-ES-CT_2018-01-17-0101.16.0.13.3.2" msd="UPosTag=DET|PronType=Art|Gender=Fem|Number=Sing" lemma="el">la</w>
   <w xml:id="ParlaMint-ES-CT_2018-01-17-0101.16.0.13.3.3" msd="UPosTag=NOUN|Gender=Fem|Number=Sing" lemma="senyora">senyora</w>
   <name type="PER">
      <w xml:id="ParlaMint-ES-CT_2018-01-17-0101.16.0.13.3.4" msd="UPosTag=PROPN" lemma="ribas">Ribas</w>
   </name>
   <pc xml:id="ParlaMint-ES-CT_2018-01-17-0101.16.0.13.3.5" msd="UPosTag=PUNCT|PunctType=Peri">.</pc>
   <linkGrp type="UD-SYN" targFunc="head argument">
      <link ana="ud-syn:root" target="#ParlaMint-ES-CT_2018-01-17-0101.16.0.13.3 #ParlaMint-ES-CT_2018-01-17-0101.16.0.13.3.1"/>
      <link ana="ud-syn:det" target="#ParlaMint-ES-CT_2018-01-17-0101.16.0.13.3.3 #ParlaMint-ES-CT_2018-01-17-0101.16.0.13.3.2"/>
      <link ana="ud-syn:nsubj" target="#ParlaMint-ES-CT_2018-01-17-0101.16.0.13.3.1 #ParlaMint-ES-CT_2018-01-17-0101.16.0.13.3.3"/>
      <link ana="ud-syn:appos" target="#ParlaMint-ES-CT_2018-01-17-0101.16.0.13.3.3 #ParlaMint-ES-CT_2018-01-17-0101.16.0.13.3.4"/>
      <link ana="ud-syn:punct" target="#ParlaMint-ES-CT_2018-01-17-0101.16.0.13.3.1 #ParlaMint-ES-CT_2018-01-17-0101.16.0.13.3.5"/>
   </linkGrp>
</s>
```

Figure 4: Example of the linguistic annotation of one `<seg>` 'Mrs. Ribas starts' with lemma, POS, morphosyntactic description, named entity, and Grammatical Relations markup following the UD tagset (Nivre et al., 2017).

languages Catalan and Spanish, we created a DataFrame[17] to order the data like CoNLL format files. Finally, the DataFrame with the morphosyntactic analysis data was merged with the DataFrame of the TEI .xml file creating the new .ana.xml file using the Python XML library.

Finally, to assess the reliability of linguistic annotation we validated the quality of the annotations on the Parliamentary transcription texts, which can be said to constitute a particular domain. A subset of sentences amounting about 1000 tokens and randomly selected were reviewed by three expert annotators. As for PoS tagging and NER, the quality of the annotation is very high, close to 100% accuracy for NER. All the 23 named entities in the test set were identified and only 2 of them were wrongly classified. For PoS, only 46 PoS tagging errors were found, mostly grammatical words that occur with apostrophe in the text and are tagged as abbreviations. The evaluation of annotation of universal dependencies was done in terms of accuracy for attachment and labeling. The results were 96.9% of tokens that have been assigned the correct dependency and 94.4% the correct label. The results of our evaluation are in line with declared UAS and LAS of the model.

6 Conclusions

In this paper, we have described the ParlaMintCAT corpus, a newly created corpus that contains the transcriptions of the plenary assembly sessions of the Parlament de Catalunya from January 2015 to August 2022. We have also described the most important technical details for the conversion of a collection of word documents in .docx files into a well formed, enriched corpus, following the ParlaMint project requirements. We have reported about how the ParlaMint TEI-based guidelines have been interpreted and applied and the issues we had to solve because of the nature of the texts and of the language.

Summing up, we haven given details about the major tasks that had to be undertaken for the conversion of .docx files to TEI .xml, ParlaMint-compliant files:

- Identification and classification of all types of notes and transcribers' comments.
- Language identification and annotation for bilinguality and code switching phenomena.
- Identification of speakers, mentioned with position or by name solving the issues derived of Catalan vs. Spanish different spellings.

We have also reported about the choice of tools for the linguistic processing of the files and why and how FreeLing and UDpipes for Catalan were used for getting PoS tagging, morphosyntactic descriptions, NER, and grammatical dependencies annotation.

Because of the amount of data enriched with high quality metadata, ParlaMintCAT corpus can be of interest to researchers in humanities and social sciences for discourse, genre, sociology, politics, etc., but also to NLP developers interested in areas like political orientation detection, speaker identification, etc. The relation of ParlaMintCAT with all the ParlaMint corpora in more than 30 different languages, all of them covering similar topics like COVID19 pandemics, is also of great interest for studies related to multilinguality.

[17] https://github.com/IULATERM-TRL-UPF/ParlaMint_ES-CT/blob/main/src/util_freeling.py

Acknowledgments

We are very thankful to Ivan Antiba, who started this project, and to the Departament d'Edicions of the Parlament de Catalunya, specially to its director Nei Torrell Camps, and to Neus Pinart Bartolí for their friendly collaboration. This work was supported with CLARIN.eu and Project PID2019-104512GB-I00, Ministerio de Ciencia, Innovación y Universidades and Agencia Estatal de Investigación (Spain) funding. Rodolfo Zevallos was supported with a FI grant of the Departament de Recerca i Universitats, Generalitat de Catalunya.

Bibliografía

Abercrombie, G. and R. Batista-Navarro. 2020. ParlVote: A corpus for sentiment analysis of political debates. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5073–5078, Marseille, France, May. European Language Resources Association.

Erjavec, T., M. Ogrodniczuk, P. Osenova, N. Ljubešić, K. Simov, A. Pančur, M. Rudolf, M. Kopp, S. Barkarson, S. Steingrímsson, et al. 2022. The parlamint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 02.

Honnibal, M. and I. Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Jongejan, B., D. Hansen, and C. Navarreta. 2022. Enhancing clarin-dk resources while building the danish parlamint corpus. In *Selected Papers from the CLARIN Annual Conference 2021, Virtual Event, 2021, 27–29 September / Monica Monachini and Maria Eskevich (eds.)*. Linköping Electronic Conference, Linköping, Sweden.

Külebi, B., C. Armentano-Oller, C. Rodriguez-Penagos, and M. Villegas. 2022. Parlamentparla: A speech corpus of catalan parliamentary sessions. In *Proceedings of The Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 125–130, Marseille, France, June. European Language Resources Association.

Ménard, P. A. and D. Aleksandrova. 2022. A French corpus of Québec's parliamentary debates. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 25–32, Marseille, France, June. European Language Resources Association.

Naderi, N. and G. Hirst. 2018. Automatically labeled data generation for classification of reputation defence strategies. In D. Fišer, M. Eskevich, and F. de Jong, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Nivre, J., Z. Agic, L. Ahrenberg, L. Antonsen, M. J. Aranzabe, M. Asahara, L. Ateyah, M. Attia, A. Atutxa, L. Augustinus, E. Badmaeva, M. Ballesteros, E. Banerjee, S. Bank, V. B. Mititelu, J. Bauer, K. Bengoetxea, R. A. Bhat, E. Bick, V. Bobicev, C. Börstell, C. Bosco, G. Bouma, S. Bowman, A. Burchardt, M. Candito, G. Caron, G. Eryiğit, G. G. A. Celano, S. Çetin, F. Chalub, J. Choi, S. Cinková, Çagri Çöltekin, M. Connor, E. Davidson, M.-C. de Marneffe, V. C. V. de Paiva, A. D. de Ilarraza, P. Dirix, K. Dobrovoljc, T. Dozat, K. Droganova, P. Dwivedi, M. Eli, A. M. Elkahky, T. Erjavec, R. Farkas, H. F. Alcalde, J. Foster, C. Freitas, K. Gajdoova, D. Galbraith, M. García, M. Gärdenfors, K. Gerdes, F. Ginter, I. Goenaga, K. Gojenola, M. Gökirmak, Y. Goldberg, X. G. Guinovart, B. G. Saavedra, M. Grioni, N. Gruzitis, B. Guillaume, N. Habash, J. Hajic, L. H. My, K. Harris, D. T. T. Haug, B. Hladká, J. Hlavácová, F. Hociung, P. Hohle, R. Ion, E. Irimia, T. Jelínek, A. Johannsen, F. Jørgensen, H. Kaşıkara, H. Kanayama, J. Kanerva,

[18] http://creativecommons.org/licenses/by/4.0/"

T. Kayadelen, V. Kettnerová, J. Kirchner, N. Kotsyba, S. Krek, V. Laippala, L. Lambertino, T. Lando, J. Lee, P. L. Hong, A. Lenci, S. Lertpradit, H. Leung, C. Y. Li, J. Li, K. Li, N. Ljubesic, O. A. Loginova, O. Lyashevskaya, T. Lynn, V. Macketanz, A. Makazhanov, M. Mandl, C. D. Manning, C. Maranduc, D. Marecek, K. Marheinecke, H. M. Alonso, A. Martins, J. Masek, Y. Matsumoto, R. T. McDonald, G. Mendonça, N. Miekka, A. Missilä, C. Mititelu, Y. Miyao, S. Montemagni, A. More, L. M. Romero, S. Mori, B. Moskalevskyi, K. Muischnek, K. Müürisep, P. Nainwani, A. Nedoluzhko, G. Nespore-Berzkalne, L. N. Th, H. T. M. Nguyen, V. Nikolaev, H. M. Nurmi, S. Ojala, P. N. Osenova, R. Östling, L. Ovrelid, E. O. Pascual, M. Passarotti, C.-A. Perez, G. Perrier, S. Petrov, J. Piitulainen, E. Pitler, B. Plank, M. Popel, L. Pretkalnina, P. Prokopidis, T. Puolakainen, S. Pyysalo, A. Rademaker, L. Ramasamy, T. Rama, V. Ravishankar, L. Real, S. Reddy, G. Rehm, L. Rinaldi, L. Rituma, M. D. Romanenko, R. Rosa, D. Rovati, B. Sagot, S. Saleh, T. Samardić, M. Sanguinetti, B. Saulite, S. Schuster, D. Seddah, W. Seeker, M. Seraji, M. Shen, A. Shimada, D. V. Sichinava, N. Silveira, M. Simi, R. Simionescu, K. I. Simkó, M. Simková, K. I. Simov, A. Smith, A. Stella, M. Straka, J. Strnadová, A. Suhr, U. Sulubacak, Z. Szántó, D. Taji, T. Tanaka, T. Trosterud, A. A. Trukhina, R. Tsarfaty, F. M. Tyers, S. Uematsu, Z. Uresová, L. Uria, H. Uszkoreit, S. Vajjala, D. R. van Niekerk, G. van Noord, V. Varga, E. V. de la Clergerie, V. Vincze, L. Wallin, J. N. Washington, M. Wirén, T. sum Wong, Z. Yu, Z. Žabokrtský, A. Zeldes, D. Zeman, and H. Zhu. 2017. Universal dependencies 2.1.

Nivre, J., M.-C. de Marneffe, F. Ginter, J. Hajič, C. D. Manning, S. Pyysalo, S. Schuster, F. Tyers, and D. Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France, May. European Language Resources Association.

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

Ogrodniczuk, M., P. Osenova, T. Erjavec, D. Fišer, N. Ljubešić, Ç. Çöltekin, M. Kopp, and M. Katja. 2022. Parlamint ii: The show must go on. In *Proceedings of The Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 1–6, Marseille, France, June. European Language Resources Association.

Onur Gungor, M. T. and Çağıl Sönmez. 2018. A corpus of grand national assembly of turkish parliament's transcripts. In D. Fišer, M. Eskevich, and F. de Jong, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Osnabrügge, M., E. Ash, and M. Morelli. 2023. Cross-domain topic classification for political texts. *Political Analysis*, 31(1):59–80.

Padró, L. and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *LREC2012*.

Pisani, M. 2022. *Árboles, Gráficos y Matrices de Datos. Codificación en TEI de un Corpus de Interacciones Parlamentarias con Python.* Final Master Thesis. Máster en Humanidades y Patrimonio Digitales. Universidad Autónoma de Barcelona. `https://github.com/marilinapisani/docx2tei_ParlaMint/blob/main`.

Qi, P., Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Straka, M. 2022. Universal dependencies 2.10 models for UDPipe 2 (2022-07-11). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Truan, N. and L. Romary. 2021. Building, Encoding, and Annotating a Corpus of Parliamentary Debates in XML-TEI: A Cross-Linguistic Account. *Journal of the Text Encoding Initiative*, (14).