*Proceedings of ParlaCLARIN III @ LREC2022* pages 117–124
Marseille, 20 June 2022

# Making Italian Parliamentary Records Machine-Actionable: The Construction of the ParlaMint-IT Corpus

**Tommaso Agnoloni**[1], **Roberto Bartolini**[2], **Francesca Frontini**[2], **Carlo Marchetti**[3]
**Simonetta Montemagni**[2], **Valeria Quochi**[2], **Manuela Ruisi**[3], **Giulia Venturi**[2]
[1]CNR-IGSG, Firenze Italy; [2]CNR-ILC, Pisa Italy; [3]Senato della Repubblica, Roma Italy
[1]tommaso.agnoloni@igsg.cnr.it, [2]name.surname@ilc.cnr.it, [3]name.surname@senato.it

### Abstract

This paper describes the process of acquisition, cleaning, interpretation, coding and linguistic annotation of a collection of parliamentary debates from the Senate of the Italian Republic covering the COVID-19 pandemic emergency period and a former period for reference and comparison according to the CLARIN ParlaMint prescriptions. The corpus contains 1199 sessions and 79,373 speeches for a total of about 31 million words, and was encoded according to the ParlaCLARIN TEI XML format. It includes extensive metadata about the speakers, sessions, political parties and parliamentary groups. As required by the ParlaMint initiative, the corpus was also linguistically annotated for sentences, tokens, POS tags, lemmas and dependency syntax according to the universal dependencies guidelines. Named entity annotation and classification is also included. All linguistic annotation was performed automatically using state-of-the-art NLP technology with no manual revision. The Italian dataset is freely available as part of the larger ParlaMint 2.1 corpus deposited and archived in CLARIN repository together with all other national corpora. It is also available for direct analysis and inspection via various CLARIN services and has already been used both for research and educational purposes.

**Keywords:** parliamentary debates, CLARIN ParlaMint, corpus creation, corpus annotation

## 1. Introduction

Parliamentary data is an interesting source of data for various types of investigations and analyses, in addition to the obvious applications in political studies. Due to their richness and uniqueness, parliamentary records have in fact been a fundamental resource for several research questions in different disciplines of the humanities and social sciences for the last half century (see Fišer and Lenardič (2018) for a brief overview of the different fields of studies). With the recent push towards open data and open participation, it becomes increasingly important to release actionable data sets of parliamentary debates and records in order to support empirical research and development of integrated analytical tools. Projects have recently flourished in many countries and for many languages on the construction of corpora of parliamentary debates as language resources to be used in language- and content-based web applications in order to support political discourse studies. An overview of existing projects in this sense can be found by consulting the dedicated CLARIN Resource Family of Parliamentary Corpora[1]. One of the most widely used is the Hansard Corpus, providing historical and contemporary data for the British Parliament[2].

Building on this landscaping work, the CLARIN infrastructure has decided to fund the ParlaMint project for fostering the creation of a harmonised, comparable multilingual corpus of parliamentary data in order to boost the field of comparative studies and the uptake of digital language technologies into political social sciences and cultural studies.

In this work we describe the steps taken and the challenges faced for the construction of the Italian section of the ParlaMint corpus, following the common guidelines and approach defined by the ParlaMint community. The Italian corpus is therefore constructed so as to be interoperable with all other ParlaMint corpora and is therefore comparable with the growing collection of national data sets.

The paper is structured as follows: in section 2 we first provide the background of our work. Sections 3 and 4 describe the steps that led to the creation of the corpus from the original data obtained from the Senate and its conversion and structuring according to the ParlaMint format. Section 5 describes the automatic linguistic annotation of the texts of the debates and discusses related issues, while section 6 reports on the conversion tool developed to transform the CoNLL-U annotations into the required ParlaMint.ana format. Finally, section 7 concludes by summarising the main difficulties encountered with an indication of the potential improvements for future works; furthermore, a brief mention is made to some of the current applications of the corpus.

## 2. Background and Context

The Italian Parliament is a perfect bicameral system and consists of the Senate (www.senato.it) and the Chamber of Deputies (www.camera.it). The two chambers are autonomous and independent also as regards technical, administrative and organisational aspects.

The Senate is organised into parliamentary groups according to the political party each senator belongs to; a

[1] `https://www.clarin.eu/resource-families/parliamentary-corpora`

[2] `https://www.clarin.ac.uk/hansard-corpus`

mixed group is foreseen for those senators whose formations do not reach at least 10 members and for senators not enrolled in any component. Senators representing linguistic minorities are allowed to form a Group composed of at least five members. Senators-for-life, in the autonomy of their legitimacy, may not become part of any Group[3].

The CLARIN ParlaMint project is a recent initiative, financially supported by CLARIN ERIC, that aims at the creation of a machine-actionable multilingual set of corpora of parliamentary debates, i.e. which can be directly analysed by online tools and technology for dealing with language-based content, esp. in the fields of political social sciences and cultural studies. Started as a small pilot project for 4 languages and parliaments, the project now comprises 17 languages and is in the process of expanding[4]. Given its initial focus on emergencies, the corpora focus on the COVID-19 pandemic period and include data from a previous period to be used as a reference; details for the Italian corpus will be given in section 3 below; for common issues on the multilingual corpus see Erjavec et al. (2022). An important achievement of the project was the definition of a common TEI format[5] which follows the ParlaCLARIN recommendations (Erjavec and Pančur, 2019)[6] but further constrains the schema for ensuring full comparability of the corpora across languages. All language datasets come in two variants: the ParlaMint TEI corpora (Erjavec et al., 2021a), which contain fully marked-up text of the transcribed speeches, and the linguistically annotated corpora (Erjavec et al., 2021b), which add linguistic annotation to the marked-up version of the texts.

## 3. Construction of the ParlaMint-IT Corpus

This section describes the creation of the Italian ParlaMint TEI corpus, that is the XML version containing the marked-up transcriptions, which constitute the core part of the work.

The very first and fundamental step in the creation of the Italian corpus of parliamentary debates is the acquisition, cleaning and structuring of the stenographic verbatim records of the required parliamentary sittings. As a consequence of the total independence of the two Italian parliament chambers, the processes for the production, digitisation and publication of the stenographic transcriptions of the sessions of the Chamber of Deputies and of the Senate are different and still not interoperable. For time constraints and practical reasons, in this work we deal with data from the Senate only. The covered time-span ranges from March 15 2013 (i.e. the beginning of 17th legislative term) to November 18 2020 (i.e. the date of the last data retrieval, corresponding to the current 18th legislative term). The whole corpus consists of 1199 files, one for each plenary session.

The COVID-19 sub-corpus contains sessions starting from November 1 2019 (as conventionally agreed for all ParlaMint corpora) and contains 115 sessions/files; the reference sub-corpus instead contains 1084 sessions/files, covering the preceding period. The source documents containing the transcriptions were made available in bulk by the Information Technology Service of the Senate. The same documents can however also be retrieved directly from the Senate website[7], so that the whole process should be reproducible. Although, starting from 2018, the transcripts of the plenary sessions of the Senate are also published in the Akoma Ntoso format (Palmirani and Vitali, 2011)[8], in order to uniformly cover transcripts from all the required time spans (thus including years before 2018), the HTML format was chosen as the source format for the whole corpus. Moreover, the HTML files contain additional annotations (for speeches, speakers, etc.), expressed by means of proprietary XML tags "embedded" into the HTML encoding; particularly useful is the original segmentation into utterances (tag <INTERVENTO>) and speaker annotation (tag <ORATORE>), which had to be ported to the TEI ParlaMint encoding.

Before the actual TEI encoding could start, the original HTML corpus was therefore pre-processed with the goal of extracting all the "embedded" XML annotation and discarding (almost) every HTML annotation. This produced an intermediate XML corpus which retained only paragraphs, <p>, and italic formatting, <i> HTML tags, as they represent textual segments and (potentially) parenthetical expressions. A small fraction of the intermediate files (62/1199) required manual correction in order to force XML well-formedness for their subsequent DOM (Document Object Model) parsing.

At this stage, only the transcripts of the speeches are kept from the original HTML corpus for the subsequent ParlaMint encoding (tag <RESSTEN> , i.e. *Resoconto Stenografico* 'verbatim report'). Annexed documents for the session, if any, are discarded in the current release, as these are not in focus within the project. Plenary verbatim records in fact usually are attached with annexes such as the texts of control and policy-setting documents. In particular, the texts of the motions, questions and interpellations spoken during the plenary are published in Annex B when they are submitted to the House, whereas resolutions and other documents introduced during the discussion are published in Annex A

[3] https://www.senato.it/en/

[4] https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora

[5] https://github.com/clarin-eric/ParlaMint/blob/main/README.md

[6] https://clarin-eric.github.io/Parla-CLARIN/

[7] https://www.senato.it/Leg17/3809?testo_generico=911

[8] www.akomantoso.org

of the report of the sitting.

ParlaMint additionally requires that corpora have extensive metadata (speaker name, gender, party affiliation, MP or guest status) and that each speech is marked with its speakers and their role(s) (chair, regular speaker, etc.).

In our case, metadata about the members of the Senate and the political groups were obtained from the open data portal http://dati.senato.it, and stored into structured metadata tables, as described in section 4 below. The main purpose of the Senate data portal is to make available, in open and freely reusable formats, most of the data already published on the institutional website of the Senate; this in order to ensure greater transparency on the work of the institution and to encourage the concrete participation of citizens in decision-making processes. In particular, the data referring to the composition of the Senate, to the bills, and to the activity of the senators (presentation of documents and bills, interventions and electronic voting) are openly published starting from the XIII legislature, and are updated on a daily basis. The project is coordinated with a similar initiative by the Chamber of Deputies (http://dati.camera.it).

## 4. Data Encoding in the ParlaMint TEI Format

The corpus of speeches was encoded in the ParlaMint format by developing a specific transformer that reads the input documents and data, transforms them into the required target structure and writes the ParlaMint-XML output. The input of this transformer consists in the aforementioned intermediate XML corpus and structured metadata tables (in comma separated values format). The reading, transformation and writing is implemented by means of Java XML DOM manipulation[9].

The expected output, as specified by the ParlaMint documentation, consists of a corpus root file, and a set of XML documents with the transcriptions of the plenary sessions, one session per file. The corpus root file must contain the general corpus header providing all corpus-level metadata (such as edition, funding, contributors, etc.) and includes the list of files that encode the actual transcriptions of the parliamentary sittings. It must also include all controlled vocabulary terms encoded in the form of taxonomies, i.e. metadata about speakers, political groups, parties, government in charge; these are referred to in the actual transcription files via the appropriately created term ids.

For the encoding of the corpus root file, the required metadata about speakers and political groups were mostly automatically obtained by querying the Senate Data portal dati.senato.it where data is represented in RDF according to the Ontologia Senato della Repubblica ontology (OSR)[10] and exposed both through a SPARQL endpoint and via up-to date structured open data for the most common predefined queries[11]. In particular, we collected the list of senators and the composition of parliamentary groups with all the changes that have occurred during the legislature.

For speakers who are not members of the Senate (mostly members of the government in charge who might either be members of the Chamber of Deputies or not members of the Parliament at all) manual insertion of their metadata was necessary and done by accessing their personal pages from the Senate website.

For all speakers we were thus able to populate data about: gender, date and place of birth (reconciled with persistent URIs from the GeoNames dataset[12]), affiliation to political groups over time, and link to the personal web page on the institutional website. Metadata about governments in charge over time, role of speakers in governments, coalition of political groups supporting or opposing the governments, also required manual encoding using institutional web pages as sources.

The collected data was then appropriately transformed in the target structures required by the ParlaMint Schema. Consistent interlinking of speakers with speeches was guaranteed thanks to the use of the same identifiers in all the source data and documents, appropriately transformed following the ParlaMint naming conventions for identifiers (*i.e.* human readable ids).

The rest of the corpus root is composed and structured by hard-coding in the Java source code the desired output for the different XML elements. The encoding of the document corpus of transcriptions was accomplished by parsing into a DOM the intermediate XML documents, traversing the documents and applying the appropriate transformations from the source elements to the target elements. Text not belonging to speeches was mapped onto <note> elements. Whenever possible, the type of the note is assigned via the @type attribute, with the following possible values: “role”, “speaker”, “time”, “summary”, “voting”. Incident annotations are detected among the italics <i> HTML annotation in the source files, based on a heuristic applied to the content of the tagged text (i.e. a list of keywords triggering incident text). In a similar way, the type of parenthetical clauses (kinesic, vocal, parenthesis and their type attribute) are annotated based on a heuristic analysis of their textual content. For example, if the text originally marked in italics contains words like *brusio* ‘buzz’, *commenti* ‘comments’, *ilarità* ‘hilarity’, *proteste* ‘complaints’ or *richiami* ‘admonitions’, then a <vocal> tag with type “noise”, “speaking”, “laughter” or “shouting” is used for the annotation of that portion of text. Particular attention is paid to the removal

[9] https://github.com/atomm/ResAulaSenato2ParlaMintTEI

[10] http://dati.senato.it/osr/; for a graphic visualisation see also https://dati.senato.it/DatiSenato/browse/19?testo_generico=15

[11] https://dati.senato.it/DatiSenato/browse/scarica_i_dati

[12] http://www.geonames.org/

of those HTML tags that are useless for the purpose of a TEI encoding (e.g. <i> tags not denoting parenthetical expressions or <a> HTML links) without breaking the resulting text segments with carriage returns or useless punctuation, which would result in a noisy input to the subsequent linguistic analysis pipeline. The speakers' identifiers, available in the source documents, were mapped to the identifiers used in the corpus root and kept consistent.

## 5. Automatic Linguistic Annotation

The automatic linguistic annotation of the corpus has been articulated in two stages. The first one includes the following levels of analysis: sentence splitting, tokenisation, part-of-speech tagging, lemmatisation and dependency parsing. Annotation was performed by the STANZA neural pipeline[13] which is reported to achieve state-of-the-art or competitive performance for different languages (Qi et al., 2020). The choice was motivated by the fact that the pipeline uses the annotation formalism devised in the Universal Dependency (UD) initiative (Nivre, 2015), which was a project requirement for guaranteeing interoperability and comparability with all other ParlaMint corpora. Among the different Italian available models, we used the *italian-isdt-ud-2.5* model, trained on the Italian Stanford Dependency Treebank, which represents the biggest UD Treebank for Italian covering different textual genres (Bosco et al., 2013).

The second stage consisted in the automatic Named Entity Recognition (NER). Since the STANZA package did not include a NER model for the Italian language at the time we performed the annotation, and NER annotation was a mandatory requirement, it was carried out by running, on the same raw data, the ItaliaNLP NER module (Dell'Orletta et al., 2014) [14], which assigns three standard named entity tags – i.e. Person, Organisation, Location – and achieves state-of-the-art performance.

Both tools output the annotated texts in CoNLL format, but follow different tokenisation approaches: STANZA tokenises according to UD principles, namely subtokenises agglutinated forms such as complex prepositions (e.g. *della* 'of+the.fem' becomes di la 'of the.fem') or verbs with enclitic pronouns (e.g. *farlo* 'to-do+it' becomes fare lo), while the *ItaliaNLP NER* does not (*della* and *farlo* are considered simple tokens). The outputs of the linguistic and Named Entity annotation therefore had to be post-processed for re-alignment in order to produce a unified annotation.

For this last step, a number of alignment rules were defined specifically devoted to handling mismatches. This step turned out to be cyclic, as conversion errors revealed exceptional cases of misalignment that needed to be tackled with new heuristics.

Even though both the linguistic annotation pipeline and the NER module used here achieve state-of-the-art performance for Italian, it is well known that Machine Learning algorithms suffer from a drop of accuracy when tested on domains outside of the data on which they were trained (Gildea, 2001). Speech transcriptions of parliamentary debates represent a language variety which differs from the written language testified in the used training corpora: we can thus look at them as Out-of-Domain texts for which the results of the automatic linguistic annotation need to be carefully assessed.

For this reason, we felt that the impact of the linguistic peculiarities of the language variety of the corpus on the performance of automatic linguistic and NE annotation, both from a quantitative and qualitative perspective, needed to be investigated. With this goal in mind, we started manually revising the automatic annotation of speech transcriptions of parliamentary debates: the result of this process, still ongoing at the time of writing, will be used as an evaluation benchmark. Preliminary results achieved so far show that language-specific features of the debates from the COVID pandemic period negatively affect the performance of automatic annotation, more than features from the debates of the earlier period. We hypothesise that this follows from the fact that the earlier debates belong to a specific variety of language use, which Nencioni (1976) identifies as 'spoken-written', i.e. a variety characterised by an hybrid nature featuring a co-occurrence of traits typical of both written and spoken language. Thus, they are linguistically more similar to the written texts which the linguistic annotation tools were trained on. In addition, they contain several normative references (e.g. *article 5 of law n. 184, paragraph 2, states [...]*) that make the transcriptions more similar to a written legal text. On the contrary, the debates of the COVID-19 period are mostly characterised by traits specific to the spontaneous speech (such as rethorical questions in interrogative forms to convey illocutionary force to an assertion, e.g. *is it ever possibile [...]*, interruptions), since the debates deal with issues that are more emotionally engaging given the historical period, such as the prison riot which happened in March 2020 calling for better anti-COVID measures.

## 6. Data Encoding in the TEI ParlaMint *.ana* Format

The unified CoNLL-U format obtained with the post-processing described above had finally to be back-converted to the ParlaMint TEI *.ana* format, i.e. the final format for the encoding of linguistic annotation. For this task a converter was developed in C++[15] which takes in input both the original ParlaMint-IT.xml cor-

[13] https://stanfordnlp.github.io/stanza/index.HTML

[14] http://www.italianlp.it/demo/t2k-text-to-knowledge/

[15] https://github.com/cnr-ilc/conllu2Parlamint_TEI

pus and the unified CoNLL-U dataset and outputs a valid ParlaMint-IT.*ana* corpus.

Similarly to the main corpus version, the expected output of the linguistic annotation consists in a corpus root file and a number of `XML` files encoded according to the *ParlaMint.ana* format. Each file represents one parliamentary session and contains the linguistically annotated transcriptions of all the speeches occurring in that session. The root.*ana* file instead adds two taxonomies to the corresponding root file of the not annotated base corpus: i.e. a taxonomy for Named Entities and a taxonomy for the syntactic dependency relations, which explicitly define the tags and categories used in the annotation files.

The ParlaMint format encodes the basic linguistic annotation in-line, according to the TEI Lightweight Linguistic Annotation guidelines (Bański et al., 2018) and therefore encodes sentences (<`s`>), word tokens (<`w`>), punctuation symbols (<`pc`>) as `XML` elements, while the rest of the basic linguistic information is encoded in the form of attributes of <`w`>. All the original morpho-syntactic information is concatenated into the `@msd` attribute[16]. Agglutinated forms that are treated as multi-token forms in UD are represented as nested words <`w`>[17].

Named Entities are represented in a similar way, with the outer element being <`name`>. Syntactic dependencies are represented instead by a <`linkGrp`> element under <`s`> which groups all dependency relations as <`link`>. Each link must specify the relation type, and refer to head and dependent tokens. The relation type itself is a pointer to the categories defined in the previously mentioned <`taxonomy`> in the root file.

The non-annotated input corpus is structurally divided into folders by years, from 2013 to 2020; each folder/year contains all the sessions of that year in TEI `XML` format. Each session contains sections, motions, information on the speakers and their roles, all encoded with the appropriate tag; the text of the speeches is divided into small narratives, i.e. segments, encoded with the <`seg`> tag (as described in section 3 above). Each segment may contain several utterances/sentences, and was automatically annotated linguistically as described in section 5 above. Each session therefore contains a variable number of segments each of which has a unique identifier.

The input unified CoNLL-U dataset has a parallel structure: each session has associated a folder of linguistically analysed files, in CoNLL-U tabular format, which correspond exactly to the number of segments <`seg`> the session contains. The names of the files maintain the semantics of the sessions: i.e. the files corresponding to the session `YYY.xml` of a certain year, which contains $N$ segments, correspond to $N$ CoNLL-U files contained in the $YYY$ folder and are identified by the file names `YYY.seg1.udner`, `YYY.seg2.udener`, `...YYY.segN.udner`. The content of the CoNLL-U file consists of a list of linguistically annotated word forms, one for each line, corresponding to the tokenisation of the text; a blank line separates the sentences[18].

In this format, in the case of multi-token items the numeric identifier can be a range; for Italian this occurs very frequently in the representation of agglutinated forms such as complex prepositions and enclitic particles attached to verbs as in the example below:

```
5     .........
6-7 dei _ _ _ _ _ _ _
start_char=178|end\char=181
6 di di ADP E _ 8 case _ _
7 i il DET RD  Definite=def|Gender=Masc|
Number=Plur|PronType=Art 8 det _ _
8      ..........
```

The production of the linguistically annotated TEI Corpus takes place through a Manager that moves within the CoNLL-U dataset structure described above by applying a conversion code from the CoNLL-U format to the ParlaMint TEI `XML`. The Manager handles one year (folder) at the time and takes as input the list of `XML` files that compose each year of parliamentary sessions. Each file, which represents a session, is parsed and the text contained in each <`seg`> is replaced by the result of the actual CoNLL-U to `XML` converter which parses the corresponding *.udner* file and produces an `XML` sub-tree with <`seg`> as the parent node. In this way there is a one-to-one correspondence between the starting (non-annotated) files and the linguistically annotated ones.

The heart of the production process of the linguistically annotated parliamentary corpus is represented by the transformation of the CoNLL-U files (representing linguistically annotated segments) into the corresponding segment of the `XML` file. All annotation contained in the CoNLL-U files was therefore converted to the TEI ParlaMint linguistic annotation common format[19] according to the following conversion algorithm:

1. First, a <`seg`> node is generated which replaces

[16] For details and examples see `https://clarin-eric.github.io/parla-clarin/#sec-linguistic`

[17] Although this creates mixed content and thus may not be the ideal representation, we adhered to the project specification. The interested reader might want to follow the discussion that led to this decision here `https://github.com/clarin-eric/ParlaMint/issues/61`

[18] For practical purposes, the fields of the CoNLL-U data structure are stored in an variable *entry* and are the well known *id*, *form*, *lemma*, *upos*, *xpos*, *feats*, *head*, *deprel*, plus an additional field for encoding *named entities*. For details on Universal Dependencies see also `https://universaldependencies.org/format.HTML`

[19] More details can be found at `https://clarin-eric.github.io/parla-clarin/#sec-linguistic`

the homologous node of the input XML file. This node has an @xml:id attribute obtained by concatenating the name of the file with the year, and the session and segment indices.

2. The CoNLL-U (*.udner*) file is parsed one line at a time and every time an empty line is encountered a new sentence <s> is generated as a child node of the current <seg>. The node <s> also receives an @xml:id attribute generated by concatenating a progressive numeric index (a sentence counter) to the father's id. All sentence nodes then will always be children of the current segment node.

3. For each input line a word <w> or a punctuation node <pc> is generated as an ordered child of the current node[20]. The *entry.form* value is stored as the content of <w>[21] while the attributes of the node are populated as follows:

   - @xml:id: by concatenating the id of the parent node with the value of *entry.id*;
   - @lemma: with the *entry.lemma* value;
   - @pos: with the *entry.upos* value;
   - @msd: by concatenating the value of *entry.xpos* and *entry.feat*;

4. At the end of each sentence, a child <linkGrp> node of the current sentence is also generated and will have as many <link> child nodes as there are functional relations in the sentence. The <link> node has the following attributes:

   - @ana: is the entry value *deprel*;
   - @target: contains references to the xml:id of both the head and the dependent;

In addition to being required, this conversion step was also useful for identifying errors in the alignment of the outputs of the linguistic and Named Entity annotations described in section 5 above, and entered the cyclic revision process that led to the final clean version.

## 7. Conclusions and Future Work

In this paper we have described in detail the far from trivial process that produced the Italian section of the ParlaMint corpora (Erjavec et al., 2021a), and we have learnt that: 1) the most onerous part of the work is the structuring of the base parliamentary debates corpus which involved cleaning, transformation and interpretation via heuristics of the transcripts in their original, non-standard and loosely structured format; 2) because of the differences in tokenisation, annotating linguistic features and NEs with different tools generated a substantial processing overhead for performing a good re-alignment. The experience of using different tools for linguistic annotation and NER thus proved too time-consuming and error prone.

Currently, we are taking part in the second phase of the ParlaMint project, in which we will extend the COVID-19 sub-corpus with new sessions starting from December 2020 onwards. In this context, we will explore the possibility of deriving the data from the Akoma Ntoso format, which might provide a useful result for a wider community due to its becoming good practice in a number of government bodies in several countries all over the world. As regards corpus annotation, in order to avoid tokenisation mismatches, we plan to employ the same pipeline for performing all token-based annotations, with STANZA still being the best candidate. Ideally, we would like to train a specific NER model for Italian in the neural pipeline; however, since a model has become available in the meantime, we will first experiment with it, and assess the quality of the results first.

As regards exploitation, the corpus has already been used in preliminary political studies. For instance, Cavalieri and Del Fante make use of the ParlaMint-IT corpus to compare topics discussed in Senate plenary sessions with the expenditure across budget categories. According to the authors, combining the analysis of parliamentary debates carried out by means of quantitative text analysis techniques with the analysis of final expenditures trade-offs "helps to better grasp dynamics which public budgeting is subject to and constitute a very promising venue for future research both for political science and linguistic scholars" (Del Fante and Cavalieri, 2021).

The availability of the Italian corpus, together with the other ParlaMint corpora, is also greatly beneficial for pedagogical applications. The corpus for instance was selected by a team of students of the 2021 Helsinki hackathon[22], who plotted timelines of COVID-word frequencies using relative occurrences, and added a curve indicating the number of COVID cases, thus illustrating the relation between the parliamentary debates and the epidemiological situation in four countries, including Italy. Furthermore, the corpus is currently being used in Italian universities, for various purposes. For classes in computational linguistics, the interest lies especially in the linguistically annotated sources. From this perspective, it has been used in the framework of the teaching activities of two of the authors in a Computational Linguistics course addressed to Master students within the "Digital Humanities" degree program at the University of Pisa. By manually

[20] The sentence node can also be the parent of <name> nodes, which capture information on NEs. The specific conversion algorithm for NEs is reported in the Appendix to the paper.

[21] In cases of multi-token items the word node can also include other <w> sub-nodes.

[22] *Parliamentary Debates in COVID Times* https://www.clarin.eu/impact-stories/helsinki-digital-humanities-hackathon-2021-parliamentary-debates-covid-times

revising the output of automatic linguistic annotation, students are confronted with the real problems connected with the automatic analysis of specific varieties of language use. Students from the "Master in Gestione e Conservazione e dei Documenti Digitali" at the University of Calabria are querying the ParlaMint corpus via the NoSketch Engine platform[23], thus learning basic notions of corpus annotation, usage of metadata for sub-corpus selection and querying. Also, a tutorial in Italian aimed at non-computationally savvy researchers in political and social sciences with no prior knowledge of corpus linguistics is now available (Del Fante, 2022), and might encourage further studies.

The availability of machine-actionable transcripts of parliamentary debates are indeed an important asset for many disciplines, but, especially for political studies, they are not enough. Future work might thus include not only an extension of the corpus with debates of the other parliamentary chamber and from other time periods, but also with other types of data connected with the parliamentary sittings, such as the annexed documents mentioned in section 3 above: e.g. bills under discussion, voting records, and so on.

Finally, the possibility will be explored to start a project dedicated to adding audio-video files of the sittings, possibly linked and aligned to the transcripts.

## 8. Acknowledgements

This work was supported by CLARIN-ERIC as part of the ParlaMint project, as well as by CLARIN-IT, CNR-IGSG and CNR-ILC as part of their missions. Special thanks go to Andrea Cimino for his valuable help with the alignment and merging of linguistic and NER annotations.

## 9. Bibliographical References

Bański, P., Haaf, S., and Mueller, M. (2018). Lightweight grammatical annotation in the TEI: New perspectives. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Del Fante, D. and Cavalieri, A. (2021). Politics in between crises. A political and textual comparative analysis of budgetary speeches and expenditure. In *Proceedings of the 1st Workshop on Computational Linguistics for Political Text Analysis*.

Del Fante, D. (2022). ParlaMint – IT – Il corpus del Senato Italiano. Una guida pratica per l'interrogazione del corpus ParlaMint-IT con NoSketch Engine, a supporto dell'analisi del discorso politico. `https://doi.org/10.5281/zenodo.6526914`.

Dell'Orletta, F., Venturi, G., Cimino, A., and Montemagni, S. (2014). T2K^2: a System for Automatically Extracting and Organizing Knowledge from Texts. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2062–2070, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Erjavec, T. and Pančur, A. (2019). Parla-CLARIN: TEI guidelines for corpora of parliamentary proceedings, September.

Erjavec, T., Ogrodniczuk, M., Osenova, P. N., Ljubesic, N., Simov, K. I., Pancur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çagri Çöltekin, de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Pérez, M. C., Macedo, L. D., Navarretta, C., Luxardo, G., Coole, M., Rayson, P., Morkevicius, V., Krilavicius, T., Dargis, R., Ring, O., van Heusden, R., Marx, M., and Fiser, D. (2022). The parlamint corpora of parliamentary proceedings. *Language Resources and Evaluation*, pages 1 – 34.

Fišer, D. and Lenardič, J. (2018). Clarin resources for parliamentary discourse research. In Darja Fišer, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Gildea, D. (2001). Corpus variation and parser performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*.

Nencioni, G. (1976). Parlato-parlato, parlato-scritto, parlato-recitato. *Strumenti critici*, (29).

Nivre, J. (2015). Towards a universal grammar for natural language processing. In *Computational Linguistics and Intelligent Text Processing - Proceedings of the 16th International Conference, CICLing 2015, Part I*, pages 3–16, Cairo, Egypt, April.

Palmirani, M. and Vitali, F. (2011). Akoma-ntoso for legal documents. In Giovanni Sartor, et al., editors, *Legislative XML for the Semantic Web.*, volume 4 of *Law, Governance and Technology Series*. Springer, Dordrecht.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (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*, pages 101–108, Online, July. Association for Computational Linguistics.

## 10. Language Resource References

Bosco, Cristina and Montemagni, Simonetta and Simi, Maria. (2013). *Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank.* Association for Computational Linguistics, PID https://aclanthology.org/W13-2308.

Erjavec, Tomaž and Ogrodniczuk, Maciej and Osenova, Petya and Ljubešić, Nikola and Simov, Kiril

[23] Available from `https://www.clarin.si/noske/index.HTML`

and Grigorova, Vladislava and Rudolf, Michał and Pančur, Andrej and Kopp, Matyáš and Barkarson, Starkaður and Steingrímsson, Steinh̑ór and van der Pol, Henk and Depoorter, Griet and de Does, Jesse and Jongejan, Bart and Haltrup Hansen, Dorte and Navarretta, Costanza and Calzada Pérez, María and de Macedo, Luciana D. and van Heusden, Ruben and Marx, Maarten and Çöltekin, Çağrı and Coole, Matthew and Agnoloni, Tommaso and Frontini, Francesca and Montemagni, Simonetta and Quochi, Valeria and Venturi, Giulia and Ruisi, Manuela and Marchetti, Carlo and Battistoni, Roberto and Sebők, Miklós and Ring, Orsolya and Darģis, Roberts and Utka, Andrius and Petkevičius, Mindaugas and Briedienė, Monika and Krilavičius, Tomas and Morkevičius, Vaidas and Diwersy, Sascha and Luxardo, Giancarlo and Rayson, Paul. (2021a). *Multilingual comparable corpora of parliamentary debates ParlaMint 2.1.* PID http://hdl.handle.net/11356/1432.

Erjavec, Tomaž and Ogrodniczuk, Maciej and Osenova, Petya and Ljubešić, Nikola and Simov, Kiril and Grigorova, Vladislava and Rudolf, Michał and Pančur, Andrej and Kopp, Matyáš and Barkarson, Starkaður and Steingrímsson, Steinh̑ór and van der Pol, Henk and Depoorter, Griet and de Does, Jesse and Jongejan, Bart and Haltrup Hansen, Dorte and Navarretta, Costanza and Calzada Pérez, María and de Macedo, Luciana D. and van Heusden, Ruben and Marx, Maarten and Çöltekin, Çağrı and Coole, Matthew and Agnoloni, Tommaso and Frontini, Francesca and Montemagni, Simonetta and Quochi, Valeria and Venturi, Giulia and Ruisi, Manuela and Marchetti, Carlo and Battistoni, Roberto and Sebők, Miklós and Ring, Orsolya and Darģis, Roberts and Utka, Andrius and Petkevičius, Mindaugas and Briedienė, Monika and Krilavičius, Tomas and Morkevičius, Vaidas and Bartolini, Roberto and Cimino, Andrea and Diwersy, Sascha and Luxardo, Giancarlo and Rayson, Paul. (2021b). *Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1.* PID http://hdl.handle.net/11356/1431.

## A. Appendix: Named Entity Recognition Management

As illustrated in the description of the conversion algorithm from CoNLL-U to TEI ParlaMint, between the Sentence node and the Word nodes there can be an intermediate degree of kinship if one or more entries are Named Entities. In this case the Word nodes that form the Named Entity are grouped as children under a Node `<name>` which in turn will have the Sentence Node as the parent Node. The entry.named field is in IOB format, that is the type of the Named Entities is prefixed by a prefix that assumes values, “B-“, “I-“ or “O-“ which respectively stand for Begin Intermediate and Outside. Algorithmically, the management of Named Entities is achieved by implementing a two-state automaton:

- Stato_0 :
  ```
  label{tag == B-} : Action A1 → Stato_1
  Label{ } : Action A2 → Stato_0
  ```
- Stato_1 :
  ```
  label {tag == B-} : Action A3 → Stato_1
  Label{tag == I-} : Action A4 → Stato_1
  Label{ } : Action A5 → Stato_0
  ```

The labels correspond to logical conditions and are executed in the order in which they are written, so empty labels correspond to the complementary condition of the previous. The Actions are described in pseudo programming code:

- Action A1 : {
  ```
  NodeName = generateNewNode( <name>);
  setParentOfNode(NodeName,NodeSentence);
  setAttributeNodeName(type);
  CurrentNode = NodeName;}
  ```
- Action A2 : { no action; }
- Action A3 : {
  ```
  CloseNode(NodeName)
  CurrentNode = NodeSentence
  OtherNodeName = generateNewNode( <name>);
  setParentOfNode(OtherNodeName,NodeSentence);
  setAttributeNodeName(type);
  CurrentNode = OtherNodeName; }
  ```
- Action A4 : { no action; }
- Action A5 : {
  ```
  CloseNode(NodeName);
  CurrentNode = NodeSentence; }
  ```

The part of the code that implements the automaton is within the cycle in which the linguistically annotated entries are scrolled, something like:

```
while(getline(entry)) do:
    ...
    //CurrentNode == SentenceNode
    Tag = IOBof(entry);
    execAutoma(tag);
    analize(entry);
    ...
```

Therefore, the `XML` Nodes corresponding to the entries can be children of the Node Sentence or of a Node Name set by the automaton.