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

Tommaso Agnoloni

Istituto di Informatica Giuridica e Sistemi Giudiziari (CNR-IGSG)
Roberto Bartolini, Francesca Frontini, Simonetta Montemagni, Valeria Quochi, Giulia Venturi

Istituto di Linguistica Computazionale “A. Zampolli” (CNR-ILC)
Carlo Marchetti, Manuela Ruisi

Senato della Repubblica

June, 20 2022 - ParlaCLARIN III at LREC2022 - Workshop on Creating, Enriching and Using Parliamentary Corpora
ParlaMint-IT Corpus

Corpus of Plenary speeches transcriptions of the Italian Senate (High Chamber of the Italian Parliament)

- reference subcorpus: from Mar. 2013 to Oct. 2019 (1084 sessions)
- COVID-19 subcorpus: from Nov. 2019 to Nov. 2020 (115 sessions)

Plenary verbatim reports:
- verbatim report <RESSTEN>
- Annex A (containing the texts under scrutiny)
- Annex B (containing the texts of the motions, questions and interpellations submitted in the same sitting, if any, qualified voting tables, and what are called "announcements" of documents received by the Senate.)
Data Sources

- HTML (as published on the website)
- AkomaNtoso XML (since 2018)
- dati.senato.it
  - RDF
  - CSV
Encoding into ParlaMint XML

ROOT
- from CSV obtained from dati.senato.it
- + manual compilation (governments, coalitions, fixes on groups affiliation over time, non MP speakers)

TRANSCRIPTIONS
- from HTML
- limited to <RESSTEN> (no Annexes)
- available tags: <INTERVENTO> (speech) <ORATORE> (speaker)
- incidents marked-up with heuristics applied on text in italics (<i>)

Java code

https://github.com/atomm/ResAulaSenato2ParlaMintTEI
<ListPerson>
  <head xml:lang="it">Lista dei relatori</head>
  <head xml:lang="en">List of speakers</head>
  <person xml:id="BongiornoGiulia">
    <firstName>Bongiorno</firstName>
    <lastName>Giulia</lastName>
  </person>
  <person xml:id="Femminina">
    <firstName>Femminina</firstName>
  </person>
  <person xml:id="Palermo">
    <placeName ref="https://www.geonames.org/2523918">Palermo</placeName>
  </person>
</ListPerson>
La seduta è aperta (ore 11,09). Invito il senatore Segretario provvisorio a dare lettura dei documenti.
Linguistic annotation and Named Entity Recognition

sentence splitting, tokenisation, part-of-speech tagging, lemmatisation, dependency parsing

Stanza

CONLL-U

CONLL-U

CONLL-U

alignment
(different tokenisations)

T2K
(Text-to-Knowledge)

Named Entity Recognition
### Linguistic annotation and Named Entity Recognition

| ID | Form |Lemma| UPOS| XPOS| FEATS| HEAD| DEPREL|
|----|------|-----|-----|-----|------|------|-------|
| 1  | det  | det | DET | DET | det  | 0    | DET   |
| 2  | tribunale | tribunale | NOUN | S | Gender=Masc|Number=Sing|Case=Nom|       |
| 3  | di | di | APP | E | - | 0 | APP   |
| 4  | ministri | ministero | NOUN | S | Gender=Masc|Number=Plur|Case=Nom|       |
| 5  | come | come | V | V | - | 0 | verb   |
| 6  | ricorda | ricordare | V | V | - | 0 | verb   |
| 7  | prima | prima | ADJ | A | - | 0 | adverb |
| 8  | a PUNTO | PUNTO | PUNCT | FB | - | 0 | punctuation |
| 9  | ha | avere | V | V | - | 0 | verb   |

**Stanza** (Italian-isdt-ud-2.5 model)

**CONLL-U**

**Named Entity Recognition**

 alignment (different tokenisations)

**CONLL-U**

**unified**
### Linguistic Annotation and Named Entity Recognition

| # | Il | DET | RD | Definite | Gender | Number | PronType | Art | det | start_char | end_char |
|---|----|-----|----|----------|--------|--------|---------|-----|-----|-------------|----------|
| 2 | tribunale | tribunale | NOUN | S | Masc | Sing | 12 | nsubj | start_char=1339 | end_char=1348 | B-ORG |
| 3-4 | dei | di | ADP | E | 5 | case | | | start_char=1349 | end_char=1352 | I-ORG |
| 4 | i | il | DET | RD | Definite | Masc | Plur | | 5 | det | I-ORG |
| 5 | Ministri | Ministri | PROPN | SP | 2 | nmod | | | start_char=1353 | end_char=1361 | I-ORG |
| 6 | - | - | PUNCT | FC | 8 | punct | | | start_char=1362 | end_char=1363 | - |
| 7 | come | come | SCONJ | CS | 8 | mark | | | start_char=1364 | end_char=1368 | - |
| 8 | ricordare | ricordare | VERB | V | Masc | Sing | Tense | Past | VerbForm | Part | 12 | parataxis | |
| 9 | prima | prima | ADV | B | 8 | advmod | | | start_char=1379 | end_char=1384 | - |
| 10 | - | - | PUNCT | FC | 8 | punct | | | start_char=1385 | end_char=1386 | - |
| 11 | ha | avere | AUX | VA | Mood | Ind | Number | Sing | Person | 3 | Tense | Pres | VerbForm | Fin | 12 | aux | start_char=1387 | end_char=1392 | - |
| 12 | richiedere | richiedere | VERB | V | Masc | Sing | Tense | Past | VerbForm | Part | 0 | root | |
| 13 | l' | il | DET | RD | Definite | Masc | Sing | PronType | Art | 14 | det | start_char=1488 | end_char=1492 | - |
| 14 | autorizzazione | autorizzazione | NOUN | S | Fem | Sing | 12 | | | start_char=1402 | end_char=1416 | - |
| 15 | a | a | ADP | E | 16 | mark | | | start_char=1417 | end_char=1418 | - |
| 16 | procedere | procedere | VERB | V | Inf | 14 | acl | | | start_char=1419 | end_char=1428 | - |
| 17-18 | nei | in | ADP | E | 19 | case | | | start_char=1429 | end_char=1432 | - |
| 18 | i | il | DET | RD | Definite | Masc | Plur | PronType | Art | 19 | det | - |
| 19 | confronti | confronto | NOUN | S | Masc | Plur | obl | | | start_char=1433 | end_char=1442 | - |
| 20-21 | del | - | ADP | E | 22 | case | | | start_char=1443 | end_char=1446 | - |
| 21 | il | il | DET | RD | Definite | Masc | Sing | PronType | Art | 22 | det | - |
| 22 | senatore | senatore | NOUN | S | Masc | Sing | 19 | nmod | | start_char=1447 | end_char=1455 | - |
| 23 | Matteo | matteo | PROPN | SP | 22 | nmod | | | start_char=1456 | end_char=1462 | B-PER |
| 24 | Salvini | Salvini | PROPN | SP | 23 | flat:name | | | start_char=1463 | end_char=1470 | I-PER |
Encoding in ParlaMint .ana Format

ParlaMint XML

<seg>

CONLL-U

unified

conllu2Parlamint_TEI Converter

ParlaMint .ana Root

<taxonomy>

ParlaMint .ana.XML

<seg> .ana

C++ code https://github.com/cnr-ilc/conllu2Parlamint_TEI
Linguistic annotation
<name type="ORG">
  <w xml:id="ParlaMint-IT_2019-03-19-LEG18-Sed-99.seg363.1.22" lemma="tribunale" pos="S" msd="UPosTag=NOUN|Gender=Masc|Number=Sing">tribunale</w>
  <w xml:id="ParlaMint-IT_2019-03-19-LEG18-Sed-99.seg363.1.23-24" norm="di" lemma="di" pos="E" msd="UPosTag=ADP"/>
  <w xml:id="ParlaMint-IT_2019-03-19-LEG18-Sed-99.seg363.1.24" norm="il" lemma="il" pos="RD" msd="UPosTag=DET|Definite=Def|Gender=Masc|Number=Plur|PronType=Art"/>
  <w xml:id="ParlaMint-IT_2019-03-19-LEG18-Sed-99.seg363.1.25" lemma="Ministri" pos="SP" msd="UPosTag=PROPN">Ministri</w>
</name>
Resource in use

Teaching resource

- Master students within the “Digital Humanities” degree program at the University of Pisa.
  - by manually 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.
- “Master in Gestione e Conservazione e dei Documenti Digitali" at the University of Calabria
  - querying the ParlaMint corpus via the NoSketch Engine platform thus learning basic notions of corpus annotation, usage of metadata for sub-corpus selection and querying.

Dissemination

- tutorial in Italian aimed at non-computationally savvy researchers in political and social sciences with no prior knowledge of corpus linguistics

Computational Linguistics for Political Text Analysis

- Politics in between crises. A political and textual comparative analysis of budgetary speeches and expenditure.
Conclusions and Future Plans

Corpus
- temporal extension for ParlaMint II
  - enhance reproducibility from sources
- *AkomaNtoso XML* as data source?
  - Annexed documents?
- extension to the Chamber of Deputies

Annotation
- evaluation of annotation accuracy
- test STANZA Italian NER model

Exploitation
- connect to Italian (computational) Political Sciences community