



# **ParlaCLARIN Workshop: Creating and Using Parliamentary Corpora**

**Miyazaki, 7 May 2018**

# SlovParl 2.0: The Collection of Slovene Parliamentary Debates from the Period of Secession

Andrej Pančur<sup>1</sup>, Mojca Šorn<sup>1</sup>, Tomaž Erjavec<sup>2</sup>

<sup>1</sup>Institute of Contemporary History (DARIAH-SI)

<sup>2</sup>Jožef Stefan Institute (CLARIN.SI)

Ljubljana, Slovenia

ParlaCLARIN@LREC2018

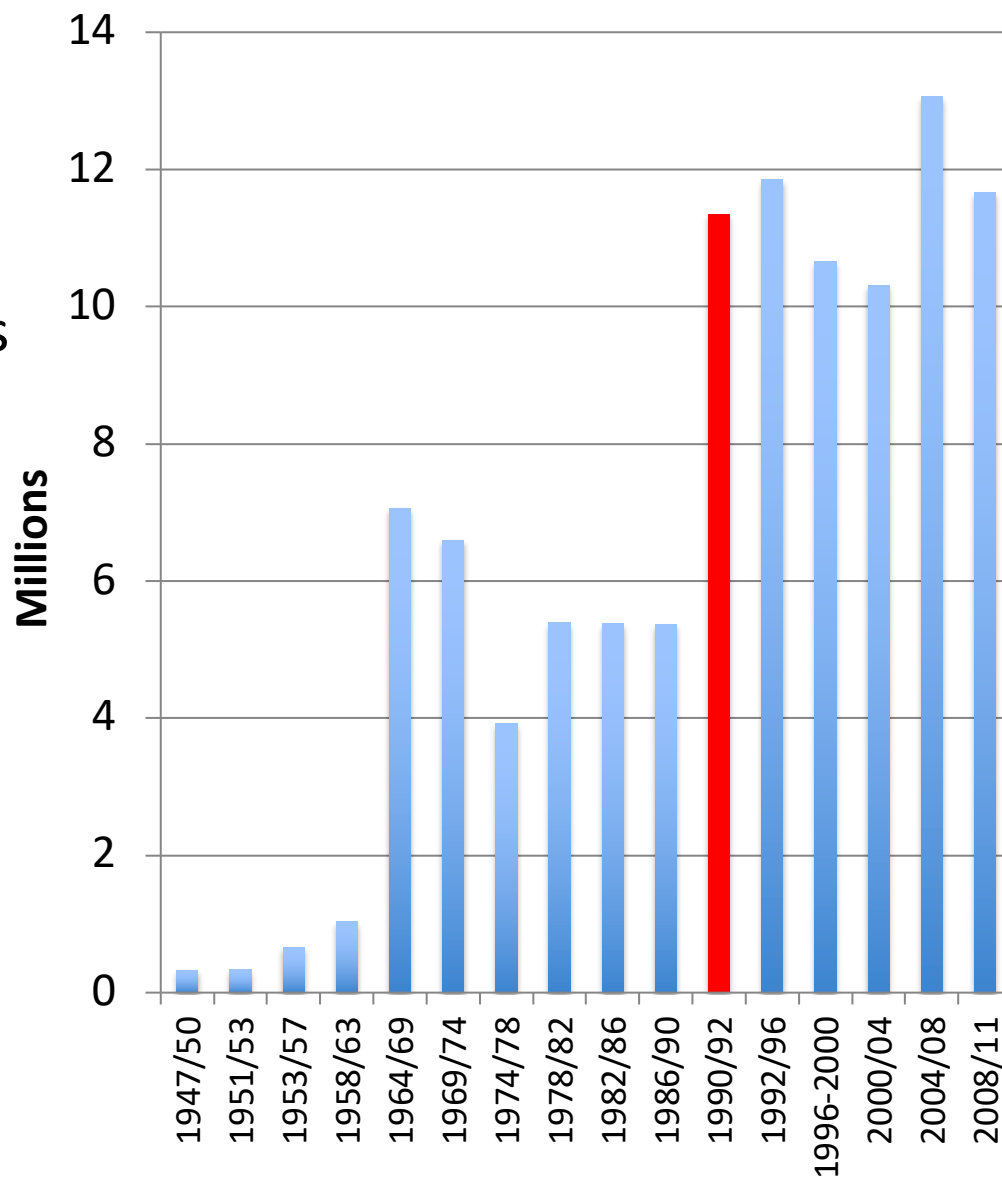
# Overview of the talk

1. Introduction
2. Building the corpus
3. Annotation: TEI drama, speech, analysis
4. Availability and maintenance

# Introduction

- Parliamentary debates as historical sources
- Large amounts of text:  
> 1945 = 170 mil. words
- Capturing as much contextual information as possible
- So far 1990 - 1992:  
before, during, and after Slovenia became an independent country

## Words per parliamentary term



# Building the Corpus

- 1990-05-05 to 1992-12-23: 3 chambers, 231 sessions, 58,813 speeches, 10.8 million words, and 121,055 non-verbal descriptions
- Basic principles:
  1. **Multidisciplinary**: not only for historians: DARIAH-SI and CLARIN.SI cooperation
  2. **All-inclusive**: in addition to parliamentary debates, other types of parliamentary papers will be included
  3. **Long-term**: not a short-term research project, but a long-term research infrastructure
  4. **Open science**: complete corpus openly available

# Transcription of source files

- Source HTML files from the Web pages of the Slovenian parliament
- Digitized analogue publications
  - some OCR errors;
  - document structure in HTML files is not clearly marked
- HTML to XML semi-automatic conversion, transcription and annotation
- Several steps, each contains:
  - XSLT for automatic annotation;
  - XPath and RE search for annotation errors;
  - additional manual annotation.

HTML pages of the Slovenian parliament,  
<https://www.dz-rs.si/wps/portal/Home/>



The screenshot displays the website of the Slovenian Parliament (Državni Zbor). The page title is "Evidenca zapisa seje" (Meeting Record). The main content area shows the details of a meeting held on May 7, 2013, at 14:30, led by Dr. France Sušnar. The agenda includes the reading of the minutes from the previous meeting, the verification of the mandate, and the reading of the minutes from the previous meeting. The page also features a calendar on the right side and a search bar at the top.


# Structure of parliamentary proceedings

- Document (1 ..  $n$ )
  - Table of contents (0 .. 1)
  - List of speakers (0 .. 1)
  - Index (0 .. 1)
  - Annex (0 ..  $n$ )
  - Meeting (1 ..  $n$ )
    - Non-verbal content (0 ..  $n$ )
    - Topic (1 ..  $n$ )
      - Non-verbal content (0 ..  $n$ )
      - Speech (1 ..  $n$ )
        - » Non-verbal content (0 ..  $n$ )
        - » Paragraph (1 ..  $n$ )
          - Non-verbal content (0 ..  $n$ )

# Annotation: Text Encoding Initiative Guidelines

## 1. TEI module for Performance text (TEI drama)

```
<teiHeader>
  <!-- Metadata -->
</teiHeader>
<text>
  <front>
    <div type="contents">
      <list><!-- Table of contents -->
        <item></item>
      </list>
    </div>
    <div>
      <castlist><!-- List of speakers -->
        <castItem><actor></actor></castItem>
      </castlist>
    </div>
  </front>
  <body>
    <div>
      <!-- Meeting -->
    </div>
  </body>
  <back>
    <div type="appendix">
      <!-- Annex -->
    </div>
    <!-- Possible Table of contents,
      List of speakers -->
  </back>
</text>
```



```
<body>
  <div><!-- Meeting -->
    <stage>Non-verbal content</stage>
    <timeline>
      <!-- Set of ordered point in time which
        are linked to the <stage type="time"> -->
      <when/><!-- from -->
      <when/><!-- to -->
    </timeline>
    <div><!-- Topic -->
      <stage>Non-verbal content</stage>
      <sp who="#reference_to_the_actor">
        <speaker>Speaker's name, affiliation</speaker>
        <stage>Non-verbal content</stage>
        <p>Speech <stage>Non-verbal content</stage>,
          speech, speech.</p>
      </sp>
```



# TEI documents included in <teiCorpus>: additional list of speakers and the index of topic

## List of speakers:

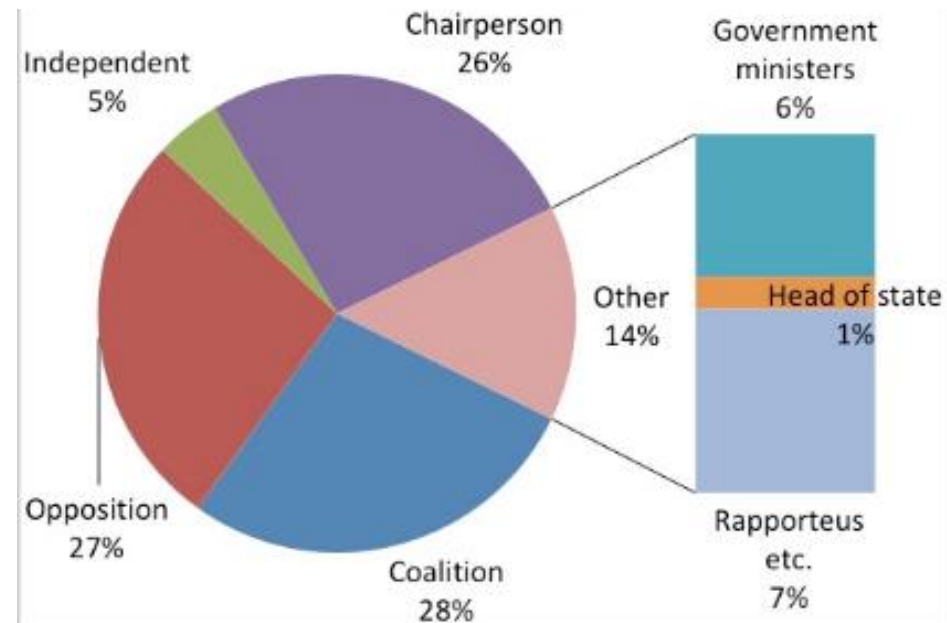
523 persons with personal data:

- personal name, gender
- date and place of birth & death
- education, profession, residence
- organization membership:
  - political party
  - parliamentary chamber
  - government membership
  - employer

## Topic index:

- 5,147 original topics classified in
  - 491 classified topic
  - 364 nested categories of <taxonomy> element

**Example:** Number of words spoken in the Assembly of the Republic of Slovenia (May 1990 – May 1992) by organization membership:



# Conversion from source TEI drama to target TEI speech (XSLT stylesheet)

## TEI drama

```
<stage type="time">(Seja je bila <time  
  to="1990-05-07T17:30:00"  
  xml:id="DZ-1990-DruzPolZb-1.stage.t.4"  
>prekinjena ob 17.30 uri</time> in se je <time  
  from="1990-05-07T21:15:00"  
  xml:id="DZ-1990-DruzPolZb-1.stage.t.5"  
>nadaljevala ob 21.15 uri</time>.)</stage>  
</div>  
<div xml:id="DZ-1990-DruzPolZb-1.div.1.1.2.2"  
  corresp="#DZ-1990-DruzPolZb-1.toc.2.2">  
<sp who="sp:BucarFrance1923"  
  corresp="#DZ-1990-DruzPolZb-1.sp.BucarFrance">  
<speaker>PRESEDUJOČI DR. FRANCE BUČAR:</speaker>  
<p>Lahko pričnemo sejo. <stage type="quorum">(68  
  prisotnih.)</stage> Zbor je sklepčen.</p>  
<p>Prosim poročevalca skupine, ki je delala pri  
<title>USKLAJEVANJU PREDLOGA POSLOVNIKA, DA  
  ZBORU POROČA O DELU</title> oziroma o uspehu  
  pri tem delu.</p>  
</sp>
```

## TEI speech

```
<note xml:id="DruzPolZb.1990-05-07.s001-01.sp-101.5"  
  type="time">Seja je bila prekinjena ob 17.30 uri  
  in se je nadaljevala ob 21.15 uri.</note>  
<anchor xml:id="DZ-1990-DruzPolZb-1.stage.t.4"/>  
</div>  
<div type="sp">  
<anchor xml:id="DZ-1990-DruzPolZb-1.stage.t.5"/>  
<note xml:id="DruzPolZb.1990-05-07.s001-01.sp-102.1"  
  type="speaker">PRESEDUJOČI DR. FRANCE BUČAR:</note>  
<u xml:id="DruzPolZb.1990-05-07.s001-01.sp-102.2"  
  who="#BucarFrance1923"  
  ana="#taxonomy.root #topic.1">Lahko pričnemo sejo.</u>  
<note xml:id="DruzPolZb.1990-05-07.s001-01.sp-102.3"  
  type="quorum">68 prisotnih.</note>  
<u xml:id="DruzPolZb.1990-05-07.s001-01.sp-102.4"  
  who="#BucarFrance1923"  
  ana="#taxonomy.root #topic.1">Zbor je sklepčen.</u>  
<u xml:id="DruzPolZb.1990-05-07.s001-01.sp-102.5"  
  who="#BucarFrance1923"  
  ana="#taxonomy.root #topic.1">Prosim poročevalca skupine,  
  ki je delala pri USKLAJEVANJU PREDLOGA POSLOVNIKA, DA  
  ZBORU POROČA O DELU oziroma o uspehu pri tem delu.</u>  
</div>
```

# Linguistic annotation of the corpus

- MSD tagged and lemmatised with ReLDI tagger  
<https://github.com/clarinsi/reldi-tagger>  
Output formatted in TEI, using the „Basic linguistic analysis“ (TEI analysis) module

```
<s>  
  <w lemma="2." ana="msd:Mdo">2.</w><c> </c>  
  <w lemma="verifikacija" ana="msd:Ncfsn">Verifikacija</w>  
  <c> </c>  
  <w lemma="mandat" ana="msd:Ncmsg">mandata</w>  
  <c> </c>  
  <w lemma="v" ana="msd:Sl">v</w><c> </c>  
  <w lemma="zbor" ana="msd:Ncmsl">zboru</w>  
  <pc ana="msd:Z">.</pc>  
</s>
```

# Availability and maintenance: open data in

- GitHub repositories:
  - DOCX to TEI drama conversion and annotation, Phase 1, [https://github.com/Slstory/Sejni\\_zapiski](https://github.com/Slstory/Sejni_zapiski)
  - HTML to TEI drama conversion and annotation, Phase 1, [https://github.com/Slstory/Seje\\_DZ](https://github.com/Slstory/Seje_DZ)
  - TEI drama additional annotation, Phase 2, <https://github.com/Slstory/SlovParl>
  - TEI speech, <https://github.com/DARIAH-SI/CLARIN.SI>
- CLARIN.SI repository  
(TEI speech + ling. annotated files)  
<http://hdl.handle.net/11356/1167>

# Linguistically annotated version of the corpus @ CLARIN.SI concordancers

## KonText

## noSketch Engine

The screenshot displays the KonText web interface. At the top, there are navigation links for 'Repository', 'About', and 'Contact', along with a 'Login' button and the 'noSketch Engine' logo. The main search area shows a query for 'neodvisnost' with 782 results. The interface is divided into several sections:

- Left sidebar:** Contains navigation options like 'Home', 'Search', 'Word list', 'Corpus info', 'My jobs', 'User guide PDF', 'Save', 'View options', 'KWIC', 'Semantic', 'Sort', 'Left', 'Right', 'Node', 'References', 'Shuffle', 'Snippet', 'Filter', 'Sub info', 'Full hit in doc', 'Frequency', 'Node tags', 'Node forms', 'Doc ID', 'Text type', 'Collocations', and 'Visualize'.
- Main content area:** Displays a list of search results. Each result includes a date (e.g., '2023rudi.1990-07-18.s004-01.1990'), a snippet of text with the search term highlighted in red, and a label 'neodvisnost'.
- Right sidebar:** Shows a 'Query' section with 'neodvisnost' and '782 hits (3 per screen)'. It also includes a 'Page 1 of 40' indicator and a 'Full hit in doc' button.

# Conclusions

- Presented SlovParl, a corpus of Slovene Parliamentary Debates from the period of secession (1990 – 1992)
- Comprehensive, clean transcriptions, rich metadata on speakers, linguistically annotated
- Freely available for download and searching via powerful concordancers
- Nice example of DARIAH-SI and CLARIN.SI cooperation
- Future work: other time periods, accessible also as „digital library“, undertake research based on this corpus