Korpusi parlamentarnih razprav

Tomaž Erjavec

Odsek za tehnologije znanja, Institut "Jožef Stefan"
Inštitut za slovenski jezik Frana Ramovša, ZRC SAZU

Predavanje na doktorskem študiju ZRC SAZU
24. 4. 2023

Pregled predavanja

Uvod

Značilnosti korpusov parlamentarnih razprav

- Široka uporaba: analiza parlamentarnih razprav je lahko zanimiva za politologe, družboslovce, zgodovinarje, za analizo diskurza, (socio)lingvistiko, študije kultur, občansko znanost
- Enostaven zajem: za večino držav so zapiski sej parlamenta dostopni na spletu
- Odprtost virov: besedila niso podvržena avtorskim pravicam ali varovanju zasebnosti
- Zato ni presenetljivo, da so v večini evropskih držav že naredili korpuse (svojih) parlamentarnih razprav.

Slovenski parlamentarni korpusi

Korpusi slovParl in siParl

- Vodilna vloga INZ:
  - PANČUR, Andrej. Označevanje zbirke zapisnikov sej slovenskega parlamenta s smernicami TEI. Zbornik konference Jezikovne tehnologije in digitalna humanistika. 2016.
- Vsi izdelani korpusi odprto objavljeni:
  - Pančur, Andrej et al. Slovenian parliamentary corpus SlovParl 1.0 (1990-1992). 2016. Repository CLARIN.SI.
  - SlovParl 2.0 (1990-1992). 2017.
  - siParl 1.0 (1990-2018). 2019.
  - siParl 2.0 (1990-2018). 2020.
  - Pančur, Andrej et al. + Katja Meden: siParl 3.0 (1990-2022). 2022. http://hdl.handle.net/11356/1748.

Značilnosti korpusov siParl

- Dostopni tako za prevzem prek repozitorija (CC BY-SA), za analizo prek konkordančnikov CLARIN.SI
- Kodirani po priporočilih TEI (XML)
- Na voljo v dveh različicah:
  strukturirano besedilo + jezikoslovno označeni
- Jezikoslovne oznake (orodje CLASSLA): leme, oblikoskladnja (MULTEXT-East, Universal Dependencies), skladnja (UD), imenske entitete
- Bogata struktura: kazalo, naslovi, opombe
- Bogati metapodatki o govorcih in političnih strankah
- Veliki: siParl 3.0 =
  32 let, 11 tisoč sej, milijon govorov, 200 milijonov besed

Transkripti

```
<note type="speaker">PREDSEDNIK IGOR ZORČIČ:</note>
<u who="#ZorčičIgor" xml:id="SDZ8-Redna-30-2022-03-23.u7"
   ana="#chair">
  <seg xml:id="SDZ8-Redna-30-2022-03-23.seg30">Kolega Koprivc,
  poglejte, bom takoj odreagiral, vam bomo dali 5 minut nazaj.</seg>
  <vocal type="interruption">
    <desc>nemir v dvorani</desc>
  </vocal>
  <seg xml:id="SDZ8-Redna-30-2022-03-23.seg31">Glejte, ni bilo
  nobenega predloga, da se to ...</seg>
  ...
</u>
```

Jezikoslovne oznake

```
<u who="#ZorčičIgor" xml:id="SDZ8-Redna-30-2022-03-23.u7"
   ana="#chair">
   <seg xml:id="SDZ8-Redna-30-2022-03-23.seg30">
      <s xml:id="SDZ8-Redna-30-2022-03-23.seg30.1">
        <w xml:id="SDZ8-Redna-30-2022-03-23.seg30.1.1"
           msd="UPosTag=NOUN|Case=Nom|Gender=Masc|Number=Sing"
           ana="mte:Ncmsn" lemma="kolega">Kolega</w>
        <name type="per">
          <w xml:id="SDZ8-Redna-30-2022-03-23.seg30.1.2"
             msd="UPosTag=PROPN|Case=Nom|Gender=Masc|Number=Sing"
             ana="mte:Npmsn" lemma="Koprivc" join="right">Koprivc</w>
        </name>
        <pc xml:id="SDZ8-Redna-30-2022-03-23.seg30.1.3"
            msd="UPosTag=PUNCT" ana="mte:Z">,</pc>
        ...
    </s>
    ...
  </seg>
  ...
</u>
```

Metapodatki: govorci

```
<person xml:id="KnežakSoniboj">
   <persName>
      <surname>Knežak</surname>
      <forename>Soniboj</forename>
   </persName>
   <sex value="M"/>
   <birth when="1962-05-13">
      <placeName ref="https://www.geonames.org
      /3188915/trbovlje.html">Trbovlje</placeName>
   </birth>
   <affiliation role="MP" ref="#DZ"
   from="2018-06-22" to="2022-05-12" ana="#DZ.8"/>
   <affiliation role="member" ref="#party.SD"
   from="2018-06-22" to="2022-05-12" ana="#DZ.8"/>
</person>
```

Metapodatki: stranke

```
<org xml:id="party.SD" role="political_party">
   <orgName full="yes" xml:lang="sl">Socialni demokrati</orgName>
   <orgName full="yes" xml:lang="en">Social Democrats</orgName>
   <orgName full="init">SD</orgName>
   <event from="2005-04-02">
      <label xml:lang="en">existence</label>
   </event>
   <idno type="URI" subtype="wikimedia" xml:lang="sl">https://
         sl.wikipedia.org/wiki/Socialni_demokrati</idno>
   <idno type="URI" subtype="wikimedia" xml:lang="en">https://
         en.wikipedia.org/wiki/Social_Democrats_(Slovenia)</idno>
</org>
```

Metapodatki: taksonomije

```
<category xml:id="parla.bi">
   <catDesc xml:lang="en">
      <term>Bicameralism</term>
   </catDesc>
   <catDesc xml:lang="sl">
      <term>Dvodomen</term>
   </catDesc>
   <category xml:id="parla.upper">
      <catDesc xml:lang="en">
         <term>Upper house</term>
      </catDesc>
      <catDesc xml:lang="sl">
         <term>Zgornji dom</term>
      </catDesc>
   </category>
   <category xml:id="parla.lower">
      <catDesc xml:lang="en">
         <term>Lower house</term>
      </catDesc>
      <catDesc xml:lang="sl">
```

Praktikum korpusnega jezikoslovja

- Fišer in Pahor de Maiti. 2021. Voices of the Parliament. *A Corpus Approach to Parliamentary Discourse Research*.
- Oziroma:
  "Prvič, sem političarka in ne politik, drugič pa . . . ", Korpusni pristop k raziskovanju parlamentarnega diskurza
- Dvojezični učbenik za korpusno analizo
- Analiza siParl 2.0 na noSketch Engine
- INZ digitalna knjižnica, učbenik napisan v TEI

Parlamentarni korpusi in CLARIN

Pobude raziskovalne infrastrukture CLARIN

- CLARIN Travelling Campus "Talk of Europe" (2014, 2015): izdelava korpusa transkriptov Evropskega parlamenta
- CLARIN-PLUS cross-disciplinary workshop "Working with parliamentary records" (2017)
- CLARIN Key Resource Families: "Parliamentary corpora" (2018–) (Darja Fišer, Jakob Lenardič)
- ParlaCLARIN LREC Workshop series on creating and using parliamentary corpora (2018, 2020, 2022) (Darja Fišer et al.)
- CLARIN ParlaFormat workshop (2019) (Tomaž Erjavec in Andrej Pančur)

Parla-CLARIN

- CLARIN Interoperability Committee 2019 organiziral "CLARIN ParlaFormat workshop", Amersfoort
- Motivacija: obstaja veliko število parlamentarnih korpusov, vendar je vsak drugače kodiran; otežena izmenjava, ponovna uporaba in primerjava
- Erjavec & Pančur predlagala uporabo TEI, več ali manj kot se je uporabljal za kodiranje siParl 2.0
- Rezultat delavnice: "Parla-CLARIN recommendations for encoding parliamentary corpora"
- Priporočila dostopna na GitHub: `https://clarin-eric.github.io/parla-clarin/`
- ERJAVEC, Tomaž, PANČUR, Andrej. The Parla-CLARIN recommendations for encording corpora of parliamentary proceedings. Journal of the Text Encoding Initiative, 2019 TEI Conference. 2021.

ParlaMint

Projekt ParlaMint I (2020–2021)

- Grajen na uspehu Parla-CLARIN
- Prvi t. i. CLARIN flagship project
- Ideja: izdelati parlamentarne korpuse evropskih držav, ki vsebujejo primerljive podatke in so enotno kodirani
- Tako kot siParl korpusi odprto dostopni na repozitoriju in konkordančnikih CLARIN.SI
- Erjavec et al., 2021, Multilingual comparable corpora of parliamentary debates ParlaMint 2.1, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1432.
- Erjavec et al., 2021, Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1431.

Objava: ParlaMint I

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darģis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer.
The ParlaMint corpora of parliamentary proceedings.
Language Resources & Evaluation. 2022.
https://doi.org/10.1007/s10579-021-09574-0

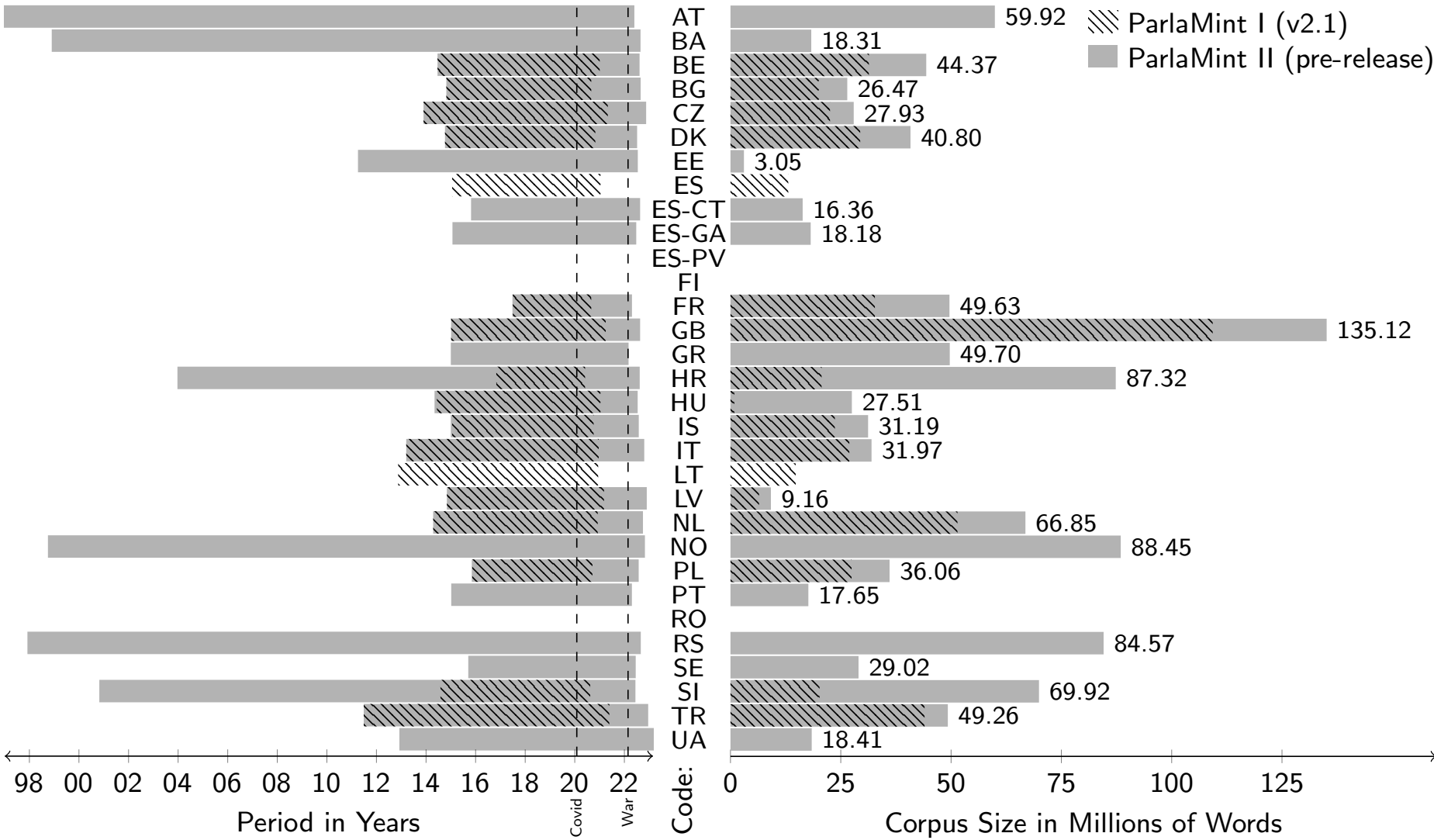
ParlaMint I

- 17 korpusov, 16 jezikov, 2015–2021:
  BE, BG, CZ, DK, ES, FR, GB, HR, HU, IS, IT, LT, LV, NL, PL, SI, TR
- Pol milijarde besed, 11 tisoč govorcev
- Struktura in oznake podobne siParl (Parla-CLARIN)
- Večinoma je vsak partner izdelal svoj korpus
- Enotna validacija in pretvorba v druge formate in objava
- Dokumentacija, sheme, programi in vzorci korpusov na GitHub
- Repozitorij: korpus v formatih Parla-CLARIN (TEI/XML), TSV, CoNLL-U
- ParlaMint korpusi uporabljeni v več raziskavah, mdr. na Helsinki Digital Humanities Hackathon 2021.

ParlaMint II (2022–2023)

- Izboljšanje kodirnih shem in navodil:
  Tomaž Erjavec (IJS), Matyáš Kopp (UFAL), Katja Meden (IJS, INZ)
- Avtomatska validacija na GitHub
  Matyáš Kopp
- 17 + 14 dodatnih korpusov:
  - SI: Andrej Pančur (INZ), Katja Meden (IJS, INZ)
  - BA, HR, RS: Nikola Ljubešić, Peter Rupnik (IJS)
  - UA: Matyáš Kopp, Anna Kryvenko (INZ)
  - Že končani: AT, BE, BG, CZ, DK, ES-CT, ES-GA, FR, GB, GR, HU, IS, IT, LV, NL, NO, PL, PT, SE, TR
  - Še čakamo: EE, ES, ES-PV, FI, LT, RO
- Korpusi pokrivajo vsaj obdobje 2015–2022-06

Časovno pokritje in velikost korpusov



Dodatki

- Novi metapodatki: ministri, politična orientacija strank
  Tomaž Erjavec, Katja Meden, Jure Skubic (INZ)
- Strojno prevajanje v angleščino
  Nikola Ljubešić, Taja Kuzman
- Govorni podatki za PL, HR, CZ
  Nikola Ljubešić et al.
- Uporaba / promocija
  Çagrı Çöltekin (Tübingen), Darja Fišer (INZ)

Git

Ključna za uspeh projekta je (bila) uporaba Git oz. GitHub:

- Javna dostopnost
- Hranjenje predhodnih različic programov, shem, dokumentacije, ...
- Možnost vzdrževanja več vej (dokumentacija, podatki, ...)
- Poročila o napakah (zahtevki)
- Kontrolirano dodajanje novih podatkov
- Izvedba validacije ob dodajanju novih podatkov
- Spletna dokumentacija

Strojno prevajanje

- Prevod vseh korpusov v angleščino nudi možnost neposredne primerjave parlamentarnih govorov
- Odprtokodni pristopi za prevajanje 32 jezikov seveda delajo tudi napake, dostikrat je problem prevajanje osebnih imen, npr. Moon (Mesec), Carrot (Trček), God (Bože)
- Nikola Ljubešić, Taja Kuzman: OpenNMT + Opus + NER `https://github.com/TajaKuzman/Parlamint-translation#sample-analysis`

Npr.:

- Na koncu pa ne velja spregledati možnosti, ki jih ponuja za reševanje tovrstnih situacij že obstoječa zakonodaja in praksa.
  Finally, it is not appropriate to overlook the possibilities already existing legislation and practice to address such situations.
- Besedo dajem predstavnikom poslanskih skupin za predstavitev stališč.
  I shall give the text to the representatives of the groups of Members for the presentation of their views.
- Evo molio bih odgovor. *Here's the answer.*

Uporaba / promocija

- CLARIN impact stories:
  - ParlaMint – A Resource for Democracy (Dario Del Fante and Virginia Zorzi)
  - Networks of Power – Gender Analysis in European Parliaments (Jure Skubic, Alexandra Bruncrona, Jan Angermeier, Bojan Evkoski, Larissa Leiminger)
- DHH 2023 (Political Polarization in the Parliament), Helsinki, 24. 5.–2. 6. 2023
- DH 2023 Pre-conference, Gradec (11. 6. 2023)

Zaključki

Zaključki

- Predstavljen zanimiv korpus pan-evropskih korpusov
- Naslednji koraki:
  1. še malo počakati na manjkajoče korpuse, končati prevajanje
  2. postaviti korpuse in prevode na konkordančnike
  3. različica 3.0: verjetno junij
  4. mogoče razširiti podaljšati korpuse v 2023 (sedaj 2022-06), dodati metapodatke, semantično označevanje, dokumentacija/sheme
  5. različica 3.1: september/oktober
- Nadaljnje delo:
  - novi parlamenti, časovna obdobja
  - poenostaviti dodajanje novih korpusov
  - nove oznake, modalnosti

Korpusi parlamentarnih razprav

Tomaž Erjavec

Odsek za tehnologije znanja, Institut "Jožef Stefan"
Inštitut za slovenski jezik Frana Ramovša, ZRC SAZU

Predavanje na doktorskem študiju ZRC SAZU
24. 4. 2023