Označevanje zbirke zapisnikov sej slovenskega parlamenta s smernicami TEI

Andrej Pančur*

* Inštitut za novejšo zgodovino
Kongresni trg 1, 1000 Ljubljana
andrej.pancur@inz.si

Povzetek

Inštitut za novejšo zgodovino sodeluje z Državnim zborom RS pri digitalizaciji sejnih zapisov. Besedila iz zbirke sejnih zapiskov zasedanj Skupščine (socialistične) republike Slovenije so v okviru raziskovalne infrastrukture Slovensko zgodovinopisje kodirane po smernicah Text Encoding Initiative (TEI). Avtor se jim v prispevku posveča s posebno pozornostjo, saj so v digitalni humanistiki *de facto* standard za kodiranje tekstovnih besedil. Podatki, ki jih je mogoče na ta način pridobiti, pa omogočajo odgovore na številna raziskovalna vprašanja.

Encoding the Slovenian Parliament Session Minutes in Line with the TEI Guidelines

The Institute of Contemporary History cooperates with the National Assembly of the Republic of Slovenia in the process of digitising its session minutes. In the context of the Research Infrastructure of Slovenian Historiography, the texts from the collection of session minutes from the (Socialist) Assembly of the Republic of Slovenia are encoded in accordance with the Text Encoding Initiative (TEI) guidelines. In his contribution, the author pays special attention to these guidelines, as they are the *de facto* standard of text encoding in digital humanities. The information that can be acquired in this manner provides answers to many research questions.

1 Uvod

V zadnjih letih se v demokratičnih državah na spletu vedno pogosteje javno objavljajo podatki (open data), ki so jih ustvarile različne državne in javne službe. Dolgo, tudi sto in večletno tradicijo javnega objavljanja podatkov o svojem poslovanju, imajo zlasti različne parlamentarne ustanove. Splošna javnost, mediji in raziskovalci so vedno kazali velik interes predvsem za zapisnike sej različnih parlamentarnih teles. To gradivo uporabljajo raziskovalci z različnih področij: zgodovinarji, sociologi, politologi, komunikologi, jezikoslovci, psihologi ipd.

Prvotno je bilo to gradivo dano na voljo javnosti v analogni obliki (večinoma kot tiskane knjige), v zadnjem času pa je vedno pogosteje objavljeno na spletu kot dokumente PDF, HTML in XHTML. Obenem vedno več organizacij poudarja prednosti odprtih formatov XML (Global Centre for ICT, 2014).

XML se je kot zelo primeren format uveljavil tudi v različnih raziskovalnih projektih, v katerih so obdelovali to gradivo. V formatu XML so tako mdr. dostopna zasedanja britanskega parlamenta (Hansard) od leta 1803,[1] nizozemskega od leta 1803 (Marx in Schuth, 2010), španskega od leta 1977 (Martin-Dancausa in Marx, 2010), češkega od leta 1993 (Jakubíček in Kovář, 2010), poljskega od leta 1993 (Ogrodniczuk, 2012) in bolgarskega (v okviru korpusa političnih govorov) od leta 2006 (Osenova in Simov, 2012).

2 Zapisniki sej zakonodajnih teles v Sloveniji

Javnosti in raziskovalcem je na voljo tudi gradivo parlamentarnih ustanov z ozemlja današnje Slovenije oziroma parlamentarnih ustanov, člani katerih so bili tudi poslanci iz Slovenije. Veliko gradiva je sicer še vedno dostopna le v analogni obliki, vedno večji del pa je tudi že digitaliziran in dan na voljo javnosti na različnih portalih:

- ALEX, Historische Rechts- und Gesetzestexte Onilne: avstrijski državni zbor (1861–1918);[2]
- Landtag Steiermark: štajerski deželni zbor (1848–1914);[3]
- Zgodovina Slovenije – SIstory:
  - kranjski deželni zbor 1861–1869;[4]
  - jugoslovanska zakonodajna telesa 1919–1939, 1942–1953;[5]
  - Ljudska skupščina Ljudske republike Slovenije (1947–1963);[6]
  - Skupščina Socialistične republike Slovenije (1963–1990);[7]
- Državni zbor Republike Slovenije od leta 1990 do danes.[8]

Razen slednjih, ki so objavljeni v formatu HTML, so vsi ostali obljavljeni kot PDF.

Na sliki 1 je glede na posamezen sklic Skupščine oziroma mandat Državnega zbora prikazano število besed govorov, ki se nahajajo v PDF publikacijah (zasedanja skupščine) na portalu SIstory (skupaj 36 milijonov besed) in kot HTML na spletni strani Državnega zbora (skupaj 69 milijonov besed). Do leta 1974 so sejni zapiski vsebovali še obsežne priloge (skupaj 10,5 milijonov besed), katere so kasneje začeli objavljati v posebni publikaciji (Poročevalec). Le-ti so po letu 2006 dostopni na spletni strani Državnega zbora. Raziskovalna infrastruktura Slovenskega zgodovinopisja, ki upravlja portal SIstory, se je v sodelovanju z Državnim zborom odločila v naslednjih

[1] Hansard archive (digitised debates from 1803), http://www.hansard-archive.parliament.uk/.

[2] http://alex.onb.ac.at/sachlichegliederung.htm.

[3] http://www.landesarchiv.steiermark.at/cms/ziel/111284715.

[4] http://sistory.si/publikacije/?menu=719.

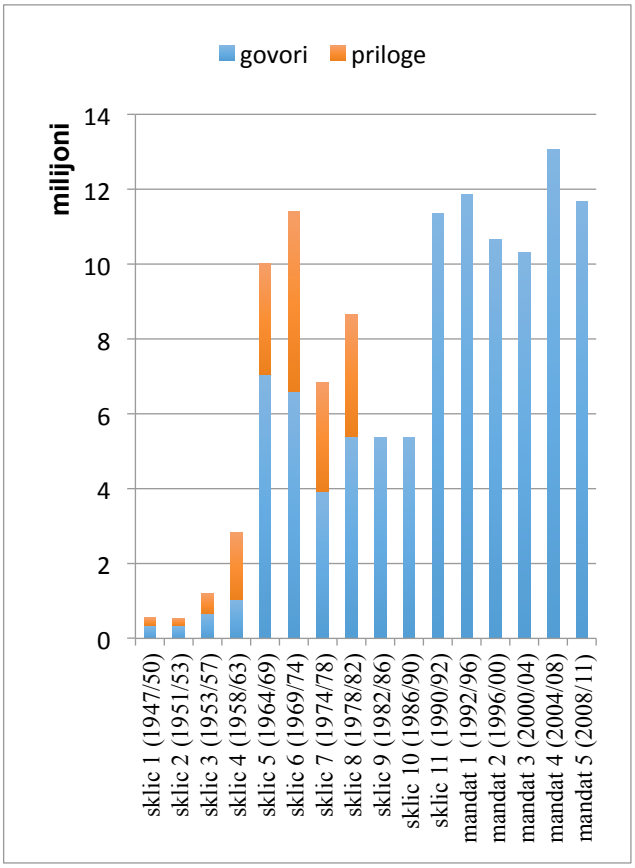
[5] http://sistory.si/publikacije/?menu=396

[6] http://sistory.si/publikacije/?menu=407.

[7] http://sistory.si/publikacije/?menu=408.

[8] Republika Slovenija, Državni zbor, https://www.dz-rs.si/wps/portal/Home/deloDZ/seje/sejeDrzavnegaZbora/PoDatumuSeje/.

treh letih digitalizirati še vse manjkajoče Poročevalce. Trenutno so digitalizirani za leta 1974/82 (skupaj 6,2 milijona besed). Če poleg zapisnikov sej, prilog in poročevalcev upoštevamo še seje različnih delovnih teles ter morebitno ostalo gradivo, pridemo do tako velikih količin besedila, ki jih noben raziskovalec ne more obdelati na klasičen način – z branjem. Zato je v teh primerih nujna strojna obdelava vsebine.



Slika 1: Število besed parlamentarnih govorov v sejnih zapiskih, objavljenih na portalu SIstory (1947–1990) in na spletni strani Državnega zbora (1990–2011).

Za nadaljnjo uporabo v raziskovalne namene se je do sedaj uporabljalo samo zapisnike sej državnega zbora, ki so v HTML formatu. Večje količine tega gradiva so uporabili v raziskovalnem projektu SloParl. Pisni del korpusa, ki je nastal v okviru tega projekta tako vsebuje 23 milijonov besed iz obdobja 1996–2005 (Žgank et al., 2006). To gradivo uspešno uporabljajo tudi različne iniciative, ki se zaradi učinkovitejšega nadzora državljanov nad delom Državnega zbora, zavzemajo za čim lažji dostop do tega gradiva.[9] Za razliko od zgoraj naštetih tujih raziskovalnih projektov pa nobeden od dosedanjih slovenskih projektov pri označevanju zapisnikov sej ni uporabil XML.

3 Označevanje in kodiranje v XML

Označevalne sheme teh korpusov se ponavadi med seboj razlikujejo. Največja korpusa (britanski in nizozemski) uporabljata lastno XML shemo, ki je prilagojena različni strukturi besedil obeh korpusov. Španski korpus je shemo prevzel po nizozemskem. Češki korpus je bil kodiran v skladu s potrebami jezikoslovnih raziskav. Delno bolgarski in predvsem poljski korpus pa sta uporabila Smernice Text Encoding Initiative (TEI) (TEI Consortium, 2015).[10]

Smernice TEI so predvsem v digitalni humanistiki *de facto* standard za kodiranje tekstovnih besedil. Zato smo te smernice uporabili tudi pri označevanju zbirke sejnih zapiskov zasedanj Skupščine (socialistične) republike Slovenije, ki jih izvajamo v okviru raziskovalne infrastrukture Slovensko zgodovinopisje na Inštitutu za novejšo zgodovino.

3.1 Vsebinska struktura zapisnikov sej

Pri pretvorbi zapisnikov sej iz formatov PDF in HTML v XML je potrebno izluščiti vsebino strukture besedila. Sejni zapisniki imajo namreč povsem enotno standardizirano strukturo besedila, kar omogoča avtomatično prepoznavanje sledeče strukture: dokument → seje → govori (govorci) → odstavki.

Na začetku vsake seje so zapisani podatki o vrsti seje, številki seje, datumu seje, predsedujočemu seje in o času začetka seje. Govori so med seboj praviloma ločeni z nekoliko večjim razmikom med zadnjim odstavkom predhodnega in prvim odstavkov naslednjega govora. Hkrati se prvi odstavek govora vedno začne z navedbo imena in priimka govornika (skupaj z morebitnimi dodatnimi informacijami o govorniku), ki je od besedila govora ločena z dvopičjem. Kljub tako jasni strukturi, pa je potrebno upoštevati še nekatere odklone. Do leta 1984 so bila npr. imena in priimki govornikov pisana z malimi, kasneje z velikimi črkami. Največ težav pa pri avtomatični pretvorbi povzroča neenoten razmik med govori. Lahko se zgodi, da razmika med govori sploh ni. Obenem se lahko zgodi, da se razmik nahaja pred začetkom novega vsebinskega sklopa. Hkrati so tudi naslovi vsebinskih sklopov lahko zapisani v veliki črkami.

Iz strukture besedila je mogoče pridobiti tudi podatke o časovnem poteku seje, številu prisotnih članov (kvorum), izidu glasovanj ter opise različnih dejanj med govori (ploskanje, vzkliki, govorjenje iz klopi ipd.). Ti podatki so kot komentarji zapisnikov magnetogramov sej ponavadi navedeni v oklepajih. Namesto oklepajev so lahko uporabljali tudi poševno črto. Hkrati je potrebno zelo paziti, da ne bomo vseh besedil v oklepajih avtomatično označili kot komentarje, saj so bili v oklepajih lahko zapisani tudi deli govorov (predvsem v okviru daljših, strokovnih poročil).

Zaradi vseh teh nedoslednosti ni nujno, da bo avtomatično označevanje strukture besedila povsem pravilno, kar je potrebno upoštevati pri naslednjih korakih označevanja besedila v XML.

[9] V okviru Kiberpipe je potekal projekt *Delajo zate!*, s katerim so želeli delo državnega zbora narediti bolj transparentno. V primerjavi s spletno stranjo Državnega zbora je na spletni strani http://www.delajozate.si/ objavljeno gradivo mogoče še dodatno filtrirati glede na poslance. Podobnim ciljem sledi tudi projekt Parlamentaria, ki je trenutno v fazi izdelave. Gl. Parlameter, https://parlameter.si/.

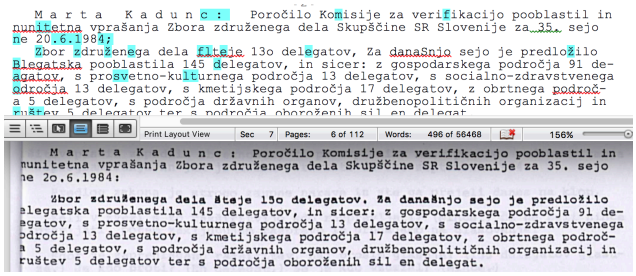
[10] Pri bolgarskem korpusu so TEI uporabili samo za označevanje dokumentov in odstavkov, drugače pa so uporabili Text Corpus Format (TCL). Pri izdelavi poljskega korpusa so govore označili v skladu z modulom za transkribiranje govorov.

3.2 Pretvorba iz PDF in HTML v osnovni TEI

Zaradi deloma zelo različne kvalitete OCR zapisa poteka pretvorba v XML na tri načine, ki zahtevajo različno stopnjo dodatnega urejanja:

1. PDF → DOCX → TEI,
2. PDF → XML (Adobe Acrobat) → TEI,
3. HTML → TEI.

Glede količine vloženega dela je najbolj zahteven prvi postopek. Ta postopek se najbolje obnese v primerih, ko z digitalizacijo ni bilo mogoče zajeti celotnega besedila. Državni zbor namreč hrani analogne zapisnike sej 1984-1990 kot tipkopis samo v enem izvodu. Pri vezavi tipkopisov je bil lahko levi rob besedila tako ozek, da je pri digitalizaciji odrezalo črko ali dve. Obenem so lahko bile črke na začetku vrstice delno rezane. Posledično so zato v OCR zapisu zelo pogoste napake. Na sliki 2 je spodaj prikazan izsek iz dokumenta PDF in zgoraj adekvatni izsek iz dokumenta DOCX, pri katerih so jasno razvidne napake, ki so nastale zaradi manjkajočega dela levega robu besedila. Po naših izkušnjah je takšne napake najlažje popravljati tako, da se PDF pretvori v DOCX z ohranjeno postavitvijo strani. Pri ročnem popravljanju napak se nato dopolnjuje manjkajoče črke na levem robu in obenem še popravlja morebitne druge napake. Za pretvorbo DOCX v TEI uporabljamo zadnjo verzijo XSLT (TEI) stilov, za katere je Tomaž Erjavec napisal dodaten profil. Pred pretvorbo je potrebno govore označiti v Wordu s slogom tei:sp.[11]



Slika 2: Rezan levi rob digitaliziranega besedila. Primerjava med PDF spodaj in DOCX zgoraj.

Veliko manj napora zahteva pretvorba zapiskov sej iz let 1947/84. Ker so bili ti zapiski tiskani v knjižnih izdajah, se je do danes ohranilo več kopij tega gradiva. Državni zbor nam je zato lahko odstopil odvečne dvojnike, ki so bili med digitalizacijo razrezani. Posledično je tudi OCR zapis teh zapisnikov skoraj brez napak. Najbolj pogoste napake kodiranja znakov se popravi po pretvorbi v XML s XSLT stilom, ki vsebuje adekvatne regularne izraze (npr. na začetku stavkov *Zeli* = *Želi*, *Ce* = *Če*; besede s številko *2* namesto z *Ž*). Vsebinsko sporne primere (npr. namesto *m* pravilno *in* ali *ni*) se popravi z iskalnikom najdi-zamenjaj, ki vsebuje regularne izraze.

Sprva se je v teh primerih za pretvorbo iz PDF v TEI kot vmesni XML zapis uporabljala pretvorba iz PDF v Adobe Acrobat XML. Vendar se je kmalu kot bolj enostaven in zanesljivejši izkazal postopek, pri katerem se z ABBYY FineReader že ob digitalizaciji poleg PDF datoteke hkrati naredil še HTML. Tako nastali HTML ima dve pomembni prednosti pred ABBYY FineReader XML:

- Ker so v ABBYY FineReader XML shranjene informacije celo o posameznih znakih, so HTML datoteke veliko manjše in preglednejše, hkrati pa še vedno dokaj verno odražajo postavitev originalnih strani.
- V primeru morebitnih kasnejših ročnih popravkov je HTML datoteka zelo pregledno in lahko berljivo referenčno besedilo.

HTML se pretvori v osnovni TEI z različnimi XSLT stili, napisanimi posebej za ta projekt. Nekoliko drugačne verzije teh stilov se uporablja tudi za pretvorbo HTML zapisnikov sej iz spletnih strani Državnega zbora. Besedila sejnih zapisnikov smo pridobili z luščenjem podatkov (Beautiful Soup)[12], v primeru širše zastavljenega projekta pa jih raziskovalci lahko pridobijo neposredno od Državnega zbora. Ker ti zapisniki sej nimajo napak v kodiranju znakov, je njihova pretvorba sorazmerno najhitrejša. To velja zlasti za zapisnike sej po letu 1996, katere se je sproti objavljajo v elektronski obliki.

3.3 Kazala vsebine in seznami govornikov

Sejni zapisniki vse do leta 1996 vsebujejo tudi kazala vsebine in sezname govornikov. Kazala vsebine kasneje zamenjajo dnevni redi, z uvedbo elektronskega glasovanja pa namesto seznama govornikov veliko večji pomen pridobijo kvorumi glasovanja.

Zlasti kazala vsebine prinašajo dodatne informacije (dnevni red in potek seje), katere iz same strukture besedila govorov niso jasno razvidne. Prehod med enim in drugim vsebinskim sklopom (točko dnevnega reda) vedno najavi predsedujoči. Če točki dnevnega reda ne sledi razprava, lahko znotraj istega govora najavi novo točko itn. Če se v vsebinsko razpravo vmešajo ostali govorniki, predsedujoči v enem od svojih naslednjih govorov zaključi razpravo in nato znotraj istega govora najavi novo vsebinsko točko. Predsedujoči lahko razpravo tudi prekine, začne znova ali zaključi. Ti vsebinski sklopi so v najboljšem primeru lahko znotraj govora predsedujočega označeni s poudarjenim naslovom (znotraj odstavka ali med dvema odstavkoma). Avtomatsko označevanje strukture teh vsebinskih sklopov je zato podvrženo pogostim napakam, zaradi česar je nujno potrebno opraviti ročno korekcijo vseh zapisnikov.

Dodatno vsebinsko vrednost prinašajo še seznami govornikov. Po naših izkušnjah je njihova glavna vrednost v bolj preglednem in pravilnejšem označevanju govornikov. Zaradi relativno pogostih napak in drugih nedoslednosti v zapisovanju imen in priimkov govornikov, je avtomatska dvojna kontrola imen in priimkov (imena v seznamih se morajo ujemati z imeni govornikov v govorih) zelo koristna, saj opozori na morebitne sporne primere. Ker smo na podoben način kot seznam govornikov označili še seznam predsedujočih posamezne seje, lahko pri analizah tako označenih govorov primerjamo (pogosto pomembne) razlike med govori predsedujočih in ostalih govornikov.

[11] DOCX to TEI to HTML Conversion, http://nl.ijs.si/tei/convert/. Kot JSI profil je vključena tudi v zadnje verzije oXygen XML urejevalnika. Pri nas uporabljamo nekoliko prilagojeno verzijo JSI profila.

[12] Beautiful Soup, https://www.crummy.com/software/BeautifulSoup/.

3.4 Text Encoding Initiative (TEI)

Pri kodiranju sejnih zapisnikov v TEI smo najprej nameravali uporabiti modul za transkribiranje govorov. Toda pravkar opisana struktura zapisnikov sej je v resnici zelo podobna elementom dramskih besedil: scena, govori in didaskalije (stage-direction) (Marx, 2009, 3). Zato smo raje uporabili TEI modul za dramska besedila.

Zapisniki sej so bili objavljani v različnih publikacijah (monografije s prilogami ali brez njih, vezani tipkopisi, spletne strani) z različnim obsegom vsebine. Razlike med temi publikacijami smo morali upoštevati tudi pri osnovni strukturi delitve besedila v dokumentu TEI. Ker smo pri tem upoštevali tudi strukturo zapisov sej starejših izvršnih in zakonodajnih teles na lokalnih (dežele, republike) in širših državnih ravneh (Jugoslavija, Habsburška monarhija), bo ta poenotena struktura primerna tudi za njih.

Vsak TEI dokument ustreza eni enoti izvornega gradiva. Zato lahko nekateri TEI dokumenti vsebujejo več sej različnih zborov, drugi pa samo eno sejo enega dne. Metapodatki v teiHeader so narejeni avtomatično pri pretvorbah z različnimi XSLT stili. Vedno vsebujejo titleStmt (naslov dokumenta title, podatke o ustvarjalcih dokumenta respStmt), publicationStmt s krajem objave in avtorsko pravico (licenca Creative Commons Priznanje avtorstva 4.0), čim bolj natančne podatke o izvirnem besedilu in njegovih avtorskih pravicah (sourceDesc) ter ob vsaki novi pretvorbi v revisionDesc še natančne podatke o opravljenih dodatnih kodiranjih.

```
<text>
   <front>
      <!-- možen titlePage, docImprint -->
      <div type="contents">
         <!-- kazalo vsebine; lahko tudi v back -->
      </div>
      <div>
         <!-- seznam govornikov; lahko tudi v back -->
      </div>
   </front>
   <body>
      <!-- govori -->
   </body>
   <back>
      <div type="appendix">
         <!-- priloge -->
      </div>
      <div>
         <!-- seznam govornikov; lahko tudi v front -->
      </div>
      <div type="contents">
         <!-- kazalo vsebine; lahko tudi v front -->
      </div>
      <!-- možen div[@type='colophon'] -->
   </back>
</text>
```

Slika 3: Osnovna delitev besedila zapisnikov sej v front, body in back.

Znotraj telesa besedila body so lahko kodirani samo govori (glej sliko 3). Ker se kazalo vsebine in seznam govornikov lahko nahajata pred ali za govori jih lahko kodiramo znotraj uvodnega razdelka front ali znotraj zaključnega razdelka back. Kazalo vsebine se mora nahajati v okviru razdelka div, ki ima atribut type z vrednostjo contents. Kazalo je kodirano kot seznam list. Sezname govornikov je potrebno zapisati v sklopu posebnega div razdelka, kateremu ni potrebno dajati atributa type, saj morajo biti seznami kodirani v okviru elementa castList (seznam nastopajočih). Vse morebitne priloge so kodirane v back v posebnem div razdelku z vrednostjo atributa type appendix.[13]

Podatki o sejah so kodirani v telesu besedila body. Ker so se sestave parlamentov tekom zgodovine lahko precej spreminjale, je te spremembe potrebno upoštevati tudi pri kodiranju v TEI. Današnji slovenski parlament (po letu 1992) se tako npr. deli na Državni zbor in na Državni svet. Skupščina (socialistične) republike Slovenije pa je med leti 1963–1992 v različnih kombinacijah sestavljalo kar 11 različnih zborov in 6 različnih (samoupravnih skupščin).

```
<body><!-- govori -->
   <div><!-- vrsta zbora oz. skupščine -->
      <div><!-- seja oz. zborovanje -->
         <docDate><date when="1990-07-02">dan zasedanja</date>,
            <date>možen drugi dan zasedanja</date></docDate>
         <timeline>
            <!-- navezava na tei:stage[@type='time'] -->
         </timeline>
         <castList>
            <!-- predsedujoči seje -->
         </castList>
         <stage type="time">
            <!-- čas začetka govorov -->
         </stage>
         <div><!-- pred dnevnim redom -->
            <div>
               <!-- vsebinski sklopi -->
            </div>
         </div>
         <div><!-- dnevni red -->
            <div>
               <!-- vsebinski sklopi oz. točke -->
            </div>
         </div>
      </div>
   </div>
</body>
```

Slika 4: Kodiranje podatkov o sejah.

Razdelek div znotraj telesa besedila body zato vsebuje informacijo o vrsti zbora oz. skupščine (glej sliko 4). Naslovi so kodirani v elementu head. Če je bila kot vir uporabljena knjiga, je teh razdelkov lahko več. Ta razdelek div nujno vsebuje nov razdelek div s podatki o seji oz. zborovanju. Z atributom n je označena številka seje. Vsaka seja vsebuje podatek o dnevu seje (kodirano v okviru datuma dokumenta docDate). Ker so seje lahko trajale več dni, pri čemer je med posameznimi dnevi zasedanja lahko minilo precej časa, je vsak datum kodiran kot date/@when. Vedno je zapisan tudi podatek o začetku seje. Ta podatek je zapisan v elementu didaskalija stage (več o tem glej spodaj). Pred začetkom govorov so vedno navedeni tudi predsedujoči (ena ali več oseb). Podatke o teh osebah se kodira v skladu s seznamom govornikov (glej sliko 5). Vsak govornik ima unikatni identifikator @xml:id, ki je avtomatično skonstruiran iz njegovega imena in priimka. Če je predsedujoči naveden tudi v seznamu govornikov, dobi atribut sameAs, ki ga navezuje na seznam predsedujočih.

Posamezne seje so nato označene v skladu s kazalom oziroma dnevnim redom (glej sliko 4). Ta se ponavadi deli na dva večja razdelka: Pred dnevnim redom in Dnevni red. Prvi div razdelek je opcijski, drugi je nujni. Vsak od teh razdelkov mora nujno vsebovati enega ali več

[13] Kodiranje različnih prilog še nismo povsem poenotili, zato v tem prispevku tega ne predstavljam. Pri tem načrtujemo, da bomo to lahko dokončno storili šele v sklopu poskusnega kodiranja Poročevalca.

razdelkov, ki zajemajo posamezne vsebinske sklope. Ti vsebinski sklopi so ponavadi usklajeni z originalnim kazalom vsebine, katere so izdelale osebe, zadolžene za izdelavo magnetogramov. Oseba, ki izvaja kodiranje v TEI, lahko vsebino govorov označi drugače iz dveh razlogov:

- če presodijo, da določena vsebina ni bila označena v originalnem kazalu;
- če presodijo, da so podpostavke v kazalu preveč nadrobne, zaradi česar označijo samo skupno postavko.

```
<front>
   <!-- ... -->
   <div>
      <!-- seznam govornikov -->
      <castList>
         <castItem>
            <actor xml:id="sp.DolinšekDrago">Dolinšek
               Drago</actor> 26, 46</castItem>
         <!-- ... -->
         <castItem>
            <actor sameAs="#sp.ZupančičJože">Jože
               Zupančič</actor> 56, 59</castItem>
         <!-- ... -->
      </castList>
   </div>
</front>
<body><!-- govori -->
   <div><!-- vrsta zbora oz. skupščine -->
      <div><!-- seja oz. zborovanje -->
         <!-- ... -->
         <castList>
            <!-- predsedujoči -->
            <castItem>
               <roleDesc>Sejo je vodil</roleDesc>
               <actor xml:id="sp.ZupančičJože">Jože Zupančič</actor>,
               <role>predsednik Zbora združenega dela</role>
            </castItem>
         </castList>
```

Slika 5: Seznam govornikov in seznam predsedujočih.

Oseba, ki izvaja kodiranje, vse razdelke div tudi naknadno označijo z globalnim atributom ana, ki se navezuje na skupno taksonomijo taxonomy v kolofonu dokumenta TEI. Razdelki, ki vsebujejo vsebinske sklope, se preko atributa corresp navezujejo na vsebinsko ustrezen item element v kazalu vsebine.

```
<div><!-- vsebinski sklop -->
   <sp who="#sp.PriimekIme">
      <speaker>Ime in priimek govornika in
         morebitne oznake govornika</speaker>
      <p>Besedilo govora, besedilo <title>naslov
         sklopa</title> besedilo govora,
         <stage>komentar zapisnikarja magnetograma
            razprave</stage> besedilo.</p>
   </sp>
   <stage type="time">
      <!-- čas prekinitve in ponovnega začetka razprave -->
   </stage>
   <sp who="#sp.PriimekIme">
      <speaker>Ime in priimek</speaker>
      <ab>Govor iz klopi, ki ga je komentator zapisal v sklopu
         govora govornika iz govorniškega odra.</ab>
   </sp>
   <stage type="time">
      <!-- čas konca razprave -->
   </stage>
</div>
```

Slika 6: Kodiranje govorov.

Znotraj posameznega vsebinskega sklopa so govori označeni v skladu s TEI modulom za dramska besedila (slika 6): Govor je označen z elementom sp, govorec z elementom speaker, govorjeno besedilo z elementi odstavek p ali anonimni blok ab. Atribut sp/@who se navezuje na omembo tega govorca v seznamu govornikov castList. Anonimni bloki se nahajajo samo znotraj tistih govorov, ki niso bili opravljeni z govorniškega odra, temveč kot medklici iz poslanskih klopi. V magnetogramih so zapisnikarji takšne govore kot medklice označili znotraj odstavkov govorov iz govorniškega odra. Kot vse druge komentarje so jih od ostalega govora zamejili z oklepaji. Zato mora oseba, ki opravlja kodiranje v TEI, takšne govore kodirati na roko. Vsi ostali komentarji znotraj odstavkov so označeni kot didaskalije stage. Znotraj odstavkov se z elementom title kodira še morebitne naslove vsebinskih blokov. To so naslovi, ki jih je ob napovedi novega vsebinskega sklopa napovedal predsedujoči.

```
<!-- čas prekinitve in ponovnega začetka razprave -->
<stage type="time">(Seja je bila <time to="1990-07-02T11:30:00"
   xml:id="stage.t.2">prekinjena ob 11.30 uri</time> in se je <time
      from="1990-07-02T19:45:00" xml:id="stage.t.3">nadaljevala ob
      19.45 uri</time>.)</stage>
```

Slika 7: Kodiranje časa začetka in konca govorov.

Komentarji zapisnikov megnetogramov, ki vsebujejo podatke o času, ko so se govori začeli oz. končali, so kodirani kot element stage, ki ima vrednost atributa type time (glej sliko 7). Seja je bila lahko večkrat prekinjena. Prekinitve so bile lahko le krajši predahi ali nekajdnevne preložitve. Znotraj elementa stage je začetek časovnega bloka kodiran s time/@from in konec s time/@to. Na unikatni identifikator elementa time se navezuje časovnica timeline/when. Slednja je narejena avtomatsko iz kodiranih podatkov v time.

Delovni proces kodiranja besedila poteka v skladu s smiselno predvidenimi koraki, ki so prilagojeni posebnostim izvornega besedila zapisnikov sej. Oseba, ki opravlja kodiranje, si za vsak novi korak najprej izbere XSLT stil, s katerim naredi avtomatsko pretvorbo. Po končani pretvorbi popravi morebitne nedoslednosti in ročno kodira nekatere dele besedila. Na ta način se krajše seje lahko kodira v pol ure, za daljše seje se običajno porabi do dve uri, za najdaljše (tudi več kot 200000 besed) pa do 4 ure.

4 Dostopnost, uporaba in načrti

Delovne verzije TEI zapisnikov sej so dostopni na GitHub.[14] Trenutno smo kodirali 53 sej zapisnikov sej do leta 1990, ki vsebujejo 4382 govorov 828 različnih govornikov, ki so skupaj izgovorili 1.150.000 besed, v prilogah pa smo kodirali še 200.000 besed. Te seje so naključno izbrani vzorci, s pomočjo katerih smo preizkušali različne načine kodiranja. Toda za uporabo različnih historičnih analiz so ti vzorci povsem neprimerni. Zato smo se odločili, da v celoti kodiramo vsaj en sklic. Izbrali smo zgodovinopisno zelo zanimiv 11. sklic ”osamosvojitvene” skupščine (1990–1992). Trenutno smo kodirali skoraj celotno besedilo (manjka Zbor občin): 41.131 govorov, 7.574.000 besed.

To besedilo nato shranimo v nov GitHub repozitorij SlovParl,[15] kjer osnovne dokumente TEI še dodatno kodiramo v skladu z nameni zgodovinopisne raziskave. Trenutno imamo v tem repozitoriju dodatno kodirano celotno besedilo Zbora združenega dela (2.739.000 besed). Največ dela smo vložili v izdelavo novih

[14] Sejni zapiski, https://github.com/SIstory/Sejni_zapiski; Seje Državnega zbora, https://github.com/SIstory/Seje_DZ.
[15] SlovParl, https://github.com/SIstory/SlovParl.

dokumentov TEI, ki omogočajo dodatno analizo TEI korpusa.

Obstoječi seznami govornikov v castList so namreč neprimerni za resno historično analizo. Ker smo njihove unikatne identifikatorje narejeni z avtomatsko pretvorbo zapisanih imen in priimkov, so iste osebe, ki imajo v drugih primerih drugače zapisano svoje ime, označene kot različne osebe. Obenem se različne osebe z istim imenom in priimkom[16] avtomatično smatra za isto osebo. Zato je na podlagi različnih zgodovinskih virov potrebno sezname poslancev, ministrov, poročevalcev in drugih govornikov ustrezno preveriti in prečistiti. To je tudi idealna priložnost, da tako narejenemu seznamu dodamo čim več javno dostopnih osebnih podatkov. Te podatke smo zapisali v ločenem dokumentu TEI speaker.xml.

Osebe smo kodirali v seznamu oseb listPerson/person. Znotraj person elementa smo označili njihova imena (persName), spol (sex), datum in kraj rojstva (birth), datum in kraj smrti (death), izobrazbo (education), poklic (occupation), bivališče (residence) in različna službovanja (affiliation) v službah, na funkcijah, v političnih strankah in drugih organizacijah, nenazadnje v parlamentu. Ker se je pripadnost osebe organizacijam čez čas spreminjala, smo veliko pozornost posvetili prav kodiranju teh sprememb. Zato so elementi affiliation preko atributov ref in ana navezani na ustrezno kodirane ustanove org v seznamu organizacij listOrg. Te organizacije (zbori, stranke, ministrstva) smo kodirali na način, ki ne omogoča le njihove analize na podlagi obstoja teh organizacij skozi čas, temveč tudi na podlagi njihovega razvoja skozi čas (preimenovanja, združevanja in razdruževanja).

Kodirani govori se preko atributa sp/@who ne navezujejo več na seznam govornikov v castList, temveč na poenoten seznam oseb listPerson. Na castList so po novem navezane preko atributa corresp. Na podlagi teh povezav lahko zastavljamo različna bolj ali manj kompleksna raziskovalna vprašanja (Pančur in Šorn, 2016). Tako npr. hitro izvemo, da sta bila v Zboru združenega dela najbolj zgovorna poslanca Jože Zupančič (735.000 besed) in Bogo Rogina (114.000), kar niti ni presenetljivo, saj je bil prvi predsednik in drugi podpredsednik tega zbora. Med navadnimi poslanci je zato rekorder Jože Arzenšek (106.000), najmanj zgovoren pa je bil Jože Košak, kateremu je uspelo izreči le 14 besed. Z le nekoliko bolj zapleteno poizvedbo lahko tudi ugotovimo, da so poslanci, ki so ob začetku mandata pripadali koaliciji DEMOS, spregovorili 22 % vseh besed, poslanci iz opozicijskih strank 23 % in neodvisni poslanci kar 55 %.

Odgovore na bolj zapletena raziskovalna vprašanja omogoča tudi na novo izdelan dokument TEI, ki v gnezdenem elementu list vsebuje tematsko kazalo točk dnevnega reda. Pri izdelavi tega kazala se je uporabilo podatke iz obstoječih vsebinskih sklopov in kazal vsebine. Postavke item tematskega kazala se pri tem navezujejo na ustrezne unikatne identifikatorje vsebinskih sklopov govorov. V enotnem kazalu smo najprej povezali vse točke dnevnega reda, ki so bile pred tem lahko razbite na različne dneve zasedanj. Potem smo med seboj povezali sorodne vsebinske sklope, npr. sprejem določenega zakona ter njegovih sprememb. Nato smo na podlagi poslovnika skupščine izdelali posebno vsebinsko shemo. Najbolj obsežen sklop *Akti in postopki* smo razvrstili v skladu s tematskim kazalom pravnega reda Republike Slovenije.[17] S pomočjo tako povezanih podatkov lahko hitro izvemo, da je bil zakon, o katerem so poslanci Zbora združenega dela najbolj razpravljali, Zakon o lastninskem preoblikovanju podjetij (103.870 besed). Med sprejemom tega zakona se je zvrstilo 520 govorov, obravnavali pa so ga v štirinajstih terminih. Kot zanimivost lahko navedem še podatek, da je v povprečju razprava nepretrgoma (do prvega odmora ali konca seje) trajala 100 minut.

Že na podlagi te stopnje kodiranja lahko torej odgovorimo na zelo različna raziskovalna vprašanja. Z dopolnjevanjem obstoječega seznama poslancev in ostalih govornikov, lahko ta vprašanja še dodatno razširimo. Z uporabo drugačnega tematskega kazala lahko raziskovalna vprašanja prilagodimo svojim potrebam. V načrtu imamo še kodiranje imenskih entitet v govorih. Poskusno smo pri tem že uporabili Stanford NER za slovenščino (Ljubešić et al., 2013). Trenutno za analizo v glavnem uporabljamo XSLT stile, ko pa bo zbirka dokumentov TEI narasla na več deset milijonov besed, se bo v glavnem uporabljalo (XQuery) aplikacije NoSQL baze dokumentov eXist, ki med drugim omogoča tudi indeksacijo (Siegel in Retter, 2015). Ko se bo naposled zbirka sejnih zapisnikov v celoti pretvorila v XML, bo primerna za uporabo povsem novih metod v historičnih raziskavah kot so različne raziskave zgodovine konceptov ali raziskave glede sprememb v odnosu do druge svetovne vojne v povojnih parlamentarnih razpravah (Piersma, 2014).

Zapisnike sej, ki smo jih kodirali v skladu s TEI modulom za dramska besedila, smo naknadno pretvorili še v dokumente TEI, kjer je besedilo označeno v skladu s TEI modulom za transkribiranje govorov. Ta pretvorba je shranjena v GitHub repozitoriju CLARIN.SI.[18] Pri tem smo za govore uporabili sledeče mapiranje:

- sp/p → div[@type='sp']/u
- sp/ab → div[@type='inter']/u
- stage → div/note.

Ker smo se pri tem odločili, da elementi u ne vsebujejo nobenih drugih elementov, temveč samo besedilo, jih lahko naknadno avtomatsko jezikovno označimo. Trenutno so partnerji iz Inštituta Jožefa Stefana že izvedli tokenizacijo, oblikosladenjsko označevanje in lematizacijo. Korpus so uvozili v spletni konkordančnik noSketchEngine (Erjavec, 2013), vsi dokumenti TEI pa so dostopni v repozitoriju CLARIN.SI (Erjavec et al., 2014).

5 Literatura

Tomaž Erjavec. 2013. Korpusi in konkordančniki na strežniku nl.ijs.si. *Slovenščina 2.0*, 1(1): 24-49. http://slovenscina2.0.trojina.si/arhiv/2013/1/Slo2.0_2013_1_03.pdf.

Tomaž Erjavec, Jan Jona Javoršek in Simon Krek. 2014. Raziskovalna infrastruktura CLARIN.SI. V: Ninth Language Technologies Conference, str. 19-24. Ljubljana: IJS. http://nl.ijs.si/isjt14/proceedings/isjt2014_03.pdf.

Global Centre for ICT (2014). Technological Options for Capturing and Reporting Parliamentary Proceedings.

[16] V 11. sklicu skupščine sta bila npr. dva Jožeta Smoleta.

[17] PIS: Pravno-informacijski sistem, http://www.pisrs.si/Pis.web/pravniRedRSDrzavniNivoKazalaTematskoKazalo.

[18] https://github.com/DARIAH-SI/CLARIN.SI.

http://www.ictparliament.org/sites/default/files/handbook-proceedings_1.pdf.

Miloš Jakubíček in Vojtěch Kovář. 2010. CzechParl: Corpus of Stenographic Protocols from Czech Parliament. V: Petr Sojka in Aleš Horák, ur., *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2010*, str. 41-46, Tribun EU. http://www.muni.cz/research/publications/914313.

N. Ljubešić, M. Stupar, T. Jurić, Ž. Agić. 2013. Combining Available Datasets for Building Named Entity Recognition Models of Croatian and Slovene. *Slovenščina 2.0*, 1(2): 35-57. http://www.dlib.si/details/URN:NBN:SI:DOC-VSWXF4CE.

Carlos Martin-Dancausa in Maarten Marx. 2010. Parliamentary documents from Spain. V: Proceedings of the International Conference on Language Resources and Evaluation, LREC, Valetta, Malta.

Maarten Marx. 2009. Advanced Information Access to Parliamentary Debates. *Journal of Digital Information*, 10(6): 1-11. https://journals.tdl.org/jodi/index.php/jodi/article/view/668.

Maarten Marx in Anne Schuth. 2010. DutchParl: The Parliamentary Documents in Dutch. V: Proceedings of the International Conference on Language Resources and Evaluation, LREC, Valetta, Malta.

Maciej Ogrodniczuk. 2012. The Polish Sejm Corpus. V: *LREC 2010, Eight International Conference on Language Resources and Evaluation*, str. 2219-2223, Istanbul, Turkey. http://www.lrec-conf.org/proceedings/lrec2012/index.html.

Petya Osenova in Kiril Simov. 2012. The Political Speech Corpus of Bulgarian. V: *LREC 2010, Eight International Conference on Language Resources and Evaluation*, str. 1744-1747, Istanbul, Turkey. http://www.lrec-conf.org/proceedings/lrec2012/index.html.

Andrej Pančur in Mojca Šorn. 2016. Digitalni pristop k parlamentarni zgodovini: uporaba gradiva Državnega zbora v digitalni humanistiki. V: *Četrt stoletja Republike Slovenije – izzivi, dileme in pričakovanja*, str. 115-126. Inštitut za novejšo zgodovino, Ljubljana.

Hinke Piersma, Ismee Tames, Lars Buitinck in Maarten Marx. 2014. War in Parliament: What a Digital Approach Can Add to the Study of Parliamentary History. *DHQ: Digital Humanities Quarterly*, 8(1). http://www.digitalhumanities.org/dhq/vol/8/1/000176/000176.html.

Erik Siegel in Adam Retter. 2015. eXist: A NoSQL Document Database and Application Platform. O'Reilly, Cambridge.

TEI Consortium. 2015. TEI P5: Guidelines for Electronic Text Encoding and Interchange. http://www.tei-c.org/Guidelines/P5.

Andrej Žgank, Tomaž Rotovnik, Matej Grašič, Marko Kos, Damjan Vlaj in Zdravko Kačič (2006). Slovenska govorna in tekstovna baza parlamentarnih razprav za avtomatsko razpoznavanje govora. V: Mednarodna multi-konferenca Informacijska družba IS, str. 115-118, Ljubljana. http://nl.ijs.si/isjt06/proc/22_Zgank_2of2.pdf.