

“PARLAMENT JE PO TEORIJI POLJE KONTROLIRANEGA KONFLIKTA”: SLOVENSKI PARLAMENTARNI KORPUS SIPARL 4.0

Katja MEDEN,^{1,2,3} Tomaž ERJAVEC,¹ Andrej PANČUR³

¹ Odsek za tehnologije znanja, Inštitut Jožef Stefan

² Mednarodna podiplomska šola Jožefa Stefana

³ Inštitut za novejšo zgodovino

Razprave nacionalnih parlamentov in drugih zakonodajnih teles, ter s tem parlamentarni korpusi, predstavljajo pomemben vir podatkov za raziskovanje političnega in predvsem parlamentarnega govora ter drugih povezanih pojavov na področju humanističnih in družboslovnih ved. Slovenski parlamentarni korpus siParl vsebuje nabor parlamentarnih razprav slovenskih parlamentarnih teles, ki pokriva obdobje med 1990 in 2022, ter s tem zajema ne le zgolj različna obdobja zgodovine Slovenije, temveč tudi prehod med različnima političnima sistemoma. Korpus ima bogato zgodovino razvoja in je tekom svojega nastajanja močno vplival na razvoj sorodnih korpusov, iniciativ in projektov. V članku predstavimo novo različico 4.0 korpusa siParl, podrobno opišemo razvoj korpusa in novosti, strukturo in označevanje podatkov v korpusu, ter predstavimo naslednje korake pri razvoju korpusa siParl.

Ključne besede: parlamentarni korpus, parlamentarne razprave, slovenski parlament, TEI XML, Parla-CLARIN

1 UVOD

Parlamentarne razprave predstavljajo pomembno povezavo med predstavniki parlamenta (poslanci) in splošno javnostjo, njihova prosta dostopnost pa igra pomembno vlogo v zagotavljanju transparentnosti parlamentarnega procesa in uresničevanju pravice do splošne obveščenosti javnosti. V tem pogledu parlamentarni korpusi predstavljajo bogat in obsežen vir za analizo političnega in parlamentarnega diskurza v okviru različnih raziskovalnih področij. Parlamentarni korpusi obsegajo tako prepise govornega jezika kot tudi bogate socio-demografske metapodatke govorcev (Fišer in Pahor de Maiti, 2020, 2021), kot je na primer njihova pripadnost različnim političnim strankam in zakonodajnim

organom. Poleg tega parlamentarne razprave v večini držav niso predmet zakonodaje o avtorskih pravicah, oziroma so iz njih izvzete, saj veljajo za javne dobrine. To je v zadnjem obdobju povzročilo hitro rast števila parlamentarnih korpusov, zlasti nacionalnih. Eden izmed njih je tudi *Slovenski parlamentarni korpus siParl*, katerega začetki segajo v leto 2016 pod takratnim imenom *Slov-Parl*. Skupaj z razvojem sorodnih iniciativ, kot je projekt ParlaMint¹ (Erjavec in sod., 2022), je siParl postavljaj pomembne mejnike pri označevanju in razvoju parlamentarnih korpusov.

V tem prispevku predstavljamo novo različico korpusa siParl (4.0), ki obsega transkripte parlamentarnih razprav in razprav delovnih teles Državnega zbora Republike Slovenije v obdobju 1990 – 2022 (1. – 8. mandatno obdobje). V primerjavi s predhodno verzijo korpusa 3.0 (Pančur in sod., 2022) so bile v novo različico dodane manjkajoče razprave delovnih teles za 8. mandatno obdobje. Poleg razširitve korpusa z dodatnimi podatki smo se v tej različici posvetili tudi posodobitvi delotoka priprave korpusa. V sklopu posodobitve smo poskušali zagotoviti čim bolj avtomatiziran proces označevanja in priprave korpusa, od začetne ekstrakcije podatkov do priprave jezikoslovno označene različice korpusa.

Prispevek je strukturiran na naslednji način: poglavje 2 se posveča pregledu zgodovine razvoja korpusa siParl in ob tem opiše in poudari pomembnost sorodnih iniciativ, na katere je korpus siParl pomembno vplival, medtem ko so te initiative prav tako pomembno prispevale k razvoju korpusa siParl. Poglavje 3 predstavi delotok razvoja korpusa, v poglavju 4 pa predstavimo njegov rezultat, novi korpus siParl 4.0. V zaključku se (poglavje 5) posvetimo naslednjim korakom pri razvoju in uporabi korpusa siParl.

2 ZGODOVINA PARLAMENTARNEGA KORPUSA SIPARL

Razvoj slovenskih parlamentarnih korpusov ima dokaj dolgo tradicijo, ki se je začela z prvimi koraki postavitve prve različice korpusa siParl. V naslednjem delu tako predstavljamo osnovne značilnosti slovenskega parlamentarnega sistema in parlamentarnih razprav, zgodovino razvoja korpusa ter predstavimo sorodne iniciative in projekte.

¹<https://www.clarin.eu/parlamint>

2.1 Slovenski parlamentarni sistem

Slovenski parlament, ki predstavlja večino parlamentarnih razprav v korpusu siParl, je sestavljen iz dveh domov, pri čemer *Državni zbor* (DZ) predstavlja najvišje predstavniško in zakonodajno telo Republike Slovenije. Poleg zakonodajne funkcije Državni zbor opravlja tudi volilno in nadzorno funkcijo, sestavlja pa ga 90 poslancev, od katerih sta dva poslanca predstavnika italijanske in madžarske narodne skupnosti.

Plenarnim zasedanjem Državnega zbora so v korpusu že skoraj od samega začetka razvoja pridružena tudi zasedanja *delovnih teles Državnega zbora*, ki so bila ustanovljena za spremljanje razmer na posameznih področjih, pripravo političnih odločitev na teh področjih in obravnavo predlogov zakonov ali zakonodajnih aktov. Kadar Državni zbor nima plenarnega zasedanja, poslanci delujejo v teh delovnih telesih (Državni zbor Republike Slovenije, 2020), ki vključujejo odbore in komisije za posamezna področja (npr. odbor za zunanje zadeve ali komisija za narodni skupnosti) ter preiskovalne komisije.

Nazadnje podatke za obdobje modernega slovenskega parlamentarizma dopolnjujejo tudi zasedanja Klegija predsednika Državnega zbora, ki je posvetovalno telo predsednika DZ in ki ima v določenih primerih tudi pristojnost odločanja. Sestavljajo ga predsednik in podpredsedniki Državnega zbora, predsedniki poslanskih skupin in poslanci narodnostnih skupin (tj. poslanci italijanske in madžarske narodnostne skupine).

V korpus so že od samega začetka poleg zasedanj dobro poznanega modernega parlamenta vključene tudi seje 11. mandatnega obdobja Skupščine Republike Slovenije. Seje skupščine predstavljajo zelo specifičen in pomemben zgodovinski kontekst za razvoj sodobnega parlamentarnega sistema v Sloveniji, ki je bila med letoma 1945 in 1991 del zvezno urejene Jugoslavije in v kateri je bil parlament odraz socialističnega sistema. Prve večstrankarske volitve so potekale aprila 1990, leta 1992 pa so poslanci skupščine sprejeli novo ustavo, ki je formalno končala obdobje Socialistične skupščine Slovenije in uvedla nov klasični parlament. Seje parlamenta tako ponujajo edinstven vpogled v obdobje pred, med in po osamosvojitvi Slovenije (Pančur in sod., 2018).

2.2 Od SlovParl do siParl

Razvoj korpusa siParl se je začel leta 2016, ko se je v okviru Raziskovalne infrastrukture Slovenskega zgodovinopisja (RI INZ) na Inštitutu za novejšo zgodovino pričel projekt digitalizacije obstoječih parlamentarnih dokumentov, ki bi zgodovinarjem omogočil boljši dostop do (in posledično omogočil tudi analizo) čedalje večjega števila digitalnih zgodovinskih virov v Sloveniji (Pančur in Šorn, 2016).

Celotni proces in posledični razvoj prvih različic korpusa siParl je že od samih začetkov temeljil na upoštevanju naslednjih načel (Pančur in Erjavec, 2020):

1. *Multidisciplinarnost*: Korpus mora biti uporaben ne le za zgodovinarje, temveč tudi za druge sorodne discipline.
2. *Inkluzivnost* dokumentov: Poleg parlamentarnih razprav je v bodoče potrebno vključevati tudi druge vrste parlamentarnih dokumentov.
3. *Dolgoročnost*: Načelo dolgoročnosti se v tem primeru nanaša predvsem na zagotavljanje stabilnega in dolgoročnega financiranja takšnih projektov (saj takšnih načrtov ni mogoče uresničevati v obdobju kratkoročnih raziskovalnih projektov), predvsem v okviru raziskovalnih infrastruktur.
4. *Upoštevanje načel odprte znanosti*: Vsa predhodna načela se lahko optimalno uresničuje le v skladu z načeli načeli odprte znanosti (npr. odprtost in zagotavljanje prostega dostopa podatkov, upoštevanje FAIR načel itd.)

V skladu z zgoraj naštetimi načeli korpus siParl že od svojih začetkov nastaja v tesnem sodelovanju med slovenskima raziskovalnima infrastrukturama DARIAH-SI in CLARIN.SI ter njunima evropskima infrastrukturama, DARIAH-EU in CLARIN.EU.

Prvi rezultat omenjenega procesa je bil korpus *SlovParl 1.0* (Pančur in sod., 2016), eden izmed prvih parlamentarnih korpusov slovansko govorečih držav. Korpus je vseboval 2,7 milijonov besed parlamentarnih razprav v Zboru združenega dela Skupščine Republike Slovenije (1990 – 1992). To obdobje je v slovenski zgodovini še posebej zanimivo, saj zajema obdobje pred, med in po

osamosvojitvi Slovenije leta 1991 ter s tem povezan prehod iz socializma v demokracijo, ki se je odražal tudi v strukturi in delovanju parlamenta. SlovParl 1.0 je bil kodiran v skladu s smernicami za označevanje besedil Text Encoding Initiative Guidelines (TEI Consortium, 2020), prvotno z uporabo modula TEI za izvedena dela² (TEI Drama), ki je bil nato preoblikovan v uporabo elementov modula TEI za transkribirani govor³ (TEI Speech) (Pančur, 2016), da bi bolje omogočil kasnejše jezikoslovno označevanje.

SlovParl 1.0 je tako postavil temelje za razvoj slovenskih parlamentarnih korpusov, ki so mu kmalu sledile druge različice:

- *SlovParl 2.0* (Pančur in sod., 2017), 1990 - 1992, dodani so bili prepisi zapisnikov Družbeno političnega zbora in Zbora Občin Skupščine Republike Slovenije (Pančur in sod., 2018).
- *siParl 1.0* (Pančur in sod., 2019), 1990 - 2018, dodani so bili prepisi Državnega zbora Republike Slovenije, delovnih teles in Kolegija predsednika Državnega zbora RS za obdobje od 1. do 7. mandata.
- *siParl 2.0* (Pančur in sod., 2020), 1990 - 2018, korpus vsebuje identične podatke kot *siParl 1.0*, vendar so pri tem bile odpravljene marsikateri nepravilnosti v označevanju in (ročno preverjenih) metapodatkih, podatki pa so ponovno označeni s takrat najsodobnejšimi orodji za jezikoslovno označevanje (Pančur in Erjavec, 2020).
- *siParl 3.0* (Pančur in sod., 2022; Meden in sod., 2024), 1990 - 2022, dodani so bili prepisi plenarnih sej Državnega zbora za 8. mandat.

Do korpusa *siParl 2.0* je razvoj potekal precej linearno, kjer je vsaka različica dopolnila korpus z novimi podatki in odpravljala morebitne napake. Sočasno z razvojem *siParl 2.0* pa se je v tem času vzpostavil mednarodni projekt ParlaMint (bolj detaljno predstavljen v poglavju 2.3), kjer se je priprava korpusa razdelila na dva sorodna korpusa, *siParl* in *ParlaMint-SI*.⁴ Korpusa sta si med seboj sicer precej podobna, vendar se med seboj razlikujeta predvsem v načinu označevanja parlamentarnih podatkov. Hkrati podatki, pripravljeni za razvoj korpusa *siParl*, služijo neposredno tudi kot osnova za pripravo korpusa *ParlaMint-SI*.

²Performance Texts TEI module

³Transcriptions of Speech TEI module

⁴V času pisanja prispevka je trenutna najnovejša različica ParlaMint-SI 4.1, del nabora primerljivih korpusov ParlaMint 4.1.

Korpus	Leto	Vsebina	Kodiranje
SlovParl 1.0	2016	Zbor združenega dela (1990 – 1992)	TEI Drama, TEI Speech
SlovParl 2.0	2017	Skupščina Republike Slovenije (1990 – 1992)	TEI Drama, TEI Speech
SiParl 1.0	2019	Skupščina RS (1990 – 1992) Državni zbor RS (1990–2018)	TEI Drama, TEI Speech
SiParl 2.0	2020	Skupščina RS (1990 – 1992) Državni zbor RS (1992 – 2018) Delovna telesa DZ (1996 – 2018) Kolegij predsednika DZ (1996–2018)	Parla-CLARIN
ParlaMint-SI 1.0	2020	Državni zbor RS (2015–2020)	Parla-CLARIN, ParlaMint
ParlaMint-SI 2.0/2.1	2021	Državni zbor RS (2015 – 2020)	Parla-CLARIN, ParlaMint
SiParl 3.0	2022	Skupščina RS (1990 – 1992) Državni zbor RS (1992 – 2018) Delovna telesa DZ (1996 – 2022) Kolegij predsednika DZ (1996 – 2018)	Parla-CLARIN
ParlaMint-SI 3.0	2023	Državni zbor RS (2000–2022)	ParlaMint

Tabela 1: Zgodovina razvoja korpusa siParl od svojih začetkov leta 2016 (SlovParl 1.0) do predzadnje različice korpusa (siParl 3.0) ter sočasen razvoj sorodnega korpusa ParlaMint-SI (besedilo, označeno z *ležečo pisavo*).

Celotna zgodovina razvoja korpusa siParl (in posredno korpusa ParlaMint-SI) do predzadnje različice korpusa (siParl 3.0, ki predstavlja izhodiščno točko nadaljnega razvoja nove različice), je predstavljena v Tabeli 1.

Nova, trenutna različica korpusa siParl (4.0) je omenjeni korpus razširila še z manjkajočimi prepisi sej delovnih teles in Kolegija predsednika Državnega zbora za 8. mandat, ki ni bil vključen v različico 3.0. Poleg tega v na tej točki zelo dobro uveljavljen delovni postopek uvaja več posodobitev, povezanih z avtomatizacijo in standardizacijo določenih delov postopka. Nova različica in posodobljen delotok sta podrobneje predstavljena v poglavju 3 in 4.

2.3 siParl4.0 in povezane iniciative

Sočasno z razvojem korpusa siParl je vzporedno potekal tudi razvoj različnih sorodnih iniciativ in priporočil za označevanje parlamentarnih podatkov. V tem pogledu je potrebno omeniti predvsem priporočila za označevanje parlamentarnih razprav Parla-CLARIN, prilagoditev smernic *Text Encoding Initiative* (TEI) (TEI Consortium, 2020), ki omogočajo označevanje širokega nabora korpusov parlamentarnih razprav in drugih sorodnih dokumentov (Erjavec in Pančur, 2022). Korpus siParl 2.0 je tako postal prvi korpus, ki je bil kodiran v skladu s priporočili Parla-CLARIN.

Poleg priporočil Parla-CLARIN, ki so se kmalu uveljavile kot referenčna shema za označevanje parlamentarnih podatkov, je v letu 2020 temu sledil začetek projekta ParlaMint I. Primarni cilj projekta ParlaMint je bil zagotoviti ustrezne mehanizme za pretvorbo posameznih nacionalnih parlamentarnih korpusov v nabor medsebojno primerljivih in poenotenih parlamentarnih korpusov. Ob zaključku prve faze projekta (ParlaMint I, 2019 - 2021) je nabor obsegal 17 nacionalnih parlamentarnih korpusov. Kot osnova za enotno kodiranje parlamentarnih razprav pa so bila uporabljena prej omenjena priporočila Parla-CLARIN, ki jih je bilo potrebno prilagoditi (in predvsem zaostriti) za zagotavljanje uniformiranosti kodiranja podatkov. Projekt se je nadaljeval v letih od 2022 do 2023, kjer se je končni nabor korpusov razširil na 29 evropskih korpusov, tako nacionalnih parlamentov, kot tudi parlamentov avtonomnih pokrajin. V tej fazi so bili korpusi tudi dodatno obogateni z razširjenim naborom ne le podatkov, temveč tudi metapodatkov (Ogrodniczuk in sod., 2022). Kot omenjeno v poglavju 2.2, je v sklopu projekta ParlaMint nastajal tudi korpus ParlaMint-SI, kjer korpus siParl predstavlja osnovo pri razvoju vsake različice ParlaMint-SI.

Poleg ParlaMint-SI pa je v sklopu pregleda sorodnih iniciativ in korpusov potrebno omeniti tudi druge nedavno objavljene korpuse, ki pokrivajo starejša obdobja slovenske parlamentarne zgodovine. *Parlamentarni korpus prve Jugoslavije Yu1Parl 1.0* (Kavčič in sod., 2023b), ki pokriva obdobje Kraljevine SHS v letih 1919 – 1939 vsebuje večjezične prepise Narodnega predstavništva Kraljevine Jugoslavije. Dokumenti so zapisani v srbohrvaškem in slovenskem jeziku (odvisno od govorca), kjer je bil srbohrvaški jezik zapisan v cirilici. Drugi zgodovinski korpus je *Korpus kranjske deželne skupščine Kranjska 1.0* (Kavčič in sod., 2023a), ki vsebuje prepise z obdobja 1861 – 1913. Dokumenti so prav

tako dvojezični, v slovenščini in nemščini, pri čemer je bila nemška pisava najprej zapisana v gotici in šele pozneje v latinici. Omenimo lahko tudi (s korpusom siParl precej povezan) *Korpus programov političnih strank za slovenske parlamentarne volitve leta 2022 Programi2022* (Polanič in Dobranič, 2022), ki vsebuje 19 programov slovenskih političnih strank, in s katerim je bil narejen prvi korak k izpolnjevanju v poglavju 2.2 omenjenega načela inkluzivnosti oziroma dodajanju različnih vrst parlamentarnih dokumentov.

3 RAZVOJ KORPUSA SIPARL4.0

siParl 4.0 pokriva parlamentarne razprave Državnega zbora in njegovih delovnih teles za obdobje 1990 – 2022, pri katerem so bile v primerjavi z verzijo 3.0 dodane razprave delovnih teles DZ za 8. mandatno obdobje (2018 – 2022) ter posodobljen delotok priprave korpusa. Celoten postopek gradnje korpusa je v sestavljen iz naslednjih faz:

1. Ekstrakcija podatkov posameznih sej parlamentarnih razprav
2. Čiščenje in priprava podatkov
3. Označevanje podatkov v skladu s priporočili Parla-CLARIN:
 - (a) Dvostopenjsko označevanje podatkov (po modulih TEI Drama in TEI Speech)
 - (b) Razločevanje govorcev
 - (c) Popis in dodajanje metapodatkov
 - (d) Priprava kolofonov korpusa (ang. corpus header) in posameznih mandatov (ang. term headers) za dokumentiranje korpusa
4. Jezikoslovno označevanje korpusa za pripravo jezikoslovno označene različice korpusa
5. Generiranje izpeljanih formatov korpusa in dodajanje korpusa v konkordančnike.
6. Objava korpusa v repozitoriju CLARIN.SI.

Čeprav je bil potek dela pri sestavljanju korpusa siParl dobro uveljavljen že pri gradnji starejših različic korpusov siParl (in SlovParl), se je skozi leta pokazala

potreba po bolj avtomatiziranem pristopu obdelave korpusa in s tem minimiziranju ročnih intervencij. Te so zaradi narave podatkov do določene mere neizogibne, predvsem pri popisu in dodajanju metapodatkov. Posodobljena verzija delotoka sicer ohranja zgoraj orisano strukturo izdelave korpusa, vendar pri posameznih fazah doda določene novosti, ki poskušajo pripomoči k prej omenjenim glavnim ciljem te posodobitve (tj. čim večja avtomatizacija in standardiziranost postopka, seveda ob ustreznem ročnem pregledu pravilnosti podatkov za čim višjo kakovost metapodatkov).

V tem razdelku se tako posvečamo posameznim fazam gradnje, izpostavimo posodobitve ter predstavimo primere in rezultate vsake posamezne faze.

3.1 Ekstrakcija in čiščenje podatkov

Začetno točko gradnje korpusa predstavlja postopek ekstrakcije podatkov. V primeru parlamentarnega korpusa so to posamezne seje parlamenta, v primeru korpusa siParl, so to seje Državnega zbora, ki so za obdobje 1992 – 2022⁵ na voljo na spletni strani Državnega zbora⁶ v formatu HTML. Ena izmed večjih novosti delotoka je avtomatizacija faze ekstrakcije podatkov, ki je bil do sedaj zgolj pol-avtomatiziran (določen del ekstrakcije je bil opravljen ročno, predvsem iz vidika kompleksne strukture spletne strani). Ravno zaradi nenavadne strukture spletne strani smo za ekstrakcijo sej razvili skripte za ekstrakcijo podatkov na podlagi Python knjižnice *Selenium*. Rezultati ekstrakcije so tako dokumenti v formatu HTML, ter ločena datoteka, ki vsebuje vse spremljajoče metapodatke posameznih dokumentov (npr. naslov seje, ime delovnega telesa, datum, tip seje). Ti dokumenti vsebujejo marsikatero nezaželeno (HTML) oznake ali celo strukturne napake, ki onemogočajo nadaljnje procesiranje podatkov.⁷ Pri čiščenju podatkov se tako v tem delu osredotočamo na odpravo strukturnih napak, pretvorbo HTML dokumentov v XML format ter druge problematične vidike originalnih podatkov, da s tem podatke pripravimo na nadaljnjo obdelavo.

⁵Prepisi so na spletni strani Državnega zbora na voljo le za obdobje 1992 - 2022, medtem ko so seje Skupščine Republike Slovenije (1990 – 1992) rezultat digitalizacijskega projekta z leta 2016.

⁶<https://www.dz-rs.si/wps/portal/Home>

⁷Večina nadaljnje obdelave temelji na skriptah XSLT, ki zahtevajo pravilno strukturirane XML dokumente (ang. valid XML structure).

3.2 TEI kodiranje podatkov – TEI Drama in TEI Speech

Naslednja faza delotoka je namenjena označevanju parlamentarnih podatkov v skladu s priporočili za označevanje parlamentarnih podatkov Parla-CLARIN (Erjavec in Pančur, 2019), ki temelji na *de-facto* standardu za označevanje strukture besedil, Text Encoding Initiative (TEI Consortium, 2020). Ta med drugim ponuja veliko število modulov, ki vsebujejo nabor označevalcev za različne namene. V primeru označevanja parlamentarnih razprav se v večini primerov uporabljata dva specifična modula: modul za označevanje izvedbenih del (*Performance Texts*, pogosteje poznan kot TEI Drama), ter modul za označevanje prepisov govora (*Transcriptions of Speech*, oziroma TEI Speech). Priporočila Parla-CLARIN tako vsebujejo elemente modula za označevanje prepisov, z dodanimi nabori elementov iz drugih modulov, kot je na primer modul za označevanje imen, datumov, ljudi in krajev (Names, Dates, People, and Places).

V TEI skupnosti že vrsto let poteka obsežna razprava o uporabi modulov TEI Drama in TEI Speech za kodiranje parlamentarnih razprav, predvsem v smeri tega, kateri izmed modulov je bolj primeren kot osnova kodiranja parlamentarnih razprav. Modul TEI Drama predstavlja parlamentarne razprave kot gledališki performans. Parlament je v tem pogledu predstava, kjer vsi poslanci in drugi akterji v parlamentu igrajo določeno vlogo oziroma predstavljajo igralce, njihovi govori pa dramsko besedilo. Ta logika je v večini primerov vezana na poslanske govore, ki so lahko (in v večini primerov tudi so) predhodno pripravljene, vendar pa se jih poslanci ne nujno tudi strogo držijo in se v svojih govorih pogosto odmaknejo od njih. Modul TEI Drama tako vključuje elemente za kodiranje govorcev kot seznam (<castList>), ime govorca (<speaker>) in "odrske napotke" (<stage>) (Pančur in sod., 2018). V nasprotju z dramskim modulom se modul TEI Speech osredotoča na poslanski govor, kjer se kot glavni argument uporablja dejstvo, da parlamentarne razprave (predvsem poslanski govori) v parlamentu predstavljajo govorjeni jezik (ang. spoken language). Čeprav se v korpusu nahajajo prepisi parlamentarnih razprav, te vsebujejo značilnosti govorjenega jezika (npr. premori med govorom, prekinitve govora, itd) (Truan in Romary, 2022).

Kljub temu, da za oba vidika obstajajo argumenti, je v večini primerov označevanja parlamentarnih korpusov na koncu prevladal modul TEI Speech, predvsem zaradi lažjega končnega jezikoslovnega označevanja parlamentarnih razprav.

V primeru korpusa siParl smo tako, kot že omenjeno v poglavju 2 v svojih začetkih (še pod imenom SlovParl 1.0 in 2.0) podatke označevali tako v obliki TEI Drama kot v TEI Speech. S prehodom v različico 2.0 pa je siParl postal prvi korpus, ki ima podatke označene v skladu s priporočilom Parla-CLARIN. Parla-CLARIN je postal zelo uveljavljen način kodiranja (slovenskih in tujih) parlamentarnih razprav, vendar pa v primeru korpusa siParl zaradi obsežne količine podatkov označevanje poteka v dveh stopnjah. Prva stopnja obsega označevanje podatkov v skladu z modulom TEI Drama, medtem ko v drugi stopnji podatke pretvorimo iz TEI Drama v format Parla-CLARIN.

Na to odločitev je vplivala tudi zgoraj omenjena razprava o naravi in ustreznem kodiranju parlamentarnih podatkov (TEI Drama proti TEI Speech). Ker je bil SlovParl primarno namenjen zgodovinarjem in humanistiki, je v tem oziru prevladal pogled na parlament kot performans in s tem povezano označevanje podatkov z modulom TEI Drama (Pančur in Šorn, 2016). Temu je kasneje sledila pretvorba v TEI Speech, ko se je korpus začelo tudi jezikoslovno označevati, da bi se s tem razširilo uporabnost korpusa tudi za druge vede (predvsem jezikoslovje). Pretvorba podatkov v oba formata je tako omogočilo ohranjanje obeh pogledov na parlamentarne podatke.⁸

Ker poleg ohranjanja narave podatkov takšen način kodiranja omogoča tudi postopno kodiranje relativno kompleksnih podatkov, smo ta način dvostopenjskega označevanja ohranili tudi v trenutni različici korpusa. V nadaljevanju bolj podrobno predstavimo potek kodiranja podatkov v obeh stopnjah.

3.2.1 KODIRANJE PODATKOV V SKLADU S TEI DRAMA

Prva stopnja kodiranja podatkov vključuje transformacijo podatkov (s pomočjo skript XSLT) iz osnovnega formata podatkov v kodiranje v skladu z modulom TEI Drama. V tej stopnji tako zajamemo in kategoriziramo (primerno označimo) vse vzorce, ki se pojavljajo v podatkih:

- imena in priimki govorcev (npr. DR. MIRO CERAR:)
- točke dnevnega reda (npr., 1. TOČKA DNEVNEGA REDA)
- vrzeli v prepisih, ki so posledica nerazumljivega dela govora ali iz drugih vzrokov manjkajočega dela besedila v prepisu (označeni kot "...")

⁸Pretvorbe podatkov so na voljo v [GitHub repozitoriju](#) korpusa.

- posamezni govori in segmenti transkripta
- prekinitve in pripombe (označene kot besedilo v oklepaju /.../ ali (...)).

Potrebno je poudariti, da so to zgolj najbolj pogosti primeri vzorcev, ki se pojavljajo v prepisih in da se zaradi pogostih napak in napačnega označevanja v originalnih prepisih lahko pojavljajo v najrazličnejših oblikah (ki jih najpogosteje zajamemo z različnimi oblikami regularnih izrazov). To pomeni, da ostajajo podatki tudi po preliminarnem čiščenju in strukturiranju še vedno relativno neurejeni.

V posodobljenem delotoku v to fazo obdelave podatkov vključimo tudi proces razločevanja govorcev, ki je v preteklosti temeljil predvsem na ročnem pregledu in urejanju podatkov. Ta del procesa smo zato do določene mere uspeli avtomatizirati s pomočjo knjižnic v programskem jeziku Python. Proces razločevanja govorcev tako sedaj temelji na računanju Levenštajnovе razdalje (ang. *Levenshtein distance*):

Levenštajnova razdalja (ang. *Levenshtein distance*) med dvema nizoma je število modifikacij, potrebnih za pretvorbo enega niza (s1) v drugi niz (s2). Pokriva enoznakovne spremembe, kot so brisanje, vstavljanje in zamenjavo.

Na primeru parlamentarnih razprav smo tako na podlagi Levenštajnovе razdalje ter frekvence, s katero se posamezna oblika pojavlja v besedilih, lahko oblikovali seznam vseh oblik imen in priimkov govorcev in najverjetnejšo pravilno obliko govorca (npr. za razločevanje oblik imen "*Alenka Bratušek*", ki se v besedilih pojavlja 1000-krat, "*Alena Bratušek*" ki se v tej obliki zapisa pojavi trikrat, in "*Alenkra Bratušek*", ki se pojavi zgolj enkrat, je kot najverjetnejša oblika pravičnega zapisa imena ponujena "*Alenka Bratušek*"). S tem uspešno minimiziramo čas, potreben za ročni pregled podatkov⁹. S tem oblikujemo osnovni seznam govorcev (<listPerson>), ter v razpravah posamezne govore opremimo z ustreznim identifikatorjem govorca (npr. <sp who="#ŽižkaFelice">).

Končni produkt prve transformacije podatkov obsega dokumente, ki vsebujejo vse (v skladu z TEI Drama) označene vzorce, ki se pojavljajo v parlamentarnih razpravah, ter ročno pregledan seznam govorcev in njihove identifikatorje. Ta

⁹Ročni pregled je seveda še vedno nujen, predvsem v primeru manj pogostih oblik imena, govorcev z enakim imenom in priimkom itd.

```
<head>Vsebina zapisa seje</head>
<stage type="title">REPUBLIKA SLOVENIJA,
DRŽAVNI ZBOR, KOLEGIJ PREDSEDNIKA
DRŽAVNEGA ZBORA</stage>
<stage type="session">1. seja</stage>
<stage type="date">(29. junij 2018)</stage>
<stage type="chairman">Sejo je vodil mag. Matej Tonin,
predsednik Državnega zbora.</stage>
<stage type="time">Seja se je pričela ob 10.01.</stage>
<sp who="ToninMatej">
  <speaker>PRESEDNIK MAG. MATEJ TONIN:</speaker>
  <p>Dober dan, spoštovani vodje poslanskih skupin.</p>
  <p>Začenjam 1. sejo Kolegija predsednika
Državnega zbora.</p>
  <p>S tem lahko preidemo kar na
  <title>1. TOČKO DNEVNEGA REDA</title>.</p>
</sp>
```

(a) TEI Drama

```
<list type="agenda">
  <item xml:id="KPDZ-001-Redna-2018-06-29.toc-item0">
    <title>Pred dnevnim redom</title>
    <ptr target="#KPDZ-001-Redna-2018-06-29.seg1"/>
    <ptr target="#KPDZ-001-Redna-2018-06-29.seg2"/>
    <ptr target="#KPDZ-001-Redna-2018-06-29.seg3"/>
  </item>
  ...
<note type="time">Seja se je pričela ob 10.01.</note>
<note type="speaker">PRESEDNIK MAG. MATEJ TONIN:</note>
<u who="#ToninMatej"
  xml:id="KPDZ-001-Redna-2018-06-29.u1"
  ana="#chair">
  <seg xml:id="KPDZ-001-Redna-2018-06-29.seg1">Dober dan,
  spoštovani vodje poslanskih skupin.</seg>
  <seg xml:id="KPDZ-001-Redna-2018-06-29.seg2">
  Začenjam 1. sejo Kolegija predsednika
  Državnega zbora.</seg>
  <seg xml:id="KPDZ-001-Redna-2018-06-29.seg3">S tem lahko
  preidemo kar na 1. TOČKO DNEVNEGA REDA</seg>
</u>
```

(b) TEI Speech

Slika 1: Primer razlik med TEI Drama in TEI Speech kodiranimi parlamentarnimi razpravami.

pretvorba tako služi kot osnova za pretvorbo podatkov v skladu z priporočili Parla-CLARIN.

3.2.2 KODIRANJE PODATKOV V SKLADU Z MODULOM TEI SPEECH OZ. PRIPOROČILI PARLA-CLARIN

Druga stopnja kodiranja podatkov temelji na pretvorbi podatkov iz kodiranja TEI Drama v TEI Speech. V tej stopnji podatke s pomočjo XSLT (in nekaj Python) skript pretvorimo v končno obliko, ki je v skladu s priporočili Parla-CLARIN (ki vsebuje predvsem nabor oznak modula TEI Speech).

V tem delu poteka tudi združevanje posameznih govorov in njihovih segmentov (odstavkov). Del delotoka predstavlja tudi dodajanje identifikatorjev, označevanje vlog govorcev (*chair* za predsedujoče govorce in *regular* za poslance in druge govorce), označevanje opomb (*kinesic*, *vocal*, *incident* in *gap*) in priprava seznama točk dnevnega reda (ang. *agenda*), ki vsebujejo povezave na določene dele seje, ki se nanašajo na specifično točko dnevnega reda. Slika 1 prikazuje primer istega besedila, označenega z naborom oznak TEI Drama (Slika 1a) in TEI Speech (Slika 1b).

Ta faza tako predstavlja zadnjo stopnjo kodiranja podatkov in jih pripravi na naslednje stopnje gradnje korpusa – pripravo tako imenovanih korenskih dokumentov (ang. *root*, bolj podrobno predstavljenih pri opisu strukture korpusa v razdelku 4) ter jezikoslovno označevanje.

3.3 Priprava metapodatkov

Čeprav že sami transkripti parlamentarnih razprav predstavljajo bogat vir podatkov za raziskave parlamentarnega diskurza, z dodajanjem metapodatkov govorcev parlamentarni korpusi omogočajo raziskovanje drugih socialno-demografskih vidikov. V tem delu je velika večina dela sestavljena iz ročnega pregledovanja različnih virov, da tako zagotovimo natančen popis govorcev in drugih metapodatkov, ki v grobem obsegajo:

1. Metapodatke o govornikih (<listPerson>):
 - (a) ime in priimek govornika
 - (b) spol govornika (generirano avtomatično na podlagi vzorcev ženskih in moških imen, skupaj z ročnim pregledom)
 - (c) datum rojstva (če je le ta na voljo)
 - (d) afiliacija govornika z različnimi parlamentarnimi (ali drugimi političnimi organizacijami, npr. politično stranko ali v primeru poslanca, njegovo/njeno afiliacijo s poslansko skupino)

2. Metapodatke o organizacijah (<listOrg>):

- (a) podatki o Državnem zboru (ali v primeru SSK11, Skupščini, `xml:id="SK"`, `xml:id="DZ"`)
- (b) seznam delovnih teles (`xml:id="workingBodies"`)
- (c) seznam političnih strank, poslanskih skupin, nepovezanih (poslanskih skupin ali posameznih poslancev), ter drugih organizacij (npr. koalicija DEMOS)
- (d) seznam Vlad RS (`xml:id="GOV"`)
- (e) Vsak seznam vsebuje nadaljnje metapodatke:
 - i. ime organizacije (<orgName>)
 - ii. podatek o časovnem obdobju, v katerem organizacija obstaja (<event>)
 - iii. povezava na Wikipedijo ali drug vir podatkov, ki je bil uporabljen za sestavo tega seznama (idno).
 - iv. seznam "dogodkov" (<listEvent>/<event>), ki označuje posamezna časovna obdobja obstoja organizacije.

Metapodatki tako obsegajo ne le osnovne identifikacijske podatke govorca, ki jih pridobimo že pri obdelavi prepisov (npr. ime in priimek govorca), temveč tudi informacije o govorčevih afiliacijah (npr. poslanske skupine, politične stranke) ter natančnih obdobjih v katerih je deloval. Slednje je potrebno pridobiti iz zunanjih virov, kot so mandatna poročila Državnega zbora Republike Slovenije¹⁰ ali popisa sestav preteklih vlad.¹¹ Hkrati so ti metapodatki na voljo tako v korenskih dokumentih posameznih mandatov in celotnega korpusa, kot tudi v datotekah, ki spremljajo korpus. Te so primerne predvsem za analizo podatkov v tabelaričnih formatih, kot sta CSV in TSV ter predstavljajo pomemben del za analizo podatkov v konkordančnih. Slika 2 predstavlja primer metapodatkov o govorcu, kodiranih v korpusu.

3.4 Jezikoslovno označevanje korpusa

Korpus siParl je poleg ravnokar predstavljene osnove različice na voljo tudi v jezikoslovno označeni različici (*TEI.ana*). Ta različica je po vsebini, metapodatkih in kodiranju enaka osnovni, vendar besedilu doda jezikoslovne označbe (anotacije). Dodane jezikoslovne oznake obsegajo tokenizacijo besedila, segmentacijo na povedi, oblikoskladenjsko označevanje (ang. *PoS tagging*, lema-

¹⁰Mandatna poročila Državnega zbora RS

¹¹Pretekle Vlade Republike Slovenije

```
<person xml:id="BandelliMarko">
  <persName>
    <surname>Bandelli</surname>
    <forename>Marko</forename>
  </persName>
  <sex value="M"/>
  <birth when="1967-11-19">
    <placeName ref="https://www.geonames.org/3165185/trieste.html">Trst</placeName>
  </birth>
  <affiliation role="MP"
    ref="#DZ"
    from="2018-06-22"
    to="2018-09-12"
    ana="#DZ.8"/>
  <affiliation role="member"
    ref="#party.SAB"
    from="2018-06-22"
    to="2018-09-12"
    ana="#DZ.8"/>
  ...
  <affiliation role="minister"
    ref="#GOV"
    from="2018-09-13"
    to="2018-11-19"
    ana="#GOV.13"/>
</person>
```

Slika 2: Primer kodiranja metapodatkov poslanca, ki vsebuje osnovne osebne podatke in govorčevu afilijacijo z Državnim zborom, politično stranko in vlado.

tizacijo, skladenjske odvisnosti (ang. *syntactic dependencies*) in prepoznavanja imenskih entitet (ang. *named-entity recognition*). Besedilo v tej verziji korpusa je bilo jezikoslovno označeno s pomočjo orodja CLASSLA-Stanza (verzija 2.1)¹² (Ljubešić in Dobrovoljc, 2019; Terčon in Ljubešić, 2023). Slika 3 prikazuje primer jezikoslovno označenega stavka "Hvala za besedo."

4 SLOVENSKI PARLAMENTARNI KORPUS SIPARL4.0

Korpus siParl 4.0 zajema parlamentarne razprave in razprave delovnih teles Državnega zbora Republike Slovenije med leti 1990 in 2022 ter obsega več kot

¹²<https://github.com/clarinsi/classla>

```
<s xml:id="KPDZ-redna-1-2014-08-22.seg10.1">
<w xml:id="KPDZ-redna-1-2014-08-22.seg10.1.1"
  msd="UPosTag=NOUN|Case=Nom|Gender=Fem|
  Number=Sing" ana="mte:Ncfsn"
  lemma="hvala">Hvala</w>
<w xml:id="KPDZ-redna-1-2014-08-22.seg10.1.2"
  msd="UPosTag=ADP|Case=Acc"
  ana="mte:Sa" lemma="za">za</w>
<w xml:id="KPDZ-redna-1-2014-08-22.seg10.1.3"
  msd="UPosTag=NOUN|Case=Acc
  |Gender=Fem|Number=Sing" ana="mte:Ncfsa"
  lemma="beseda" join="right">besedo</w>
<pc xml:id="KPDZ-redna-1-2014-08-22.seg10.1.7"
  msd="UPosTag=PUNCT" ana="mte:Z">.</pc>
<linkGrp type="UD-SYN" targFunc="head argument">
  <link ana="ud-syn:amod"
    target="#KPDZ-redna-1-2014-08-22.seg10.1.2
    #KPDZ-redna-1-2014-08-22.seg10.1.1"/>
  <link ana="ud-syn:root"
    target="#KPDZ-redna-1-2014-08-22.seg10.1
    #KPDZ-redna-1-2014-08-22.seg10.1.2"/>
  <link ana="ud-syn:cc"
    target="#KPDZ-redna-1-2014-08-22.seg10.1.4
    #KPDZ-redna-1-2014-08-22.seg10.1.3"/>
</linkGrp>
</s>
```

Slika 3: Primer jezikoslovno označenega stavka "Hvala za besedo."

13 tisoč sej, milijon govorov in 230 milijonov besed. Vsebinsko, korpus obsega parlamentarne razprave naslednjih parlamentarnih teles:

- Skupščina Republike Slovenije, 1990 - 1992, 11. mandatno obdobje
- Državni zbor Republike Slovenije, 1992 - 2022, 1. - 8. mandatno obdobje
- Delovna telesa Državnega zbora Republike Slovenije, 1996 - 2022, 2. - 8. mandatno obdobje
- Kolegij predsednika Državnega zbora, 1996 - 2022, 2. - 8. mandatno obdobje

V primerjavi z drugimi sorodnimi parlamentarnimi korpusi (npr. korpusi Parla-Mint) ima siParl 4.0 bolj kompleksno strukturo, razlog za to pa je delno povezan tudi s strukturo priporočil za označevanje parlamentarnih razprav Parla-CLARIN:

1. Korenski dokument korpusa (ang. *corpus root*, `<teiCorpus>`):
 - (a) Metapodatki za celoten korpus (`<teiHeader>`):
 - i. Bibliografski podatki o celotnem korpusu (`<fileDesc>`)
 - ii. Taksonomije za organizacije, odbore, izraze, vrste sestankov, vrste govornikov (`<taxonomy>`)
 - iii. Seznam govorcev za celoten korpus (`<listPerson>`)
 - iv. Seznam organizacij za celoten korpus (`<listOrg>`)
 - (b) XInclude datoteke posameznih sej za celoten korpus (`<xi:include>`)
2. Korenski dokument posameznih mandatov
(ločeno po vrstah parlamentarnih teles – DZ, delovna telesa) (`<teiCorpus>`):
 - (a) Metapodatki za posamezen mandat (`<teiHeader>`):
 - i. Enaka struktura kot za celoten korpus, vendar vsebuje samo informacije, ki so relevantne za posamezen mandat.
 - (b) XInclude datoteke posameznih sej za posamezen mandat (`<xi:include>`)
3. Posamezne seje, ločeno po tipu sej
(npr. redne, izredne) in dnevu (`<TEI>`)
 - (a) Metapodatki posamezne seje (`<teiHeader>`):
 - i. Bibliografski podatki posamezne seje (`<fileDesc>`)
 - ii. Točke dnevnega reda (`<agenda>`)
 - (b) Transkript seje (`<text>`):
 - i. bibliografske opombe, npr. naslov seje, datum, ura, predsednik (`<head>`, `<note>`)
 - ii. Transkript govora (`<u>`):
 - A. Opombe prepisovalca (`<note>`, `<gap>`, `<vocal>`, `<kinesic>`, `<incident>`)
 - B. Paragrafi (`<seg>`)

Struktura vsebuje korenske dokumente korpusa (ang. *corpus root*) in mandatorov (ang. *term root*) na isti ravni, dokumenta sta medsebojno neodvisna. Z drugimi besedami, bodisi celoten korpus bodisi posamezen mandat se lahko obravnava kot ločen dokument XML. Kar zadeva datoteke in strukturo direktorijev in datotek, korenski dokument korpusa in vsak korenski dokument mandata ustrezata eni datoteki. Opozoriti je potrebno, da imamo za vsak mandat dva korenska dokumenta, odvisno od tega, katero parlamentarno telo je zasedalo, bodisi Državni zbor (*SDZ*), bodisi eno izmed delovnih teles Državnega zbora (*SDT*).

Mandat		Obdobje	Dnevi/ Mandat		Dnevi- Skupaj	Org.	Govor.	Posl.
SSK11		1990-05-05 - 1992-12-23	608			3	512	242
1	DZ	1992-12-23 - 1996-11-28	462			1	315	101
2	DZ	1996-11-28 -	430	2.141	1	359	105	
	DT	2000-10-27	1711		28	1268	108	
3	DZ	2000-10-27 -	303	1708	1	296	106	
	DT	2004-10-22	1405		26	1291	102	
4	DZ	2004-10-22 -	283	1583	1	237	103	
	DT	2008-10-15	1300		26	1886	101	
5	DZ	2008-10-15 -	233	1541	1	191	103	
	DT	2011-12-15	1308		26	2144	102	
6	DZ	2011-12-16 -	196	1321	1	202	115	
	DT	2014-08-01	1125		29	1994	111	
7	DZ	2014-08-01 -	288	1987	1	207	101	
	DT	2018-06-22	1699		23	2697	100	
8	DZ	2018-06-22 -	269	1882	1	264	109	
	DT	2022-05-13	1613		27	2077	104	
Skupaj		1990-05-05 - 2022-05-13	13233			69	9568	720

Tabela 2: Osnovne kvantitativne značilnosti posameznih mandatov (**Mandat**), ki vsebujejo zajeta zakonodajna (oziroma mandatna) obdobja (**Obdobje**). Tu se podatki ločijo še na tip zakonodajnega telesa (**SSK11** – Skupščina Republike Slovenije, **DZ** – Državni zbor, **DT** – delovna telesa), število dni, ko so potekala zasedanja po posameznih zakonodajnih organih (**Dnevi/Mandat**) in skupno število dni za celotno obdobje (**Dnevi - Skupaj**). Nazadnje so predstavljena še števila drugih zakonodajnih organov ali organizacij (**Org.**), govornikov (**Govor.**) in poslancev v parlamentu (**Posl.**). Državni zbor sicer sestavlja 90 poslancev, vendar lahko spremembe poslancev med zakonodajnim obdobjem povzročijo, da skupno število preseže 90 poslancev. Poleg tega je SSK11 (Skupščina Republike Slovenije) zaradi drugačne strukture parlamentarnega sistema štela 242 poslancev (oziroma delegatov).

Posamezne seje se nato razlikujejo glede na njihovo vrsto in datum, ko so te potekale. Vrste sej se naprej delijo na *redne*, *izredne* ali *skupne* seje, pri čemer izredne seje skliče predsednik Državnega zbora na zahtevo najmanj četrtnine poslancev ali predsednika republike in se običajno ukvarjajo s časovno občutljivi-

vimi in nujnimi zadevami (Državni zbor Republike Slovenije, 2020). Vsaka seja ima dnevni red, ki vsebuje točke, o katerih se bo razpravljalo, in o katerih se razpravlja v obliki govorov (ki jih včasih lahko prekine neverbalna vsebina, kot so informacije o izidu glasovanja, dejanja, kot je aplavz, itd.). Nazadnje sledijo označeni prepisi razprave, vključno z vsemi prekinitvami in drugimi (verbalnimi in neverbalnimi) dejavniki govora.

Številčno celoten korpus obsega 13.233 besedil, 1.221.133 posameznih govorov, 3.736.988 segmentov oz. odstavkov, 13.122.555 stavkov, 230.585.189 besed in 273.023.945 pojavnic (ang. *tokens*). Tabela 2 prikazuje bolj podrobno statistiko posameznih mandatov glede na posamezen tip zakonodajnega telesa ter celotnega korpusa.

Celotni korpus je na voljo v različnih formatih. Osnovna oblika korpusa vključuje TEI XML (oziroma v skladu z Parla-CLARIN označene) parlamentarne razprave. Poleg omenjene različice je korpus na voljo tudi v jezikoslovno označeni različici (TEI.ana). Nazadnje pa je korpus dostopen tudi v drugih izpeljanih formatih, in sicer kot navadno besedilo (.txt), v formatu CoNLL-U (.conllu) ter v vertikalnem formatu (.vert), ki je namenjen uporabi korpusa v konkordančnikih.

Korpus siParl je na voljo na repozitoriju CLARIN.SI¹³ (Pančur in sod., 2024), GitHub repozitoriju,¹⁴ prav tako pa je dostopen tudi na CLARIN.SI konkordančnih NoSketch Engine.¹⁵

5 ZAKLJUČKI

V prispevku smo predstavili novo različico Slovenskega parlamentarnega korpusa siParl 4.0, ki obsega parlamentarne razprave Republike Slovenije med leti 1990 in 2022 in s tem pokriva različna obdobja zgodovine Slovenije ter politične sisteme. Kot tak predstavlja bogat vir podatkov za analizo političnega in parlamentarnega diskurza. Korpus siParl je že v svojih začetkih še kot digitalizacijski projekt pomembno prispeval k usmerjenemu naboru parlamentarnih razprav in s tem k kasnejšemu razvoju slovenskih parlamentarnih korpusov. Hkrati se je korpus razvijal vzporedno z sorodnimi iniciativami, kot so priporo-

¹³<http://hdl.handle.net/11356/1936>

¹⁴<https://github.com/DARIAH-SI/siParl>

¹⁵<https://www.clarin.si/ske/#dashboard?corpname=siparl40>

čila za označevanje parlamentarnih razprav Parla-CLARIN in projektom Parla-Mint, na katere je tudi (ne)posredno vplival.

V primerjavi s prejšnjo različico smo obstoječi korpus siParl 3.0 razširili z manjkajočimi razpravami delovnih teles in razpravami Kolegija predsednika Državnega zbora Republike Slovenije za 8. mandat (2018 – 2022). Poleg omenjene razširitve podatkov, ki jih korpus pokriva, smo se v tej verziji posvetili tudi posodobitvi delotoka razvoja korpusa, z namenom poenostavitve in čim večje avtomatizacije procesa.

V prihodnosti se želimo posvetiti razširitvi korpusa s plenarnimi razpravami in razpravami drugih delovnih teles 9. mandata, ki je trenutno še v teku. Poleg vsebinskih razširitev si želimo korpus nadgrajevati tudi na metapodatkovni in tehnični ravni. V tem pogledu na dolgi rok načrtujemo konsistentno dodajanje novih prepisov in povezovanje le teh z zunanjimi podatkovnimi bazami ali drugimi dodatnimi parlamentarnimi gradivi. Tu primarno izpostavljamo povezovanje parlamentarnih razprav s Poročevalcem Državnega zbora¹⁶, ki vsebuje besedila zakonov in drugih dodatkov ki jih sami prepisi razprav ne vsebujejo, in ki je prosto dostopen v okviru portala Zgodovina Slovenije - SIstory.¹⁷ Nazadnje želimo v prihodnosti postopno razširiti nabor metapodatkov, ki trenutno opisujejo parlamentarne razprave v korpusu.

Poleg nadaljnega tehničnega in vsebinskega razvoja korpusa želimo v prihodnosti pripraviti poglobljeno analizo vsebine Slovenskega parlamentarnega korpusa siParl. Kot je navedeno v citatu v naslovu tega prispevka,¹⁸ parlament in parlamentarne razprave predstavljajo nenehne kontrolirane polemike oziroma razprave med poslanci in drugimi govorniki – v ta namen želimo korpus siParl v prihodnosti obogatiti z označevanjem sentimenta, ki nam bodo skupaj s prepisi in njihovimi metapodatki omogočile poglobljen vpogled v “polje kontroliranega konflikta”, ki ga parlament predstavlja.

¹⁶Poročevalec Državnega zbora

¹⁷Portal Zgodovina Slovenije - SIstory

¹⁸Citat pripada poslanki Maši Kociper, izrečen na 90. izredni seji Državnega zbora (22. 12. 2021) v 8. mandatu.

6 ZAHVALA

Avtorji se zahvaljujemo anonimnim recenzentom za koristne pripombe ter dr. Nikoli Ljubešiću za jezikoslovno označevanje korpusa siParl 4.0. Prispevek je nastal v okviru slovenskih raziskovalnih infrastruktur DARIAH-SI in CLARIN.SI, raziskovalnega programa "Tehnologije znanja (P2-0103)" ter infrastrukturnega programa "Raziskovalna infrastruktura Slovenskega zgodovinopisja (I0-0013)", financiranih s strani Javne agencije za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije.

LITERATURA

- Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., ... Fišer, D. (2022). The Parlamint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 1–34. <https://doi.org/10.1007/s10579-021-09574-0>
- Erjavec, T. in Pančur, A. (2022). The Parla-CLARIN Recommendations for Encoding Corpora of Parliamentary Proceedings. *Journal of the Text Encoding Initiative*(14). <https://doi.org/10.4000/jtei.4133>
- Erjavec, T. in Pančur, A. (2019). *Parla-CLARIN: A TEI Schema for Corpora of Parliamentary Proceedings*. <https://clarin-eric.github.io/parla-clarin/>
- Fišer, D. in Pahor de Maiti, K. (2020). Voices of the parliament. *Modern Languages Open*. <https://doi.org/10.3828/mlo.v0i0.295>
- Fišer, D. in Pahor de Maiti, K. (2021). *Voices of the parliament: A corpus approach to parliamentary discourse research*. <https://sidih.github.io/voices/index.html>
- Kavčič, A., Mundjar, A. in Marolt, M. (2023a). *Carniolan provincial assembly corpus kranjska 1.0*. Faculty of Computer and Information Science, University of Ljubljana. <http://hdl.handle.net/11356/1824>
- Kavčič, A., Mundjar, A. in Marolt, M. (2023b). *Parliamentary corpus of first yugoslavia (1919-1939) yu1Parl 1.0*. Faculty of Computer and Information Science, University of Ljubljana. <http://hdl.handle.net/11356/1845>
- Ljubešić, N. in Dobrovoljc, K. (2019, August). What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. V *Proceedings of the 7th workshop on Balto-Slavic Natural Language Processing* (str. 29–34). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W19-3704>
- Meden, K., Erjavec, T. in Pančur, A. (2024). Slovenian parliamentary corpus siParl. *Language Resources and Evaluation*, 1–21.
- Ogrodniczuk, M., Osenova, P., Erjavec, T., Fišer, D., Ljubešić, N., Çöltekin, Ç., Kopp, M.,

- & Meden, K. (2022). ParlaMint II: The show must go on. V D. Fišer, M. Eskevich, J. Lenardič in F. de Jong (Ur.), *Proceedings of the workshop parlaclarin iii wi-thin the 13th language resources and evaluation conference* (str. 1–6). Marseille, France: European Language Resources Association. <https://aclanthology.org/2022.parlaclarin-1.1>
- Pančur, A. (2016). Označevanje zbirke zapisnikov sej slovenskega parlamenta s smernicami TEI. V *Zbornik konference jezikovne tehnologije in digitalna humanistika* (str. 142–48).
- Pančur, A., Erjavec, T., Meden, K., Ojsteršek, M., Šorn, M. in Blaj Hribar, N. (2022). *Slovenian parliamentary corpus (1990-2022) siParl 3.0*. Institute of Contemporary History. <http://hdl.handle.net/11356/1748>
- Pančur, A., Erjavec, T., Ojsteršek, M., Šorn, M. in Blaj Hribar, N. (2019). *Slovenian parliamentary corpus (1990-2018) siParl 1.0*. Institute of Contemporary History. <http://hdl.handle.net/11356/1236>
- Pančur, A., Erjavec, T., Ojsteršek, M., Šorn, M. in Blaj Hribar, N. (2020). *Slovenian parliamentary corpus (1990-2018) siParl 2.0*. Institute of Contemporary History. <http://hdl.handle.net/11356/1300>
- Pančur, A., Meden, K., Erjavec, T., Ojsteršek, M., Šorn, M. in Blaj Hribar, N. (2024). *Slovenian parliamentary corpus (1990-2022) siParl 4.0*. Institute of Contemporary History. <http://hdl.handle.net/11356/1936>
- Pančur, A. in Šorn, M. (2016). Smart big data: Use of Slovenian parliamentary papers in digital history. *Contributions to Contemporary History*, 56(3), 130–146.
- Pančur, A., Šorn, M. in Erjavec, T. (2016). *Slovenian parliamentary corpus (1990-1992) SlovParl 1.0*. Institute of Contemporary History. <http://hdl.handle.net/11356/1075>
- Pančur, A., Šorn, M. in Erjavec, T. (2017). *Slovenian parliamentary corpus (1990-1992) SlovParl 2.0*. Institute of Contemporary History. <http://hdl.handle.net/11356/1167>
- Pančur, A., Šorn, M. in Erjavec, T. (2018). Slovparl 2.0: The collection of Slovene parliamentary debates from the period of secession. V D. fišer, & ad j. maria eskevich (eds.) *parlaclarin 2018 workshop proceedings may* (Zv. 7, str. 2018).
- Pančur, A. in Erjavec, T. (2020). The siParl corpus of Slovene parliamentary proceedings. V *Proceedings of the second parlaclarin workshop* (str. 28–34).
- Polanič, P. in Dobranič, F. (2022). *Corpus of political party programs Programi2022*. Institute of Contemporary History. <http://hdl.handle.net/11356/1734>
- Sestava in organiziranost*. (2020). Pridobljeno 25. februar 2023: <https://www.dz-rs.si/wps/portal/Home/odz/pristojnosti/organiziranost>.

TEI P5: Guidelines for Electronic Text Encoding and Interchange. (2020). Pridobljeno 27.2.2023: <https://tei-c.org/guidelines/p5/>.

Terčon, L. in Ljubešić, N. (2023). *CLASSLA-Stanza: The next step for linguistic processing of South Slavic Languages*. <https://doi.org/10.48550/arXiv.2308.04255>

Truan, N. in Romary, L. (2022). Building, Encoding, and Annotating a Corpus of Parliamentary Debates in TEI XML: A Cross-Linguistic Account. *Journal of the Text Encoding Initiative*(14).

“ACCORDING TO THEORY, PARLIAMENT IS A FIELD OF CONTROLLED CONFLICT”: SLOVENIAN PARLIAMENTARY CORPUS SIPARL 4.0

The debates of national parliaments and other legislative bodies, and thus the parliamentary corpora, represent an important source of data for the study of political and especially parliamentary speech and other related phenomena in the humanities and social sciences. The siParl Slovenian Parliamentary Corpus contains a series of parliamentary debates of Slovenian parliamentary bodies from the period between 1990 and 2022, covering not only different periods of Slovenian history, but also transitions between different political systems. The corpus has a long history of development and has influenced the development of related corpora, initiatives and projects. In this article, we present the new version of siParl 4.0, describe the development of the corpus, the structure and the encoding of parliamentary data in the corpus. We also outline the next steps in the further development of siParl.

Keywords: parliamentary corpora, parliamentary debates, Slovenian Parliament, TEI-XML, Parla-CLARIN

To delo je ponujeno pod licenco Creative Commons: Priznanje avtorstva-Deljenje pod enakimi pogoji 4.0 Mednarodna.

This work is licensed under the Creative Commons Attribution-ShareAlike 4.0 International.

<https://creativecommons.org/licenses/by-sa/4.0/>

