

Digitalna humanistika, jezikovne tehnologije in parlamentarni podatki

Katja Meden^{1,2}

¹Odsek za tehnologije znanja, Institut Jožef Stefan,

²Raziskovalna infrastruktura slovenskega zgodovinopisja, Inštitut za novejšo zgodovino

Predavanje 19. 5. 2025 na Oddelku za bibliotekarstvo, informacijsko znanost in založništvo

Vsebina

- Digitalna humanistika in jezikovne tehnologije
- Raziskovalne infrastrukture
 - DARIAH-SI
 - CLARIN.SI
 - OPERAS-SI
- Korpusi in parlamentarni podatki
 - Kaj je korpus?
 - Slovenski parlamentarni korpusi: siParl in ParlaMint korpusi
 - Priprava parlamentarnega korpusa
- Obogatitev parlamentarnih podatkov:
 - Raziskava – motivacija in raziskovalni problem
 - Novi metapodatki in anotacije sentimenta
- **Namen:** *znanja informacijske znanosti odpirajo vrata tudi v številna, morda manj očitna področja, kot so digitalna humanistika, jezikovne tehnologije in raziskovalne infrastrukture.*

Digitalna humanistika in jezikovne tehnologije

Področja, rezultati in študije

Digitalna humanistika

Definicije zagotavlja ponudnik [Oxford Languages](#) · [Več o tem](#)



digital humanities

noun

an academic field concerned with the application of computational tools and methods to traditional humanities disciplines such as literature, history, and philosophy.

"the unit will advance scholarship in both classical studies and **the digital humanities**"

Jezikovne tehnologije

- Jezikovne tehnologije: področje, ki se ukvarja z računalniško obdelavo naravnega jezika.
- Vedno bolj raziskovalno/komercialno zanimivo področje (Google Translate, ChatGPT)
- Glavna paradigma: nadzorovano strojno učenje
- Programi so večinoma jezikovno neodvisni
 - potrebujejo pa učne podatke za učenje modela, in
 - testne podatke za evalvacijo

Digitalna humanistika

- **Korpusi:** Strukturno označeni korpusi (npr. z oznakami za stavčno analizo, semantiko, sentiment).
 - Zgodovinski ali specializirani korpusi (npr. parlamentarni, pravni, literarni).
- Digitalne znanstevnokritične izdaje:
 - Npr. Kritične izdaje besedil z digitalnimi prepisi/pripisi, faksimili in večjezičnimi prevodi.
- **(Meta)Podatkovne baze:**
 - Zaključene/dinamične zbirke (raziskovalnih) podatkov
- Interaktivne vizualizacije
- Zemljevidi in GIS podatki
- Digitalni arhivi
- Ontologije in taksonomije:
 - Strukture znanja, ki kategorizirajo informacije za boljše iskanje in povezovanje

Jezikovne tehnologije

- **Korpusi**
- **Konkordančniki**
- **Veliki jezikovni modeli**
- (Digitalni) Leksikoni
- (Meta)Podatkovne zbirke
- Računalniški modeli in simulacije

Korpusi

kórpús -a m (ô)

SSKJ² 

1. *vojaška enota iz divizij in posebnih enot*: formirati korpus; poveljnik korpusa / ekspedicijski korpus
2. jezikosl. *zbirka besedil v elektronski obliki s standardizirano obliko zapisa in izbranimi podatki*: analiza korpusa; delo s korpusom; besedila za korpus; korpus in slovar / referenčni korpus *ki je glede na vključena besedila zvrstno kar se da uravnotežen / besedilni korpus*
3. publ., navadno s prilastkom *skupina, zbor*: nastopili so najboljši korpusi; baletni, godalni korpus / diplomatski korpus *diplomatski zbor*; korpus miru zlasti v ameriškem okolju *organizacija iz prostovoljcev, ki skrbi zlasti za gospodarski razvoj in pomoč nerazvitim deželam*
4. knjiž. *zbirka*: korpus srednjeveških fresk
 - ekspr. *ta ženska je pa korpus je zelo debela*

Vzporedni prikazi

Vrsta prikaza (strani) ▾

Način vzporednega prikaza (facs-dipl-crit) ▾

Prelom vrstice:

Da

Ne

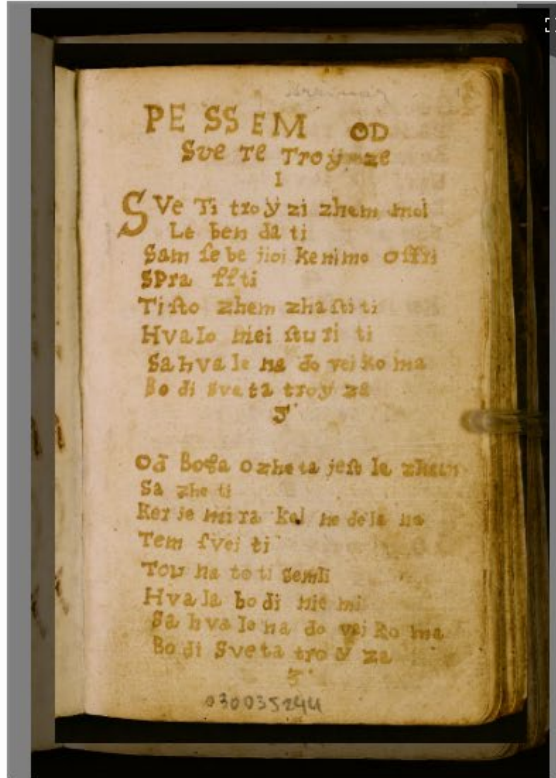
01v <<

stran 02r

>> 02v

Digitalne znanstveno-
kritične izdaje

Ditmajer, N., et al. (2021).
Foglarjeva pesmarica ([Elektronska
znanstvenokritična izdaja](#)).



PE SS EM OD Sve Te Troj ze

1

Sve Ti tro y zi zhem moi
Le ben da ti
Sam fe be jioi kenimo offri
Spra ffiti
Tisto zhem zha fti ti
Hvala niei ftu ri ti
Hvalo niei ftu ri ti

Sahva le na do vei ko ma
Bo di sve ta troj za

2

Od Boga o zhe ta jest le zhem
Sa zhe ti
Ker je mira kel ne dela na
Tem fvei ti
Tou na to ti semli
Hva la bo di nie mi

Sa hva le na do vei ko ma
Bo di sveta tro y za

3

Pesem od Svete trojce

1

Sveti trojci čem moj leben dati,
sam sebe joj k enimo ofri sprafti.
Tisto čem častiti,
hvalo njej sturiti.

Zahvalena do vejcoma
bodi Sveta trojca.

2

Od Boga Očeta jest le čem začeti,
ker je mirakelne dela na tem svejti.
Tou na toti zemli
hvala bodi njemi.

Zahvalena do vejcoma
bodi Sveta trojca.

3

Celotno besedilo

Celotno besedilo

Besedilo dokumenta

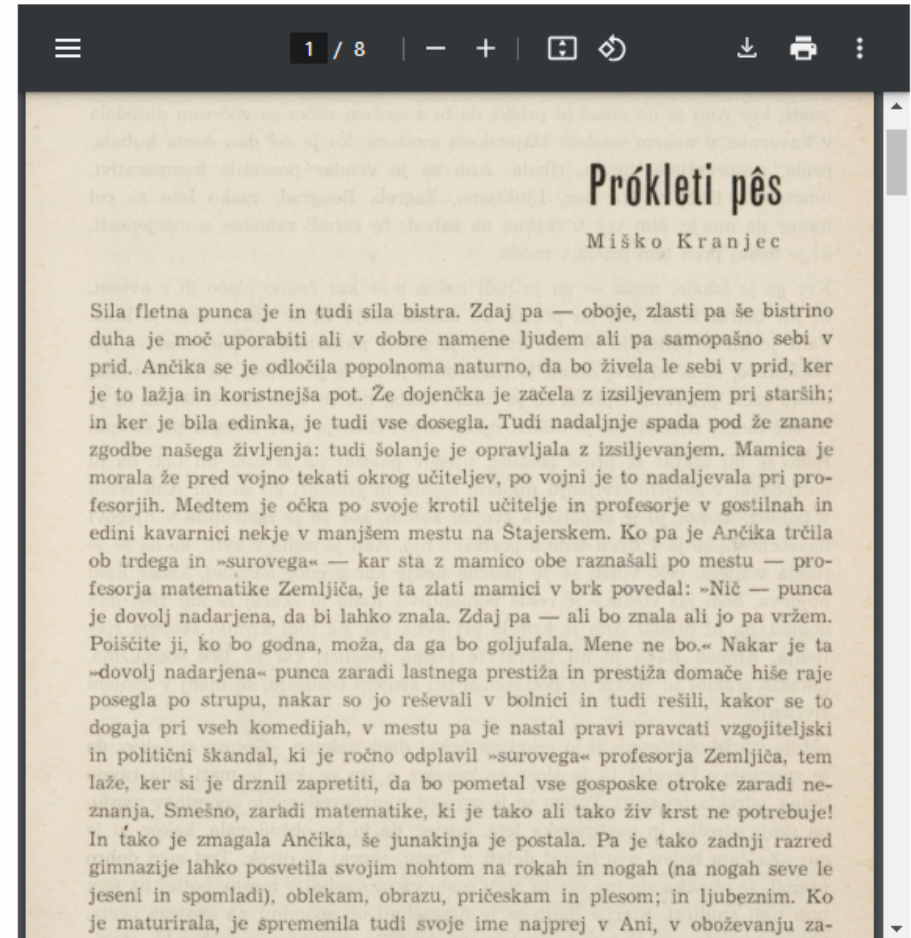
Miško Kranjec (1908): Prókleti pês (1964)

>>

Digitalne znanstveno-
kritične izdaje

Sila fletna punca je in tudi sila bistra. Zdaj pa — oboje, zlasti pa še bistrino duha je moč uporabiti ali v dobre namene ljudem ali pa samopašno sebi v prid. Ančika se je odločila popolnoma naravno, da bo živela le sebi v prid, ker je to lažja in koristnejša pot. Že dojenčka je začela z izsiljevanjem pri starših; in ker je bila edinka, je tudi vse dosegla. Tudi nadaljnje spada pod že znane zgodbe našega življenja: tudi šolanje je opravljala z izsiljevanjem. Mamica je morala že pred vojno tekati okrog učiteljev, po vojni je to nadaljevala pri profesorjih. Medtem je očka po svoje krotil učitelje in profesorje v gostilnah in edini kavarnici nekje v manjšem mestu na Štajerskem. Ko pa je Ančika trčila ob trdega in »surovega« — kar sta z mamico obe raznašali po mestu — profesorja matematike Zemljiča, je ta zlati mamici v brk povedal: »Nič — punca je dovolj nadarjena, da bi lahko znala. Zdaj pa — ali bo znala ali jo pa vržem. Poiščite ji, ko bo godna, moža, da ga bo goljufala. Mene ne bo.« Nakar je ta »dovolj nadarjena« punca zaradi lastnega prestiža in prestiža domače hiše raje posegla po strupu, nakar so jo reševali v bolnici in tudi rešili, kakor se to dogaja pri vseh komedijah, v mestu pa je nastal pravi pravcati vzgojiteljski in politični škandal, ki je ročno odplavil »surovega« profesorja Zemljiča, tem laže, ker si je drznil zapreti, da bo pometal vse gosposke otroke zaradi neznanja. Smešno, zaradi matematike, ki je tako ali tako živ krst ne potrebuje! In tako je zmagala Ančika, še junakinja je postala. Pa je tako zadnji razred gimnazije lahko posvetila svojim nohtom na rokah in nogah (na nogah se ve le jeseni in spomladi), oblekam, obrazu, pričeskam in plesom; in ljubeznim. Ko je maturirala, je spremenila tudi svoje ime najprej v Ani, v oboževanju za-

Juvan, M., Žejn, A., Šorli, M.,
Mandić, L., Tomažin, A., Jež, A.,
Balžalorsky Antić, V., Erjavec,
T. (2021). *Korpus Maj68*
([Elektronska znanstvenokritična
izdaja](#), [korpus](#))



Slovenian Twitter hate speech dataset IMSyPP-sl



“ Za citiranje vnosa uporabite naslednjo referenco ali jo izvozite v prednastavljeno obliko:

BIBTEX

CMDI

Kralj Novak, Petra; Mozetič, Igor and Ljubešić, Nikola, 2021, *Slovenian Twitter hate speech dataset IMSyPP-sl*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, <http://hdl.handle.net/11356/1398>.



Delite:

CLARIN.SI Data & Tools

Avtorji Kralj Novak, Petra ; Mozetič, Igor and Ljubešić, Nikola

Identifikator vnosa <http://hdl.handle.net/11356/1398>



URL projekta <http://imsypp.ijs.si/>

Datum objave 2021-02-17

Vrsta [corpus, text](#)

Velikost 120000 items

Jezik(i) [Slovenian](#)

Opis A hand-labeled training (50,000 tweets labeled twice) and evaluation set (10,000 tweets labeled twice) for hate speech on Slovenian Twitter. The data files contain tweet IDs, hate speech type, hate speech target, and annotator ID. For obtaining the full text of the dataset, please contact the first author.

Hate speech type:

1. Appropriate - has no target
2. Inappropriate (contains terms that are obscene, vulgar; but the text is not directed at any person specifically) - has no target
3. Offensive (including offensive generalization, contempt, dehumanization, indirect offensive remarks)
4. Violent (author threatens, indulges, desires, or calls for physical violence against a target; it also includes calling for, denying, or glorifying war crimes and crimes against humanity)

Podatkovne baze



PUBLIKACIJE ŽRTVE I. SV. **ŽRTVE II. SV.** POPISI ZIC

Vnesite avtorja in/ali naslov

 / ŽRTVE II. SV. VOJNE

 KOPIRAJ POVEZAVO

 SLOJI

vse

AA

ANTON ANTONČIČ  **Inštitut za novejšo zgodovino**

priimek Antončič ime Anton oče Janez mati Jožefa priimek matere Zorec datum rojstva 21. 01. 1911 kraj rojstva Hrib ob Temenici
kraj smrti Teharje

BA

BOGOMIR ANTONČIČ  **Inštitut za novejšo zgodovino**

priimek Antončič ime Bogomir oče Gregor mati Frančiška priimek matere Mahnič datum rojstva 23. 10. 1922
kraj rojstva Dolenja vas kraj smrti Škrlovisa

DA

DUŠAN ANTONČIČ  **Inštitut za novejšo zgodovino**

priimek Antončič ime Dušan kraj smrti Mauthausen

FA

FRANC ANTONČIČ  **Inštitut za novejšo zgodovino**

priimek Antončič ime Franc oče Anton kraj rojstva Vrh, Cerknica kraj smrti Nanos

☒ Inštitut za novejšo zgodovino

☒ Prispevki drugih uporabnikov

Drago Antonič Inštitut za novejšo zgodovino

Osební podatki

priimek **Antonič**ime **Drago**kraj rojstva **Komen**naslov **Komen**stara občina **Komen**nova občina **Komen**pokrajina **Primorska**

zakonski stan ?

poklic (soc. status) **delavec**narodnost **Slovenec**spol **moški**

Podatki o smrti

vojni status **NOV in POS**država smrti **Slovenija**država pokopa **Slovenija**vzrok smrti **Oborožen spopad**

povzročitelj smrti ?

Viri

Inštitut za novejšo zgodovino

- Pričevanja 1941-1945, Sveto 1997, str. 25.

Opombe

Inštitut za novejšo zgodovino

- RMK Komen ga ne navaja



To delo je ponujeno pod [Creative Commons Priznanje avtorstva-Deljenje pod enakimi pogoji 4.0 Mednarodna](#)

Primer študije: Parlament in čas

- **Teza**: V dobi moderne je čas postajal pomembna dobrina, z njim je bilo treba ravnati *racionalno* in zato je bilo treba delovati *učinkovito*.
- Slovenski parlamentarni korpus siParl 2.0 (NoSketch Engine), korpusno jezikoslovje
- Besede, ki izražajo dožemanje poslanskega zbora (*čim prej, hitrost, učinkovito, neučinkovito, racionalen, čas je*).
- Luščenje kolokacij; seznam besed, ki se neposredno ob jedrni besedi (pred njo in za njo), povezani s časom, pojavljajo pogostejše kot naključno
- Zaključki:
 - *Neučinkovito = neracionalno = počasno*
 - »Dobro« rešitev bi sprejemali predolgo, danes pa je čas »dragocen« in »zlato«.
 - V parlamentu je čas še največkrat »skrajni«.
 - Sama beseda čas se v vseh sklonih v parlamentu pojavi kar 262.964-krat

[5] Tabela 7: Racionalen – kolokacijski kandidati

		<u>Cooccurrence count</u>	<u>Candidate count</u>	<u>logDice</u>
<u>P</u> <u>N</u>	obnašati	163	1.754	9,494
<u>P</u> <u>N</u>	gospodarno	128	1.384	9,219
<u>P</u> <u>N</u>	rabo	118	3.414	8,736
<u>P</u> <u>N</u>	učinkovito	211	14.715	8,407
<u>P</u> <u>N</u>	racionalno	107	5.462	8,301

[7] Tabela 9: Čas je ... – kolokacijski kandidati

		<u>Cooccurrence count</u>	<u>Candidate count</u>	<u>logDice</u>
<u>P</u> <u>N</u>	potekel	411	3.343	10,320
<u>P</u> <u>N</u>	Skrajni	207	628	9,761
<u>P</u> <u>N</u>	skrajni	194	3.403	9,229
<u>P</u> <u>N</u>	Ves	124	1.604	8,852
<u>P</u> <u>N</u>	zadnjem	359	22.692	8,621
<u>P</u> <u>N</u>	minilo	69	1.885	7,961

Raziskovalne infrastrukture

DARIAH-SI, CLARIN.SI in OPERAS-SI

Raziskovalne infrastrukture

- [DARIAH-SI](#): Raziskovalna infrastruktura za umetnost in humanistiko
- [CLARIN.SI](#): Infrastruktura za jezikovne vire in tehnologije
- [OPERAS-SI](#): Infrastruktura za odprto znanstveno komunikacijo v družboslovju in humanistiki
- [Slovenska skupnost odprte znanosti](#)

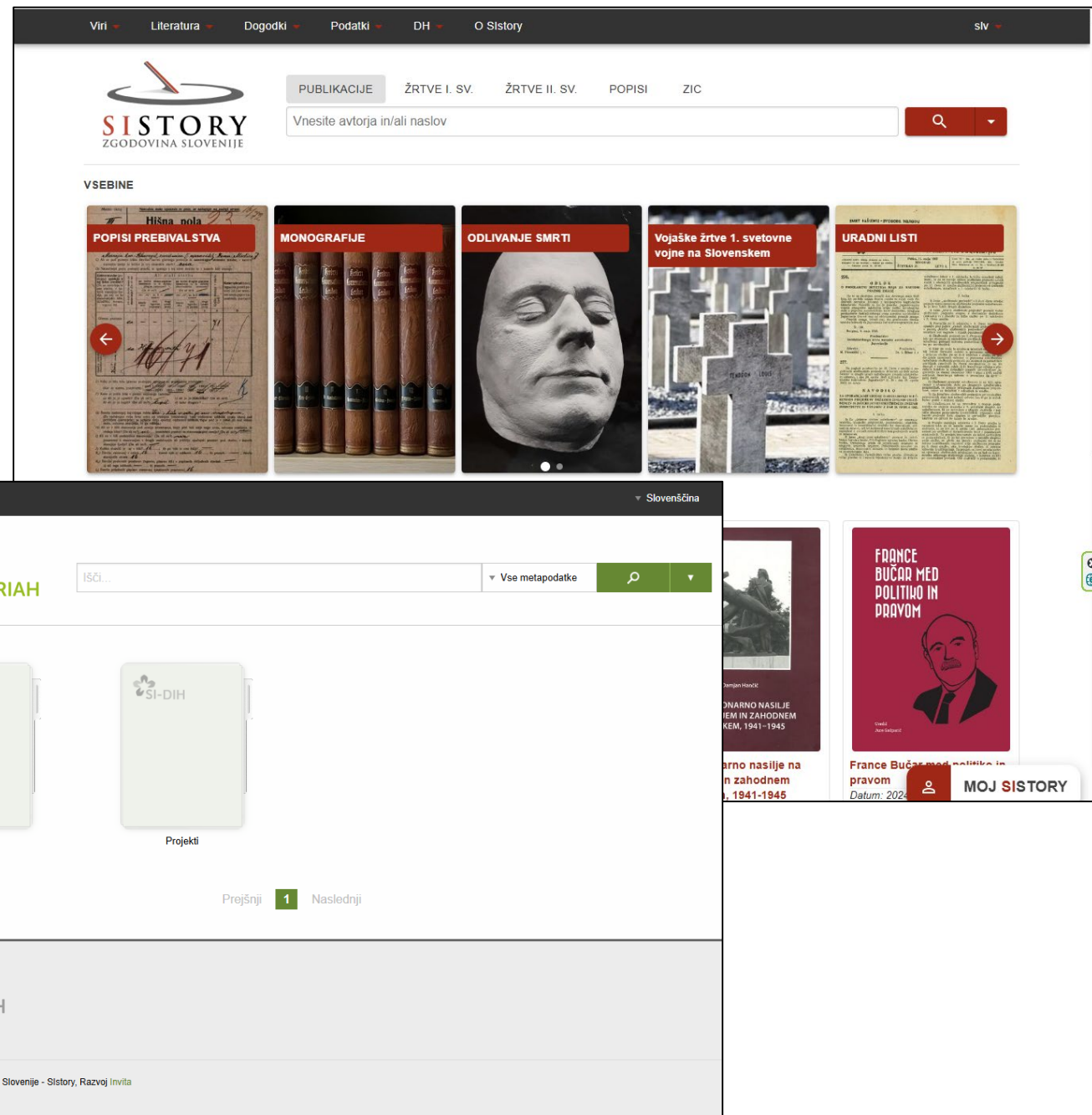


Raziskovalna infrastruktura za umetnost in humanistiko DARIAH-SI

- Digital Research Infrastructure for the Arts and Humanities
- Nacionalna veja DARIAH EU, koordinacijska inštitucija Inštitut za novejšo zgodovino
- Članice: ZRC SAZU
- Nacionalni koordinator: dr. Vojko Gorjanc
- Namen:
 - uresničevanje deklariranih nacionalnih strategij in politik,
 - poenotenje digitalne raziskovalne infrastrukture za umetnost in humanistiko na nacionalnem nivoju,
 - povezovanje digitalnih podatkovnih zbirk na nacionalni ravni pod enotno platformo,
 - povezovanje, diskurz ter homogenizacija skupnosti razvijalcev in uporabnikov sodobnih rešitev,
 - popularizacija uporabe naprednih orodij v humanistiki.

Produkti DARIAH-SI

- Portal Zgodovina Slovenije – Sistory.si
- Repozitorij digitalne humanistike SI-DIH
- PRINZ – Praksa INZ:
 - Od 2024 dalje
 - strukturiran program usposabljanja za podiplomski študij
 - Oddelek za zgodovino, FF
 - Vsebinska predavanja in praktično usposabljanje
 - Uporaba digitalnih metod



Infrastruktura za jezikovne vire in tehnologije CLARIN.SI

- CLARIN: Common Language Resources and Technology Infrastructure
- *Inštitut Jožef Stefan* + 11 konzorcijskih partnerjev:
 - 4 univerze: Ljubljana, Maribor, Nova Gorica, Primorska
 - 4 raziskovalni inštituti: ZRC SAZU, IJS, INZ, ZRS Koper
 - 1 knjižnica: NUK
 - 1 društvo: Slovensko društvo za jezikovne tehnologije
 - 2 podjetji: Amebis, Alpineon
- Nacionalni koordinator: dr. Tomaž Erjavec
- Vizija: dostopnost raznolikih digitalnih jezikovnih virov in orodij za vse (evropske) jezike
- Namenjena dolgotrajnemu in obsežnemu hranjenju ter dostopu do jezikovnih virov in tehnologij
- Prispevek k ohranjanju in podpiranju večjezične evropske kulturne dediščine
- Paradigma sodelovanja pri razvoju virov in orodij, zagotavljanje večkratne uporabnosti in prilagajanja individualnim potrebam

Repozitorij CLARIN.SI:

- Repozitorij CLARIN.SI / Prikaz vnosa

Slovenian parliamentary corpus (1990-2022) siParl 4.0

” Za citiranje vnosa uporabite naslednjo referenco ali jo izvozite v prednastavljeno obliko:

BIBTEX CMDI

Pančur, Andrej; et al., 2024, *Slovenian parliamentary corpus (1990-2022) siParl 4.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, <http://hdl.handle.net/11356/1936>.

Ta vir je integriran tudi v naslednje storitve:

Delite:

noSketch

[CLARIN.SI Data & Tools](#)

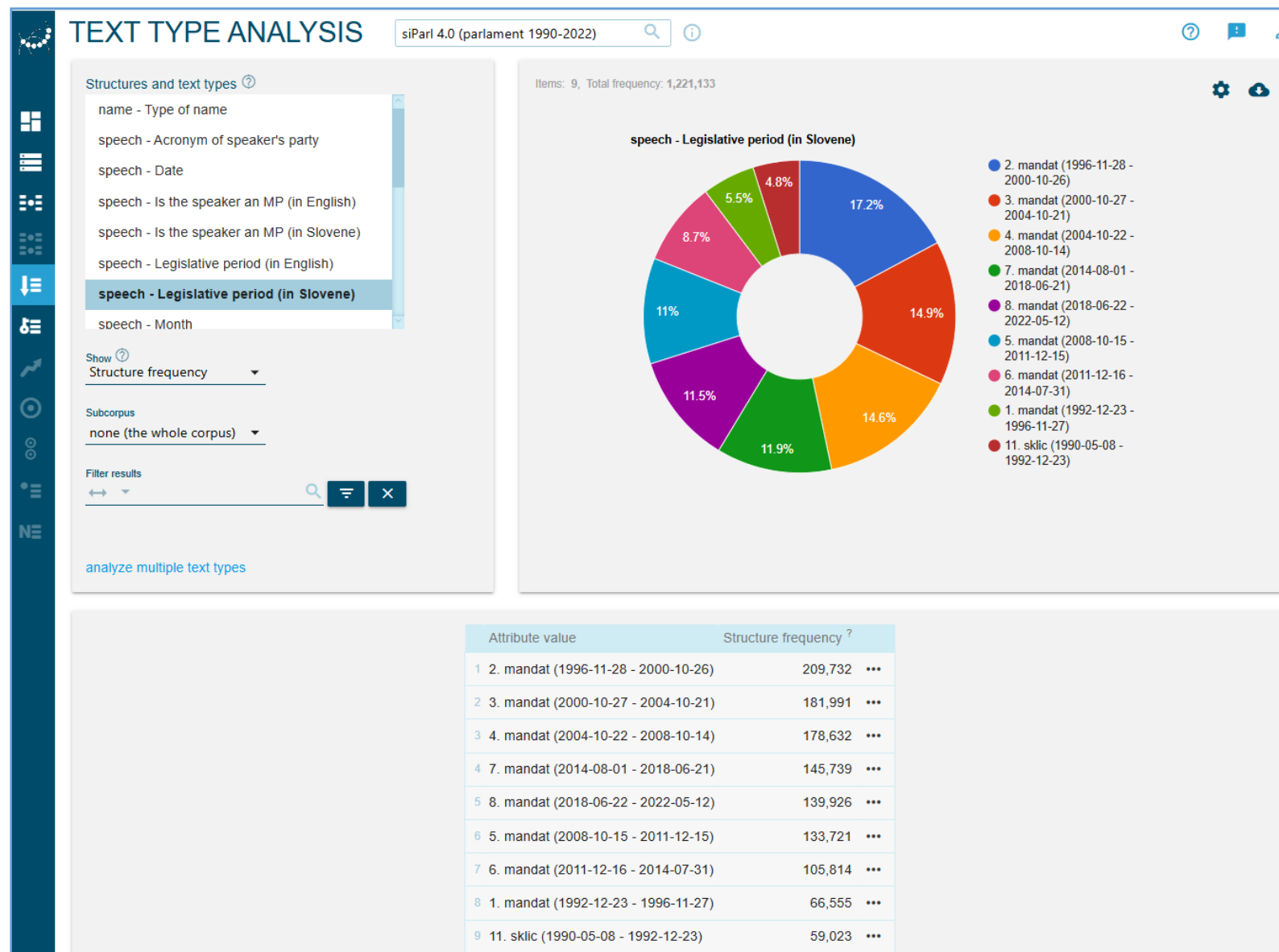
✎ Avtorji	Pančur, Andrej ; et al. ▶ prikaži vse
↶ Identifikator vnosa	http://hdl.handle.net/11356/1936
🔗 URL projekta	https://github.com/DARIAH-SI/siParl/
📄 Dokumentirano v	https://doi.org/10.1007/s10579-024-09746-8
📅 Datum objave	2024-06-05
📁 Vrsta	corpus , text
⚙ Velikost	13233 texts, 1223250 utterances, 239224540 words
🇸🇯 Jezik(i)	Slovenian
📖 Opis	The siParl 4.0 corpus contains minutes of the Assembly of the Republic of Slovenia for 11th legislative period 1990-1992, minutes of the National Assembly of the Republic of Slovenia from the 1st to the 8th legislative period 1992-2022, minutes of the working bodies of the National Assembly of the Republic of Slovenia from the 2nd to the 8th legislative period 1996-2022, and minutes of the Council of the President of the National Assembly from the 2nd to the 8th legislative period 1996-2022. The corpus comprises of over 13 thousand sessions, one million speeches and 230 million words. The corpus is encoded according to the Parla-CLARIN schema (https://github.com/clarin-eric/parla-clarin). Each mandate is in one directory, and each session in one file.

- 🔍 Brskanje
 - > Celoten repozitorij
 - 👤 Moj račun
 - ➡ Odjava
 - 👤 Profil
 - 📤 Vnosi
 - ☰ Kontekst
 - > Urejanje vnosa
 - > Izvoz vnosa
 - > Izvoz metapodatkov
 - ⚙ Administrativno
 - 🔧 Nadzorna plošča
 - 🔒 Nadzor dostopa
 - > Administracija vsebin
 - 📊 Registri
 - > Skupnosti in zbirke
 - 📈 Statistika
 - 📄 Naloge pri kuriranju
 - 🔗 Licence

Storitve CLARIN.SI II

Konkordančniki (NoSketch Engine, KonText)

- Orodja za (spletno) analizo korpusov
- Podpirajo delo z zelo velikimi korpusi (več milijard besed)
- Korpusi so lahko bogato označeni
- Ponujamo cca. 100 korpusov v 33 jezikih z 20 milijard besed



OPERAS-SI

- Raziskovalna infrastruktura, ki podpira odprto znanstveno komunikacijo v družboslovju in humanistiki (SSH) v Evropskem raziskovalnem prostoru
- Koordinacijska inštitucija ZRC SAZU (Založba ZRC SAZU);
- Članice:
 - Inštitut za novejšo zgodovino
 - Univerza v Ljubljani
 - Univerza v Mariboru
 - Univerza na Primorskem
 - Centralna tehnična knjižnica
 - Univerza v Novi Gorici (opazovalka)
- Nacionalni koordinator: mag. Aleš Pogačnik
- Odprta znanost in odprto založništvo – objavlanje znanstvenih del v okviru različnih nivojev odprtega dostopa (predvsem spodbuda za objavlanje v diamantnem dostopu)
- Open publishing systems: [Open Journal System](#) (serijske publikacije), [Open Monograph Press](#) (monografije) in [Open Preprint Systems](#)

Slovenska skupnost odprte znanosti

- **Vizija:** Medsebojno povezati deležnike odprte znanosti v Sloveniji in vzpostaviti sodoben, mednarodno pomemben in konkurenčen znanstveni sistem, ki temelji na načelih odprte znanosti ter je usklajen in povezan z evropskim raziskovalnim prostorom in ustreznimi evropskimi iniciativami.
- CLARIN.SI, DARIAH-SI in OPERAS-SI so članice skupnosti
- **Aktivnosti:**
 - izboljšanje delovnih pogojev raziskovalcev,
 - spodbude za širjenje, izmenjavo in ponovno uporabo znanja na način odprtega dostopa,
 - razvoj kompetenc za prakticiranje odprte znanosti in
 - vzpostavitev sistema usposabljanja za ciljne uporabnike.



Korpusi in parlamentarni podatki

Korpusi I

- Enovito kodirana in dokumentirana zbirka besedil
- Označena:
 - ročno ali strojno
 - Jezikovne anotacije

Attributes **ana** and **msd** both encode the morphosyntactic characteristics of the word in focus (in blue). The **ana** attribute contains Slovene-specific tags according to MULTEXT-East Specifications, while the **msd** attribute contains the Universal Dependency tagset which greatly simplifies cross-lingual comparison.

The attribute values Q and PART both stand for *particle*. The difference between the two tags is in the tagset used.

The **word form** also known as a token from the running text in the corpus.

The attribute **lemma** indicates the basic word form of the token, that is of the word in focus (in blue).

The corpus also encodes syntactic parses and named entities but since they are not used in this tutorial they were omitted from this illustrative example.

```
<s>
<w ana="mte:Q" msd="UpoSTag=PART" lemma="zlasti">Zlasti</w>
<w ana="mte:Sg" msd="UpoSTag=ADP|Case=Gen" lemma="glede">glede</w>
<w ana="mte:NcmSg" msd="UpoSTag=NOUN|Case=Gen|Gender=Masc|Number=Sing"
lemma="nadzor">nadzora</w>
<w ana="mte:Va-r3s-n"
msd="UpoSTag=AUX|Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Pres|VerbForm=Fin"
lemma="biti">je</w>
<w ana="mte:Pd-fsn" msd="UpoSTag=DET|Case=Nom|Gender=Fem|Number=Sing|PronType=Dem"
lemma="ta">ta</w>
<w ana="mte:NcfSg" msd="UpoSTag=NOUN|Case=Nom|Gender=Fem|Number=Sing"
lemma="stvar">stvar</w>
<w ana="mte:Rgp" msd="UpoSTag=ADV|Degree=Pos" lemma="zelo">zelo</w>
<w ana="mte:Agpfsn" msd="UpoSTag=ADJ|Case=Nom|Degree=Pos|Gender=Fem|Number=Sing"
lemma="kočljiv">kočljiva</w>
<pc ana="mte:Z" msd="UpoSTag=PUNCT">.</pc>
</s>
```

One pair of the opening and closing structural tags which, in our case, indicate punctuation (**pc**), word (**w**) or sentence (**s**).

Korpusi II

- Referenčni korpusi (splošen korpus, ki ponuja pregled jezika, npr. [Gigafida](#), [metaFida](#))
- Vzporedni korpusi (korpusi v različnih jezikih):
 - [Japonsko-slovenski vzporedni korpus jaSlo](#)
- Specializirani korpusi:
 - Korpusi družbenih omrežij/forumov/nestandardnih jezikov: ([Janes Tweet](#), [Janes Forum](#), [Janes News](#), [Janes TwePo](#) – tviti politikov)
 - Korpusi spleta ([CLASSLA-web.sl](#),)
 - Korpusi Wikipedij ([CLASSLA-Wikipedia](#) ...)
 - Parlamentarni korpusi

Korpusi parlamentarnih razprav

- Parlamentarne razprave: transkripcije plenarnih zasedanj, zapisniki sej
- Prepisi govornega jezika
- Vsebujejo dragocene (sociodemografske) metapodatke:
 - Ime in priimek govorca
 - Datum rojstva
 - Politične / parlamentarne stranke
 - Spol govorca...
- Korpusi parlamentarnih razprav: digitalne zbirke prepisov parlamentarnih razprav
- Bogat vir za različne raziskovalne vede, predvsem za:
 - Družboslovje in humanistiko
 - Raziskovanje parlamentarnega/političnega diskurza

Lastnosti

- Prepisi niso vedno dobesedni
- Izpuščene napake, neprimeren jezik
- Oblikovani v skladu z konvencijami parlamenta:
 - Grški parlament: prekinitve strogo prepovedane, sankcionirane
 - Britanski parlament: prekinitve so popolnoma običajne in niso sankcionirane
 - Poslanci imajo pogosto vnaprej pripravljene govore
- TEI kodiranje in jezikovna anotacija besedila

Official records

"Madam President, I think we all agreed that equal pay is, first of all, a matter of elementary fairness, and the fact that we still have this average pay gap of 16-17% also shows, among other things, the lack of respect for women, which is so deeply rooted in traditions and stereotypes. /.../."

(Věra Jourová, Member of the Commission)

Verbatim transcription

*"**Ladies and Gentlemen, yes gentlemen also are here still. Am...** I think we all agreed that equal pay is, first of all, a matter of elementary fairness, and **that** the fact that still **we** have this **pay gap am...** average 16-17% **this** also shows, among other things, the lack of respect to women, which is so deeply rooted in traditions and stereotypes. /.../."*

(Věra Jourová, Member of the Commission)

Darja Fišer in Kristina Pahor de Maiti. 2021. *Prvič, sem političarka in ne politik, drugič pa ---* : a corpus approach to parliamentary discourse research. Inštitut za novejšo zgodovino, Ljubljana. Dostopno na: <https://sidih.si/cdn/121/ch3.html>

Slovenski parlamentarni korpusi

- [Slovenski parlamentarni korpus \(1990 – 2022\) siParl 4.0](#)
- [Parlamentarni korpusi ParlaMint \(ParlaMint-SI, 2000 - 2022\)](#)
- [Parlamentarni korpus prve Jugoslavije \(1919-1939\) Yu1Parl](#)
- [Parlamentarni korpus Kranjskega deželnega zbora \(1861 – 1913\)](#)

	Kranjska 1.0	Yu1Parl 1.0	siParl 4.0	ParlaMint-SI 4.1
Obdobje	1861 – 1913	1919 – 1939	1990 – 2022	2000 - 2022
Zapisniki sej	Kranjski deželni zbor	Predstavništvo, senat in skupščina Kraljevine Jugoslavije	Državni zbor, Delovna telesa DZ, Kolegij predsednika DZ	Državni zbor
Jeziki	slovenščina (58%), nemščina (42%)	srbsščina (cirilica, 59%), hrvaščina (38%), slovenščina (3%)	slovenščina	slovenščina, angleščina (strojni prevodi)
Velikost korpusa (besede)	10.883.619	13.271.885	239.224.540	69.032.700

Kako ustvarimo takšne (in podobne) korpuse?

Surovi podatki -----> Korpus v konkordančniku (iskanje)

11	☐		ZalarAleš • 200...	Morda bi bilo po mnenju Vlade pri tem primerno upoštevati tudi znanstveno poročilo z naslovom Ključne značilnosti slovenske politike v letih 1929 do 1955, ki ga je za Državni zbor pripravil Inštitut za novejšo zgodovino .	
12	☐		GrčarSlavko • 2...	V Sloveniji imamo zelo relevantno raziskavo, ki jo opravlja Inštitut za novejšo zgodovino s poimenskim evidentiranjem vseh žrtev, slovenskih žrtev, mislim na slovenskih tleh tekom 2. svetovne vojne in veste, da je teh žrtev naštetih že 81 ali 82.000.	
13	☐		PahorBorut • 20...	Ostale razrešitve, ki jih prikazuje števec vaše stranke, so zaradi zamenjav v organih upravljanja: nadzorni svet, Eles, Kad, Sod, Javna agencija za podjetništvo in tuje investicije, Inštitut za novejšo zgodovino , Filmski sklad.	
14	☐		LavtižarBeblerD...	Zakaj se ministrstvo ni odločilo, da bi k sodelovanju povabilo obstoječe institucije. kot na primer Inštitut za novejšo zgodovino . SAZU in podobne, saj se je, kolikor je me	
15	☐		BattelliRoberto...	In de	
16	☐		BevkSamo • 2005...	N iz konteksta velikanskih morij in zločinov druge svetovne vojne, iz konteksta zločinov, ki so jih nad slovenskim in drugimi jugoslovanskimi narodi uresničevali okupatorji in njihovi sodelavci, iztrgano iz konteksta žrtev, ki so jih ti povzročili. V Sloveniji imamo zelo relevantno raziskavo, ki jo opravlja Inštitut za novejšo zgodovino s poimenskim evidentiranjem vseh žrtev, slovenskih žrtev, mislim na slovenskih tleh tekom 2. svetovne vojne in veste, da je teh žrtev naštetih že 81 ali 82.000. Razen tega še 12.000 potem po vojni pobitih domobrancev. Od teh 82.000 je devet desetih teh žrtev na račun	
17	☐		KučlerDolinarMo...	V : zg	
18	☐		ZupanJure • 200...	Nc za	

SLAVKO GRČAR: Hvala. Najprej hvala za povabilo na to sejo. Rad bi zelo na kratko obvestil to komisijo o nekaterih sicer znanih dejstvih. Namreč prva taka stvar je, da v zadnjih skoraj dvajsetih letih, številnih naših dokumentih in govorih, smo nesmiselno obsodili poboje. To posebej poudarjam zaradi tega, ker se še vedno pojavljajo ugotovitve, da smo mi zagovorniki poveljnih pobojev. To ni res. Obsodili smo jih. Druga stvar, ki bi jo rad izrečno povedal je, da podpiramo strokovno tehnično delovanje in druge strokovne aktivnosti pri realizaciji določb zakona o vojnih grobiščih. Torej tudi strokovno tehnično delo te komisije o katerem je danes pogovor. Tudi nam se zdi, da se morda določene aktivnosti preveč dolgo že vlečejo in da bi morda kakšne stvari bilo mogoče hitreje in temeljiteje opravljati v dosedanjih obdobjih pa morda bo zato v bodočih letih ta aktivnost kaj bolj intenzivna. Seveda pa ne podpiramo strupenih agitk, ki se vežejo na odkritje teh tako imenovanih prikritih grobišč. Mislimo, da je čas, da se o vseh stvareh, ki so povezane z 2. svetovno vojno, začnemo pogovarjati na kulturni način. Tudi v skladu s 63. členom slovenske Ustave, ki prepoveduje razpihovanje nestrpnosti in sovražstva. Mislimo, da je čas za to. Mislimo tudi, da je čas za kulturo dialoga, da se je treba začeti pogovarjati na človeški način. Ne tako, kot se je zdaj velikokrat dogajalo pa ne bom citiral posameznih izjav, ampak saj ste jim bili sami priča. To ni dobro. Ne za razmere v državi, ne za mednarodne odnose. Mnoge stvari, ki se pri nas dogajajo in potem prepotencirajo, potem izkoriščajo v tujini na neprimerne načine. Te primere smo recimo v Italiji, imeli smo jih na Hrvaškem pa morda še kje. Torej tu bi morali z eno odgovornostjo, bi rekel, se obravnavati vsa ta vprašanja. Predvsem se nam tudi zdi sporno, da o tem razpravljamo marsikdaj izdvojeno, iz konteksta velikanskih morij in zločinov druge svetovne vojne, iz konteksta zločinov, ki so jih nad slovenskim in drugimi jugoslovanskimi narodi uresničevali okupatorji in njihovi sodelavci, iztrgano iz konteksta žrtev, ki so jih ti povzročili. V Sloveniji imamo zelo relevantno raziskavo, ki jo opravlja **Inštitut za novejšo zgodovino** s poimenskim evidentiranjem vseh žrtev, slovenskih žrtev, mislim na slovenskih tleh tekom 2. svetovne vojne in veste, da je teh žrtev naštetih že 81 ali 82.000. Razen tega še 12.000 potem po vojni pobitih domobrancev. Od teh 82.000 je devet desetih teh žrtev na račun okupatorjev in njihovih sodelavcev.

Delotok - Faze

1. Digitalizacija in optično prepoznavanje znakov (OCR)/Ekstrakcija podatkov (HTML)
2. Čiščenje in označevanje podatkov
3. Razločevanje govorcev
4. Popis in dodajanje metapodatkov
5. Jezikovno označevanje korpusa
6. Objava korpusa (repozitorij + konkordančnik)

```

<person xml:id="DimitrovskiMaja">
  <persName>
    <surname>Dimitrovski</surname>
    <forename>Maja</forename>
  </persName>
  <sex value="F"/>
  <birth when="1975-03-11"/>
  <idno type="URI" subtype="wikimedia">https://sl.wikipedia.org/wiki/Maja_Dimitrovski</idno>
  <affiliation role="member" ref="#DZ" from="2011-12-16" to="2014-07-31" ana="#DZ.6">
    <roleName xml:lang="en">MP</roleName>
  </affiliation>
  <affiliation role="member" ref="#party.PS" from="2011-12-16" to="2014-04-29" ana="#DZ.6">
    <roleName xml:lang="en">Member</roleName>
  </affiliation>
  <affiliation role="member" ref="#party.NP" from="2014-05-05" to="2014-06-12" ana="#DZ.6">
    <roleName xml:lang="en">Member</roleName>
  </affiliation>
  <affiliation role="member" ref="#party.ZaAB" from="2014-06-13" to="2014-07-31" ana="#DZ.6">
    <roleName xml:lang="en">Member</roleName>
  </affiliation>

```

```

<org xml:id="party.ZaAB" role="parliamentaryGroup">
  <orgName full="yes" xml:lang="sl">Zavezništvo Alenke Bratušek</orgName>
  <orgName full="yes" xml:lang="en">Alliance of Alenka Bratušek</orgName>
  <orgName full="abb">ZaAB</orgName>
  <event from="2014-05-31" to="2016-05-21">
    <label xml:lang="en">existence</label>
  </event>
  <idno type="URI" subtype="wikimedia">https://sl.wikipedia.org/wiki/Stranka_Alenke_Bratusek</idno>
  <idno type="URI" subtype="wikimedia">https://en.wikipedia.org/wiki/Party_of_Allenka_Bratusek</idno>
  <state type="politicalOrientation">
    <state type="Wikipedia" source="https://sl.wikipedia.org/wiki/Stranka_Alenke_Bratusek">
      <state type="variable" ana="#ches.party_id">
        <state type="value" from="2014" to="2018" n="2913"/>
      </state>
    <state type="variable" ana="#ches.party">
      <state type="value" from="2014" to="2018" n="ZaAB"/>
    </state>
  </state>

```

phrase **Parlament je** • 1,184

4.34 per million tokens • 0.00043%



Details

sentence

- | | | |
|----|---|---|
| 1 | ☐ RajičBranislav ... | In drugo, parlament je predstavnik ljudi in mi smo tukaj tudi, da smo neka korelacija z nekim narodom s katerim smo v p bližini ali bliskosti že stoletja. |
| 2 | ☐ TrčekFranc • 20... | Parlament je tudi neka procedura, vlagajo se takšni in drugačni zakon, dobri, slabi. |
| 3 | ☐ GregorčičMonika... | Vemo, nekaj parlamentom v Evropi je že sprejelo podobne dokumente in pa konec koncev tudi Evropski parlament je že obsodil agresorsko dejanje. |
| 4 | ☐ ŽnidaričMojca •... | Tudi vi imate možnost, vsi lahko kandidirate, tako da tudi če pridete mladi v parlament je vaš glas seveda toliko bolj sliš |
| 5 | ☐ PogačnikMarko •... | In parlament je , bi rekel, zakonodajna veja oblasti. |
| 6 | ☐ ŠkrinjarMojca •... | Če vse te stvari vpletamo v svoje razprave, kajti parlament je predmet razprave ne pa operative, potem si lahko od tega da res zagotovo bo veliko mest v domovih za ostarele, ker bomo vsi prej umrli, preden nas bi kakšen ukrep dosegel. |
| 7 | ☐ SukičNataša • 2... | Da bi prezrli dejstvo, da če se zbere cel parlament je 47 poslank in poslancev proti frekingu. |
| 8 | ☐ KociperMaša • 2... | Parlament je po teoriji polje kontroliranega konflikta. |
| 9 | ☐ PrebilNik • 202... | Ne glede na to, če imamo na določeno politično situacijo povsem drugačno stališče, je to prostor – to tukaj, parlament je soočamo. |
| 10 | ☐ GolobRobert • 2... | Baje, da ista oseba, ki sprejema direktorje distribucij v parlamentu tako, da tokrat ni hodil v parlament je samo po telefon |
| 11 | ☐ StarovićVojko •... | Parlament je najnižje na lestvic. |
| 12 | ☐ SukičNataša • 2... | Tukaj je resna debata, so resni pomisleki in parlament je prostor, kjer se takšne pomisleke izrazi in to ni nobena žalitev, že gre malo počasi na živce, ker vsak to potem začne opletati, ko ni argumenta postanete žrtve. |
| 13 | ☐ SukičNataša • 2... | Parlament je avtonomen. |

Obogatitev parlamentarnih podatkov

Motivacija in raziskovalni problem

- Parlamentarne razprave: pomembna povezava med predstavniki parlamenta in splošno javnostjo
- Parlamentarni korpusi: bogat podatkovni vir za različne discipline (npr. zgodovinopisje, sociologijo, jezikoslovje itd.)
- Veliki napredki pri gradnji parlamentarnih korpusov (npr. ParlaMint)

Vendar: Kljub napredku pri oblikovanju in povezljivosti korpusov ostaja ključni izziv ***kontekst***.

- Diskurza ne oblikuje le jezik, temveč tudi politična pripadnost, ideološko stališče, institucionalna vloga govorca... (Van Dijk, 2016).
- Za raziskovanje specifičnih (političnih in diskurzivnih) pojavov so potrebni dodatni metapodatki, ki jih ni mogoče razbrati zgolj iz prepisov.
- *Več metapodatkov = bolj podrobne in poglobljene raziskave in analize*

Glavna teza: **kontekst ni zgolj postranskega pomena, temveč ključen** – ne le za analizo diskurza, temveč tudi za zasnovo, sestavo in interpretativno uporabo parlamentarnih korpusov.

Cilji raziskave

- Razširitev in kodiranje slovenskih (in evropskih) parlamentarnih korpusov:
 - Slovenski parlamentarni korpus siParl
 - Parlamentarni korpusi projekta ParlaMint (SI)
- Zagotavljanje novih metapodatkov in anotacij:
 - Metapodatki o ministrih
 - Politična orientacija
 - Anotacija sentimenta (na ravni stavkov in polnih govorov)
- Predstavitev novo razvitih izboljšav za analizo parlamentarnega diskurza
 - Študija političnih identitet v slovenskem parlamentarnem govoru

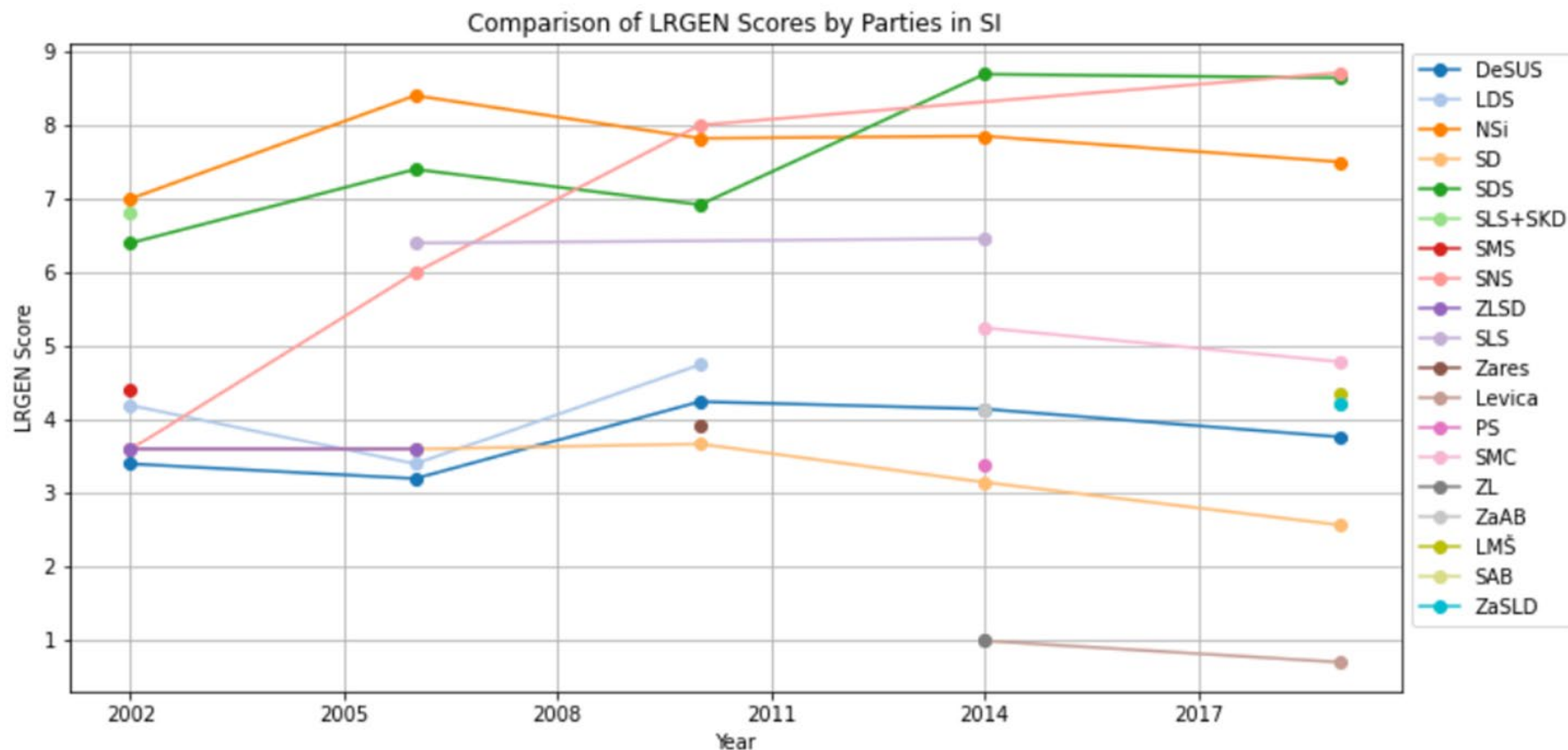
Dodajanje novih metapodatkov – **Ministri in politična orientacija**

- Ročen nabor metapodatkov o ministrskih pozicijah govorcev
- Viri:
 - [Chapel Hill Expert Survey Europe](#) (CHES Europe)
 - Wikipedia (ročni nabor podatkov)
 - Znanje kolegov iz drugih držav
- 85 CHES oznak za različne pozicije političnih strank (npr. pozicija stranke glede migracij/državljskih pravic...)

Abbreviation	Value
FL	Far-left
LLF	Left to far-left
L	Left
CLL	Centre-left to left
CL	Centre-left
CCL	Centre to centre-left
C	Centre
CCR	Centre to centre-right
CR	Centre-right
CRR	Centre-right to right
R	Right
RRF	Right to far-right
FR	Far-right
BT	Big tent ⁵
PP	Pirate Party ⁶
SY	Syncretic politics ⁷
SI	Single-issue politics ⁸
NP	Nonpartisanism ⁹

Table 1: Political orientation values, identified in the Wikipedia data.

Primer analize politične orientacije skozi čas - SI

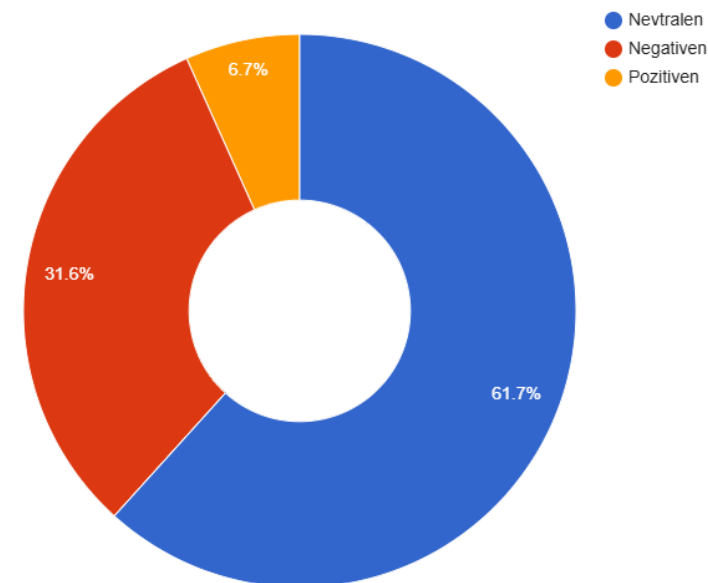


(a) Plot of the lrngen values for SI per individual year.

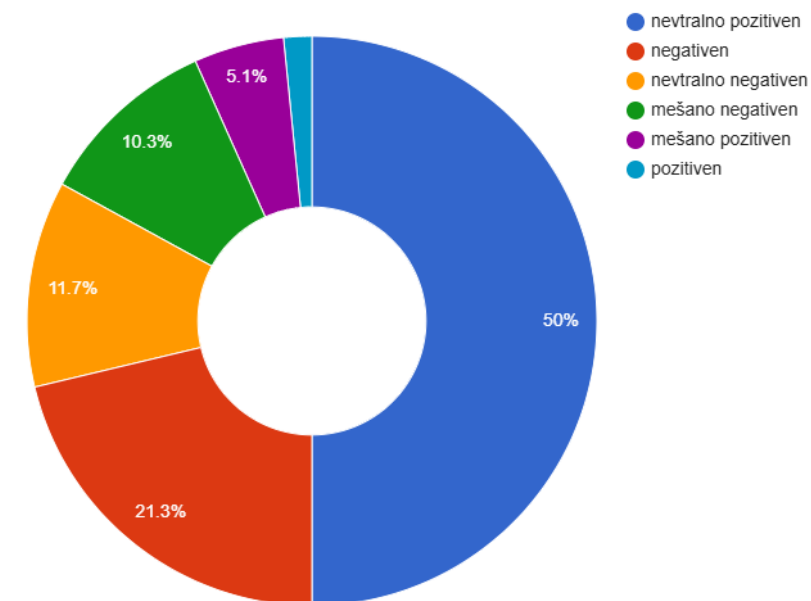
Dodajanje novih metapodatkov: Sentiment

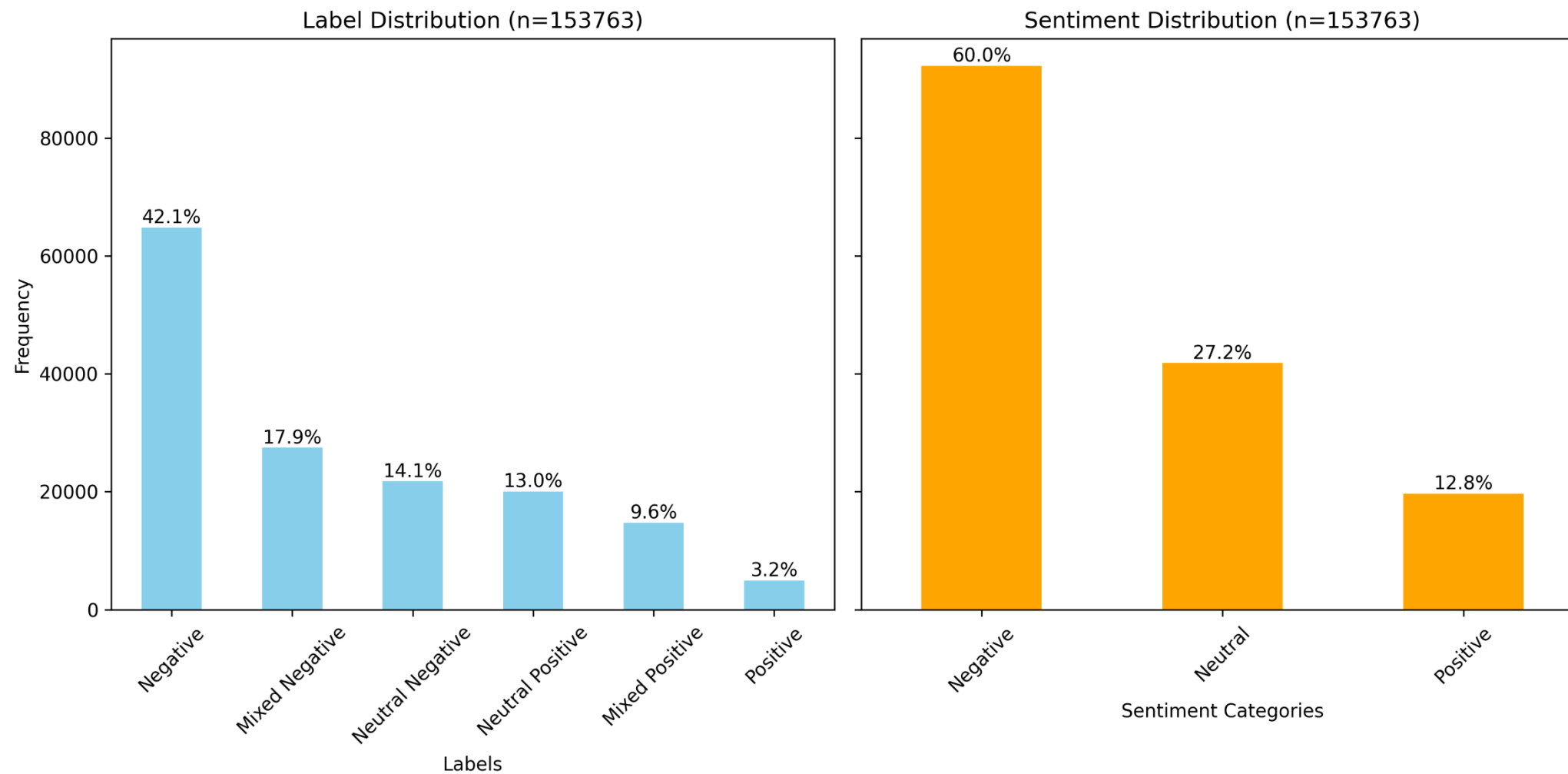
- Anotacija sentimenta na bolj podrobni lestvici kot tradicionalna delitev na negativno, pozitivno, nevtralno
- Ročne anotacijske kampanje:
 - Na ravni stavkov: 2.600 stavkov, anotiranih s strani dveh anotatorjev ([ParlaSent 1.0](#))
 - Na ravni govorov: 1.000 celotnih govorov (isti anotatorji), namenjeno agregaciji ocen sentimenta
- Oznake:
 - 6-stopenjska lestvica: pozitivno, mešano pozitivno, nevtralno pozitivno, nevtralno negativno, mešano negativno, negativno
 - 3-stopenjska lestvica: pozitivno, nevtralno, negativno

speech - senti_3



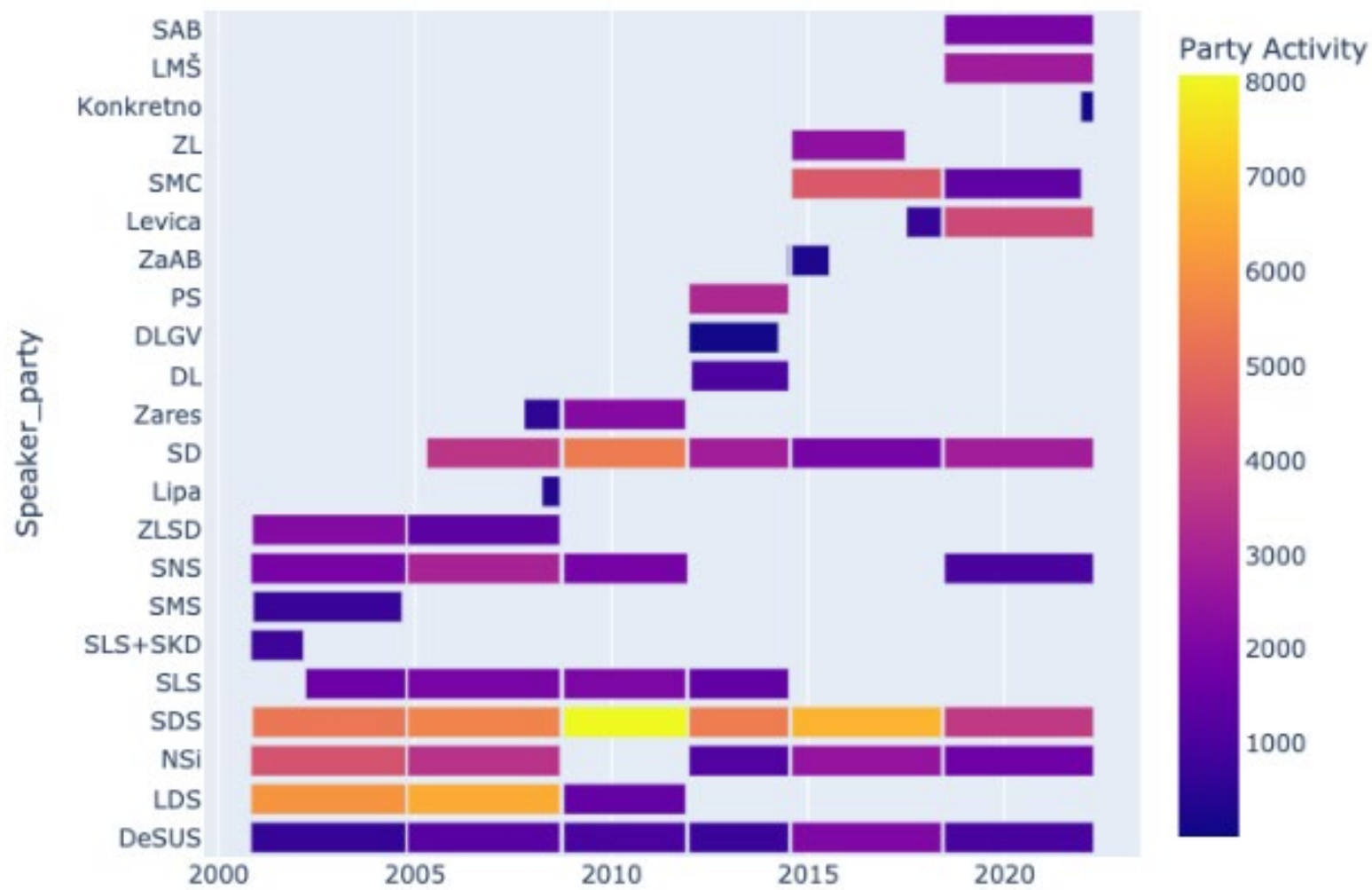
speech - senti_6



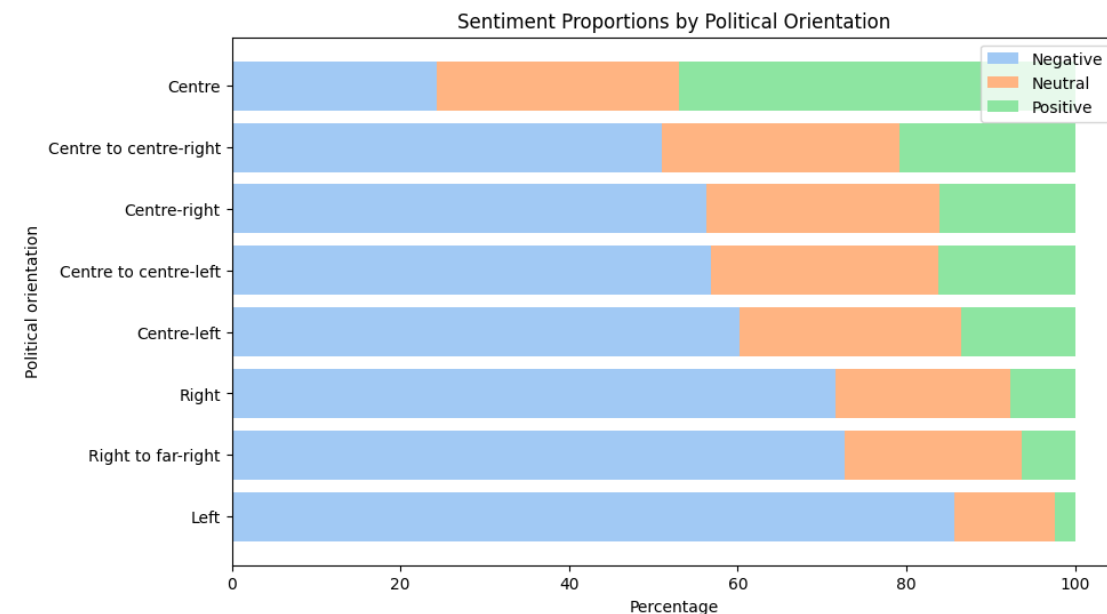
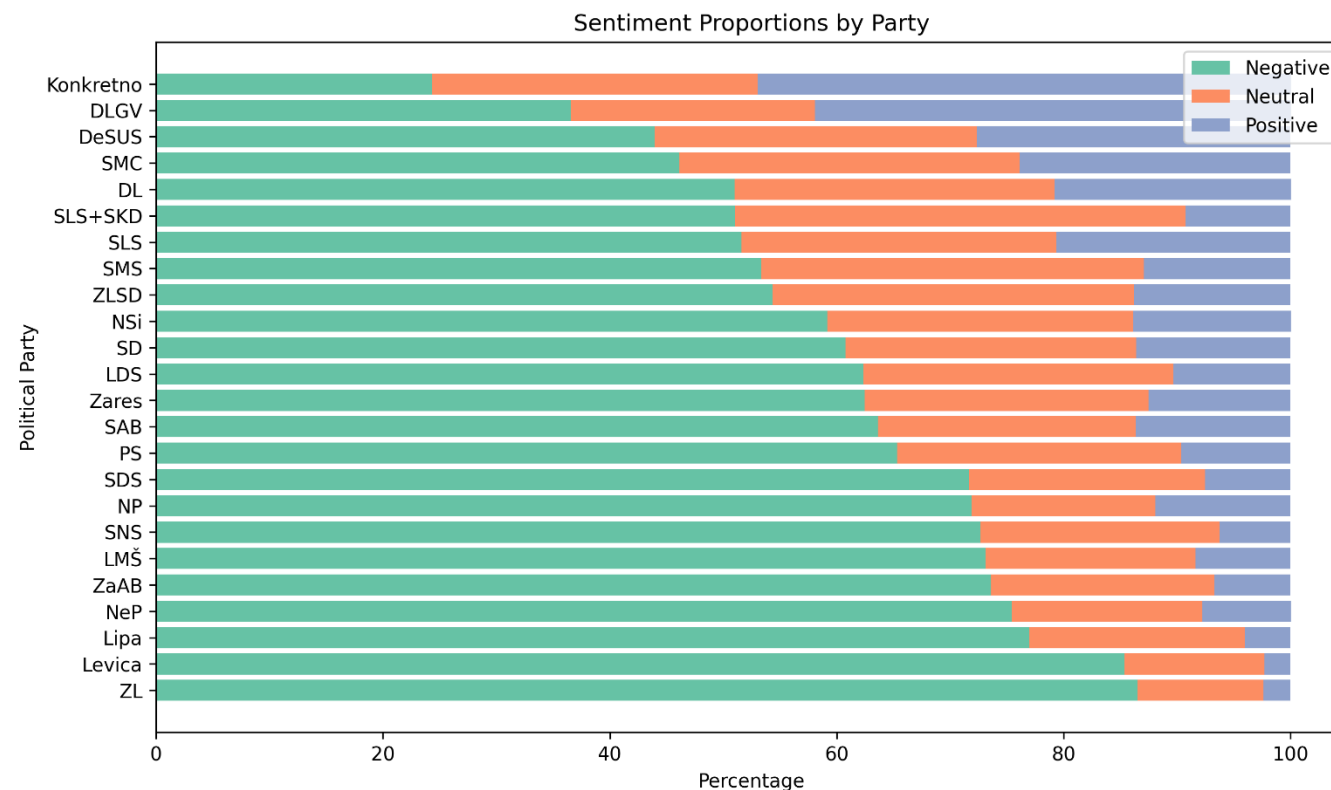


Porazdelitev sentimenta v ParlaMint-SI korpusu z **odstranjenim proceduralnim govorom**
(n=311,347, oznake 3- in 6-stopenjske lestvice)

Party Activity Across Terms (Gantt Chart)



Aktivnost poslanskih skupin oziroma političnih strank v parlamentu med letoma 2000 in 2022, prikazana skozi obdobja njihove prisotnosti v parlamentu ter število izrečenih govorov



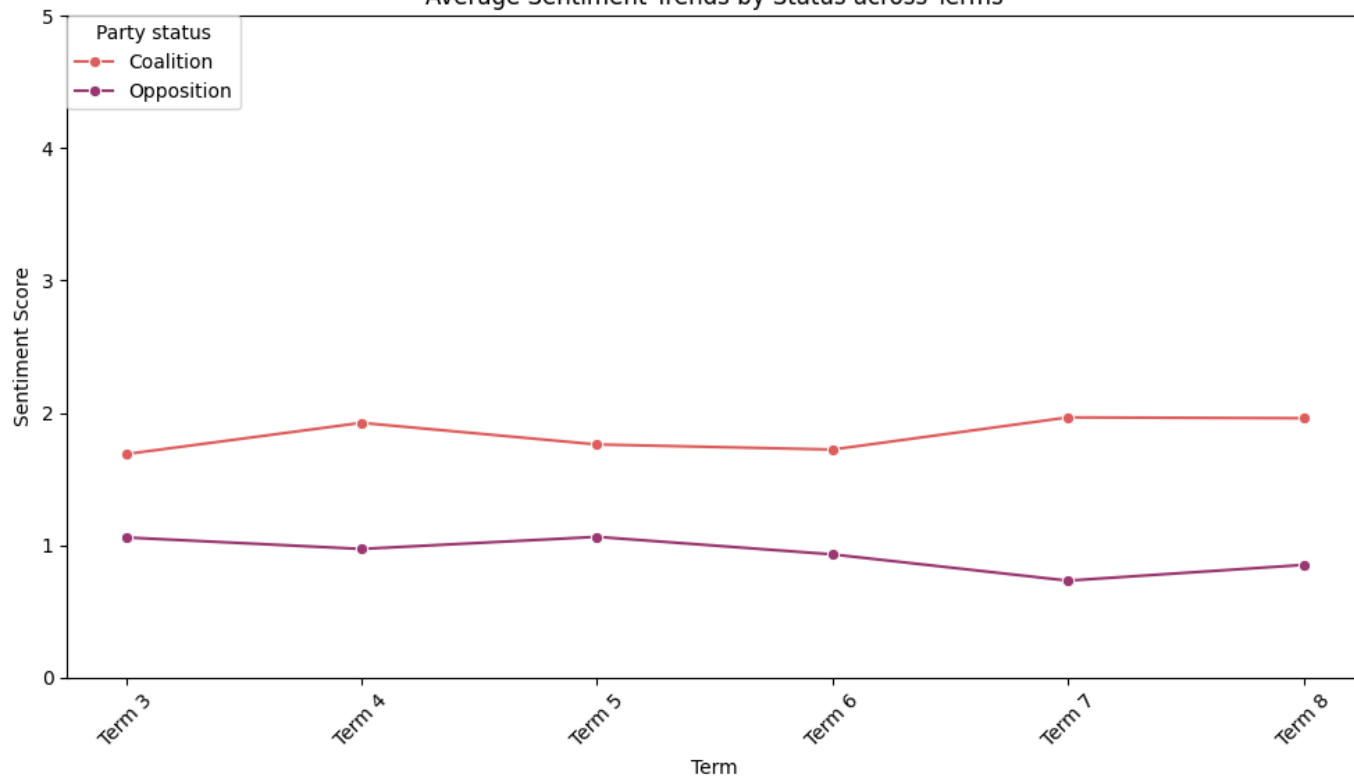
Porazdelitev sentimenta po političnih strankah:

- *Upoštevanje konteksta:*
 - Levica (in ZL) v parlamentu prisotna zgolj kratko obdobje z manjšo količino govorov (cca. 3% vseh govorov v korpusu)
 - SDS prisoten čez celotno obdobje, ki ga pokriva korpus (2000 – 2022) in hkrati kot ena največjih strank prispeva veliko količino govorov (cca. 22% vseh govorov v korpusu)

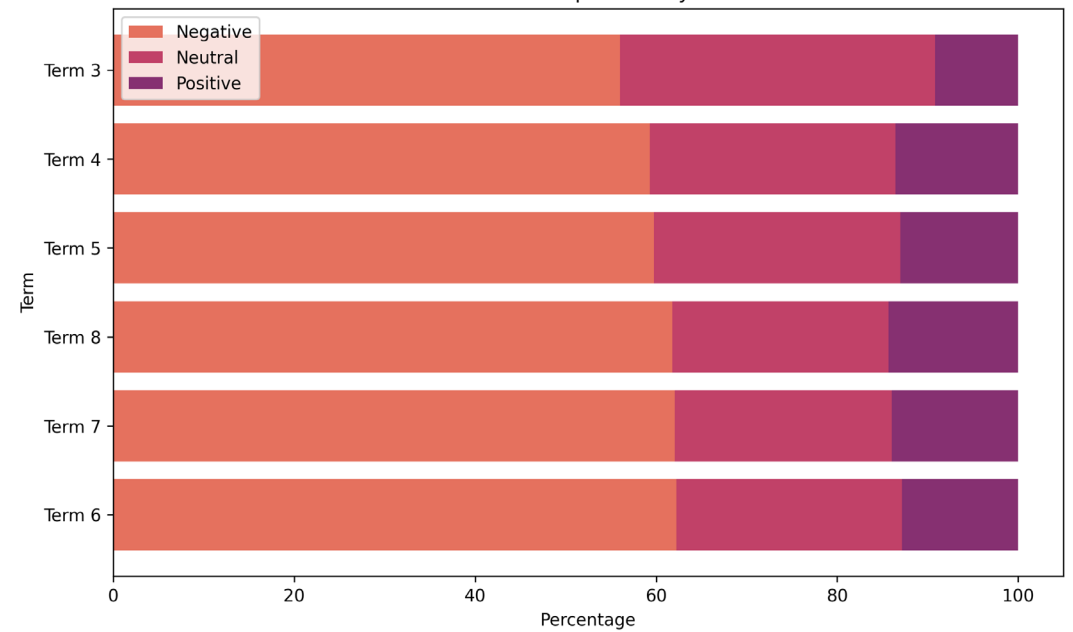
Porazdelitev sentimenta po političnih orientacijah:

- Wikipedia oznake (zatečeno stanje iz leta 2022)
- Podobna porazdelitev kot po strankah (kar ni presenetljivo, mapiranje 1:1 z Wikipedia oznakami):
 - Npr. SDS -> Right
 - Levica, ZL -> Left
 - ...

Average Sentiment Trends by Status across Terms



Sentiment Proportions by Term



Zaključki

- Področje digitalne humanistike in jezikovnih tehnologij je obsežno z različnimi vrstami storitev in ponudb.
- Raziskovalne infrastrukture omogočajo povezovanje skupnosti, delo z raziskovalnimi podatki in ponujajo podporo pri raziskovalnem delu.
- Parlamentarni podatki so (lahko) kompleksni, analize le teh pa pripomorejo k boljšemu razumevanju strategij, ki jih uporabljajo poslanci oziroma druge parlamentarne entitete.
- Bogatenje parlamentarnih podatkov z dodatnimi različnim podatki je zahtevno in zahteva razumevanje pojavov, ki jih želimo zajeti ter upoštevanje *konteksta*.
- Novi metapodatki in anotacije omogočajo nove, bolj poglobljene analize.