

***Corpus of Italian and Serbian Public
Discourse (CISPD):***
koncept, metodologija i perspektive
novog dvojezičnog resursa

dr Jovana Bazić
Univerzitet u Kragujevcu
Filološko-umetnički fakultet
Centar za naučno-istraživački rad



**Uvod: Trostruki resursni jaz u
dvojezičnim istraživanjima**

1

**Teorijski okvir: Od introspekcije do
empirijske i multimodalne lingvistike**

2

**Metodologija: Epistemološki dizajn i
arhitektura uporednog korpusa**

3

**Model anotacije: Prototip troetažnog
označavanja i rešenja**

4

**Diskusija: Evaluacija resursa i šira
primena u NLP tehnologijama**

5

**Zaključak: Perspektive razvoja i
strateški pravci faze CISP 2.0**

6

Reference

7



Struktura izlaganja

01

Istraživački problem od koga polazimo tiče se ozbiljnog nedostatka dvojezičnih resursa za jezički par italijanski–srpski u domenu savremenog javnog govora. Taj problem možemo sažeti kroz tri ključna aspekta:

02

LABLITA, BADIP, VoLIP KIParla, Caffi(2007) ili srWaC, CLASSLA V2, ParCoJour. Jednojezični. Usmereni su primarno na leksikografiju i morfosintaksu, bez dvojezične simetrije i pragmatičkih metapodataka.

03

Ograničenja književnog paralelnog korpusa: Za naš jezički par izrađen je izuzetan književni paralelni korpus *SerbltaCor3* (Moderc et al., 2023). ITALSERB (Ceković2026): Resurs za proučavanje usvajanja italijanskog (L2).

04

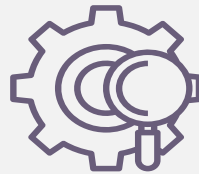
CISPD uobličen u okviru naše doktorske disertacije (Bazić, 2025), ali je od samog početka koncipiran kao trajan, metodološki ujednačen, otvoren i višenamenski resurs namenjen široj akademskoj zajednici.

Od introspekcije do empirijske i multimodalne lingvistike

- (NLP)
- Empirijska utemeljenost, reprezentativnost i proverljivost.
- Korpusna lingvistika više nije samo pomoćna metoda, već primarni epistemološki okvir za istraživanje žive jezičke realnosti.

Fokus na javni diskurs (Silaški i dr. 2009)

Istraživanje medijskog, političkog, naučnog i društvenog diskursa kao poligon za brzu verifikaciju jezičkih inovacija i suptilnih pragmatičkih strategija.



Teorijski okvir

Multimodalna teorija (Allwood, 2008)

Komunikacija je primarno multimodalna.
Neraskidiva veza između **leksičkog izbora**, **prozodijskih parametara** (intonacija, ritam), pauza/oklevanja i **neverbalnih** modaliteta (gestikulacija, mimika, pogled).

Dizajn i arhitektura uporednog korpusa (CISPD)

[PARALELNI KORPUS] (Izvor + Prevod)

✗ Efekat "Translationese"-a (prevodilački jezik)

✗ Narušavanje autentične frekvencije ublaživača

✗ Preslikavanje sintakse izvornog jezika

[UPOREDNI / CISPD] (Izvorni IT + Izvorni SR)

✓ Čista pragmatička norma izvornih govornika

✓ Identičan temporalni i tematski okvir (2023–2024)

✓ Maksimalna metodološka simetrija (342.632 reči)

Emiteri i platforme:

Italijanski izvor: Rai1/RaiPlay, Sky, Mediaset (Canale 5, Infinity), Nove, La7, te podcasti *Passa dal BSMT i Il Bazar Atomico*.

Srpski izvor: RTS, Nova S, N1, Prva, Insajder, Brainz TV, Hram, te podcasti *Agelast i Lobotomija*.

Čitanost pisanog potkorporusa

Italijanski mediji (Audiweb, Audipress):

Corriere della Sera:

Mesečna čitanost 28.748.157

La Repubblica:

Mesečna čitanost 27.732.058
korisnika.

Il Sole 24 Ore:

Mesečna čitanost 8.340.996
korisnika.

Il Messaggero:

Mesečna čitanost 20.968.213
korisnika.

Srpski mediji (Gemius, Mediana TGI)

Danas: 2.123.856 čitalaca (Gemius)
/ 1,3% dnevno (Mediana).

Politika: 853.920 čitalaca (Gemius)
/ 0,90% dnevno (Mediana).

Blic: 3.519.888 čitalaca (Gemius) /
4,63% dnevno (Mediana).

NIN: Preko 1 miliona korisnika u
digitalnom formatu / 0,95% dnevno
(Mediana).

Gledanost govornog potkorpusa

Italijanski emiteri

Rai: 7.796.602 gledalaca
Mediaset (Canale 5 / Infinity):
7.401.714 gledalaca (uz 11,8 miliona poseta platformi Infinity).
Sky: 1.599.800 gledalaca.
La7: 1.220.577 gledalaca.
Nove: 786.389 gledalaca.
Podkasti:
Passa dal BSMT (15. mesto) i
Il Bazar Atomico (69. mesto).

Srpski emiteri

RTS: Oko 2.627.000 gledalaca (mart 2024).
PrvaTV: 1.874.000 gledalaca.
N1/Nova S: Visoka gledanost potvrđena kroz izveštaje agencije TV Beat
YouTube metrike za podkaste i emisije:
Podkast Agelast sa preko 57 miliona pregleda, **Lobotomija** sa 17 miliona pregleda
Insajder, Brainz TV i Hram.

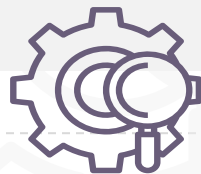
Veličina i struktura korpusa (CISPD) (Faza 1.0): 342.632 reči

Jezik	Novine	Broj reči	Procenat
Italijanski jezik	Corriere della Sera	60991	52,35%
Italijanski jezik	La Repubblica	26595	22,83%
Italijanski jezik	Il Sole 24 Ore	15402	13,22%
Italijanski jezik	Il Messaggero	13515	11,60%
Italijanski jezik	Ukupno	116503	100,00%
Srpski jezik	Danas	60256	51,89%
Srpski jezik	Politika	20480	17,64%
Srpski jezik	Blic	18395	15,84%
Srpski jezik	NIN	16991	14,63%
Srpski jezik	Ukupno	116122	100,00%

Govorni potkorpus (Ukupno: 110.007 reči / 12 sati):

Italijanski : 55.134 reči (6 sati transkribovanog materijala sa TV kuća i podkasta).

Srpski : 54.873 reči (6 sati transkribovanog materijala sa TV kuća i podkasta).



Izvor	Emisija	Trajanje (minuti)	Izvor	Emisija	Trajanje (minuti)
Rai 1	Telegiornale	30	Rts	Naučni portal	10
Rai 1	Uno mattina	35 (10+9+5+3+8)	Rts	Intervju	20
Rai Play	Porta a Porta	10	Rts	Oko magazin	15
Rai Play	Tango	10	Rts	Vesti	5
Rai Play	Geo	10	N1	Izborna debata	20
Rai Play	Report	6	N1	Novi dan	10+10
Rai Play	La Volta Buona	10	NovaS	Dnevnik	30
Sky	Vite	20	NovaS	Utisak nedelje	20
Sky	Health Progress	15	NovaS	Pokreni se	10
Sky	SkyTG 24 News	20	Prva	150 minuta	15
Mediaset Infinity	Quarta Repubblica	30	Prva	Jutro	5+5
Nove	Che tempo che fa	15	Insajder	Insajder Debata	20+15
La7	Attualità	10	Insajder	Agape	20+20
La7	L'aria che tira	10	Insajder	Tabu	15
La7	Piazza Pulita	4+3	BrainzTV	CV naučnika	10
La7	Artbox	16	BrainzTV	Vitelevizija	15
La7teche	Intanto	15	BrainzTV	Brainz Art	10
Passa dal Basement	Episodio 75	20	Hram	Vrlinoslov	15
Il Bazar Atomico	Episodio 95	20	Agelast Lobotomija	215/217 035	20+20 20
UKUPNO	Italijanski jezik	360 min=6 sati	UKUPNO	Srpski jezik	360 min=6 sati

Matrica kriterijuma i metodološki izazovi transkripcije (CISPD)

A. *Eksterni kriterijumi:*

- **Institutional Impact & Market Share:** Proverena renomiranost i široki doseg medijskih redakcija i emitera u oba jezika.
- **Synchronous Temporal Window:** Striktno omeđen period (2023 – 1. jul 2024) radi reakcije na identičan globalni tematski okvir.
- **Multi-domain Coverage:** Ravnopravna pokrivenost ključnih domena (politika, ekonomija, društvo, kultura, nauka, zdravlje, religija).

B. *Interni (lingvističko-tehnički) kriterijumi:*

- Sistematsko uzorkovanje početnih sekvenci (10–30. minut) radi eliminacije pristrasnosti.
- Primarna AI transkripcija (Whisper) uz obaveznu 100% manuelnu verifikaciju (Post-Editing).
- Ručno uklanjanje vanjezičkog šuma (reklame, impresumi, komentari), za razliku od standardnih web korpusa.

Dubinska razrada transkripcije i segmentacije:

- Tretman diskursnih markera: Čuvanje i verifikacija "suvišnih" reči tipa ***diciamo, ecco*** (IT) ili ***znači, recimo*** (SR)
- Samoispravke i false starts: Beleženje prekida u mislima i preoblikovanja rečenica kao indikatora unutrašnje pragmatičke korekcije.
- Kriterijum segmentacije: Kombinovanje sintaksičke zaokruženosti klauze i prosodijskih pauza dužih od 0,5 sekundi.

Normalizovana frekvencija (NF)

Normalizovane frekvencije na 1.000 reči:
"NF" = $(f/N) \times 1.000$ (gde je f apsolutni broj pojavljivanja, a N ukupan broj reči u potkorpusu).
Eksperimentalni primer iz istraživanja:

Italijanski govor: 120 pojavljivanja ublaživača *diciamo* na 55.134 reči $\rightarrow$ "NF" $\approx 2,18$ na 1.000 reči.

Srpski govor: 115 pojavljivanja ublaživača *recimo* na 54.873 reči $\rightarrow$ "NF" $\approx 2,10$ na 1.000 reči.

Strategija ekspanzije i model buduće anotacije (CISPD 1.0 $\rightarrow$ CISPD 2.0)

Strategija ekspanzije
(CISPD 2.0):

U toku je sakupljanje,
obrada i transkripcija
građe za period
2025–2026. godine,
sa krajnjim ciljem
dostizanja
milionskog korpusa
(1.000.000 reči).

Potencijal obrade govornog korpusa prema stepenu spontanosti

➤ Pripremljeni / Monološki govor

(TV Dnevници, vesti):

- Oslonac na pisanu podlogu (teleprompter).
- Najniža frekvencija ublaživača i pragmatičkih markera kao referentna vrednost.

Polupripremljeni / Dijaloški govor

(Debate, intervjui):

Koncipiranje rečenica u hodu pod pritiskom sagovornika i javnog konteksta.
Fokus na funkciju zaštite lica (face-saving).

➤ Spontani /

Otvoreni interakcijski govor

- (Podkasti: Agelast, BSMT):
- Slobodni formati bez striktnog vremenskog pritiska.
- Idealan teren za najvišu frekvenciju leksičkih ublaživača, samoispravki i diskursnih markera.

Troslojna matrica za buduću anotaciju korpusa

Sloj 1: Morfosintaksički nivo



POS tagovanje (vrste reči), sintaksička forma i identifikacija leksičkih markera.

Pragmatičko-funkcionalni nivo



Dodeljivanje etiketa za pragmatičke strategije (ograđivanje, zaštitu obraza i aproksimaciju).

Sloj 3: Multimodalni nivo (za govor)



Ekstrakcija neverbalnih signala: gestovi i kinezika, proksemički pomaci i prosodijske pauze duže od 0,5 sekundi.



Tehnička struktura i usklađenost sa FAIR principima

F

Findable (**Dostupnost**):

Svaki fajl treba da poseduje strukturisanu nomenklaturu sa metapodacima ([Jezik]_[Modalitet]_[Medij]_[Datum]_[ID].txt), a korpus će biti registrovan na repozitorijumima uz dodelu jedinstvenog DOI broja radi lakog lociranja.

A

Accessible (**Pristupačnost**):

Podaci pohranjeni u otvorenim tekstualnim formatima (.txt, .csv, .tsv) i dostupni javnosti i naučnoj zajednici putem otvorenih repozitorijuma (DARIAH-RS / CLARIN).

I

Interoperable (**Usklađenost**):

Podaci moraju biti u stanju da se „razumeju“ i integrišu sa drugim podacima, alatima ili programskim jezicima. Korišćenje univerzalnog UTF-8 kodiranja i izvoz u standardne formate (ELAN, UAM Corpus Tool, i AntConc) i programskim jezicima (Pajton i R-a).

R

Reusable (**Ponovna upotreba**):

Visok kvalitet obrade, jasna troetažna anotacija i očuvani dijakritici čine korpus potpuno pouzdanom polaznom tačkom za buduća istraživanja i nove naučne analize.

Anotacija Sloja 1 (Morfosintaksa) i Sloja 2 (Pragmatička uloga)
u tekstu (alat UAM Corpus Tool / XML format).

01

PRIMER 1 (Italijanski pisani diskurs – Epistemički štit):

Rečenica: „Le nuove misure economiche potrebbero avere un impatto limitato sul PIL.” (CdS, 14.5.2024)

Anotacioni kod: <target_marker lemma="potere"
pos="VERB_MODAL" mood="conditional"
pragmatic_category="HEDGING"
sub_category="EPISTEMIC_SHIELD">potrebbero</target_
marker>

Metodološko obrazloženje:

Kondicional glagola *potere* označava da novinar svesno distancira sopstvenu autorsku odgovornost od tačnosti ekonomske prognoze.



Anotacija Sloja 1 (Morfosintaksa) i Sloja 2 (Pragmatička uloga) u tekstu (alat UAM Corpus Tool / XML format).

02

PRIMER 2 (Srpski pisani diskurs — Aproksimacija):

Rečenica: „Ova odluka predstavlja gotovo nesmotren potez koji bi mogao izazvati reakciju.”(Danas,13.8.2024)

Anotacioni kod: <target_marker lemma="gotovo" pos="ADV" pragmatic_category="HEDGING" sub_category="APPROXIMATOR">gotovo</target_marker>

Metodološko obrazloženje: Prilog *gotovo* ublažava oštru vrednosnu kvalifikaciju („nesmotren potez”) i smanjuje potencijalni sukob sa političkim akterom.



Simulacija multimodalne anotacije u govornom potkorpusu

Okvir obrade: Sinhronizacija tekstualnog transkripta sa video-zapisom putem milisekundskih vremenskih kodova (alat ELAN).

PRIMER 3 (Italijanski govor — Porta a Porta):

Segment: „Noi dobbiamo, diciamo, riorganizzare le priorità del Paese.”
([00:12:15.2 – 00:12:18.8])

Multimodalna sinteza: Jezička jedinica *diciamo* na leksičkom nivou deluje kao poštapalica, ali sinhronizovani video-signal (otklon pogleda u stranu i sleganje ramenima u trajanju od 0,9s) potvrđuje svesno ublažavanje imperativa.



Simulacija multimodalne anotacije u govornom potkorpusu

PRIMER 4 (Srpski govor — *Utisak nedelje*):

Segment: „To je potpun promašaj... **mislim, možda** nije najsrećnije rešenje.”

([00:41:03.5 – 00:41:07.1])

Multimodalna sinteza: Nakon oštre ocene praćene oštrim gestom, govornik pravi pauzu od 0,6s i uz gest otvorenih dlanova (izvinjenje) uvodi korektore *mislim* i *možda* radi rebalansa javnog etosa.



(Italijanski govorni diskurs – Diskursni marker + Kinezika):

Vremenski kod (Time-Code)	Sloj 1: Transkript (Tekst)	Sloj 2: Morfosintaksa	Sloj 3: Pragmatička funkcija	Sloj 4: Multimodalni signal (iz video-zapisa)
00:12:15.200 - 00:12:16.100	Noi dobbiamo...	Pronoun + Modal Verb	Direktna obaveza	Direktan kontakt očima sa voditeljem
00:12:16.100 - 00:12:17.000	diciamo	Discourse Marker	Ograđivanje / Zaštita obraza	Otklon pogleda u stranu + blago podizanje ramena
00:12:17.000 - 00:12:18.800	riorganizzare...	Verb Infinitive	Glavna propozicija	Povratak u neutralan stav, spuštanje ruku

(Srpski govorni diskurs – Samoispravka + Prozodijska pauza + Gestikulacija):

Vremenski kod (Time-Code)	Sloj 1: Transkript (Tekst)	Sloj 2: Morfosintaksa	Sloj 3: Pragmatička funkcija	Sloj 4: Multimodalni signal (iz video-zapisa)
00:41:03.500 - 00:41:04.200	To je potpuni promasaj...	Asertivna procena	Čin ugrožavanja obraza	Oštar gest rukom nadole (kategoričnost)
00:41:04.200 - 00:41:04.800	[PAUZA 0,6s]	Prozodijska pauza	Unutrašnja preformulacija	Prekid gestikulacije, mikro-pauza
00:41:04.800 - 00:41:07.100	mislim, možda...	DM + Epistemički prilog	Samoispravka / Ublažavanje	Širenje dlanova nagore (gest otvorenih dlanova / izvinjenje)

Šira primena korpusa u jezičkim tehnologijama i NLP-u

Doprinos razvoju i finom podešavanju NLP alata:

**Ručno verifikovani i
anotirani dvojezični korpusi
kao ključni resurs za fino
podešavanje NLP alata.**

Detekcija stava i subjektivnosti

**Prevazilaženje slabosti savremenih
algoritama kod ublaženih tvrdnji, ironije i
indirektnih iskaza.**

**Treniranje modela da prepoznaju sumnju,
nesigurnost i uzdržanost govornika.**



Treniranje kontekstualnih vektorskih prikaza

**Unapređenje razumevanja prirodnog govora od strane
veštačke inteligencije kroz primenu na *BERT* i
mBERT varijante za srpski i italijanski jezik.**

Sinteza ostvarenih rezultata



Simetrična dvojezična baza (CISPD 1.0):
Obuhvata 342.632 reči i predstavlja prvi metodološki kontrolisan, podjednako zastupljen resurs za italijanski i srpski javni diskurs.
Obezbeđena je potpuna srazmernost građe po obimu i strukturi za oba jezika.

Standardizovana transkripcija govornog korpusa:
Obradeno **12 sati** audio-video materijala (po 6 sati za svaki jezik).
Očuvani pragmatički elementi: diskursni markeri, poštapalice, samoispravke, pauze.

Matematička komparabilnost
(NF_{1000}):

Normalizovana frekvencija na
1.000 reči:

$$NF_{1000} = \left(\frac{f_{aps}}{N} \right) \times 1000$$

Eliminisana statistička pristrasnost
između pisanih i govornih izvora.

Ograničenja trenutne faze istraživanja



➤ **Izazovi i opseg govornog potkorpusa:**

Uprkos izuzetnoj zahtevnosti i vremenskoj iscrpnosti ručne verifikacije 12 sati govornog materijala, govorni deo trenutno obuhvata manji udeo u ukupnoj građi u odnosu na pisani potkorpus.

Potreba za daljom ekspanzijom i balansiranjem medijskih kanala u narednoj fazi projekta.

➤ **Metodološki okvir prve faze (Odsustvo automatskog tagovanja):**

Baza je u ovoj fazi primarno verifikovana na makro-nivou: struktura teksta, usklađenost audio-video zapisa i precizno definisani vremenski kodovi.

Svesno odsustvo automatskog mašinskog tagovanja radi apsolutne kontrole tačnosti i očuvanja autentičnosti sirovih podataka.

➤ **Fokus na sledeći korak i preduslovi za nadogradnju:**

Dubinska pragmatička anotacija (strategije ograđivanja) i složena multimodalna obrada ostavljeni su za narednu fazu (CISPD 2.0).

Postignuti temeljni nivo obezbeđuje stabilnu osnovu za sve buduće nadogradnje i analize.

Perspektive i naredni koraci (CISPD 2.0)

Strateški pravci razvoja u fazi 2:

1. Širenje i ekspanzija korpusa:

Unos i obrada nove građe za period 2025–2026. godine.

Cilj: dostizanje obima od **1.000.000 reči** uz dodatnu ekspanziju govornog potkorpusa i ravnotežu medijskih kanala.

2. Implementacija troetažne anotacije:

UAM Corpus Tool: Morfosintaksički i pragmatički sloj za analizu strategija ograđivanja (*hedging*).

ELAN okruženje: Višekanalna multimodalna anotacija (povezivanje teksta, prosodijskih parametara i neverbalnih signala).

Standardizacija i FAIR publikacija:

Otvoreni pristup (*Open Access*) i trajno deponovanje resursa. Integracija u evropske mreže jezičkih resursa (*DARIAH-RS* / *CLARIN*) radi primene u lingvistici i veštačkoj inteligenciji (NLP / NMT).

Zaključak i širi naučni doprinos (CISPD 2.0)

Multidisciplinarna primena CISPD resursa:

Italijanistika i kontrastivna pragmatika: Precizno proučavanje ublažavanja, diskursnih strategija i govornih nijansi u uporednoj perspektivi.

Primenjena lingvistika i analiza diskursa:

Empirijsko upoređivanje javnog govora i medijskih formata u italijanskom i srpskom jeziku.

Jezičke tehnologije i NLP:

Unapređenje mašinskog prevođenja (*NMT*) za specifične pragmatičke ublaživače.

Razvoj i fino podešavanje dvojezičnih jezičkih modela i veštačke inteligencije.

Temelj za potonje studije:
Uspostavljeni stabilni metodološki temelji za objektivno i multimodalno izučavanje javnog diskursa.

Reference



Allwood, J. [2008]. Multimodal Corpora, in Lüdelling & Kytö [2008], chapter 12, pp. 207–225.



Bazić, J. (2025). Ograđivanje u javnom diskursu: kontrastivna analiza italijanskog i srpskog jezika (Doktorska disertacija). Univerzitet u Kragujevcu, Filološko-umetnički fakultet.



Caffi, C. (2007). Mitigation. Elsevier.



Ceković, N. (2026). *The ItalSerb Corpus: L2 Italian spoken by Serbian students*. Filološki fakultet Univerziteta u Beogradu.



Moderc, S. (2025). SerbltaCor3: Paralelni korpus srpskog i italijanskog jezika. Filološki fakultet.



Moderc, S. et al. (2023). Corpora in Contrastive Studies: Italian and Serbian. Carocci.



Philip, G. (2009). Arriving at equivalence: Making a case for comparable general reference corpora in translation studies. U A. Beeby, P. Rodríguez Inés, & P. Sánchez-Gijón (Ur.), *Corpus use and translating* (str. 59–74). John Benjamins Publishing Company. <https://doi.org/10.1075/btl.82.07phi>



Silaški, N., Đurović, T. & Radić-Bojanić, B. (2009). Javni diskurs Srbije: kognitivističko-kritička studija. Beograd: Centar za izdavačku delatnost Ekonomskog fakulteta.



Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. John Benjamins.



Utvić, M. V. (2013). Izgradnja referentnog korpusa savremenog srpskog jezika (Doktorska disertacija). Univerzitet u Beogradu, Filološki fakultet.



Wasserscheidt, P. (2023). PDRS - Parallel Discourse Repository Serbia: Concept and Methodology. Humboldt University Berlin.