Универзитет у Београду
Филолошки факултет

ITALSERB

govorni korpus srbofonih studenata italijanskog kao stranog jezika

Nevena Ceković

Katedra za italijanistiku

03.05.2023

U izlaganju se prezentuju **metodološke pretpostavke** za nastanak ITALSERB korpusa i **modaliteti prikupljanja podataka** (odabir tehnika za elicitaciju jezičkih podataka i procedura za ortografsku transkripciju tekstova, beleženje sociolingvističkih podataka), uz poseban osvrt na **aplikativne aspekte** rada na konstituisanju korpusa. U tom smislu, namera nam je i da ukratko predočimo rezultate **nekolicine naučnih istraživanja** sprovedenih na korpusnim podacima (*corpus-based*), kako bismo skrenuli pažnju na **pedagoške implikacije** i korisnost učeničkih korpusa za potrebe različitih aktera naučno-nastavnog procesa, odnosno ukazali na **značaj** ovakvih tipova korpusa u oblasti ne samo korpusne već i primenjene lingvistike, a naročito teorije usvajanja drugog jezika i glotodidaktike.

03.05.2023

Универзитет у Београду
Филолошки факултет

1. OSNOVNI POJMOVI

- **Korpusi drugog jezika (KJ2)** ili **učenički korpusi** (eng. *L2 corpora, learner corpora;* it. *corpora di L2, corpora di apprendenti*): elektronska baza pisane i/ili transkribovane usmene produkcije nematernjih govornika (NNS) (Granger 2003, Granger *et al.* 2007)
- Već par decenija su dragocen teorijsko-metodološki instrument za proučavanje procesa usvajanja drugog ili stranog jezika kod nematernjih govornika
- **Faze u kreiranju KJ2:** planiranje (*corpus design:* obim, uzorak, dinamika, naziv, …), prikupljanje sociolingvističkih podataka, elicitacija, snimanje, arhiviranje, ... revizija
- **Ortografska transkripcija**: sistematsko prezentovanje govora u pisanom obliku, preliminarna i nezaobilazna operacija u cilju formiranja upotrebljivog korpusa

(Gibbon *et al*. 1997: 79, nav. u Savy 2005: 2)

- **Anotacija**: obogaćivanje transkripta (para-/van-)jezičkim informacijama
- **Etiketiranje** ili **tagiranje** (*Part-of-Speech tagging*): obeležavanje reči prema gramatičkim kategorijama, rodu, broju, morfološkim i fonološkim karakteristikama, …

(Dobrić 2009: 51)

2. FAZE NASTANKA ITALSERB KJ2

- Jedan takav govorni (i ne samo) korpus italijanskog kao stranog jezika pod nazivom ITALSERB kreiran je **2010. na Katedri za italijanski jezik i književnost** Filološkog fakulteta Univerziteta u Beogradu (Ceković-Rakonjac 2012, 2013, Ceković 2016)
- Osim govornog sadrži i **pisani potkorpus** (uključeno više veština: usmena produkcija, pisana produkcija, usmena recepcija, pisana recepcija, poznavanje gramatičkih struktura)
- Naziv: ITALSERB ← it. ***ITALiano dei SERBofoni***, Corpus SERBo di ITALiano L2 – italijanski srbofonih govornika (prethodno ESNAKIT ili ITALBEG)
- Projekat (2010) realizovan u saradnji s Univerzitetom za strance u Sijeni i Univerzitetom u Ljubljani, s ciljem evaluacije opšteg nivoa kompetencije u italijanskom kao J2 studenata (I-IV godine) beogradske Italijanistike
- Akteri: **nastavnici i saradnici, studenti** (OAS, MAS i DAS)

- **Govorni KJ2 ITALSERB**:

- govorni subkorpus deo šireg (i pisanog) korpusa
- 25 sati audio snimaka govorne produkcije
- ortografski transkribovan materijal, očekivani obim oko 200.000 tokena
- period snimanja, 4 akademske godine: 2010/2011-2013/2014
- 170 različitih studenata Italijanistike (I-IV godina OAS)

- **Govorni subkorpusi ITALSERB:**

- transverzalni: A2 (I god.), B1 (II god.), B2 (III god.), C1 (IV god.)
- longitudinalni: ista generacija od A2 (I god.) do C1 (IV god.)
- još barem 4 subkorpusa, prema nivoima ZEO: A2, B1, B2, C1
- parcijalno (3,5h, nivo B1-B2) dostupno u disertaciji Ceković 2016

✓2010	11 h	140 informanata (I-IV god.)	100.000 tokena
✓2011	17 h	160 informanata (II-IV god.)	140.000 tokena
✓2012	23 h	170 informanata (III, IV god.)	180.000 tokena
✓2013	25 h	170 informanata (IV god.)	200.000 tokena

- N.B. 170 **različitih** informanata, jer su neki informanti više puta snimani, tj. longitudinalno praćeni

Sociolingvistički upitnik

- Upitnik = **odobrenje** od ispitanika za snimanje i obradu podataka
- Dobijeni podaci ukazuju na:

- ženski pol

- 19-25 god.

- J1 srpski

- razni J3+

- uče J2 1-6 god.

- kratki boravci u Italiji

Elicitacija

- Tipologija zadataka za usmenu produkciju: **certifikacioni testovi CILS** (*Certificazione di Italiano come Lingua Straniera,* Barni *et al*. 2009) Univerziteta u Sijeni (Università per Stranieri di Siena), decembar 2009 (sessione: dicembre 2009)
- Testovi CILS, široko korišćeni u komercijalne svrhe, upotrebljeni su i pri izradi **korpusa LIPS Univerziteta u Sijeni**, najvećeg italijanskog KJ2 (*Lessico dell'italiano parlato da stranieri*, Vedovelli *et al.* 2007)
- Cilj testova: merenje nivoa lingvističke i komunikativne kompetencije učenika italijanskog kao J2 (nivoi: A1, A2, jedan B1, dva B2, tri C1, četiri C2).
- Sačinjeni su od niza zadataka radi evaluacije **5 jezičkih veština**:
  - razumevanje slušanog teksta
  - razumevanje pročitanog teksta
  - analiza komunikacionih struktura
  - pisana produkcija
  - usmena produkcija.

- Tipologiju zadataka za evaluaciju govorne produkcije u okviru *CILS* testa čine:

  1 dijalog s ispitivačem + 1 monolog na zadatu temu

- Teme su deskriptivnog, narativnog, argumentativnog ili ekspozitornog karaktera
- Primeri zadatih tema, od ispitanika se traži da:

- odgovore na lična pitanja i da opišu omiljeni dan u nedelji (A2)
- iskažu lične preferencije u vezi sa raznovrsnom problematikom i da opišu jednu knjigu ili fotografiju (B1)
- govore o svom karakteru i da opišu život u rodnom gradu (B2)
- simuliraju razgovor za posao i da iskažu lične stavove na temu humanitarnog rada ili položaja žene u društvu (C1).

- U dijaloškoj produkciji ispitivač-nastavnik (maternji ili nematernji govornik (N)NS) postavlja studentu pitanja u cilju razvijanja dvosmerne konverzacije (tzv. dvosmerna razmena *F-to-F* s mogućnošću slobodnog uzimanja reči, up. De Mauro *et al*. 1993)
- U monološkoj produkciji (tzv. jednosmernoj razmeni u prisustvu primaoca, up. De Mauro *et al*. 1993) kandidat je pozvan da samostalno govori, bez pomoći nastavnika, na odabranu temu za koju je imao kratko vreme da se pripremi pre izlaganja, s tim da mu u slučaju javljanja poteškoća ispitivač eventualno može priskočiti u pomoć.
- Tip interakcije u korpusu:

  licem u lice *(Face-to-Face)*, **u učionici,**

  elicitirana, poluspontana, asimetrična (na relaciji nastavnik-student),

  dijaloška i monološka produkcija,

  za nivoe: A2 (I god.), B1 (II god.), B2 (III god.), C1 (IV god.)
- Testovi *CILS*, kao i nastava iz predmeta Savremeni italijanski jezik na Katedri za italijanistiku strukturirani su po nivoima jezičke kompetencije definisanim prema Zajedničkom evropskom okviru za žive jezike - ZEO (Council of Europe 2001): A2-C1
- Stoga su ispitanici testirani na nivou kompetencije predviđenom za onu godinu studija koju su u datom trenutku pohađali.

3. ORTOGRAFSKA TRANSKRIPCIJA

- **Norme za anotiranu ortografsku transkripciju** usmene produkcije: prema parametrima **CLIPS korpusa maternjih govornika italijanskog**, oformljenog 1999-2004. na Univerzitetu u Napulju (*Corpora e Lessici di Italiano Parlato e Scritto;* 100h, 300.000 tokena) (Ceković-Rakonjac 2012, Savy 2005, Albano Leoni 2006)
- Norme CLIPS **jedan su od najrazvijenijih notacionih sistema** na italijanskoj teritoriji, obogaćen anotacijom (bitno kod korpusa s audio-materijalom) i nastalim iz potrebe za usklađivanjem sa standardima sličnih evropskih projekata, a uz primenu iskustava u izradi drugih korpusa J1: AVIP i API (Savy 2005: 6; Albano Leoni 2006: 6, 2003).
- Up. sisteme transkripcije kod: LIP (De Mauro *et al.* 1993, De Palo 1993), AVIP-API (Crocco, Savy & Cutugno 2003), C-ORAL-ROM (Cresti & Moneglia 2005) i LIPS (Vedovelli *et al.* 2007).
- Sistem je izrađen u skladu sa preporukama **grupe EAGLES** (*Expert Advisory Group on Language Engineering Standards*), formirane pod okriljem Evropske komisije (Cresti & Panunzi 2013: 166).

Tehnička obrada materijala

- **Arhiviranje snimaka** (prenos podataka s diktafona, *backup*)
- **Formatiranje i imenovanje fajlova i foldera**
- **Sastavljanje zaglavlja**:
  - datum i mesto testiranja tj. snimanja
  - ID aktera: ispitanik, ispitivač, transkriptor
  - tema i tekstualni žanr
- **Strukturiranje fajla**:

  razmena turnusa (kraj jednog i početak drugog?), određivanje MRT (mesta relevantne tranzicije), preklapanja, …

- Svaki trankript u zaglavlju sadrži sledeće podatke (struktura preuzeta iz korpusa LIPS zbog istovetne tipologije testa, Vedovelli et al. 2007):
  - **datum i mesto testiranja**
  - **registarski broj upisa** tj. indeksa
  - **ime i prezime kandidata** (it. *C = candidato)*
  - **jezički nivo** prema ZEO
  - **godina studija**
  - **šifra testa** (podatak o tome po koji put kandidat učestvuje u testiranju)
  - **šifra odgovarajućeg audio snimka**
  - **ime i prezime transkriptora** (it. *trascrittore*)
  - **ime i prezime ispitivača** (it. *E = esaminatore*).

- Slede zatim informacije o komunikativnoj situaciji, tj. temi govorne produkcije i tekstualnom žanru obeleženom na jedan od sledećih načina:

  D (dijalog)

  M (monolog)

  DM ili **MD** (dijalog i monolog jednako prisutni u tekstu)

  Dm (dijalog s delovima monologa)

  Md (monolog s delovima dijaloga)

- U reprezentativnom primeru navedena je 2. po redu komunikativna situacija na temu omiljenog dana u nedelji u formi monologa s delovima dijaloga:

  SE 2 argomento: il giorno della settimana che preferisci (Prova n.2) Md

Anotirana obrada materijala

- **Znaci interpunkcije** (tačka, zarez, razne vrste zagrada)
- **Leksički elementi** (fragmenti reči, prekidi unutar reči, pogrešan start, nepostojeće reči)
- **Neleksički elementi** (pauze, produženi izgovor glasova, *back-channels* ili potvrda prijema/praćenja, neverbalne vokalne pojave, akustičke pojave, nerazumljive reči)
- **Komentari transkriptora** (varijacije u tonu, promena koda, strane reči)

- **Norme primenjene na ITALSERB →**

Simbol (INTERPUNKCIJA)	Primena	Primer
. (tačka)	Sint.-seman. granica	*Sì.*
, (zarez)	Sint.-semant. granica	*Sì, a volte quando...*
? (upitnik)	Upitni iskaz	*E la sera?*
! (uzvičnik)	Uzvični iskaz	*Buongiorno!*
- (crtica)	Alfabet. citiranje	*Un Ci-Di*
‘ (apostrof)	Afereza, elizija	*Un po’*
A..., B..., C... (vel. slovo)	Imena, toponimi, naslovi	*Ho Voglia Di Te*

Simbol (LEKSIČKI ELEMENTI)	Primena	Primer
22 (brojevi)	Slovna transkripcija	*Ho ventidue anni*
+ (plus)	Fragment reči	*obbli+ (obbligatoria)*
_ (donja crta)	Prekid unutar reči	*sa_lario*
/ (kosa crta)	Pogrešan start	*un lavoro… che ha / cerca<aa>*
*(asterisk)	Nepostojeće reči (lapsusi, omaške)	**rifaro*

Simbol (NELEKSIČKI ELEMENTI)	Primena	Primer
<sp>, <lp>	Neispunjena, prazna pauza: kratka <sp> ili duga <lp>	*sì sì sì <lp> <eeh> il mio passatempo preferito. <sp> <eeh> <sp> adesso…*
<eeh>, <ehm>	„Glasna“ pauza ispunjena vokalizacijom ili nazalizacijom	*mi alzo di solito alle <eeh> dieci o dieci e mezza <ehm>*
<vv> (v=vokal), <kk> (k=konsonant)	Produženi izgovor glasova	*Le traduzioni<ii> <ss>sono<oo>*
<eh>, <ah>, <mh>, <ahah>, <mhmh>	Potvrda prijema (*back-channels*)	*C: … era una gita* *E: <mhmh>*
<ah!>, <eh!>	Uzvici	*<tongue click> <eh!>*

Simbol (NELEKSIČKI ELEMENTI)	Primena	Primer
<breath>, <clear-throath>, <cough>, <inspiration>, <laugh>, <tongue-click>	Vokalne, neverbalne pojave (kašalj, smeh, …)	*<laugh> mi piace molto fare lo shopping*
<i.talkers>	Glasovi drugih govornika u pozadini	*CMT <i.talkers>*
<NOISE >	Prateće akustičke pojave	*<NOISE> che tipo sei?*
<unclear>	Nerazumljiva reč ili sekvenca	*mi piace molto <unclear>*
{ }	Preklapanje akustičkih pojava s tekstom	*{<NOISE> bene}*
#	Preklapanje turnusa	*E:…con #<C> chi ci sei andata# C: #<E> <eh> l'ultima# volta…*

Simbol (KOMENTARI TRANSKRIPTORA)	Primena	Primer
[whispering]	Varijacije u tonu glasa	*<ehm> [whispering]*
[dialect]	Reč iz dijalekta	*tanto [dialect]*
[foreign word]	Strana reč	*ukus [foreign word]*

1. Primer: A2

Transkript + snimak

Data esame: maggio 2010
Sede: Beograd
Numero matricola: XXXXXX
Nome candidato: XXXXXX
Livello: A2
Anno di studi: I
Codice esami: 1
Numero cassetta: VN780012
Trascrittore: Vesna Bogojević, Nevena Ceković
Esaminatore: Julijana Vučo

SE 2 argomento: il giorno della settimana che preferisci (Prova n.2) Md

E: il giorno della settimana che preferisci?

C: <eeh> il giorno della settimana che preferisco è<ee> venerdì perché vado a casa <tongue-click> <eeh> sto con la mia famiglia<aa> e<ee> poi <eeh> vedo tutti miei amici e andiamo al bar parliamo de+ / <tongue-click> di<ii> tante cose che<ee> <inspiration> <eeh> hanno passato<oo> quando<oo> <eeh> quando <ss>sono stata a Belgrado<oo> è<ee> molto interessante <tongue-click> <eeh> <eeh> poi vedo il mio ragazzo perché lui vive a Valjevo [foreign word] e<ee> ed io vivo <inspiration> qui a Belgrado non ci vediamo spesso <tongue-click> e ne sono felice quando è venerdì <laugh> <eeh> non so cosa dire

E: #<C> basterà così?#

C: #<E> <unclear>#

Problemi

- **Homogenost zapisa,** jednoznačna interpretacija („više babica …“)
- **Neleksički elementi**: pauze (neispunjene, kratke : duge? -/+ 3 sek.; ispunjene vokalizacijom ili nazalizacijom? “razmišljanja otvorenih i zatvorenih usta”); produženi izgovor? +1sek., + unutar reči (*no<oo>n, ta<aa>nto*, *li<ii>ngue)*; razlučivanje vokalnih, neverbalnih pojava (*<breath>, <clear-throath>, <cough>, <inspiration>, <tongue-click>*)
- **Leksički elementi**: nepostojeće reči (**Vènezia, *praticàno, *pèrche* - uvođenje odgovarajućeg akcenta dok CLIPS predviđa samo *); (**cossa, *univers[z]ità, *cris[z]a)* - beleženje pogrešnog izgovora glasova u reči prema ortografskim pravilima italijanskog ili, u slučaju odstupanja, fonetske transkripcije IPA*)*
- **Ometajuće akustičke, paralelne pojave** (pozadinski govor ili buka):

  rekonstruisati elemente i segmente ili svesti nerekonstruisano na minimum
- **Nepotpune informacije, greške „ispeglati“** (brojevi, „rascepani“ snimci)
- **Utrošak vremena** (najpre 1’ snimka = 60’ transkribovanja, pa 1’ = 5-10’)
- **Tehnička podrška** (skromna oprema, specif. akustička svojstva učionica)

4. PRIMENA KORPUSA U ISTRAŽIVANJIMA

- **Doprinos KORPUSNOJ LINGVISTICI**:

- **konstituisanje KJ2**: primer dobre prakse
- **izrada sistema normi** za ortografsku transkripciju i anotaciju
- **problemi** i načini njihovog prevazilaženja:
  - nužnost preciziranja i prilagođavanja normi (razlika J2 i J1),
  - odsustvo standarda na polju transkripcije, …
- **standardizacija normi**

(Ceković-Rakonjac 2012, 2013; Drljević 2012)

- **Doprinos PRIMENJENOJ LINGVISTICI, USVAJANJU J2** kroz proučavanje pojedinih aspekata komunikativne kompetencije (pragmatičko-diskursne):

- diskursni markeri (DM), progresivni razvoj (broj i obim formi i funkcija) u međujeziku srbofonih studenata (transverzalni subkorpus): A2, B1, B2, C1 (Ceković 2014, 2016) →

- repertoar formi DM, u korelaciji sa nivoima kompetencije A1-C2 (Ceković 2016) →

- **metatekstualna kompetencija** kod srbofonih na nivou B1-B2 (Ceković 2020)

- **Doprinos PRIMENJENOJ LINGVISTICI, GLOTODIDAKTICI** kroz bavljenje didaktičkim implikacijama:
- **Eksperimentalno istraživanja (corpus-based) o efektima eksplicitnog, longitudinalnog inputa s DM, na nivou B1-B2**:

faze inputa, efekti inputa kod eksperimentalne u poređenju s kontrolnom grupom, pre i posle instrukcionog perioda, upotrebljeni didaktički materijali (Ceković 2016)

(Ceković 2016, 2018, Ceković & Radojević 2015, Ceković, & Vučo 2020)

- **Silabus formi DM** A1-C2 (Ceković 2016)
- **Opis opštih odlika DM** u italijanskom kao J2 (Ceković 2016)
- **Funkcionalna taksonomija DM** u didaktičke svrhe (Ceković 2016)
- **Produktivnost izvesnih formi i funkcija DM**, otkrivenih u korpusu, u korelaciji sa subjektivnim stavovima informanata B nivoa, vezano za teškoće u usvajanju (Ceković 2018)
- **Korpusni podaci o diskursnoj fluentnosti** u vidu osnove za proveru kriterijuma pri evaluaciji govorne kompetencije na B1 i B2 nivou (Ceković & Vučo 2020)

- **Didaktička intervencija (*corpus-based*) uz upotrebu glottotehnologija,** ciljane *linguistic awareness activities*: govorna produkcija, metalingvistička autorefleksija, samoevaluacija, introspekcija, jezički dnevnik (Ceković & Radojević 2015)

- složena intervencija, 10 studenata B1-B2 po fazama radi na svojim podacima iz KJ2 (slušanje snimaka, analiza i ispravka transkripata vlastitih usmenih testiranja):

- upotreba KJ2 u nizu korisnih aktivnosti: uz korpus kao osnovu za introspekciju i metalingvističko promišljanje (analizu leksičkih, morfosintaksičkih i diskurzivnih odlika međujezika) student se osposobljava da preuzme odgovornost za vlastita postignuća (poredeći ih u različitim vremenskim periodima, otkrivajući jake/slabe tačke, …).

- **Seminar za nastavnike**:

  „Govorni korpus italijanskog kao stranog jezika u didaktici i istraživanjima” („Evropski dan jezika” 2012. Filološki fakultet i Italijanski institut za kulturu)

- **Primer**: izrada distraktora u testovima s višestrukim izborom:

√ *Quando c’è il tempo, bisogna cercare di approfondire*

gli argomenti che ci interessano. (nivo A2-B1) (Corino & Marello 2009: 282-283)

*A. Quando c’è il tempo, bisogna cercare **di** approfondire*
gli argomenti che ci interessano.

B. *Quando c’è il tempo, bisogna cercare **ad** approfondire*
gli argomenti che ci interessano.

C. *Quando c’è il tempo, bisogna cercare **per** approfondire*
gli argomenti che ci interessano.

D. *Quando c’è il tempo, bisogna cercare **da** approfondire*
gli argomenti che ci interessano.

5. ZNAČAJ

- Značaj KJ2:

✓ **Teorija usvajanja J2**: autentičan, obiman i reprezentativan, kontekstualizovan materijal za analizu; bolje razumevanje procesa usvajanja L2: međujezik, sekvencijalnost, transfer (Conrad & Levelle 2008, Granger 2003, Andorno & Rastelli 2009)

✓ **Didaktika**: uvid u "tipične greške" (kontrastivne, npr.); izrada distraktora u testovima s višestrukim izborom; promovisanje didaktike *corpus-based* (materijali, silabusi, nastavne aktivnosti) (Granger *et al.* 2007; Corino & Marello 2009; Conrad & Levelle 2008)

✓ **Leksikografija i kontrastivna didaktika**: statistička osnova za planiranje efikasnih intervencija

- Značaj transkripcije:

✓**Instrument za obradu KJ2**

✓**Instrument za proučavanje J2**

✓**Doprinos standardizaciji** normi za transkripciju

- Značaj za aktere:
- **Istraživači**:
  - ✓otkrivanje tipičnih obrazaca u međujeziku
  - ✓poređenje učenika s istim ili različitim J1
- **Nastavnici**:
  - ✓planiranje inputa (od usmenog outputa do inputa) (npr. distraktori u testovima)
  - ✓promišljanje o didaktici baziranoj na korpusima
- **Učenici**:
  - ✓metalingvističko (samo)promišljanje
  - ✓introspekcija
  - ✓autodidaktika

Govorni KJ2 (za italijanski):

- **LIPS** (Siena): 100 sati/700.000 tokena
- **Banca dati di italiano L2** (Pavia): 120 sati/600.000 tokena
- **VIP** (Verona, Modena & R. Emilia): 70 sati
- **ADIL2** (Siena): 37 sati/270.000 tokena

→ **ITALSERB (Beograd): [25 sati/200.000 tokena]**

- **INTERITA** (Stokholm): 30 sati/108.000 tokena
- **COCERIT** (Roma Tre): 12 sati
- **A.Ma.Dis** (Madrid): 6 sati/80.000 tokena
- **Corpus Chini** (Pavia): 60.000 tokena
- **Corpus parlato di italiano L2** (Perugia): 28.000 tokena
- **Corpus Rosi** (Pisa): 20.000 tokena [do 2016]

- **KJ2 velikih dimenzija**

  Cfr. „*A corpus of 200,000 words is big in the SLA field where researchers usually rely on much smaller samples*” (Granger 2003: 465)

- **prvi (italijanski) KJ2 kod nas**
- **jedan od retkih longitudinalnih KJ2** (uz: LIPS)
- **jedan od retkih KJ2 sa srbofonim informantima** (uz: LIPS, ADIL2)
- **jedan od retkih KJ2 u kontekstu učenja italijanskog kao stranog jezika**

  potencijalno najvećih dimenzija (uz: INTERITA, A.Ma.Dis)

- **jedan od retkih KJ2 u akademskom kontekstu**

  (uz: INTERITA, Corpus Chini, Corpus parlato di italiano L2, Corpus Rosi)

Aktuelna faza

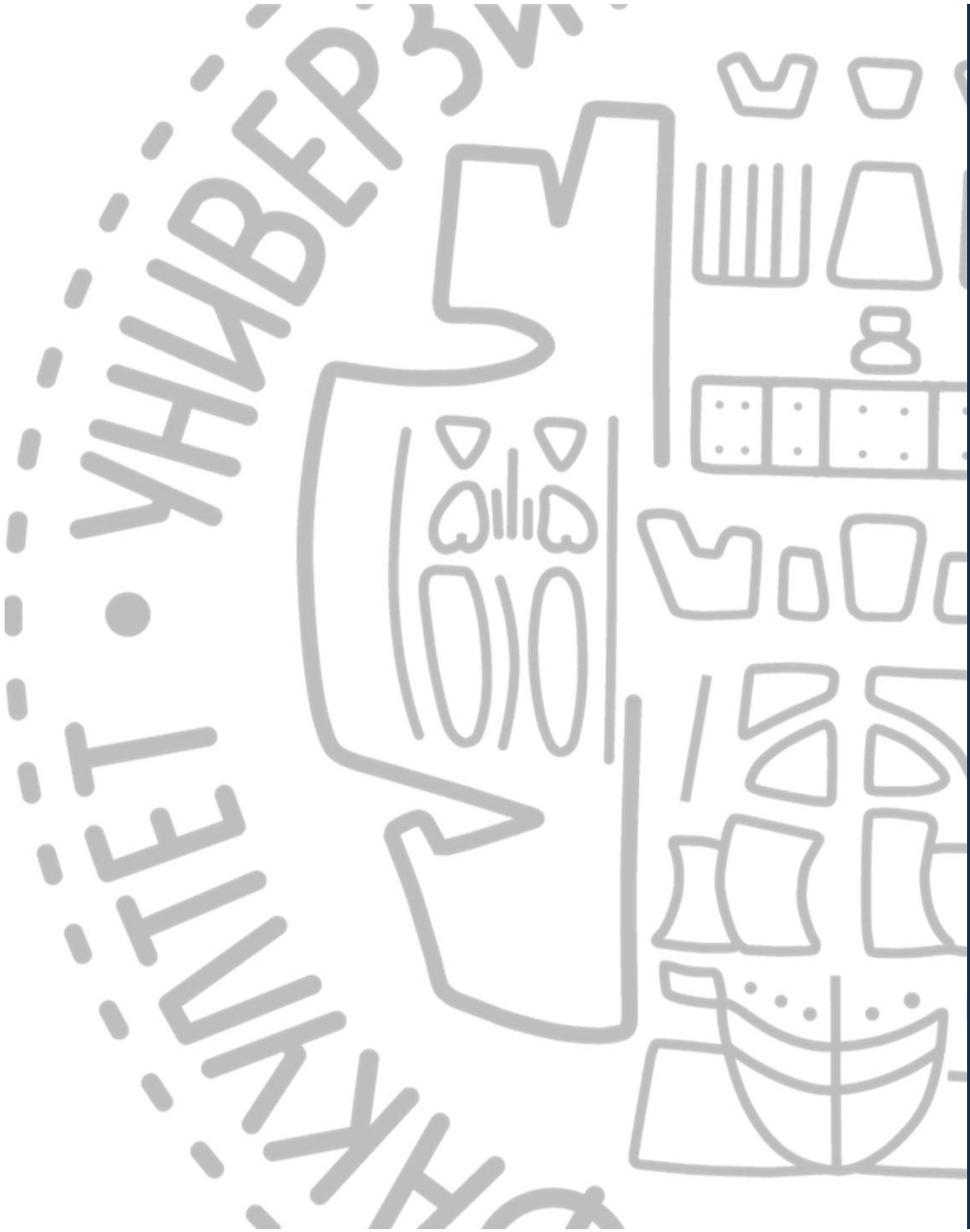
- Trenutno je ITALSERB u fazi **revizije transkripcionog procesa** (ovaj zahtevan proces odvija se po fazama i s prekidima od 2010. i dalje) koja teži tačnosti, objektivnosti, homogenosti zapisa, tj. da obezbedi:

- **sveobuhvatnost** (pokušaj da se obuhvati što veći broj različitih jezičkih fenomena)
- **pouzdanost** (izbeći nametanje lične interpretacije pri tumačenju opserviranih pojava)
- **doslednost** (internu koherentnost ili konzistentnost sistema, gde određeni simbol uvek predstavlja određenu pojavu, tj. uvek isti simbol za istu pojavu)
- **fleksibilnost** (otvorenost za inovacije, dodavanje novih simbola, ažuriranje sistema)

- Paralelno s revizijom, 2022. započeta je **anotacija materijala**, tj. tagiranje prema gramatičkim kategorijama (*PoS tagging*)

Naredni koraci

- Završetak revizije transkripcije
- Završetak tagiranja
- Digitalizacija materijala
- Pitanje dostupnosti
- Raščlanjivanje ITALSERB na subkorpuse
- Konstituisanje paralelnih korpusa



- **Analiza i etiketiranje grešaka** (*error-tagging*): analiza i popis grešaka, priprema i izrada taksonomije grešaka zajedno s nizom odgovarajućih etiketa (*tag*) za njihovu anotaciju (*Contrastive Interlanguage Analysis-CIA, Computer Error Analysis-CEA),*tj.:

- izrada kompjuterskih alatki za detekciju i korekciju grešaka
- izrada taksonomije grešaka i odgovarajućih etiketa (*tag*)
- ukazivanje na jezičku kategoriju (gramatika, vreme, glagol) i vrstu promene (izostavljanje, dodavanje, hipergeneralizacija, analogija) (Andorno i Rastelli, 2009)

- Najrasprostranjeniji nivo: *PoS-tagging* (gram. kateg., rod, broj, morfološke odlike, …)
- Italijanski KJ2: Banca dati di italiano L2, LIPS, Corpus parlato di italiano L2 uglavnom koriste *Tree Tagger* i *Italian NLP*
- Odabrani sistem mora biti: informativan, upotrebljiv, fleksibilan, konzistentan (Granger 2003)

Primer ICLE (International Corpus of Learner English)

tip greške	primer
omaška u pisanju	inly (umesto only)
transfer u vrsti reči	proud (umesto pride)
greška u morfologiji glagola	became (umesto become)
gramatička greška	much (umesto many)
leksika J1	mean (umesto means)
morfologija J1	differents (umesto different)
ortografija J1	confort (umesto comfort)
fonetika J1	Baticano (umesto Vatican)
hiperkorekcija	anorexy (umesto anorexia)

(Andorno i Rastelli, 2009:12)

Krajnji ciljevi

- **Veći obim i dostupnost:** formiranje drugih KJ2, a naročito KSJ
- **Standardizacija KJ2**: uporedivost, dostupnost i prenosivost podataka
- **Standardizacija normi** za ortografsku transkripciju (uz garanciju čitljivosti)
- **Upotrebljivost** sistema za pretraživanje KJ2 i transparentnost podataka
- Lakoća raspolaganja i pretraživanja podataka (user-friendly)
- Mogućnost kontinuiranog ažuriranja, bogaćenja KJ2, formiranja novih potkorpusa
- Eksperimentisanje i upotreba KJ2 u didaktičke svrhe
- Veća razmena između didaktičke teorije i prakse

ITALSERB:

- Instrument za proučavanje J2:

  izvor brojnih, važnih informacija o međujeziku srbofonih studenata u različitim fazama učenja i usvajanja italijanskog J2
- Kvantitet i kvalitet podataka doprinose njihovoj upotrebljivosti i korisnosti
- Sredstvo primenjivo u istraživanjima iz više oblasti:

  korpusna lingvistika, teorija usvajanja J2, glotodidaktika
- Osnova za istraživanja transverzalnog i longitudinalnog karaktera
- Dostupan različitim akterima: istraživači, nastavnici, učenici

Reference (izbor)

Andorno, C. & Rastelli, S. (eds.) (2009). *Corpora di italiano L2: tecnologie, metodi, spunti teorici.* Perugia: Guerra.

Ceković, N. (2014). I segnali discorsivi nell'interlingua degli studenti universitari di italiano L2. *Italica Belgradensia, 2*, 93-110.

Ceković, N. (2016). *Diskursni markeri u govornoj produkciji na italijanskom kao drugom jeziku* (Neobjavljena doktorska disertacija). Beograd: Filološki fakultet.

Ceković, N. (2018). Segnali discorsivi in classe di italiano Ls: uno sguardo dalla parte del docente e dell'apprendente. *Italiano a stranieri, 25,* 16-20.

Ceković, N. (2020). Metatekstualna kompetencija učenika na srednjem nivou znanja italijanskog kao drugog jezika. U V. Polovina & B. Kovačević (prir.), *Primenjena lingvistika danas* (str. 197-209). Beograd: Društvo za primenjenu lingvistiku Srbije.

Ceković-Rakonjac, N. (2012). Ortografska transkripcija govornog korpusa ESNAKIT. U A. Vraneš *et al*. (prir.), *Digitalizacija kulturne i naučne baštine, univerzitetski repozitorijumi i učenje na daljinu*, *knj*. *3* (str. 163-182). Beograd: Filološki fakultet.

Ceković-Rakonjac, N. (2013). ITALBEG corpus parlato di italiano L2. *Italica Belgradensia, 1*, 336-348.

Ceković, N. & Radojević, D. (2015). Didattica con i corpora orali di italiano L2. In C. Ramsey-Portolano (ed.), *The Future of Italian Teaching: Media, New Technologies and Multi-Disciplinary Perspectives* (pp. 96-106).Newcastle upon Tyne: Cambridge Scholars Publishing.

Ceković, N. & Vučo, J. (2020). Uno studio longitudinale come base per la verifica dei criteri di valutazione della competenza discorsiva in italiano L2. *Nasleđe*, *46*, 97-107.

Corino, E. & Marello, C. (2009). Didattica con i corpora di italiano per stranieri. *Italiano LinguaDue*, *1*, 279-285.

Dobrić, N. (2009). Korpusna lingvistika kao osnovna paradigma istraživanja jezika. *Philologia, 7,* 47-58.

Drljević, J. (2012). Ciljna grupa i tipologija zadataka u stvaranju govornog korpusa italijanskog kao J2. Projekat ESNAKIT. In A. Vraneš *et al*. (prir.), *Učenje na daljinu i interaktivna nastava* (str. 285-293). Beograd: Filološki fakultet.

Granger, S. (2003). Error-tagged Learner Corpora and CALL: A Promising synergy. *CALICO Journal, 20* (3), 465-480.

Granger, S. *et al.* (2007). Integrating learner corpora and natural language processing: [...]. *ReCALL, 19* (3), 252-268.

Pravec, N. (2002). Survey of Learner Corpora. *ICAME Journal, 26*, 81-114.

Savy, R. (2005). Specifiche per la trascrizione ortografica annotata dei testi. In F. Albano Leoni & R. Giordano (a c. di), *Italiano Parlato, Analisi di un dialogo*. Napoli: Liguori.

Vučo, J. & Ceković, N. (2019). Il corpus ITALSERB di italiano L2: implementazione e scopi. *Convegno internazionale "L'italianistica nel terzo millennio: le nuove sfide nelle ricerche linguistiche, letterarie e culturali"*, Универзитет „Св. Кирил и Методиј", 27-28. septembar 2019, Skoplje, Severna Makedonija. (usmeno saopštenje)

… hvala na pažnji!

n.cekovic@fil.bg.ac.rs