The ItalSerb Spoken Corpus of L2 Italian and a Paired Subcorpus of L1 Serbian

Nevena Ceković[1] [0000-0002-5407-5759]

[1]University of Belgrade, Faculty of Philology, Department of Italian Studies, Serbia, n.cekovic@fil.bg.ac.rs

Scientific paper DOI: 10.18485/judig.2026.2.ch17

Abstract

This paper presents **progress in the development of** a subcorpus of spoken production in the native language (L1) of Serbian-speaking students of Italian as a foreign language (L2), paired with the main ItalSerb corpus of L2 Italian. The subcorpus consists of 46 samples of dialogic and monologic production, amounting to approximately four hours of recordings, collected on two occasions from 23 students at an intermediate level of L2 proficiency. Particular attention is devoted to the methodological considerations underlying the construction of the subcorpus, including participant data, interaction type, genre and topic, with special emphasis on matching the newly collected metadata with the existing metadata included in the ItalSerb collection. The preliminary analysis of the L1 and L2 production of one participant followed longitudinally reveals similarities and differences in the patterns and frequency of use of the pragmatic elements under investigation, namely discourse markers functioning in **turn initiation (floor-taking)**. Idiosyncratic preferences for particular marker forms that are not functionally equivalent across the two languages were also identified. The significance of these findings lies in their potential contribution to research on the mutual influence of L2 and L1. The methodological experience gained in constructing the subcorpus also provides a basis for developing guidelines for the digitisation and transcription of the collected (sub)corpus data.

Keywords: ItalSerb, learner corpus, paired corpus, L2 Italian, L1 Serbian, Serbian-speaking students, discourse markers

1. The ItalSerb Corpus of L2 Italian

The ItalSerb learner corpus (It. *ITALiano dei SERBofoni*; Eng. *ITALian of SERBian-speaking learners*; Serb. *ITALijanski SRBofonih*; Ceković-Rakonjac 2012, 2013 Ceković-Rakonjac, 2012, 2013; Ceković, 2026a, in press; Janković, 2024a) is a collection of spoken learner production by Serbian-speaking students of Italian as a foreign language (L2), which has been compiled since 2010 at the Department of Italian Studies, Faculty of Philology, University of Belgrade. It constitutes a large-scale collection within the field of L2 learner corpora (Granger, 2012; Andorno and Rastelli, 2009), which represent an important methodological resource for investigating the process of L2 acquisition (Granger, 2002; Nesselhauf, 2004; Cacchione and Borreguero Zuloaga, 2017; Forti and Spina, 2019; Janković, 2024b). To the best of our knowledge, ItalSerb is the first corpus of Italian as a foreign language developed in Serbia and the only such corpus in the country.

The significance of the ItalSerb corpus is further underscored by the small number of learner corpora available for other foreign languages (cf. Radonja and Šućur, 2021; Safari and Ghaffari, 2026), as well as by the limited number of other Italian-language corpus collections developed in the Serbian research context, including corpora of literary and academic texts and public discourse (cf. Moderc, 2015, 2025; Moderc et al., 2023; Koprivica Lelićanin, 2014; Bazić, 2025). ItalSerb also occupies a notable position in the international landscape of corpora specialised in Italian as a foreign language (where, according to the available data, only the Swedish InterIta corpus is larger in terms of recording duration, with approximately 30 hours of recorded material; Bardel and Vallerossa, 2025) and of corpora of Italian as a second language (e.g. LIPS, Vedovelli et al., 2006; Gallina, 2017; *Banca dati di italiano L2*, Andorno, 2001; VIP, Pallotti et al., 2010), developed at Italian and other international university centres.

The ItalSerb corpus comprises 28 hours of spoken material collected between 2010 and 2013 and approximately 200,000 words in orthographic transcripts. It contains 315 audio samples of spoken production by 172 Serbian-speaking students of Italian, representing different years of study and corresponding curricular target levels of competence: A2+ in the first year, B1 in the second, B2+ in the third and C1 in the fourth year of study. Both dialogic and

monologic production were sampled following the model of the CILS certification tests of the University for Foreigners of Siena (*Certificazione di Italiano come Lingua Straniera*, Barni et al., 2009; December 2009 session, Centro CILS, 2009b), in the form of asymmetric, semi-spontaneous interviews.

ItalSerb is organised into several subcorpora, allowing both cross-sectional comparisons as well as longitudinal tracking of individual learners and several student cohorts across multiple academic years. A subset of this material was previously made available in a doctoral dissertation (Ceković, 2016). ItalSerb is currently in the final stage of revision of the transcription procedure (Ceković, 2026a, in press; Ceković-Rakonjac, 2012), together with the morphosyntactic annotation of a sample amounting to approximately three and a half hours (Janković, 2024a). At the same time, the collected language data are being digitised and prepared for deposition and public release via the website of the Centre for Applied Linguistics of the Faculty of Philology[1].

The present paper describes the most recent steps undertaken towards the development of a smaller paired L1 subcorpus linked to the ItalSerb collection. In this way, the existing ItalSerb corpus of spoken production in Italian as an L2 is supplemented by a paired collection of spoken production in Serbian as an L1, collected from the same participants during the same general period of study. The material was originally collected in the course of a doctoral research project (cf. Ceković, 2016: 234), but has not previously been systematically processed or presented to the scholarly community. Furthermore, to the best of our knowledge, no comparable paired L1–L2 learner collection for Italian as a foreign language has previously been developed in Serbia.

2. Paired Subcorpus of L1 Serbian

The subcorpus consists of 46 audio samples of spoken production in Serbian as the participants' native language (L1), produced by 23 students of Italian as a foreign language at the Department of Italian Studies, Faculty of Philology, University of Belgrade, all of whom had originally been randomly selected as participants in the ItalSerb collection. Almost all participants were female (22 female

[1] https://fil.bg.ac.rs/sr/pregled-fakulteta/centri-i-instituti/centar-za-primenjenu-lingvistiku/italserb

and one male), with a mean age of 22 (19-25). They reported knowledge of English as an L3 and various L4s or L5s, most commonly French or German, followed by Spanish, but also Greek, Japanese, Catalan, Norwegian, Polish and Russian. None of the participants had spent more than 90 days in Italy and their contact with the L2 occurred primarily through classroom instruction and the media.

At the time of data collection, the participants were in the third year of their undergraduate studies (fifth and sixth semesters), during which the curricular target level of Italian progressed from B1 to B2+. The first L1 production samples were collected during the winter (fifth) semester, while the second samples were collected approximately six months later, during the summer (sixth) semester of the same academic year.

The total duration of the resulting audio material is approximately four hours (3 hours, 54 minutes and 23 seconds). The first data-collection session accounts for 1 hour, 58 minutes and 5 seconds and the second for 1 hour, 56 minutes and 18 seconds. Two samples of spoken L1 production were collected from each participant, with an average duration of approximately five minutes per individual sample. In addition, during the first data-collection session, a number of partial samples were obtained from 14 other students who also contributed to the ItalSerb collection. These samples consisted either of only one of the two speaking tasks or only the first of the two recording sessions. The 14 partial samples amount to 37 minutes and 30 seconds of material; however, because the data were incomplete, they were not included in the paired subcorpus analysed in the present study.

The elicitation protocol was based on one of the CILS certification tests for the B1 level (June 2009 session, Centro CILS, 2009a) and consisted of an asymmetric semi-structured interview comprising one dialogic and one monologic speaking task. The resulting spoken interaction took place between a teacher acting as examiner and a student acting as candidate. For each task, the student selected one of the proposed topics (previously translated into Serbian) and was given a fixed period of preparation time (2-3 minutes) before the task began. The teacher, who was also a native speaker of Serbian, controlled the course and duration of the interaction by asking questions during the dialogic section while refraining from intervention during the monologic section. The interactions took place in a familiar envi-

ronment, namely the foreign-language classroom and were recorded with a visible audio recorder, with the participants' informed consent.

The sampled interactions covered a range of topics. The first speaking task concerned favourite music, food, sport or means of transport, while the second concerned a public figure, a summer holiday or a description of a given picture. The same topics were used for each individual participant in both L1 recording sessions. This design increased within-participant comparability, although repeated exposure to the same topics should be acknowledged as a possible source of topic-familiarity or practice effects. The L1 tasks were also matched in topic, task type and genre to the L2 tasks previously completed by the same students for the ItalSerb collection, first at the end of their second year of study (fourth semester), at the target B1 level of L2 Italian proficiency, and later at the end of their third year (sixth semester), at the target B2+ level. Because the first corresponding L2 and L1 samples were collected in different semesters (fourth vs fifth), the interval between them should be reported explicitly if direct L1–L2 comparisons are made.

3. Pairing the L1 Subcorpus with the L2 Corpus

This methodological choice was made to facilitate the identification of characteristic linguistic patterns in native- and foreign-language production and the systematic comparison of similarities and differences in their use. It also provides a basis for a more precise examination of the influence of L1 on L2 and, conversely, of L2 on L1. This is particularly relevant because, during the interval between the two L1 data-collection sessions, some of the participants received experimental instructional intervention in the L2, described in detail in Ceković (2016), focusing on specific features of pragmatic-discursive competence.

The instructional treatment consisted of longitudinal explicit-instruction intervention structured according to the Presentation – Practice – Production framework and focused on the forms and functions of Italian discourse markers. These are pragmatically multifunctional words and expressions that play a central role in managing the flow of spoken interaction and in organising discourse (Bazzanella, 1995; Ceković, 2026b).

The experimental group, comprising ten participants, received explicit instruction, whereas the equally sized control group did not

receive experimental instructional treatment. The pre-test and post-test results showed that, at the end of the instructional period, the experimental group outperformed the control group in their L2 use of discourse markers in terms of the range of forms used, their frequency and their multifunctionality (Ceković, 2016: 226–229).

Against this background, combining samples from the L1 sub-corpus and the main L2 corpus makes it possible to investigate whether participants changed their use of particular linguistic forms and functions over time and with the presumed development of their linguistic competence resulting from continued exposure to the foreign language. More specifically, it allows us to examine whether the targeted instruction in L2 discourse markers received by the experimental group between the two L1 recording sessions affected the use of the relevant forms not only in L2 but also in L1. The design further allows comparison of the experimental and control groups in order to compare patterns in L2 and L1 use, both with respect to discourse markers and other linguistic phenomena.

In addition, the construction of this paired L1–L2 corpus opens up possibilities for future second-language acquisition research in which variables capable of substantially affecting language data can be held constant, matched or accounted for analytically. Such variables include participant's native language, task type, genre and topic of spoken production, as well as other relevant variables recorded in the participants' sociolinguistic metadata, such as length of L2 study and degree of exposure to the foreign language, alongside gender, age, social background and knowledge of additional languages (L3+). The methodological decision to use identical speaking-task topics in both L1 recording sessions, and comparable tasks in the foreign and native languages, provides a basis for holding constant or systematically accounting for topic, task type and genre as relevant variables in the analysis of spoken production.

4. Interaction Topics

Table 1 presents task-related metadata from the first and second L1 data-collection sessions for the ten participants in the experimental group (Ceković, 2016). The metadata include the individual identification code (IIC) assigned to each participant, together with the topics of the first and second speaking tasks. Most participants

selected favourite music or food for the first speaking task and a summer holiday for the second.

Table 1: Speaking-task topics in the experimental group.

IIC	*Topic of Task 1*	*Topic of Task 2*
1. C09002	*Omiljena prevozna sredstva* (Favourite means of transport)	*Slika br. 1 (odmor)* (Picture no. 1 (holiday))
2. C09006	*Hrana koju volim da jedem* (Food I like to eat)	*Jedno moje letovanje* (One of my summer holidays)
3. C09013	*Hrana koju volim da jedem* (Food I like to eat)	*Jedno moje letovanje* (One of my summer holidays)
4. C09028	*Omiljena prevozna sredstva* (Favourite means of transport)	*Jedno moje letovanje* (One of my summer holidays)
5. C09044	*Muzika koju slušam* (Music I listen to)	*Slika br. 1 (odmor)* (Picture no. 1 (holiday))
6. C09045	*Hrana koju volim da jedem* (Food I like to eat)	*Jedno moje letovanje* (One of my summer holidays)
7. C09146	*Muzika koju slušam* (Music I listen to)	*Slika br. 1 (odmor)* (Picture no. 1 (holiday))
8. C09036	*Muzika koju slušam* (Music I listen to)	*Jedno moje letovanje* (One of my summer holidays)
9. C09222	*Muzika koju slušam* (Music I listen to)	*Slika br. 2 (biblioteka)* (Picture no. 2 (library))
10. C09224	*Hrana koju volim da jedem* (Food I like to eat)	*Jedno moje letovanje* (One of my summer holidays)

The IIC was automatically generated and assigned to participants during the digitisation of the material (Janković, 2024a). Its initial component contains information on the year in which the participants enrolled in undergraduate studies (in this case, C09 refers to *Candidates* enrolled in *2009*) and serves to pseudonymise participants and reduce their direct identifiability, while the information linking the codes to the original participant records remains accessible to the

corpus compilers. More importantly, the existence of the IIC makes it possible to integrate, within an internal JSON metadata database, information concerning all recording sessions associated with a given participant during the compilation of the ItalSerb L2 corpus and, subsequently, its L1 subcorpus, as well as to retrieve data for groups of participants, such as individual sudent cohorts. This provides a means of linking the main corpus and the subcorpus while ensuring traceability, consistency and systematic organisation of a wide range of metadata.

Table 2 presents the data collected during the two L1 recording sessions for the ten participants in the control group (Ceković, 2016), as well as three participants who belonged to neither the control nor the experimental group (C09053, C09143, C09228). Most participants represented in Table 1b selected the sport they practise as the first speaking-task topic and one of their summer holidays as the second.

5. Analysis of the Data

The initial analyses of the sampled material indicate differences between L1 and L2 production, particularly in the patterns of use of discourse markers, which are a central focus of our research. These multifunctional discourse elements, such as Italian *allora* ('well/so'), *diciamo* ('let us say'), *ecco* ('there/there you go'), *non (lo) so* ('I don't know'), *per esempio* ('for example'), or Serbian *pa* ('well/so'), *onda* ('then'), *znači* ('so/that is') and *dakle* ('therefore/so'), contribute substantially to the dynamics of dialogic interaction and the cohesion of monologic spoken discourse. They can perform a broad range of pragmatic functions, including the turn management (taking, holding or yielding the floor), interruption, checking or confirming reception, modulation and, more generally, signalling speakers' intentions at the interactional level, as well as organising and structuring discourse at the metatextual level (Bazzanella, 1995). At times, these elements may also function as a kind of verbal habit or "tic", produced without the speaker's full awareness, while helping the current speaker hold the floor and signal to the interlocutor that they are not yet ready to yield the turn while experiencing difficulties in planning and formulating their utterance.

Given the frequent and idiosyncratic use of particular marker forms by some participants in L2 production, previously identified

Table 2: Speaking-task topics in the control group and other non-treatment participants.

IIC	*Topic of Task 1*	*Topic of Task 2*
1. C09001	*Muzika koju slušam* (Music I listen to)	*Jedna ličnost iz javnog života* (A public figure)
2. C09015	*Sport kojim se bavim* (A sport I practise)	*Jedno moje letovanje* (One of my summer holidays)
3. C09023	*Sport kojim se bavim* (A sport I practise)	*Jedno moje letovanje* (One of my summer holidays)
4. C09024	*Muzika koju slušam* (Music I listen to)	*Slika br. 2 (biblioteka)* (Picture no. 2 (library))
5. C09029	*Sport kojim se bavim* (A sport I practise)	*Jedno moje letovanje* (One of my summer holidays)
6. C09040	*Sport kojim se bavim* (A sport I practise)	*Jedno moje letovanje* (One of my summer holidays)
7. C09051	*Hrana koju volim da jedem* (Food I like to eat)	*Jedno moje letovanje* (One of my summer holidays)
8. C09052	*Sport kojim se bavim* (A sport I practise)	*Jedno moje letovanje* (One of my summer holidays)
9. C09053	*Muzika koju slušam* (Music I listen to)	*Jedno moje letovanje* (One of my summer holidays)
10. C09058	*Muzika koju slušam* (Music I listen to)	*Slika br. 1 (odmor)* (Picture no. 1 (holiday))
11. C09142	*Sport kojim se bavim* (A sport I practise)	*Slika br. 2 (biblioteka)* (Picture no. 2 (library))
12. C09143	*Sport kojim se bavim* (A sport I practise)	*Jedno moje letovanje* (One of my summer holidays)
13. C09228	*Muzika koju slušam* (Music I listen to)	*Jedna ličnost iz javnog života* (A public figure)

in the experimental study (Ceković, 2016), we sought to determine whether there was a relationship between the use of this pragmatic category in L2 and L1. More specifically, the newly established subcorpus provided a basis for examining whether, during the period between the two recording sessions, changes occurred in the use of this functional category in the participants' native language as well as in the foreign language (cf. Ceković, 2016). This issue may be investigated in greater detail in future research.

This is particularly relevant because, unlike the control group, only the experimental group was exposed during this period to longitudinal explicit instruction focusing on the forms and functions of Italian discourse markers (Ceković, 2016). In addition, the material collected in the subcorpus makes it possible to identify elements, mechanisms and strategies relevant to other aspects of communicative competence, rather than solely to this particular component of pragmatic-discursive competence.

For the initial analysis presented here, we focus on one participant, identified by the code C09028. This student belonged to the experimental group and received explicit instruction. In addition to the two L1 recording sessions, her L2 production was sampled within the ItalSerb collection throughout all four years of study (cf. Ceković, 2026a, in press: § 3.4).

At the L2 recording conducted at the end of the second year of study, corresponding to the target B1 level, this participant, who belonged to the 19–25 age group and also reported knowledge of English, had previously studied Italian for two years and had spent seven days in Italy. Her contact with the foreign language occurred primarily through classroom instruction and the media. One year later, during the L2 recording at the end of the third year of study, at the target B2+ level, she reported in the sociolinguistic questionnaire that she had by then spent a total of 25 days in Italy and that her contact with Italian now occurred through classroom instructions and interaction with friends.

The experimental study (Ceković, 2016) revealed an idiosyncratic tendency in this participant's use of particular L2 forms, especially *eh sì* ('well, yes'), *sì* ('yes'), *non (lo) so* ('I don't know'), *diciamo* ('let us say'), *anche* ('also'), *ecco* ('there/there you go'), *no(n)* ('no/not') and *per esempio* ('for example'). It is also worth noting that the student had an excellent academic record: during the first three years of study, she obtained the highest possible grade (10/10), in each of the six examinations in the core language course *Contemporary Italian Language G1–G6*.

6. Preliminary Results

For the purposes of the present study, we focused on one of many pragmatic functions performed by discourse markers in interaction, namely turn initiation (i.e. taking the floor, turn-taking). By

using this type of marker, the current speaker signals the initiation of their turn to the interlocutor and may gain additional time to plan and formulate the verbal content that follows. The analysis included audio recordings and corresponding transcripts of the participant's L2 production: 4 minutes and 15 seconds of recording and 704 words at the B1 level, and 6 minutes and 34 seconds of recording and 931 words at the B2+ level, amounting to a total of 10 minutes and 49 seconds and 1,635 words across the two proficiency levels. The participant's L1 recordings were also analysed: the first recording lasted 4 minutes and 20 seconds and the second 6 minutes and 21 seconds, giving a total duration of 10 minutes and 41 seconds. The L1 recordings were not orthographically transcribed when the subcorpus was compiled.

The analysis of L2 and L1 samples of approximately equal duration revealed the following descriptive differences in the patterns and frequencies of discourse-marker use in turn initiation across the participant's longitudinal samples:

- in L2 production at the target B1 level (at the end of the second year of study, i.e. the fourth semester), the participant did not use any discourse marker to initiate a turn;
- approximately six months later, during the first L1 recording (at the beginning of the third year of study, i.e. the fifth semester), immediately before the instructional treatment, the same participant used the marker pa four times to take the floor;
- a further six months later (at the end of the third year of study, i.e. the sixth semester), at the target B2+ level of L2 proficiency and after completion of the instructional period, the participant used only one marker form for turn initiation, *eh sì*, which occurred once;
- by contrast, at approximately the same time, during the second L1 Serbian recording (in the same, sixth semester), the participant used the turn-initial marker pa more frequently, with seven occurrences in the turn-initiation function. In some contexts, Italian *allora* may function as a pragmatic equivalent of Serbian pa; however, the participant did not use *allora* in either of the L2 samples analysed.

7. Illustrative Examples

The following examples illustrate the participant's use of discourse markers for turn initiation in both L2 and L1. The annotated orthographic transcription of the corpus excerpts uses the following symbols: < > for annotations of non-lexical, paralinguistic, acoustic and other phenomena, including filled pauses, short (*sp*) or long (*lp*) unfilled pauses, back-channels, lengthened sounds, non-verbal and non-vocal events); _ for word-internal interruption; / for a false start; # # for inter-turn overlap; { } for overlap between acoustic events and speech. Turns are labelled C (It. *Candidato*, 'candidate') for the participant/student and E (It. *Esaminatore*, 'examiner') for the examiner/teacher.

These transcription conventions were adopted for the ItalSerb L2 corpus (Savy, 2005; Ceković-Rakonjac, 2012; Ceković, 2026a, in press: § 4.3). They derive from a detailed notation system developed for spoken Italian and used in the CLIPS corpus of native Italian speakers (*Corpora e Lessici dell'Italiano Parlato e Scritto*, Savy, 2005). Since the L1 subcorpus had not yet undergone orthographic transcription at this initial stage of its development, the L1 excerpts presented below were transcribed using the same conventions to ensure uniformity and comparability.

The first example illustrates the student's use of the marker *eh sì* for turn initiation at the B2+ level in L2, together with other forms performing pragmatic-discoursive functions, such as *diciamo* ('let us say'), *per me* ('for me'), *anche* ('also') and *per caso* ('by any chance'). The speaking-task prompt was: *Andare in giro per negozi, provarsi abiti e scarpe: è un'attività che ti rilassa e ti diverte o è un dovere necessario?* ('Going around shops and trying on clothes and shoes: is it an activity that relaxes and entertains you, or is it a necessary chore?'; all translations are ours).

E: prova numero uno
C: eh sì, ho scelto andare in giro per negozi, provarsi abiti e scarpe: è un'attività che ti rilassa e ti diverte o è un dovere necessario?
E: cosa rispondiamo `<sp>` *a questa domanda?*
C: `<ehm>` *diciamo che io sono una donna e* `<sp>` *per me è un'attività che mi rilassa, però può anche essere molto*`<oo>` *stressante*`<ee>` `<laugh>`*, e se per caso non ho*`<oo>` *soldi e*`<ee>` *io vedo, diciamo, un abito che mi piace tanto e io devo trovare un modo per comprarlo* `<laugh>`

(English translation:
E: task number one
C: well, yes, I chose: going around shops, trying on clothes and shoes, is it an activity that relaxes and entertains you, or is it a necessary chore?
E: how do we answer <sp> this question?
C: <ehm> let's say that I'm a woman and <sp> for me it's an activity that relaxes me, but it can also be very<lengthening> stressful<lengthening> <laugh>, and if by any chance I don't have<lengthening> money<lengthening> and<lengthening> I see, let's say, a dress that I really like, I have to find a way to buy it <laugh>)

The second example illustrates the use of the Serbian turn-opening marker *pa* in the student's L1. The form occurs during the first recording session, at the very beginning of a conversation on the topic of the participant's favourite means of transport. The marker is also attested in turn-initial position in several subsequent turns, together with other forms that appear later in the interaction, such as the filler *ovaj* ('erm/this'), the confirmation marker *da* ('yes') and the modulator *otprilike* ('approximately/more or less').

E: koje je vaše omiljeno prevozno sredstvo?
C: <laugh> <eeh> pa, u Srbiji, nijedno <laugh>
E: { <laugh> ha ha, što, što u Srbiji #C: nijedno# }
C: #E: <laugh>#
E: a #C: <laugh># gde to jeste neko drugo?
C: #E: <eeh># pa u Srbiji nijedno zato što su sva <sp> <NOISE> nepouzdana i prljava
E: <mhmh>
C: i nikakva, a van zemlje <eeh> to je, pošto sam bila u Italiji, definitivno voz <sp> i metro.
E: i metro.
C: da.
E: a jesu l' u Srbiji problem sredstva ili ljudi? <laugh>
C: <eeh>
E: mislim, ko je prljav prevozno sredstvo ili čovek? <laugh>
C: <eeh> pa sredstva su prljava zato što ljudi ne vode računa o njima, a ne bih rekla baš da su ljudi prljavi, mada u letnjim mesecima je i čak i to diskutabilno <laugh>
E: i to je pod znakom pitanja <laugh>

C: `<laugh>`
E: ne_ mislila sam prljavi figurativno
C: `<laugh>`
E: tako da oni ustvari prljaju
C: da
E: super. znači bolje je u Italiji?
C: da
E: šta je znači to bolje, u smislu kad biste uporedili prevozna sredstva naša i njihova, koje su baš te sličnosti, a koje su razlike
C: pa prvo metro ne mogu da uporedim jer mi nemamo metro, ali svakako je efikasnije prevozno sredstvo

(English translation:

E: what is your favourite means of transport?
C: `<laugh>` `<eeh>` well, in Serbia, none of them `<laugh>`
E: { `<laugh>` ha ha, why, why in Serbia #C: none of them# }
C: #E: `<laugh>`#
E: and #C: `<laugh>`# where is there one that you do like?
C: #E: `<eeh>`# well, in Serbia none of them, because they are all `<sp>` `<NOISE>` unreliable and dirty.
E: `<mhmh>`
C: and generally no good, but abroad `<eeh>` since I've been to Italy, definitely the train `<sp>` and the metro.
E: and the metro.
C: yes.
E: and in Serbia, is the problem the means of transport or the people? `<laugh>`
C: `<eeh>`
E: I mean, who is dirty, the means of transport or the person? `<laugh>`
C: `<eeh>` well, the vehicles are dirty because people don't take care of them, and I wouldn't exactly say that people are dirty, although in the summer months even that is debatable `<laugh>`
E: even that is questionable `<laugh>`
C: `<laugh>`
E: no_ I meant dirty figuratively
C: `<laugh>`
E: so they're actually the ones who make them dirty
C: yes
E: great. so it's better in Italy?
C: yes

E: what exactly is better, I mean, if you were to compare our means of transport with theirs, what are the similarities and what are the differences
C: well, first of all, I can't compare the metro because we don't have one, but it's certainly a more efficient means of transport)

Four turn-initial occurrences of the marker pa can be identified in the excerpt, all produced by the participant as she takes the floor in response to the examiner. The form occurs at the near utterance onset, either directly at the turn onset or following the filled pause eeh, before the propositional content of the response, and may help the student to introduce responses to the examiner's questions. By initiating a turn in this way, the student may gain additional time to plan the forthcoming production, that is, to formulate an answer before proceeding with the turn.

The same form occurs four times at short intervals across the student's successive turns. None of these occurrences is immediately preceded by the examiner's use of the same form, suggesting that the pattern cannot be straightforwardly attributed to form-based priming from the interlocutor (cf. Bazzanella (1995)). This does not, however, exclude the possibility of persistence or self-priming resulting from the participant's own repeated use of the form. The repeated occurrence of pa in the excerpt may therefore be indicative of an idiosyncratic preference on the part of the participant, who, as noted above, also displayed preferences for particular, functionally non-equivalent forms in L2.

Occurrences of the same marker were also recorded in the participant's second L1 sample, this time with a higher raw frequency. In only one of the seven cases did the turn-initial marker pa occur on its own. In the remaining cases, it was immediately followed by another form, producing combinations such as pa verovatno ('well, probably'), pa ne znam ('well, I don't know'), pa recimo ('well, let's say') and pa eto ('well, there you go/so'). The only form that occurred slightly more frequently than pa in this sample was the filler ovaj, while i ('and'), da ('yes'), recimo ('let's say'), eto ('there/there you go') and šta znam ('I don't know/what do I know') were also recurrently attested.

8. Conclusions and Future Directions

The analysis presented in this study revealed, on the one hand, different patterns and raw frequencies of discourse-marker use in the function of turn initiation in the observed participant's L1 and L2 production. The L1 floor-taking marker pa occurred more frequently than eh sì, which performed the same turn-initiation function in L2, despite the approximately equal total duration of the L1 and L2 interaction samples. On the other hand, the analysis also revealed certain similarities. In both languages, the later sample, recorded after the experimental instructional period, contained a higher number of occurrences of markers performing the target function than the earlier sample. Although these findings are based on a highly limited dataset, they may raise the possibility that changes in the realisation of the target function may have occurred alongside the period of L2 instruction, in both L2 and L1. Any possible relationship between these changes and the instructional treatment will need to be tested on a larger sample, particularly through comparison with the control group.

The analysis revealed further similarities in the participant's idiosycratic preferences for particular forms in both languages. However, in L1, the idiosyncratic pattern identified here involved the repeated use of pa specifically in the target function of turn initiation, whereas the participant's previously observed idiosyncratic preferences in L2 involved other forms performing different discourse functions. The possible relationship between idiosyncratic patterns of marker use in the two languages, as suggested by the observations concerning the participant analysed here, remains to be investigated in greater detail in future studies.

Building on the newly developed paired L1 subcorpus, future corpus-based research could investigate other discourse-marker forms and functions and include the remaining participants represented in the corpus project. Such research could compare patterns of forms and function in the experimental and control groups, providing a broader perspective on the effects of possible instruction-related changes and on the development of this pragmatic category not only in L2 (cf. Ceković, 2016) but also in L1. This would support more reliable conclusions concerning patterns of L1 and L2 use, particularly in relation to discourse markers, with the present study serving as a starting point.

In addition to the issues discussed here concerning the construction of the L1 subcorpus and its pairing with the main L2 corpus, including the protection of personal data and the processing of metadata and sociolinguistic information, several further methodological questions remain to be addressed. The most important concern the orthographic transcription and digitisation of the collected subcorpus material. In this respect, consideration should be given to integrating all subcorpus data into the existing internal (meta)database created for the main ItalSerb collection. This would allow the two collections to be fully aligned and, ultimately, enable parallel access to the L2 and L1 production of the same participants.

The L1 subcorpus could also be expanded to include the incomplete material from the 14 additional participants not considered in the present study, amounting to 37 minutes and 30 seconds of additional material. This would increase the total size of the collection to approximately four and a half hours of audio (4 hours, 31 minutes and 53 seconds). To distinguish the paired subcorpus from the main corpus, it could be assigned a separate name, for example *ItaL2SerbL1*.

Orthographic transcription of the recorded production represents one of the most important future steps towards completing the paired L1 subcorpus. We propose applying the same transcription conventions used for the main ItalSerb corpus, since the excerpt presented in this study demonstrates the feasibility of this approach and was itself transcribed according to these conventions. This would ensure a unified approach to the transcription of material from the two collections, L2 and L1, making them comparable in this respect as well.

The preliminary analysis of one participant's production presented in this paper has illustrated, through the analysis of a single discourse-marker function, both differences and similarities between L1 and L2 production. These findings may provide a basis for more detailed investigation of cross-linguistic influence, particularly in relation to these important discourse elements. The description of the principal methodological assumptions underlying the construction of the subcorpus, the pairing of the L1 and L2 collections, and the illustration of these procedures through representative corpus examples are intended to highlight practices relevant to the development of learner corpora and to provide guidelines for exploiting such data

in different types of research on L2 acquisition and foreign-language pedagogy.

Language Resources

Andorno, Cecilia. 2001. *Banca dati di Italiano L2. Progetto di Pavia.* Università di Pavia. CD-ROM.

References

Cecilia Andorno and Stefano Rastelli, editors. 2009. *Corpora di italiano L2: tecnologie, metodi, spunti teorici.* Guerra, Perugia.

Camilla Bardel and Francesco Vallerossa. 2025. Past, Present and Future of InterIta: Reexploring a Swedish Learner Corpus of Italian. *Italiano LinguaDue*, 17(2):49–67.

Monica Barni, Andrea Bandini, Laura Sprugnoli, Silvia Lucarelli, Anna Maria Scaglioso, Barbara Strambi, Cristina Fusi, and Anna Maria Arruffoli. 2009. *Linee guida CILS. Certificazione di Italiano come Lingua Straniera, Università per Stranieri di Siena.* Guerra, Perugia.

Jovana Bazić. 2025. *Ograđivanje u javnom diskursu: kontrastivna analiza italijanskog i srpskog jezika.* Ph.D. thesis, Filološki fakultet, Beograd. Unpublished doctoral dissertation.

Carla Bazzanella. 1995. I segnali discorsivi. In Lorenzo Renzi, Giampaolo Salvi, and Anna Cardinaletti, editors, *Grande grammatica italiana di consultazione*, volume 3, pages 225–257. Il Mulino, Bologna.

Annamaria Cacchione and Margarita Borreguero Zuloaga. 2017. I corpora di l2 come spunti per la riflessione didattica. l'interazione nativo-non nativo nel percorso formativo degli insegnanti. *RiCOGNIZIONI. Rivista di lingue, letterature e culture moderne*, 8:15–30.

Nevena Ceković. 2016. *Diskursni markeri u govornoj produkciji na italijanskom kao drugom jeziku.* Ph.D. thesis, Filološki fakultet, Beograd. Unpublished doctoral dissertation.

Nevena Ceković. 2026a. The ItalSerb corpus: L2 italian spoken by serbian students. In press.

Nevena Ceković. 2026b. Tratti generali dei segnali discorsivi in italiano l2: evidenze da un corpus. *Anali Filološkog fakulteta*, 38(1):31–44.

Nevena Ceković-Rakonjac. 2012. Ortografska transkripcija govornog korpusa ESNAKIT. In Aleksandra Vraneš, Ljiljana Marković, and G. Alexander, editors, *Digitalizacija kulturne i naučne baštine, univerzitetski repozitorijumi i učenje na daljinu*, volume 3, pages 163–182. Filološki fakultet, Beograd.

Nevena Ceković-Rakonjac. 2013. ITALBEG corpus parlato di italiano l2. *Italica Belgradensia*, 1:336–348.

Centro CILS. 2009a. Certificazione di italiano come lingua straniera. livello CILS-UNO B1. produzione orale. sessione giugno 2009.

Centro CILS. 2009b. Certificazione di italiano come lingua straniera. produzione orale. sessione dicembre 2009.

Luciana Forti and Stefania Spina. 2019. Corpora for Linguists vs. Corpora for Learners. Bridging the Gap in Italian L2 Learning and Teaching. *EL.LE*, 8(2):349–361.

Francesca Gallina. 2017. LIPS: Creating a learner corpus. In Anna Gudmundson, Laura Álvarez López, and Camilla Bardel, editors, *Romance Languages. Multilingualism and Language Acquisition*, pages 161–181. Peter Lang, Frankfurt am Main.

Sylviane Granger. 2002. A bird's eye view of learner corpus research. In Sylviane Granger, Joseph Hung, and Stephanie Petch-Tyson, editors, *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, pages 3–33. John Benjamins, Amsterdam and Philadelphia.

Sylviane Granger. 2012. Learner Corpora. In Carol A. Chapelle, editor, *The Encyclopedia of Applied Linguistics*, pages 3235–3242. Wiley-Blackwell, Oxford.

Nikola Janković. 2024a. Anotacija ITALSERB govornog učeničkog korpusa srbofonih studenata italijanskog jezika. In Snežana Gudurić, Marija Stefanović, and Jasmina Dražić, editors, *Jezici i kulture u vremenu i prostoru XI/2*, pages 298–311. Filozofski fakultet, Novi Sad.

Nikola Janković. 2024b. Tipovi korpusa i mogućnosti njihove primene u nastavi italijanskog jezika. In Nenad Tomović, editor, *Zbornik radova sa I naučne konferencije za mlade istraživače i doktorande „Savremeni tokovi u izučavanju jezika, književnosti i kulture“*, pages 198–219. Filološki fakultet, Beograd.

Marija Koprivica Lelićanin. 2014. *Metadiskurs u savremenim italijanskim akademskim i novinskim člancima*. Ph.D. thesis, Filološki fakultet, Beograd. Unpublished doctoral dissertation.

Saša Moderc. 2015. Elektronski korpus srpskih književnih dela i njihovih prevoda na italijanski jezik. *Anali Filološkog fakulteta*, 27(2):301–316.

Saša Moderc. 2025. Korpus srpskog književnog jezika SerbItaCor3. *Srpski jezik*, 30:257–267.

Saša Moderc, Ranka Stanković, Aleksandra Tomašević, and Mihailo Škorić. 2023. An Italian-Serbian Sentence Aligned Parallel Literary Corpus. *Pregled Nacionalnog centra za digitalizaciju*, 43:78–91.

Nadja Nesselhauf. 2004. Learner corpora and their potential in language teaching. In John Sinclair, editor, *How to Use Corpora in Language Teaching*, pages 125–152. John Benjamins, Amsterdam.

Gabriele Pallotti, Stefania Ferrari, Elena Nuzzo, and Camilla Bettoni. 2010. Una procedura sistematica per osservare la variabilità nell'interlingua. *Studi Italiani di Linguistica Teorica e Applicata*, 39(2):215–241.

Minja Radonja and Srđan Šućur. 2021. O korpusu studenata anglistike (KorSAng) i mogućnostima njegove softverske eksploatacije. *Infotheca – Journal for Digital Humanities*, 21(1):37–58.

Saeed Safari and Mahbod Ghaffari. 2026. Cambridge-belgrade (CamBel) persian learner corpus: Theoretical frameworks and methodological insights. In Rajna Dragićević, Miloš Đurić, Vladimir Karanović, and Ana Kolarić, editors, *Savremena filološka istraživanja*, volume 2, pages 276–289. Filološki fakultet, Beograd.

Renata Savy. 2005. Specifiche per la trascrizione ortografica annotata dei testi. In Federico Albano Leoni and Rosa Giordano, editors, *Italiano Parlato, Analisi di un dialogo*, pages 1–28. Liguori, Napoli.

Massimo Vedovelli, Carla Bagna, Monica Barni, Francesca Gallina, Lorenzo Miraglia, and Alessia Pallassini. 2006. Il LIPS –

lessico di frequenza dell'italiano parlato dagli stranieri. In *Progetto dizionario italiano svedese. Atti del I Colloquio, Stoccolma 10–12 febbraio 2005*, pages 55–78, Stockholm. Intellecta Docusys.

ItalSerb: govorni korpus L2 italijanskog i upareni subkorpus L1 srpskog

Nevena Ceković

Sažetak

U radu se predstavljaju dostignuća u izgradnji uparenog subkorpusa govorne produkcije na maternjem jeziku (L1) srbofonih studenata italijanskog kao stranog (L2). Glavni korpus govorne produkcije na L2 italijanskom, ItalSerb, time je dopunjen uparenim subkorpusom produkcije na L1 srpskom, prikupljene od istih ispitanika tokom odgovarajućeg perioda studiranja. Učenički korpus ItalSerb (ital. *ITALiano dei SERBofoni* – ITALijanski SRBofonih), prema našim saznanjima, prvi je i jedini korpus L2 italijanskog u Srbiji, gde do sada nije zabeleženo ni postojanje uparenih učeničkih zbirki za italijanski kao strani jezik. Obuhvata 28 sati govornog materijala i 200.000 reči u ortografskim transkriptima, odnosno 315 audio-uzoraka produkcije 172 srbofona studenta italijanskog, praćenih transverzalno i longitudinalno tokom četiri studijske godine i na različitim ciljnim nivoima jezičke kompetencije (od A2+ do C1). Relevantnost kolekcije naročito dolazi do izražaja s obzirom na mali broj učeničkih korpusa za druge strane jezike, kao i na mali broj drugih italijanističkih korpusnih zbirki, nastalih u Srbiji. Značajna je i u međunarodnom kontekstu korpusa specijalizovanih za italijanski kao strani (među kojima je, prema dostupnim podacima, po ukupnom trajanju snimljenog materijala veći jedino švedski InterIta, sa približno 30 sati snimaka) ili kao drugi jezik (LIPS, Banca dati di italiano L2, VIP), kreiranih u italijanskim i drugim inostranim univerzitetskim centrima.

Novoformirani subkorpus produkcije na L1 srpskom čini 46 uzoraka dijaloške i monološke produkcije, ukupnog trajanja približno četiri sata (prosečno oko pet minuta po uzorku). Uzorci su prikupljeni u dva navrata, u razmaku od šest meseci, od 23 studenta na srednjem nivou kompetencije u L2, tokom petog i šestog semestra studija, kada su ciljni nivoi kompetencije u L2 bili u rasponu od B1 do B2+. Interakcija je elicitirana, kao i u korpusu ItalSerb, po modelu certifikacionih testova, na raznovrsne teme i u vidu asimetričnog polustrukturisanog intervjua. Ispitanici su za prvi zadatak uglavnom birali omiljenu muziku ili hranu, a za drugi opis jednog letovanja. Svaki ispitanik je u oba L1 snimanja obrađivao iste teme. Uporedivost teme, tipa zadatka i žanra produkcije u L1 i L2 omogućavaju identifikovanje karakterističnih jezičkih obrazaca i sistematsko poređenje upotrebe u maternjem i stranom jeziku.

Posebna pažnja u radu posvećena je metodološkim pretpostavkama u izgradnji subkorpusa: pseudonimizaciji i zaštiti podataka, tipu, žanru i tematici interakcije, kao i uparivanju novoprikupljenih metapodataka iz subkorpusa sa postojećim podacima kolekcije ItalSerb. Preliminarna analiza produkcije jednog longitudinalno praćenog ispitanika na L1 i L2, sprovedena u radu, ukazuje na sličnosti i razlike u obrascima i broju pojavljivanja diskursnih markera u funkciji uzimanja reči (pa u L1 i funkcionalno ekvivalentan združeni oblik eh sì u L2). Uočene su i idiosinkratičke preferencije prema pojedinim oblicima markera koji u dva jezika nisu funkcionalno ekvivalentni. Analiza odabrane markerske funkcije pruža početni uvid u odnos između produkcije na L1 i L2 i otvara prostor za detaljnije proučavanje međujezičkog uticaja, kako u vezi sa markerima diskursa tako i sa drugim jezičkim fenomenima.

Pored pitanja vezanih za koncipiranje subkorpusa L1 i njegovo povezivanje sa glavnim korpusom L2, uključujući zaštitu ličnih podataka, obradu metapodataka i sociolingvističkih informacija, u daljem radu potrebno je razmotriti i postupke ortografske transkripcije i digitalizacije prikupljene građe. Dosadašnja iskustva u tom pogledu mogu poslužiti kao osnova za izradu smernica za njenu dalju obradu i korišćenje. Predstavljanje osnovnih metodoloških principa izgradnje subkorpusa i povezivanja zbirki na L1 i L2, uz njihovu ilustraciju reprezentativnim korpusnim primerima, ima za cilj da ukaže na metodološki relevantne postupke u izradi učeničkih (sub)korpusa, kao i da ponudi smernice za njihovu primenu i korišćenje u različitim vrstama istraživanja u oblasti usvajanja L2 i didaktike stranih jezika.

Ključne reči: ItalSerb, učenički korpus, upareni korpus, L2 italijanski, L1 srpski, srbofoni studenti, diskursni markeri

Works are licensed under the terms and conditions of the Creative Commons Attribution 4.0 International License.

https://creativecommons.org/licenses/by/4.0/