https://doi.org/10.31820/f.37.1.4

Nada Poropat Jeletić, Gordana Hržica, Eliana Moscarda Mirković

THE CORPUS OF SPOKEN ISTROVENETIAN/ FIUMAN AND CROATIAN (C-ORAL-IC)

dr. sc. Nada Poropat Jeletić, Sveučilište Jurja Dobrile u Puli, Filozofski fakultet
nada.poropat.jeletic@unipu.hr orcid.org/0000-0001-8787-9748

dr. sc. Gordana Hržica, Sveučilište u Zagrebu, Edukacijsko-rehabilitacijski fakultet
Gordana.hrzica@gmail.com orcid.org/0000-0001-6067-9148

dr. sc. Eliana Moscarda Mirković, Sveučilište Jurja Dobrile u Puli, Filozofski fakultet
eliana.moscarda.mirkovic@unipu.hr orcid.org/0009-0006-2397-634X

pregledni rad
UDK: 811.131.1’282(497.57)
811.163.42’282(497.57
rukopis primljen: 4. prosinca 2024; prihvaćen za tisak: 10. ožujka 2025.

Bilingual conversational corpora are invaluable for studying genuine contact phenomena in spontaneous bilingual speech. This paper presents the Corpus of Spoken Istrovenetian/Fiuman and Croatian *(C-ORAL-IC), the first corpus documenting unscripted Istrovenetian and Fiuman dialects spoken among bilinguals in the Istrian and Kvarner areas of Croatia. The region has a long history of Croatian and Italian cultural and linguistic interaction, shaping a complex sociolinguistic system with diglossic and polyglossic relations. C-ORAL-IC includes data from 87 bilingual/multilingual speakers and features over 85,000 tokens and 27,000 types. Available on TalkBank (BilingBank subsection) [https://talkbank.org, https://biling.talkbank.org/*

* Acknowledgment: The work on this paper was supported by the Croatian National Foundation, grant UIP-2017-05-6603 for the project *Multilevel Approach to Discourse in Language Development*.

access/C-ORAL-IC.html], it includes transcribed, phonologically adapted, coded, segmented and morphologically tagged recordings. Additional participant data on language history and usage are available. C-ORAL-IC provides a rich resource for exploring spontaneous bilingual speech, offering insights into conversational features, structure, and synchronic changes in Istrovenetian/Fiuman.

Keywords: *language sampling; spoken speech corpora; codeswitching; bilingual speech*

1. Introduction

The growing research interest in spontaneous bilingual speech led to the consequent development of analytical tools for building and studying corpora[1] of spoken language in informal unmonitored bilingual settings. To represent the nature and variation of discourse in naturally occurring communication within members of a speech community (Ruhi and Isik Tas 2014), spontaneous and high interaction speech is essential for investigating spoken multi-party[2] corpora (Čermak 2009). Bilingual corpora have mostly been designed for natural language processing, for example for the development of machine interpretation or speech recognition. Bilingual corpora collected for the study of contact phenomena are sparser (but see, for example – Dal Negro 2013 – for individual corpora or BilingBank in TalkBank – MacWhinney 2007). Such corpora offer many advantages: public availability allows for easy replication of studies and cumulative progress as a research community builds up around the corpus (Mair 2009), while the tools developed within the field of corpus linguistics enable easy retrieval and analysis of information. Bilingual conversational corpora represent a meaningful and the most comprehensive data source for investigating the genuine contact phenomena in non-monitored bilingual speech productions. They can be particularly useful for bilingual research

[1] As stated by Sinclair (1991:171), a corpus is "a collection of naturally-occurring language text, chosen to characterize a state or variety of a language". Speech corpora are necessary for retrieving information about spoken language, or studying different aspects of spoken language. Of course, since building written corpora is much simpler than engaging in language sampling and transcription of speech sequences, spoken corpora are sparser and smaller in size. In general corpora are lemmatized and morphologically marked.

[2] A multi-party interaction involves "three or more participants, is the 'canonical' case, around which conversational mechanisms are designed" (Haviland 1986: 249).

since some features of bilingual interaction can hardly be accessed with more traditional methodologies (e.g. elicitation tasks). The method of language sampling provides the resources for describing language interaction in a bilingual community and/or in bilingual situations (e.g. code-switching, amount of languages used, number of languages used, etc.). To capture these phenomena in genuine discourse situations, such sampling should be as close as possible to spontaneous communication. Thus, bilingual spoken corpus design is methodologically demanding because it has to fulfill the following criteria: (1) the corpus language has to be authentic, (2) representative and (3) language sampling has to be used (Tognini Bonelli 2001).

2. The sociolinguistic context of Istria and Rijeka

The Istrian peninsula is the western-most part of Croatia, surrounded by the northern Adriatic Sea. It stretches from the bays of Trieste and Venice (in the North-West) to the bay of Rijeka and Kvarner (in the North-East) and the Cape Premanture (in the South). The part of the peninsula belonging to Croatia partially coincides with the Istria County, the only statutory bilingual county in Croatia where the Croatian-Italian bilingualism is recognized *de jure* and *de facto*. The whole area is characterized by a permanent contact of Croatian and Italian cultures and language varieties and by a complex and fragmented sociolinguistic macro-system shaped by the mutual interplay of asymmetric and diglossic/polyglossic relations among two official languages (Croatian and Italian), complemented by macro-regional dialects (the Istrovenetian *koine* and the Chakavian *koine*), micro-regional dialects (Chakavian, Kaikavian, Shtokavian), with the addition of local dialects in Istria (like the Istriot dialects, the Istroromanian dialects, etc.) (Blagoni et al. 2016; Lisac 2009; Lukežić 1990; Malecki 2002, 2007; Ujčić 2015).

Particularly important is the documentation of the macro-regional minority diatopic varieties of Venetian: Istrovenetian and Fiuman. Istrovenetian (It. *istroveneto*) includes Italian dialects from the Croatian and Slovenian parts of Istria, which belong to the Eastern branch of the Venetian dialect system (ISO 639-3 code: VEC). The same type of dialect (called Fiuman dialect – It. *fiumano*) is used in Rijeka as well as in its surrounding towns and villages and on the Kvarner islands. Istrovenetian differs from the other Venetian dialects, with which it shares a common lexicon and linguistic structure, insomuch as it contains pre-Venetian Istro-romance lexical remains and linguistic elements of Croatian and Slovene languages. The two diatopic subvarieties of Venetian represent a long-lasting 'lingua franca' of the Italian Istrian and Fiuman repertoire, the mother tongue of the members of the autochtonous Italian National Community, and the primary code of informal everyday communication among the Italophone population and the primary/privileged code of their community identification (e.g. Blagoni et al. 2018; Milani Kruljac 1990, 2003; Poropat Jeletić 2015, 2017a, 2017b; Scotti Jurić 2001).

2.1. Aims

This paper has two aims. The first aim is to present the design of conversational spoken language corpus of the bilingual/multilingual Italophone minority community living in Croatia, an old autochthonous community shaped by a several century old cultural and linguistic contacts between the Croatian and Italian cultures. The second aim is to present the ancillary resources that were also collected, namely questionnaires with biographical data and language use data.

3. C-ORAL-IC Corpus Design

The Istrovenetian dialect and the Fiuman dialect were recorded within naturally occurring spontaneous conversation taking place among bilingual/multilingual local speakers, representing the most authentic way of capturing a community's language practices (Grosjean 2001). The main structural and methodological features for constructing the corpus will be presented in this paper.

3.1. Structure of the corpus

The Corpus of Spoken Istrovenetian/Fiuman and Croatian (C-ORAL-IC) consists of 39 transcripts, their corresponding source audio files (37 media files) and an accompanying participant spreadsheet (87 participants). All files are publicly available within BilingBank (MacWhinney 2019), which is a part of the TalkBank, the largest available database of spoken corpora (MacWhinney 2000). The spreadsheet contains demographic and sociolinguistic data about each speaker. According to the rules of the TalkBank (https://talkbank.org/share/rules.html), published works based upon TalkBank corpora should cite the original TalkBank publication by MacWhinney (2019), the corpus itself (https://biling.talkbank.org/access/C-ORAL-IC.html) and the present paper. C-ORAL-IC is available under a *Creative Commons Attribution-ShareA-*

like 4.0 International Public License (https://creativecommons.org/licenses/by-sa/4.0/legalcode).

3.2. Participants

When selecting the participants, the researchers made every effort to ensure that they accurately represented the bilingual community in Croatia. This was achieved by selecting participants of different ages, different education and from different geographic areas, while at the same time trying to preserve the most common characteristics of the bilingual community: command and usage of language varieties spoken in the area (namely: Croatian standard language, Croatian Chakavian dialect, Italian standard language, Italian dialect – Istrovenetian/Fiuman dialect) and early exposure to at least two language varieties. The procedure was approved by the Ethics committee of the Faculty of Education and Rehabilitation Sciences of the University of Zagreb. All participants signed an informed consent in which the data collection was described. They were informed that their data will be published anonymized and that they can withdraw from this study at any time.

3.2.1. Basic information

The original group of recorded participants consisted of 102 bilingual speakers. In total, 15 of them withdrew during sampling, or issues related to the COVID-19 pandemic (2019–2021) affected their participation. All the transcripts were annotated to include the participants' basic information regarding gender, age, and the location of the conversation.

Table 1. *Information about participants*

Gender	Age (range; average)	Education (range; M)	Income (range; M)
F: 59 M: 28	18–82 Average: 42	Elementary school – PhD Average: high school	Very low – very high Average: average

3.2.2. Relevant sociolinguistic features

The spreadsheet accompanying the transcripts and audio files includes more detailed data: participants' origin (place of birth), place of living, level of education, profession, socioeconomic status, mother tongue(s), and questions regarding linguistic proficiency and language use in their social networks.

Place of recording. Participants were recruited by the investigators in the city of Rijeka and all the statutory bilingual towns of the Istrian County characterized by the historical presence of the members of the Italian National Community, specifically the towns and the surrounding areas of Pula-Pola, Vodnjan-Dignano, Rovinj-Rovigno, Poreč-Parenzo, Novigrad-Cittanova, Buje-Buie, Umag-Umago (Figure 1), to ensure diatopic representativeness. It is important to note that not all areas of Istria are equally populated by bilingual speakers, and the number of participants in certain geographic areas roughly reflects this distribution.

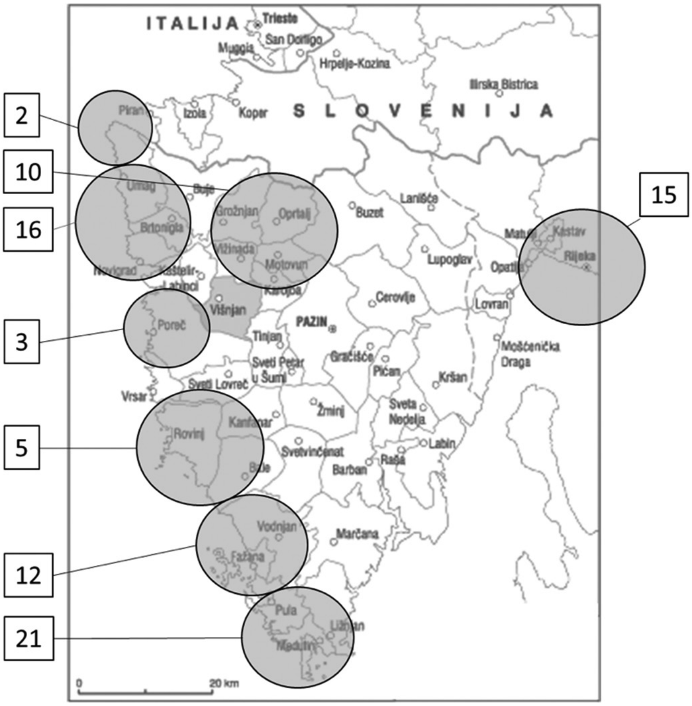


Figure 1. *Number of participants per geographic location*

Onset of bilingual language acquisition. A large majority of participants (98%) were simultaneous bilinguals, stating that they had been exposed to both Croatian and Italian language varieties by the age of three. Two participants were exposed to the other language variety (in their case, the Italian standard language) at the age of seven when starting school. Five participants reported their initial exposure to the second language at the age of three, while the others reported exposure prior to the age of three. Many participants had been exposed to both Croatian and Italian language varieties from birth.

Language use of participants. All participants use at least two varieties from two languages every day for different activities (Table 4). The majority use all four varieties: Croatian standard language, Italian standard language, Croatian dialect, and Italian dialect (59%), or a combination of Croatian standard language, Italian standard language, and Italian dialect (27%).

Table 2*. Language usage of participants*

Language varieties	No. of speakers	Percentage of speakers
Croatian, Italian	1	1%
Croatian, Italian dialect	8	9%
Croatian, Croatian dialect, Italian dialect	54	59%
Croatian, Italian, Italian dialect	25	27%
Croatian, Croatian dialect, Italian dialect	3	3%

Languages of education. From kindergarten to university, the majority of participants were educated in Italian (Table 3). The percentages are higher than 70% for Italian at all educational levels except for university. This is expected since there are many Italian kindergartens, elementary schools, and high schools available in Istria and Rijeka (due to the parallel educational system in Croatian and Italian language). However, there are fewer opportunities to study in Italian at the university level in Croatia, and studying in Italy is less accessible. Still, a majority of participants (52%) did study in Italian.

Table 3*. Language of education*

	Croatian	Italian	Croatian and Italian	Other
Kindergarten	21%	73%	6%	
Elementary school	22%	74%	3%	1%
High school	20%	76%	4%	
University	35%	52%	7%	6%

Language dominance. All participants use at least two language varieties, but the dominant language in their everyday life can differ. Based on several questions (e.g. the amount of language usage in everyday activities, languages used at work, language of dreams, amount of language used with family, etc.), we estimated the language dominance of the participants (Table 4). Some participants reported using one language variety more than 60% of the time on average, indicating they have one dominant variety, which is most often an Italian dialect. The majority of users reported using two language varieties at least 40% but less than 60% per individual variety on average, indicating they have two balanced languages. This is most often a combination of the Croatian standard language and an Italian dialect. Some participants use three languages regularly, with an average usage between 30% and 40%. This is most frequently a combination of the Croatian standard language, the Italian standard language, and an Italian dialect. A significant number of participants (21%) use all four varieties in comparable amounts, with an average usage between 20% and 30%.

Table 4. *Language dominance*

Language dominance	No. of speakers	Percentage of speakers
ONE LANGUAGE (average use >60%)		
Croatian	1	1%
Italian	1	1%
Italian dialect	21	25%
TWO BALANCED LANGUAGES (average use for each 40%–60%)		
Croatian, Italian	1	1%
Croatian, Italian dialect	24	29%
Croatian dialect, Italian dialect	4	5%
THREE BALANCED LANGUAGES (average use for each 40%–60%)		
Croatian, Croatian dialect, Italian dialect	3	4%
Croatian, Italian, Italian dialect	10	12%
Croatian dialect, Italian, Italian dialect	1	1%
THREE BALANCED LANGUAGES		
Croatian, Italian, Croatian dialect, Italian dialect	17	21%

3.3. Procedure

The methodological approach used for the creation and annotation of C-ORAL-IC was based on the four criteria presented by Fort (2012). Firstly, the preparatory process included the identification of the investigators/annotators, the experts for monitoring and the speakers who participated in the recorded conversations. Secondly, a transcription, annotation and codification guide were built. Afterwards, a pre-campaign was settled, during which the investigators/annotators were trained and the process of making the first contacts with the potential participants began. The running-in period followed and it included field work (registration of the conversations and administration of a questionnaire to the participants). After the transcription of the audio files, the coding process and the double detailed correction took place.

Data were collected from 2018 to 2021 and language sampling was performed by investigators from bilingual communities with access to groups of bilingual speakers. Investigators were recruited and trained to collect bilingual speech samples. Sampling was performed in different everyday informal interactive situations, such as informal gatherings, socializing or family meals. Participants were administered a background questionnaire providing information about the sociodemographic, sociolinguistic and socioeconomic status, language exposure and language usage in their social networks.

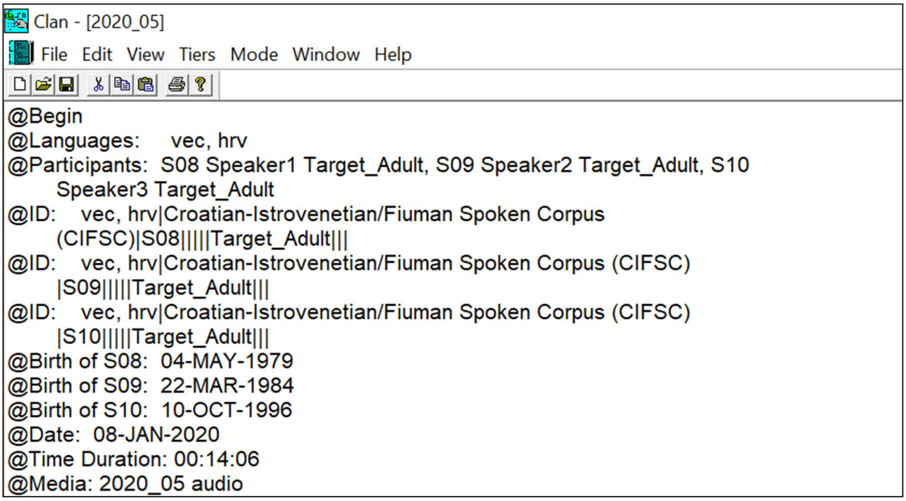
With the aim of mitigating the Observer's paradox (Labov 1972), two criteria were applied, similar as in Kuvač Kraljević and Hržica (2016). First, all the participants were informed about the research aims and speech sampling procedure. They all provided a written informed consent in which they agree to be recorded without their explicit knowledge at a random point within the period of one month after signing the consent. Second, the investigators were trained to participate in the recorded sessions as little as possible.

C-ORAL-IC was transcribed by using the TalkBank uniform transcription standard (CHAT) as a coding system, which allows data sharing. TalkBank enables data sharing in different modes (http://talkbank.org/share/irb/options.html) (MacWhinney 2000). In order to guarantee the anonymization of the speakers, each of them was given a unique pseudonymized three-letter sequence. The corpus was morphologically tagged to enhance its usability and analytical value.

3.3.1. Transcription and annotation

All corpus-building procedures adhered to the standard TalkBank rules of contribution. The transcriptions were phonologically adapted into a standardized orthographic form, then coded and segmented using the Codes for Human Analysis of Transcripts (CHAT) transcription format and the Computerized Language Analysis (CLAN) (MacWhinney 2007). The recording data were transcribed by trained investigators under the supervision of expert transcribers. Each transcript was reviewed twice by a team of experienced researchers, with careful attention to speech-stream segmentation and coding.

Each transcript begins with a header section that contains basic information about the transcript itself and the speakers participating in the transcribed conversation (Figure 2). Some speakers participated in more than one conversation/transcript and are identifiable by their three-letter code. In the C-ORAL-IC corpus, it is common for three languages to appear in the same transcript: VEC (Istrovenetian/Fiuman), HRV (Croatian), and ITA (Standard Italian). The CHAT transcription system provides robust support for bilingual interactions, including the annotation of code-switching phenomena. The languages spoken by each participant are indicated using the @Languages header tier (Figure 2).



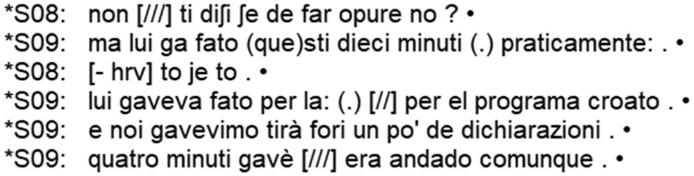
```
@Begin
@Languages:    vec, hrv
@Participants:  S08 Speaker1 Target_Adult, S09 Speaker2 Target_Adult, S10
      Speaker3 Target_Adult
@ID:   vec, hrv|Croatian-Istrovenetian/Fiuman Spoken Corpus
      (CIFSC)|S08|||||Target_Adult|||
@ID:   vec, hrv|Croatian-Istrovenetian/Fiuman Spoken Corpus (CIFSC)
      |S09|||||Target_Adult|||
@ID:   vec, hrv|Croatian-Istrovenetian/Fiuman Spoken Corpus (CIFSC)
      |S10|||||Target_Adult|||
@Birth of S08:  04-MAY-1979
@Birth of S09:  22-MAR-1984
@Birth of S10:  10-OCT-1996
@Date:  08-JAN-2020
@Time Duration: 00:14:06
@Media: 2020_05 audio
```

Figure 2. *Example of a transcript header*

The first language listed on this line is treated as the default language until a switch is indicated. Utterances that switch to a second language are marked with pre-codes, such as [- hrv] for a switch to Croatian. Individual words that switch away from the default language to the second language are marked with the @s terminator.

Time duration of the recording was 15 minutes. Header section Media shows that there is a corresponding media (audio) file to this recording available and that it has been linked to the transcript so that it can be played directly from the file.

The header is followed by the body of the transcript, which is organized into a series of rows. Each row begins with a three-letter code identifying the speaker and it represents one utterance (Figure 3). Speech was segmented into communication units (C-units – Loban 1966) based on syntactic criteria. Istrovenetian is defined as the main language of the transcript. Therefore, all C-units in Croatian are marked with the code [- hrv] at the beginning. This annotation allows for separate analysis of the two languages and supports the investigation of bilingual phenomena such as code-switching (e.g. Poropat Jeletić et al. 2021).



```
*S08:   non [///] ti diʃi ʃe de far opure no ? •
*S09:   ma lui ga fato (que)sti dieci minuti (.) praticamente: . •
*S08:   [- hrv] to je to . •
*S09:   lui gaveva fato per la: (.) [//] per el programa croato . •
*S09:   e noi gavevimo tirà fori un po' de dichiarazioni . •
*S09:   quatro minuti gavè [///] era andado comunque . •
```

Figure 3. *Example of the transcript body*

CHAT codes enable the capture of numerous morphosyntactic and discourse-structural features, such as false starts ([///]), repetitions ([/)]) or pauses ((.)). Some of these features are illustrated in Figure 3.

3.3.2. Orthographic norm

Istrovenetian lacks a universally recognized or standardized orthographic system. Therefore, a primarily phonographic transcription system was adopted for the creation and annotation of the corpus based on conventions established in previous linguistic studies. In particular, reference has been made to the models proposed by Manzini and Rocchi (1995), Pafundi (2011), Dussich (2019) and Rota (2021). Additionally, the descriptive and normative frameworks proposed by Boerio (1929), Doria (1991) and Samani (2007) were taken into account, as they provide valuable insights into orthographic variation in Istrovenetian. Although the works of Glavinić (2000), Buršić Giudici and Orbanich (2009), Filipi and Buršić Giudici (2012) and Todorović (2019, 2020, 2022) have made significant contributions to the study of Istrovenetian varieties, their linguistic atlases were not used as references for transcription because they employ a phonetic transcription system rather than a fono-orthographic approach. The transcription methodology adopted in this study primarily follows the conventions of Italian graphemic representation, though some clarifications are necessary.

Phonologically, Istrovenetian exhibits several differences from Italian (Table 5). First of all, Istrovenetian, like other Venetian dialects, shows the complete absence of consonant gemination, which is present and prescribed in Italian (Istrovenetian: *sete, tuto, pozo*; Italian: *sette, tutto, pozzo*; English: *seven, all, well*). The velar C (lat. C followed by the palatal vowel -e- or -i-) is pronounced as [ts] or [s], whereas Italian uses only [t͡ʃ] or [ʃ] in non-standard dialectal varieties. The sounds [ʃ] and [ʎ], which correspond to "sc" and "gl" in standard Italian, are entirely absent in Istrovenetian. Moreover, unlike Italian, Istrovenetian (as well as Venetian) palatalizes the Latin cluster -CL- [kl] as the affricate [t͡ʃ], whereas Italian renders it as [kj]. Istrovenetian completely lost the distinctiveness of closed and open vowels *o* and *e*, moving from a system of seven vowels to a system of five. The dialect has twenty-six phonemes: twenty-one consonantal and five vocalic. In the transcripts used in this research, the grapheme *s* represents the voiceless consonant (e.g., sufiàr = 'to blow'), while the grapheme ʃ is used for the voiced counterpart (e.g., ʃburton = 'thrust'). It is considered useful, for illustrative purposes, to report also the difference in the Istrovenetian dialect between the lemmas *caseta*=box and *caʃeta*=small house.

Table 5. *Schema of Istrovenetian/Fiuman vowel and consonantal sounds*

Grapheme	Phoneme	Explanations	Example in Istrovenetian and English translation
a	/a/	-	alba = sunrise
b	/b/	-	boto = barrel
c	/tʃ/	c + e, i	cicara = cup

	/k/	c + a, o, u, C
		ch + e, i
			cocal = seagull
			cheba = cage

d	/d/	-	dameiana = carboy
e	/e/	-	erba = grass
f	/f/	-	fameia = family
g	/dʒ/	g + e, i	girar = to turn
g	/g/	g + a, o, u, C	goma = gum
gn	/ɲ/	-	gnora = daughter-in-law
h	-	Aspirated	-
i	/i/	-	incarigar = to load
l	/l/	-	luganega = sausage

m	/m/	-
		When followed by the bilabial stops *b* and *p*, the bilabial nasal *m* changes to *n*
			man = hand
			canpo = field

n	/n/	-	netar = to clean
o	/o/	-	ocio = eye
p	/p/	-	paion = straw-mattress
q	/k/	-	quartier = neighborhood
r	/r/	-	recia = ear
s	/s/	-	scarsela = pocket
s’c	/s/ + /tʃ/	sk + e, i	s’cenʃa = splinter

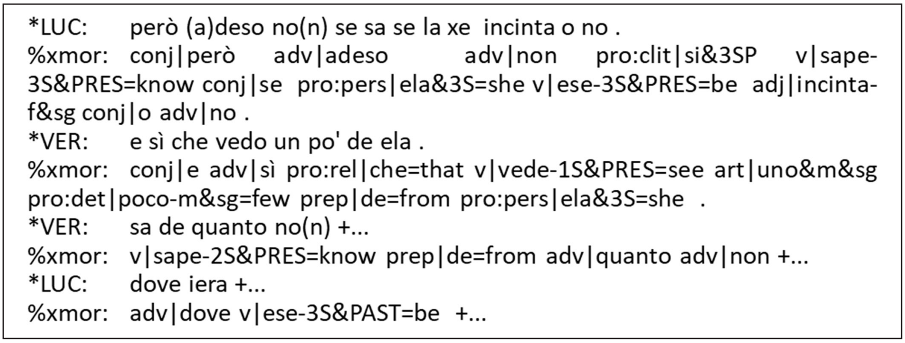
ʃ	/z/	s + b, d, g, l, m, n, v
		in third person singular present indicative form of the verb “to be” (eser)
			ʃbregar = to tear off
			ʃe = is

t	/t/	-	tovaia = tablecloth

Grapheme	Phoneme	Explanations	Example in Istrovenetian and English translation
u	/u/	-	uʃel = bird
v	/v/	-	vendema = grape harvest
z	/ts/	The voiceless [ts] and voiced [dz] dental affricates are foreign to the historical Istrovenetian dialect. When they appear in loanwords from Italian language, they are regularly replaced by one of the Istrian fricative sounds: e.g. ʃero = zero, suchero = zucchero.	-
	/dz/		

3.3.3. Morphological coding

Each word in the corpus was assigned a corresponding lemma and morphological description. This tagging process began with the use of automated tools, specifically the CLAN and the Italian MOR tool (Figure 4). However, given the significant linguistic differences between the dialect and the standard language, a substantial portion of the data required manual intervention.



```
*LUC:	però (a)deso no(n) se sa se la xe  incinta o no .
%xmor:	conj|però    adv|adeso          adv|non     pro:clit|si&3SP     v|sape-3S&PRES=know conj|se  pro:pers|ela&3S=she v|ese-3S&PRES=be  adj|incinta-f&sg conj|o adv|no .
*VER:	e sì che vedo un po' de ela .
%xmor:	conj|e adv|sì pro:rel|che=that v|vede-1S&PRES=see art|uno&m&sg pro:det|poco-m&sg=few prep|de=from pro:pers|ela&3S=she  .
*VER:	sa de quanto no(n) +...
%xmor:	v|sape-2S&PRES=know prep|de=from adv|quanto adv|non +...
*LUC:	dove iera +...
%xmor:	adv|dove v|ese-3S&PAST=be  +...
```

Figure 4. *Example of a morphologically coded transcript body*

3. Results and discussion

This paper presents the main structural and methodological features of the Corpus of Spoken Istrovenetian/Fiuman and Croatian (C-ORAL-IC), which documents the synchronic local forms of interactive exchange and conventional communicative behaviors in everyday bilingual conversation. As the final product, C-ORAL-IC encompasses language samples and sociolinguistic information of 87 multilingual speakers from bilingual areas of Istria and Rijeka. These participants produced 85.449 tokens and 27.495 types. All transcripts are morphologically coded.

C-ORAL-IC encompasses language samples from a diverse group of speakers varying in age, gender, place of residence, and socioeconomic status. Despite these differences, all participants represent the bilingual community of Istria and Rijeka. Their language use reveals a complex dynamic involving the Croatian standard language, Croatian dialects, Italian standard language, and Italian dialects (specifically, the two diatopic Venetian subvarieties: Istrovenetian and Fiuman) within the community. All participants use at least two varieties daily (one Croatian and one Italian) for various activities, and the vast majority regularly use three or even all four varieties (90%). For the speakers in C-ORAL-IC, the two diatopic subvarieties of Venetian serve as a long-standing 'lingua franca', mother tongue, and the primary mode of informal daily communication within the Italophone population (Blagoni et al. 2016, 2018; Milani Kruljac 2003).

The C-ORAL-IC is a valuable resource for exploring different issues relevant to linguistic, sociolinguistic and applied psycholinguistic research. The broader range of metadata conventions collected within this project enables the development of a more nuanced and detailed investigation of contact-induced innovation phenomena on the structural and discourse level, thus allowing for a deeper understanding of the sociolinguistic context of the specific bilingual speech community. Specifically, several main points of interest can be explored: code-switching[3], language change, lan-

[3] As stated by Gumperz: "[c]onversational code switching can be defined as the juxtaposition within the same speech exchange of passages of speech belonging to two different grammatical systems or subsystems" (1982:59), in our case the Croatian language (rarely the Chakavian dialect) and the Istrovenetian/Fiuman dialect, used by highly proficient bilingual/multilingual speakers (Myers-Scotton 1997) with collaborative communicative endeavors. The most useful way for studying bilingual practices is to investigate its communicative effects. Therefore, the C-ORAL-IC offers a great source for that goal.

guage features and the relationship between language and extralinguistic parameters.

For instance, we can investigate language change (i.e. the use of the dialect within different generations of bilingual speakers), language features (such as word frequency and usage of different morphological forms, among others) and extralinguistic features (age, gender, language exposure, education, socio-economic status etc.) and their impact on language.

Spoken language corpora of bilingual speakers are relatively rare and include limited populations (e.g. specific interest on language acquisition corpora) and pairs of languages. The design of the bilingual corpus for Istrovenetian/Fiuman speakers in Croatia could be adapted to other bilingual communities with similar diglossic features. For instance, the German-speaking community in South Tyrol, Italy, where standard German is used formally and Tyrolean dialects informally, mirrors the Istrian diglossic situation (Lanthaler 2001). Other examples include the Catalan community in Catalonia (Miller and Miller 1996), the Occitan community in southern France (Blanchet and Harold 2004), and so forth.

4. Conclusion

This paper presents the Croatian-Istrovenetian/Fiuman Spoken Corpus, which documents language use within a bilingual community in Croatia speaking the macro-regional minority diatopic Venetian varieties: Istrovenetian and Fiuman. It offers a detailed account of the corpus design, including language sampling, transcription and coding methods, participant selection, and their sociolinguistic profiles. The corpus is made available to researchers interested in this specific population, as well as those studying bilingualism more broadly, with information on how to access it. Additionally, the paper suggests several potential research applications to inspire future users of the corpus.

References

Blagoni, Robert; Poropat Jeletić, Nada; Blecich, Kristina (2016) "The Italophone Reefs in the Croatophone Sea", *Bilingual Landscape of the Contemporary World*, ed. Grucza, S.; Olpińska-Szkiełko, M.; Romanowski, P., Peter Lang Verlag, Warszawa – Wien, pp. 11–36.

Blagoni, Robert; Poropat Jeletić, Nada; Blecich, Kristina (2018) “Istroveneto e italofonia in Istria: prospettive e visioni di un’insularità etnolinguistica”, *Le isole linguistiche dell’Adriatico*, ed. Šimičić, L.; Škevin, I.; Vuletić, N., Gioacchino Onorati Editore, Roma, pp. 69–92.

Blanchet, Philippe; Schiffman, Harold (2004) “Revisiting the sociolinguistics of Occitan: a presentation”, *International Journal of the Sociology of Language*, 169, pp. 3–24.

Boerio, Giuseppe (1929/1856) *Dizionario del dialetto Veneziano*, Andrea Santini e figlio, Venezia.

Buršić Giudici, Barbara; Orbanich, Giuseppe (2009) *Dizionario del dialetto di Pola*, Centro di ricerche storiche, Trieste – Rovigno.

Čermak, František (2009) “Spoken corpora design”, *International Journal of Corpus Linguistics*, 14(1), pp. 113–123.

Dal Negro, Silvia (2013) “Dealing with bilingual corpora: Parts of speech distribution and bilingual patterns”, *Revue Française de Linguistique Appliquée*, 18(2), pp. 15–28.

Doria, Mario (1991) *Grande dizionario del dialetto triestino storico etimologico fraseologico*, Edizioni “Trieste Oggi”, Trieste.

Dussich, Marino (2019) *Dizionario Italiano-Buiese*, Centro di ricerche storiche, Rovigno – Trieste.

Filipi, Goran; Buršić Giudici, Barbara (2012) *Istromletački lingvistički atlas/ Atlante Linguistico Istroveneto/Istrobeneški lingvistični atlas*, Naklada Nediljko Dominović, Zagreb.

Glavinić, Vera (2000) *Vocabolario del dialetto istroveneto di Pola*, Filozofski fakultet, Pula.

Grosjean, Francois (2001) “The bilingual’s language modes”, *One Mind, Two Languages*, ed. Nicol, J., Blackwell, Oxford, pp. 1–22.

Gumperz, John Joseph (1982) *Discourse strategies,* Cambridge University Press, Cambridge.

Haviland, John Beard (1986) “Pointing, Gesture Spaces, and Mental Maps”, *Language*, pp. 231–246.

Hunt, Kellogg (1966) “Recent measures in syntactic development”, *Elementary English*, 43, pp. 732–739.

Kuvač Kraljević, Jelena; Hržica, Gordana (2016) “Croatian Adult Spoken Language Corpus (HrAL)”, *Fluminensia*, 28(2), Rijeka, pp. 87–102.

Labov, William (1972) *Sociolinguistic Patterns*, University of Pennsylvania Press, Philadelphia.

Lanthaler, Franz (2001) "Zwischenregister der deutschen Sprache in Südtirol", *Die deutsche Sprache in Südtirol. Einheitssprache und regionale Vielfalt*, ed. Egger, K.; Lanthaler, Franz, Folio Verlag, Wien-Bozen, pp. 137–152.

Lisac, Josip (2009) *Hrvatska dijalektologija 2. Čakavsko narječje*, Golden marketing – Tehnička knjiga, Zagreb.

Loban, Walter (1966) *Language Ability: Grades Seven, Eight, and Nine*. Washington, DC: Government Printing Office.

Lukežić, Iva (1990) *Čakavski ikavsko-ekavski dijalekt*, Izdavački centar Rijeka, Rijeka.

MacWhinney, Brian (2000) *The CHILDES Project: Tools for Analyzing Talk*. Third Edition. Mahwah, Lawrence Erlbaum Associates, New York.

MacWhinney, Brian (2007) "The TalkBank Project", *Creating and Digitizing Language Corpora: Synchronic Databases, Volume 1*, ed. Beal, J. C.; Corrigan, K. P.; Moisl, H. L., Palgrave-Macmillan, Houndmills, pp. 163–180.

MacWhinney, B. (2019) "TalkBank and SLA", *The Handbook of SLA and Corpora*, ed. Tracy-Ventura, N.; Paquot, M., Routledge, New York.

Mair, Christian (2009) "Corpus linguistics meets sociolinguistics: The role of corpus evidence in the study of sociolinguistic variation and change", *Language and Computers,* 69(1), pp. 7–32.

Małecki, Mieczysław (2002) *Slavenski govori u Istri*, Hrvatsko filološko društvo – Graftrade, Rijeka.

Małecki, Mieczysław (2007) *Čakavske studije*, Maveda, Rijeka.

Manzini, Giulio; Rocchi, Luciano (1995) *Dizionario storico fraseologico etimologico del dialetto di Capodistria*, Centro di ricerche storiche, Trieste – Rovigno.

Milani Kruljac, Nelida (1990) *La Comunità Italiana in Istria e a Fiume fra diglossia e bilinguismo*, Centro di Ricerche storiche, Rovigno.

Milani Kruljac, Nelida (ed.) (2003) *L'italiano fra i giovani dell'Istroquarnerino*, Pietas Iulia – Edit, Pola – Fiume.

Miller, Henry; Miller, Kate (1996) "Language policy and identity: the case of Catalonia", *International Studies in Sociology of Education*, 6(1), pp. 113–128.

Myers-Scotton, Carol (1997) "Code-switching", *The Handbook of Sociolinguistics*, ed. Coulmas, F., Oxford, pp. 217−237.

Pafundi, Nicola (2011) *Dizionario fiumano-italiano, italiano-fiumano*, Associazione Libero Comune di Fiume in Esilio, Padova.

Poropat Jeletić, Nada (2015) "Italian Language in Istria: Status Planning, Corpus Planning and Acquisition Planning", *Mediterranean Journal of Social Sciences*, 6(2), pp. 385–392.

Poropat Jeletić, Nada (2017a) "Italofona dijasistemska raslojenost u hrvatskoj Istri: jezični i komunikcijski status, korpus i prestiž", *Annales – Anali za Istrske in Mediteranske Studije – Series Historia et Sociologia*, 27(1), pp. 191–204.

Poropat Jeletić, Nada (2017b) "O hrvatsko-talijanskoj dvojezičnosti u Istri i ishodima jezične doticajnosti", *Annales – Anali za Istrske in Mediteranske Studije – Series Historia et Sociologia*, 27(3), pp. 629–639.

Poropat Jeletić, Nada; Moscarda Mirković, Eliana; Bortoletto, Anna (2021) "Incidenza e implicazioni di alcuni tratti formali pertinenti tipici del discorso bilingue istriano: i casi di commutazione di codice". *Annales – Anali za Istrske in Mediteranske Studije – Series Historia et Sociologia*, 31(2), pp. 329–340.

Poropat Jeletić, Nada; Moscarda Mirković, Eliana; Hržica, Gordana (2024) *BilingBank C-ORAL-IC Corpus*. https://biling.talkbank.org/access/C-ORAL-IC.html

Rosamani, Enrico (1999) *Vocabolario giuliano dei dialetti parlati nella Venezia Giulia, in Istria, in Dalmazia, a Grado e nel Monfalconese*, Lint, Trieste.

Rota, Vlado (2021) *Vocabolario del dialetto di Umago e del suo territorio*, Comunità degli Italiani "Fulvio Tomizza", Umag.

Ruhi, Şukriye; Işik Taş, Elvan Eda (2014) "Constructing General and Dialectical Spoken Corpora for Language Variation Research: Two Case Studies for Turkish". In: Ruhi, Ş.; Haug, M.; Schmidt, T.; Wörner, K. (ed.) *Best Practices for Spoken Corpora in Linguistic Research*, Cambridge Scholar Publishing, Newcastle upon Tyne, pp. 36–56.

Samani, Salvatore (2007) *Dizionario del dialetto fiumano I–III*, Società di Studi Fiumani, Roma.

Scotti Jurić, Rita (2001) "Il dialetto istroveneto: può diventare un problema?", *Tabula*, 4, Pula.

Sinclair, John (1991) *Corpus Concordance Collocation*, Oxford University Press, Oxford.

Todorović, Suzana (2019) *Istrskobeneški jezikovni atlas severozahodne Istre. 1, Vremenske razmere, geomorfologija, običaji in institucije, telo in bolezni = Atlante linguistico istroveneto dell'Istria nordoccidentale. 1, Fenomeni atmosferici, configurazione del terreno, tradizioni ed istituzioni, corpo e malattie*, Libris – Unione Italiana, Koper/Capodistria.

Todorović, Suzana (2020) *Istrskobeneški jezikovni atlas severozahodne Istre. 2, Števniki in opisni pridevniki, čas in koledar, življenje, poroka in družina, hiša in posestvo = Atlante linguistico istroveneto dell'Istria nordoccidentale. 2, Numerali e aggettivi qualificativi, scorrere del tempo e calendario, vita, matrimonio e famiglia, casa e podere*, Libris – Unione Italiana, Koper/Capodistria.

Todorović, Suzana (2022) *Istrskobeneški jezikovni atlas severozahodne Istre. 3, Garderoba in dodatki, hrana in pijača, čustva in občutki, oljkarstvo in oljarstvo, perjad, zelenjava, sadje in sadno drevje, živali = Atlante linguistico istroveneto dell'Istria nordoccidentale. 3, Vestiario e accessori, cibi e bevande, sentimenti ed emozioni, olivicoltura e torchiatura, pollame, verdura, frutta e alberi da frutto, animali*, Libris – Unione Italiana, Koper/Capodistria.

Tognini-Bonelli, Elena (2001) *Corpus Linguistics at Work*, John Benjamins, Amsterdam – Philadelphia.

Ujčić, Rudolf (2015) *Istarske čakavske dijalektološke teme*, Matica Hrvatska, Pula.

SAŽETAK

Nada Poropat Jeletić, Gordana Hržica, Eliana Moscarda Mirković

KORPUS GOVORENOG ISTROMLETAČKOG/FIJUMANSKOG I HRVATSKOG (C-ORAL-IC)

Dvojezični konverzacijski korpusi od neprocjenjive su važnosti za proučavanje autentičnih jezičnih kontakata u spontanom dvojezičnom govoru. Ovaj rad predstavlja Korpus govornog istromletačkog/fijumanskog i hrvatskog jezika (C-ORAL-IC), prvi korpus koji dokumentira spontane razgovore na istromletačkom i fijumanskom dijalektu dvojezičnih osoba iz istarskog i kvarnerskog područja Hrvatske. Regija ima dugu povijest kulturne i jezične interakcije hrvatskog i talijanskog jezika, što je oblikovalo složen sociolingvistički sustav s obilježjima diglosije i poliglosije.

Korpus C-ORAL-IC obuhvaća podatke 87 dvojezičnih i višejezičnih govornika, s više od 85.000 pojavnica i 27.000 različitih riječi (oblićnica). Dostupan je na platformi TalkBank, u podsekciji BilingBank (https://talkbank.org, https://biling.talkbank.org/access/C-ORAL-IC.html). Sadrži transkribirane, fonološki prilagođene, kodirane, segmentirane i morfološki označene transkripte uparene sa snimkama. Dostupni su i dodatni podaci o jeziku i upotrebi jezika sudionika.

C-ORAL-IC predstavlja bogat izvor za proučavanje spontanoga dvojezičnog govora te nudi uvid u konverzacijska obilježja, strukturu i sinkrone promjene u istromletačkom i fijumanskom.

Ključne riječi: *jezično uzorkivanje; govoreni korpusi; kodno prekljućivanje; dvojezični diskurs*