Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

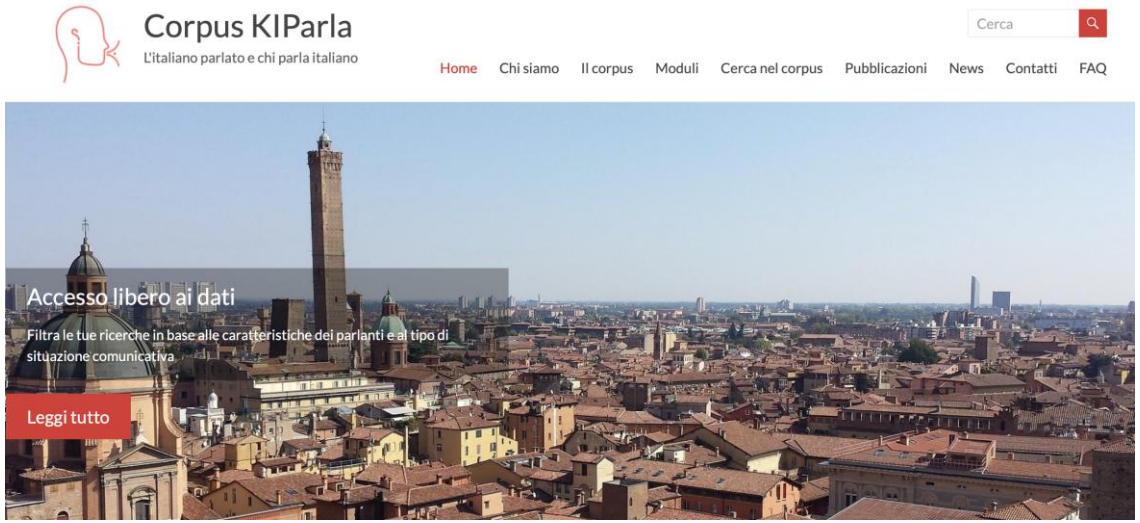
Stra-ParlaBO: documenting Italian spoken by plurilingual communities in the city of Bologna

Caterina Mauri, Eleonora Zucchini
University of Bologna, Masaryk University

Developing new languages in migration contexts. Brno, 30 September – 1 October 2025

** The research leading to these results has received funding from Project "DiverSIta-Diversity in spoken Italian", prot. P2022RFR8T, CUP J53D23017320001, funded by EU in NextGenerationEU plan through the Italian "Bando Prin 2022 - D.D. 1409 del 14-09-2022"*

The KIParla corpus



The KIParla is a corpus of spoken Italian that offers:

- **freely accessible** transcripts aligned with audio files

www.kiparla.it

- **search interface** based on an **international standard on** *NoSketchEngine*

(Mauri et al. 2019)

- **transparent metadata system** regarding conversations and speakers

- **modular, incremental, replicable infrastructure** allowing for the expansion of the corpus over time

Modularity and incrementality

KIParla is made of different **modules**, small/medium sized **corpora**

- built for different, often complementary **purposes**
- data from different geographical areas, types of interaction and types of communities
- while maintaining **substantial underlying comparability in structure and accessibility**.

The **comparability** and at the same time the **specificity** of each module are what can make the KIParla corpus **representative of spoken Italian over time**:

the more modules are added

the more dimensions of variation can be explored

Data collection and transcription

Most of the data have been collected and **transcribed** by students.

2018 – today: around 100 students (universities of Bologna and Turin) took part in the corpus building.

→ **internship**

Training

- Selected books and papers;
- Training sessions;
- Trial transcription and feedback.

Supervision

- Weekly meetings;
- Frequent updates (via e-mail).

Data transcription

Data are **manually** transcribed using ELAN.

→ flexible system that incorporates both orthographic and conversational conventions (from Jefferson, 2004).

→ Very limited set of clear rules shared among all transcribers.

All transcriptions are then revised (for consistency and anonymization) by a single person.

Symbol	Meaning
,	Slight rising intonation
?	Sharp rising intonation
.	Final falling intonation
:	Prolonged vowel or consonant (one or two colons common, three or more colons only in extreme cases)
(.)	Micropause
~	Interrupted word
=	End of one transcription unit and beginning of next begin with no gap/pause in between
°hello°	Syllables or words distinctly quieter than surrounding speech by the same speaker
HELLO	Syllables or words louder than surrounding speech by the same speaker
<hello>	Decreased speaking rate
>hello<	Increased speaking rate
[hello]	Speech overlaps
(hello)	Uncertain syllables or words
x	Inaudible syllable
((laughs))	Non-verbal behavior

KIParla corpus today

Four modules:

- KIP: educated speakers, different interactional settings (69 hours)
- KIPasti: kitchen-table conversations, different regions (42 hours)
- ParlaTO and ParlaBO: semi-structured interviews, speakers with different profiles (49 and 65 hours)

Corpus size:

- Over 200 hours
- 2.328.193 tokens

Two new modules - *Stra-ParlaBO* and *Stra-ParlaTO*:

- people with a background of migration
- aim: 128 hours of recording

KIParla corpus today

Four modules:

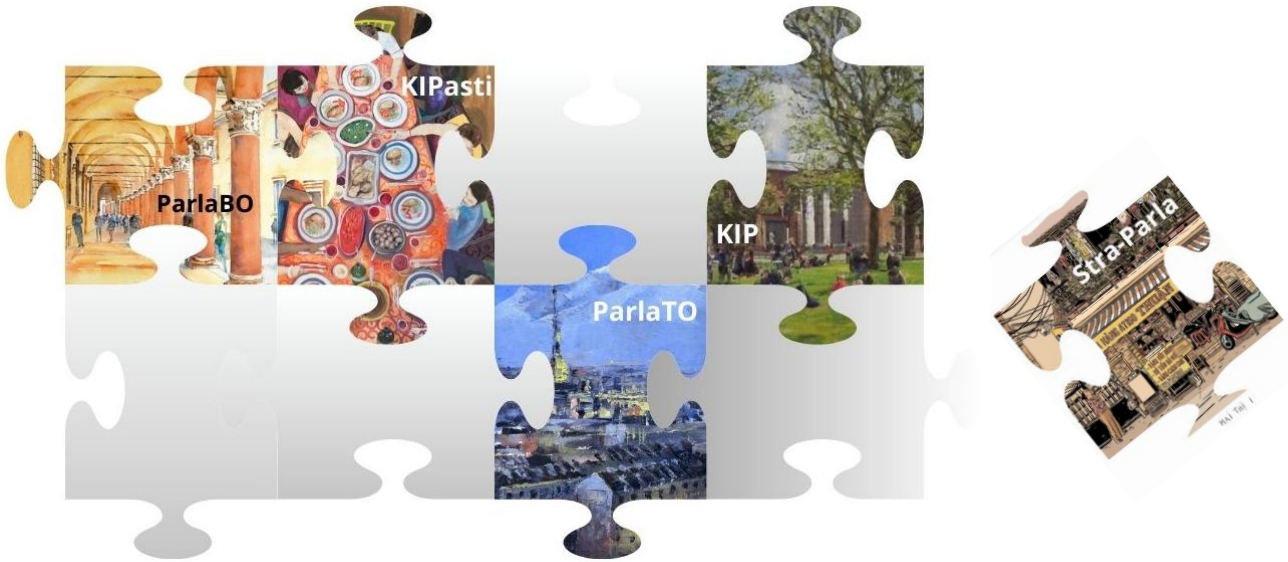
- KIP: educated speakers, different interactional settings (69 hours)
- KIPasti: kitchen-table conversations, different regions (42 hours)
- ParlaTO and ParlaBO: semi-structured interviews, speakers with different profiles (49 and 65 hours)

Corpus size:

- Over 200 hours
- 2.328.193 tokens

Two new modules - ***Stra-ParlaBO*** and *Stra-ParlaTO*:

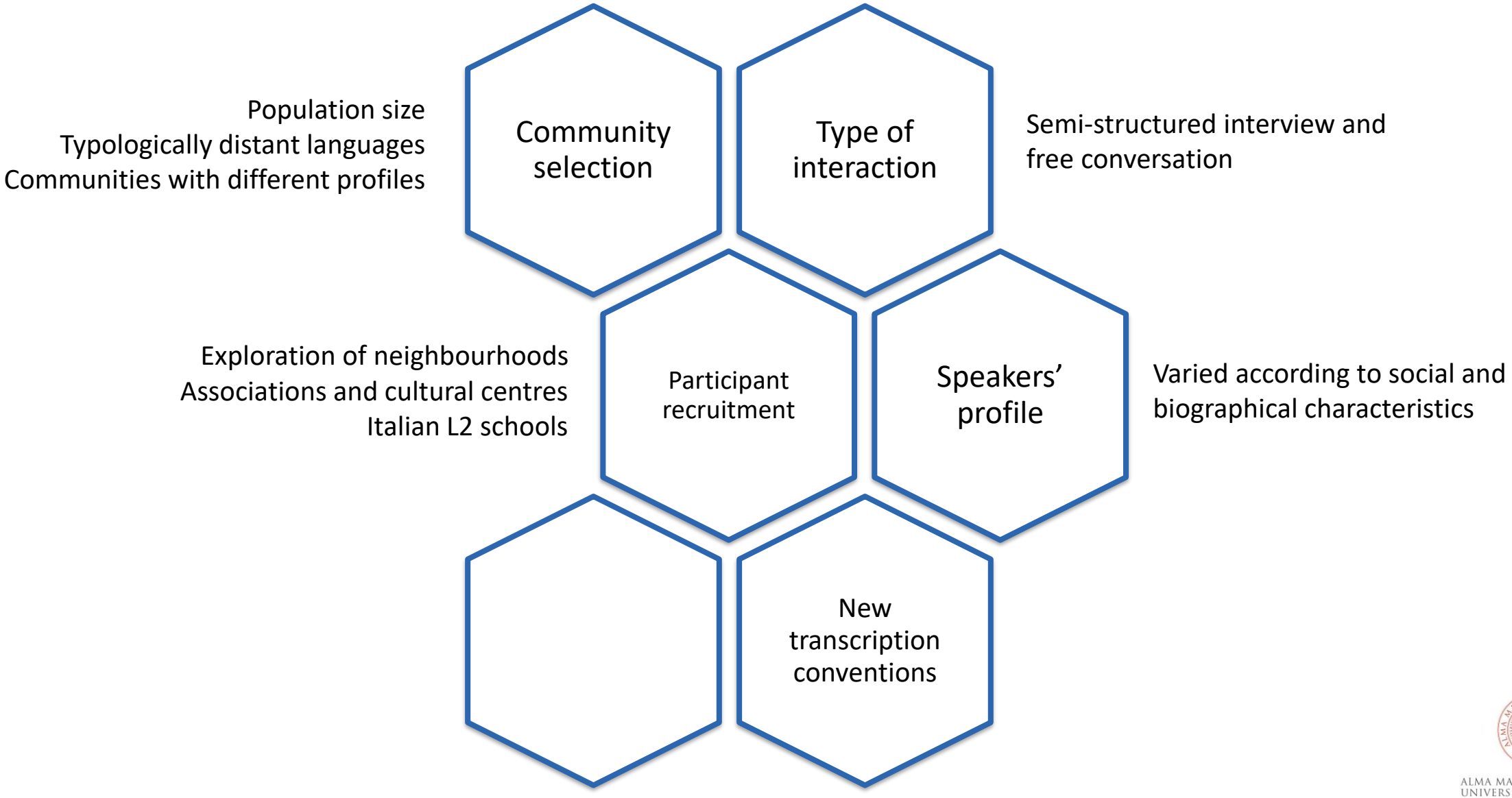
- people with a background of migration
- aim: 128 hours of recording



Aims of the *Stra-Parla* modules

- Better represent urban contexts, which are multicultural
- Wider look on linguistic phenomena, compare learners and native speakers' varieties (similarities and differences)
- Investigate the role of learners and speakers with plurilingual repertoires on the emergence of innovative constructions

Stra-ParlaBO: corpus design



Stra-ParlaBO: some details

Communities involved:

Ukranian

Moroccan

Chinese

Bangladeshi

Stra-ParlaBO: some details

Communities involved:

Ukranian

Moroccan

Chinese

Bangladeshi

Target:

- 8 hours of semi-structured interviews and
- 8 hours of free conversation for each community

Total: 64 hours of recording

Stra-ParlaBO: some details

Communities involved:

Ukranian

Moroccan

Chinese

Bangladeshi

Target:

- 8 hours of semi-structured interviews and
- 8 hours of free conversation for each community

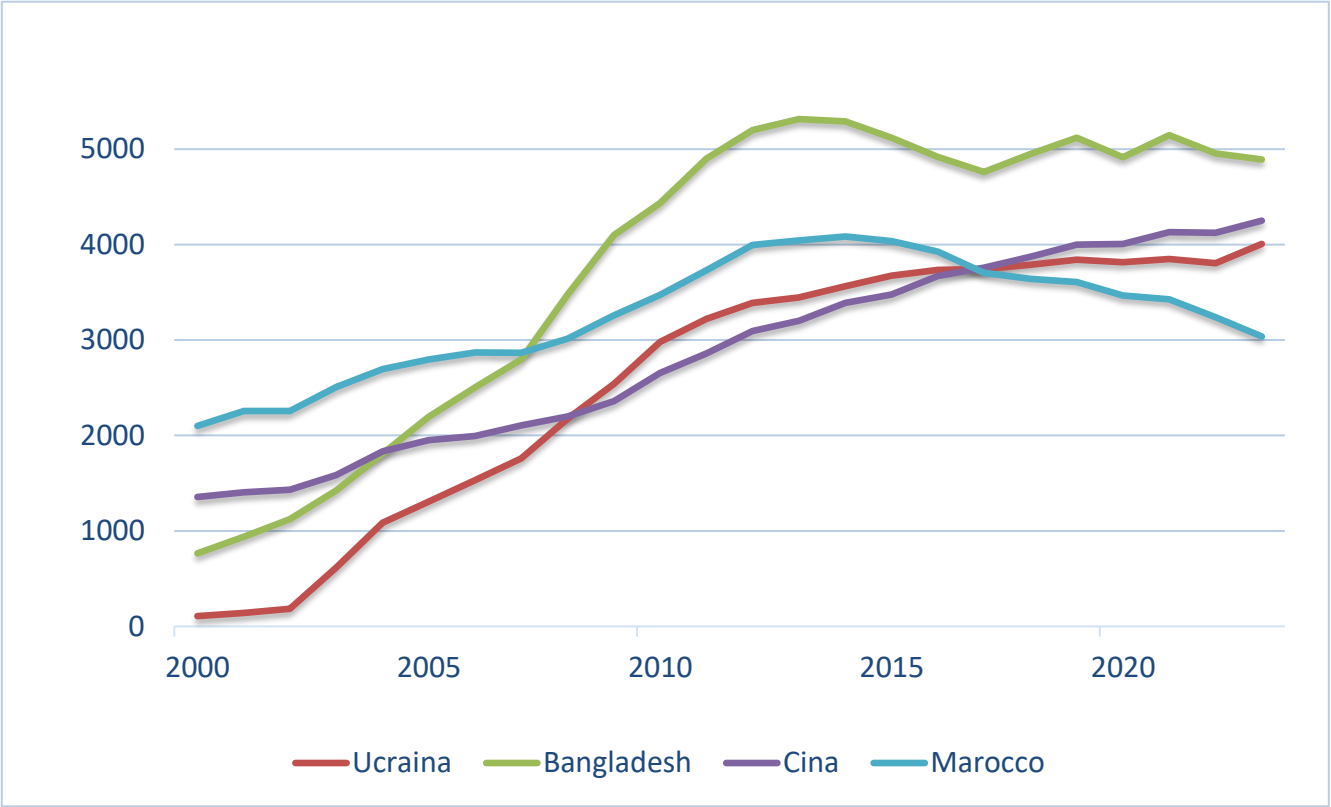
Total: 64 hours of recording

Metadata required:

- Age, gender
- Profession
- Degree
- Mother tongue
- Age at arrival in Italy (5-year ranges)
- Years spent in Italy (2-year ranges)
- Italian learned in academic setting (yes/no)

New and old waves of migration in Bologna

Year	Ukraine	Bangladesh	China	Morocco
2000	108	756	1.356	2.100
2023	4890	3039	4.251	4.007



Dati da [Home | I numeri di Bologna (inumeridibolognametropolitana.it)]

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Data collection methodology: semi-structured interviews

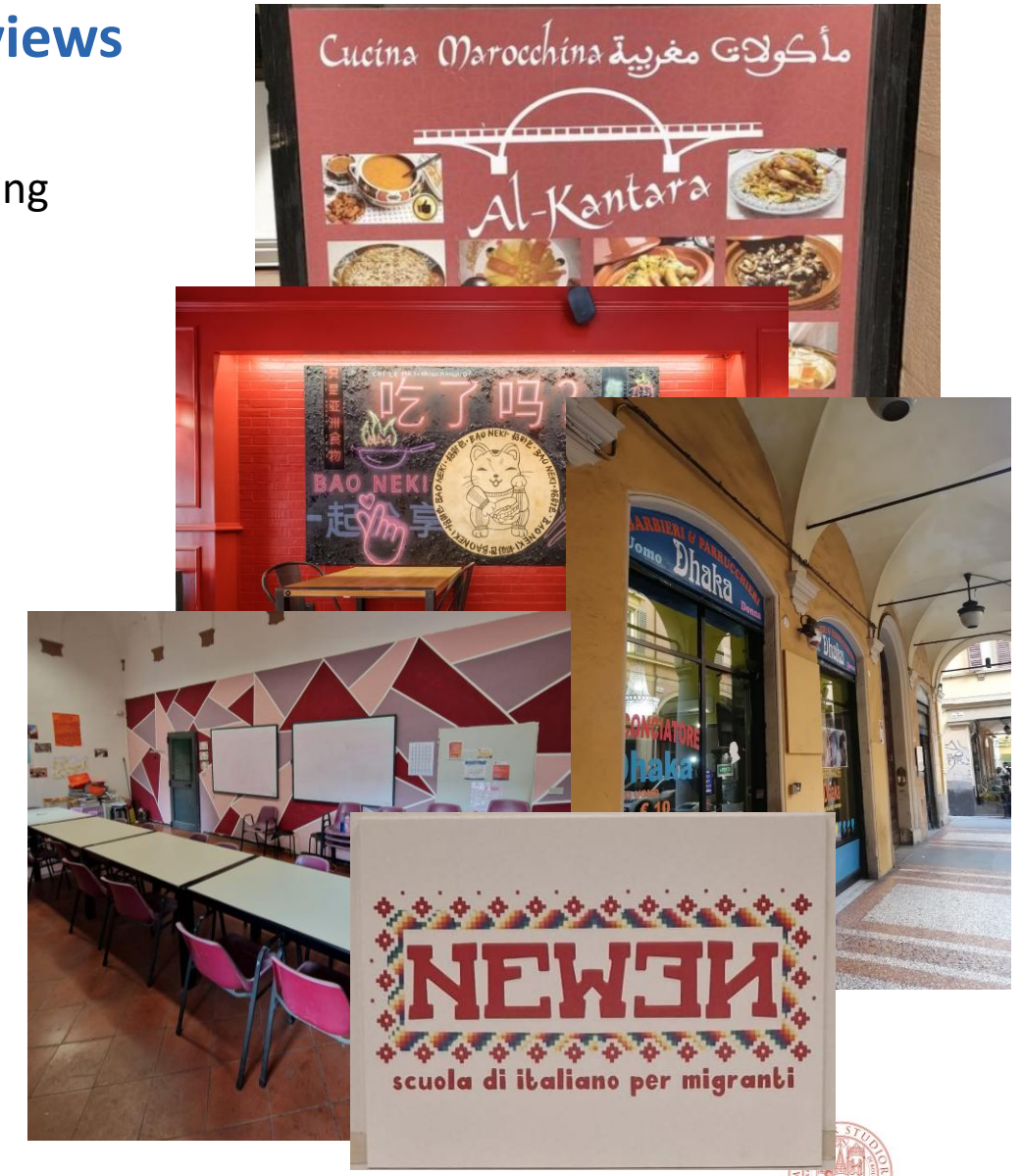
Interviews conducted with Labovian methodology (Labov 1984), adapting the interview outline used for ParlaBO module:

- Culture of origin
- Language practices
- Learning of Italian

Participant rectruitment:

1. Personal contacts (friend of a friend)
2. Exploration of the city neighbourhood where community members have shops and other activities
3. Contacts with association and participation to events

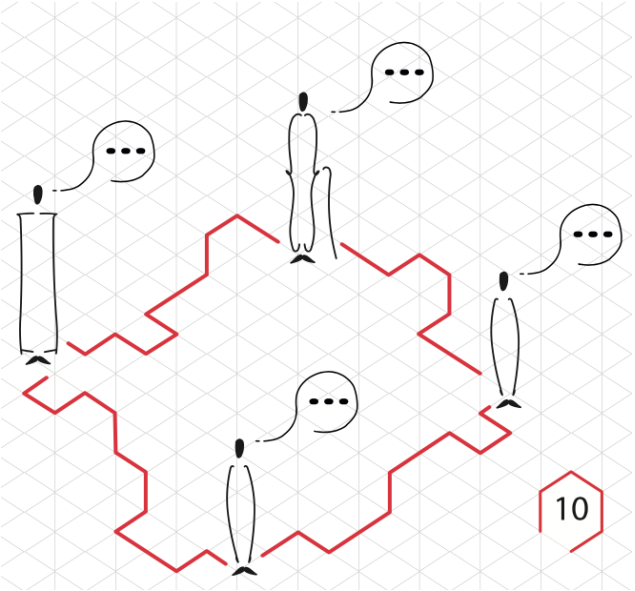
Important to reach people outside personal social network



Data collection methodology: free conversations

Main features (and criteria):

- Communication event outside research design (naturally-occurring)
- Conversation topic undefined
- No predictable asymmetry between participants
- Language shouldn’t be imposed



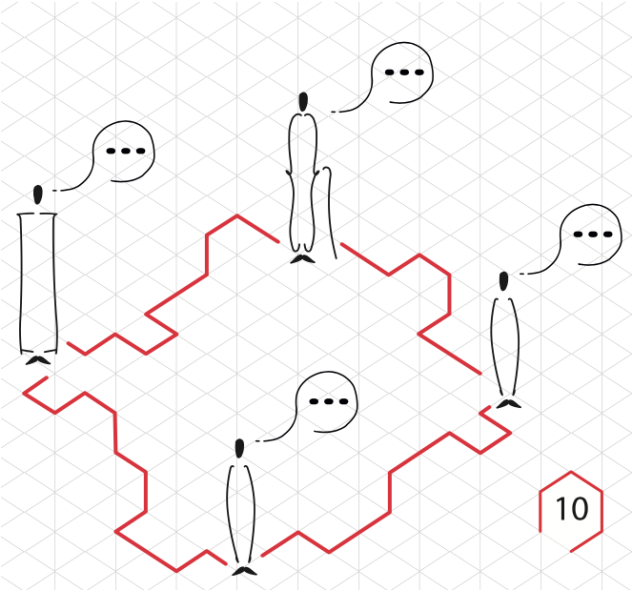
Data collection methodology: free conversations

Main features (and criteria):

- Communication event outside research design (naturally-occurring)
- Conversation topic undefined
- No predictable asymmetry between participants
- Language shouldn’t be imposed

→ Contexts where Italian was used spontaneously

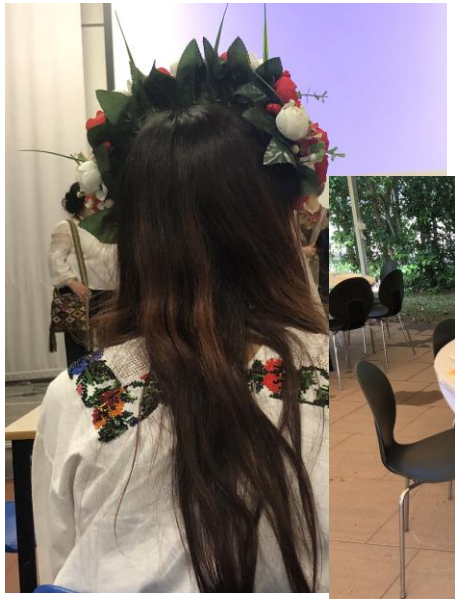
→ Often collected by community members, self-recording



Data collection methodology: free conversations

Key strategies:

- Contacts and friendships with community members
- Participant observation during events and activities



Data collection methodology: free conversations

Complications:

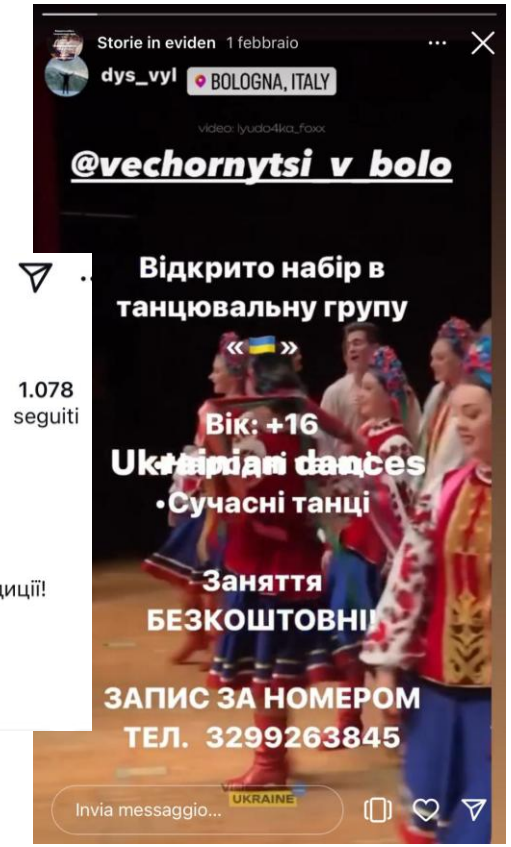
- Communities are not homogeneous
- Many associations are informal and non-istitutionalised

→ They are invisible (no official hq, no website ...)

- Comunication is conducted in language (and alphabet) of origin

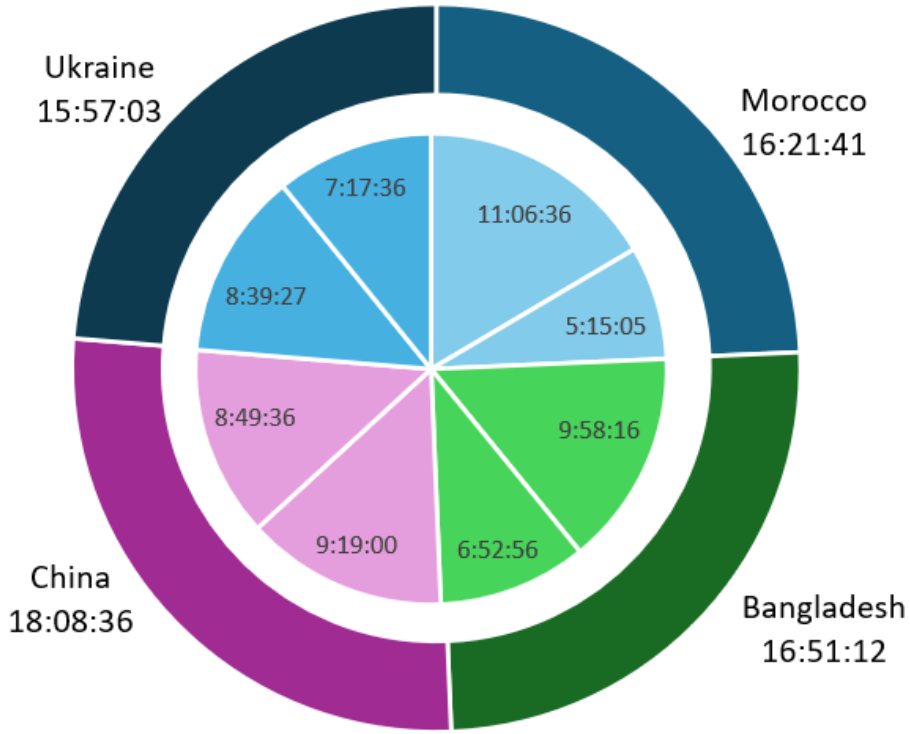
- Active involvement of community members is very hard because research aims are incomprehensible

- Research requirements clash with real-life informal scenarios:
  - Conversation duration
  - Unstable participants' configuration
  - Language choice (*plus* awareness of language practices)

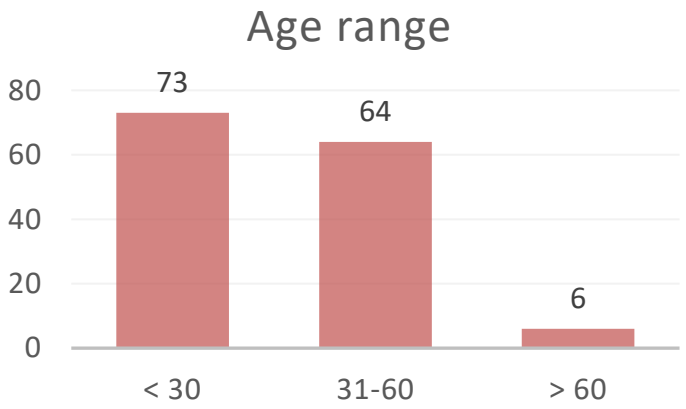


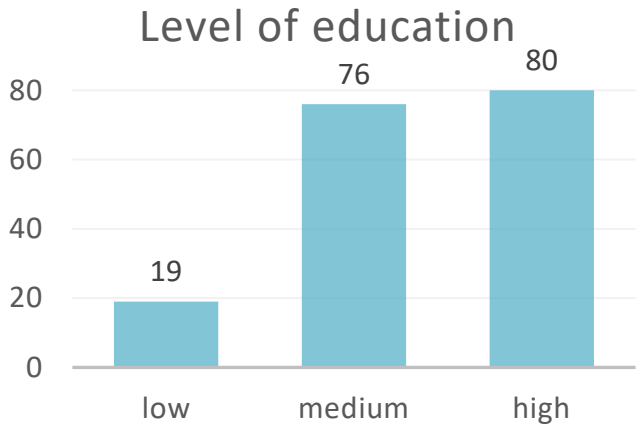
StraParlaBO: corpus size

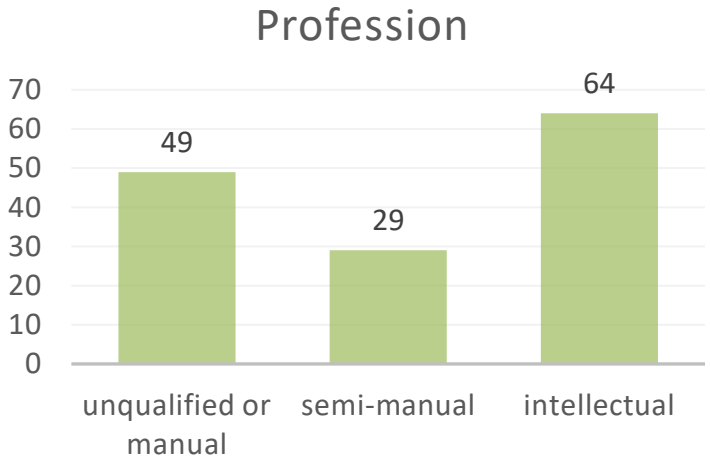
Community	Interviews	Free conversation	Total
Moroccan Arabic	11:06:36	5:15:05	**16:21:41**
Bangla	9:58:16	6:52:56	**16:51:12**
Chinese	9:19:00	8:49:36	**18:08:36**
Ukranian	8:39:27	7:17:36	**15:57:03**
Total	**39:03:19**	**28:15:13**	**67:18:32**

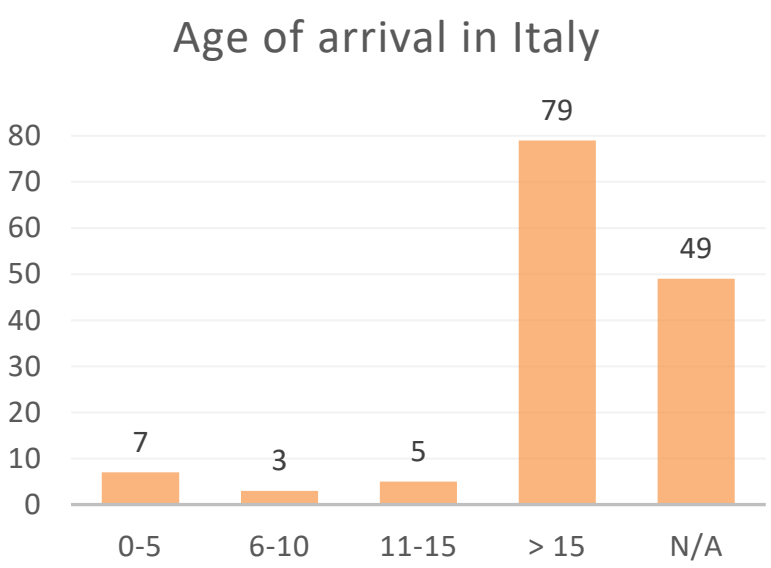


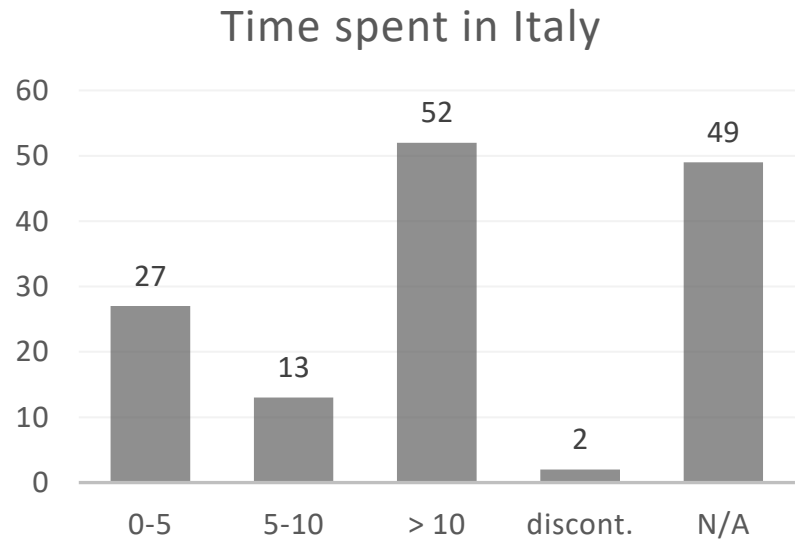
StraParlaBO: participants involved











Tot: 143 participants

Transcription: problems and solutions

KIParla methodology:

– **Main goal**: balance between accuracy and ‘searchability’ (homogeneity and processing)
– systematically disregard phonetic variation and **transcribe** phenomena that involve **morphology**

New modules only sharpened pre-existing problems deriving from informal and spontaneous communication contexts

Main issues:

- Code-switching
- Language contact
- Inexistant or unclear productions

Transcription: code-switching

(1) Free conversation, Morocco (SBCA006)

PSB125 [#had] #l3am #had #l3am #inshallah #nhabto #lmaghrib @sara #ghadi #dir #liha matrimonio casa una festic~ [una festi~ festicciola]

this year if god wants we go to morocco and sara will have her wedding at home a small party

PSB126 [#_wash e:h?] // #_wash fel maghreb oula [fi talien?]

but // but in morocco or in italy?

PSB125 [no #fi] #lmaghreb #fi #talien già #darto #fi #talien #talien [già #darto]

no in morocco in italy she had it already in morocco

PSB126 #_[lmaghrib] // #fi #dar #dial casa #tata @miriam [#oula]

in morocco // in the house of untie miriam or

PSB125 eh // casa #dial #tatak @miriam // una piccola fe~ fe:sta

eh // house of your untie miriam // a small party

Transcription: code-switching

(1) Free conversation, Morocco (SBCA006)

PSB125 [#had] #l3am #had #l3am #inshallah #nhabto #lmaghrib @sara #ghadi #dir #liha matrimonio casa una festic~ [una festi~ festicciola]

this year if god wants we go to morocco and sara will have her wedding at home a small party

PSB126 [#_wash e:h?] // #_wash fel maghreb oula [fi talien?]

but // but in morocco or in italy?

PSB125 [no #fi] #lmaghreb #fi #talien già #darto #fi #talien #talien [già #darto]

no in morocco in italy she had it already in morocco

PSB126 #_[lmaghrib] // #fi #dar #dial casa #tata @miriam [#oula]

in morocco // in the house of untie miriam or

PSB125 eh // casa #dial #tatak @miriam // una piccola fe~ fe:sta

eh // house of your untie miriam // a small party

*indicates code-switching:*

- *#_ when the whole TU is in a different language*
- *# before tokens in a foreign language is a mixed-language TU*

Transcription: what language?

(2)

a. naturale che (è) fatto bianco io **#*ch'**#aggia **#*fa** è questo è fashion
obviously I make it white what should I do this is fashion

b. e ho iniziato trovare qualche **#*informazia** e ho trovato quelli (.) studenti
and I started to find some information and I found those students

c. sì sì arrivato: e:h **#*conferenze** di: sicurezza come si chiama
yes yes came conference of security how do you call it

*We use **#*** when:*

- *The pronunciation of the word in two languages match*
- *We don't know if it's a non-standard pronunciation of an Italian word or a word in a foreign language*
- *Pronunciation is unclear and we can't distinguish languages*

Transcription: when pronunciation hides morphology

(3)

a. anche (.) abita da **sol(o)** sempre. → *solo.MS* 'alone'
(he) also lives on his own always

b. e io non **$mang(e)** carne → *mangio.3S* 'eat'
and i don't eat meat

- *We transcribe recordings orthographically*
- *Whenever normalization hides some morphological aspects, we transcribe phonetically*
- *If the sound pronounced is unclear, we transcribe the closest sound and put it into brackets*

Transcription: words which cannot be assigned to any language

(4)

a. e io non **$mang(e)** carne → *mangio.3S* ‘eat’
i don’t eat meat

b. per **$fornare** pizza → *infornare* ‘put in the oven’
to put pizza in the oven

c. se tu quello che **$extraordinario** che fai → *straordinario*
if that overtime work that you do

Speakers produce tokens that are not part of any language lexicon

Issue for lemmatization →
$ symbol to keep track of these forms

Publication

Stra-ParlaBO will be published by the end of this year or early 2026

We will keep you posted!


CONCORDANCE
KIPARLA
simple grazie • 772
475.15 per million tokens • 0.048%
Details
Left context
KWIC
Right context
1 KPN015 ion è uguale dappertutto // mh xx // però eh però lì partivan dal secco // grazie // beh allora forse cambia
2 KPN015 biglietto // eh infatti volevo scriverci un altro biglietto con scritto non so grazie ben fatto // ti tengo d'occh
3 KPN015 o' di agrumato nel profumo ecco // cosa vuoi? ah // un'altra clementina grazie // c'ha del profumo c'ha d
4 BOA3004 enticinque // posso invitarvi a questa cortesia? // lasciol volantino // grazie // che cos' è maria? // è ic
5 BOA3004 i // che però non è // ma lui ha parlato al // xxx beautiful girl // ciao // no grazie niente // don't worry be ha
6 BOA3004 xxx beautiful girl // ciao // no grazie niente // don't worry be happy // no grazie non abbiamo niente // no
7 BOA3004 sso a posto tutto l' armadio // ciao xx // ciao // a posto? // niente guarda grazie // sì non abbiamo niente /
8 BOA3004 ie // sì non abbiamo niente // xx xx orologi // eh lo so ma abbiamo tutto grazie // mi spiace // xx xx c' ho
9 BOA3004 enza nostra // ma per cosa? // è un proge~ è una ricerca di linguistica // grazie per avere accettato di par
10 BOA3004 untamento // ah è vero // adesso vediamo // buon pranzo // a domani // grazie ciao emi // ciao // ciao // b
11 BOA3004 embrava un pezzo di bacon // no no no xxx // // vuoi un pezzetto? // no grazie // cos' hai mangiato tu? //
grazie!
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

www.unibo.it