Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA | ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

The KIParla corpus of Spoken Italian

LingCor2024

25/07/2024

Caterina Mauri & Eleonora Zucchini with

Silvia Ballarè (Bologna), Eugenio Goria, Massimo Cerruti, Beatrice Bernasconi (Torino)

** 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"*

Overview

1. Corpus design and implementation

2. The modules

3. Using the corpus: perspectives and examples

4. Next steps and future challenges

Overview

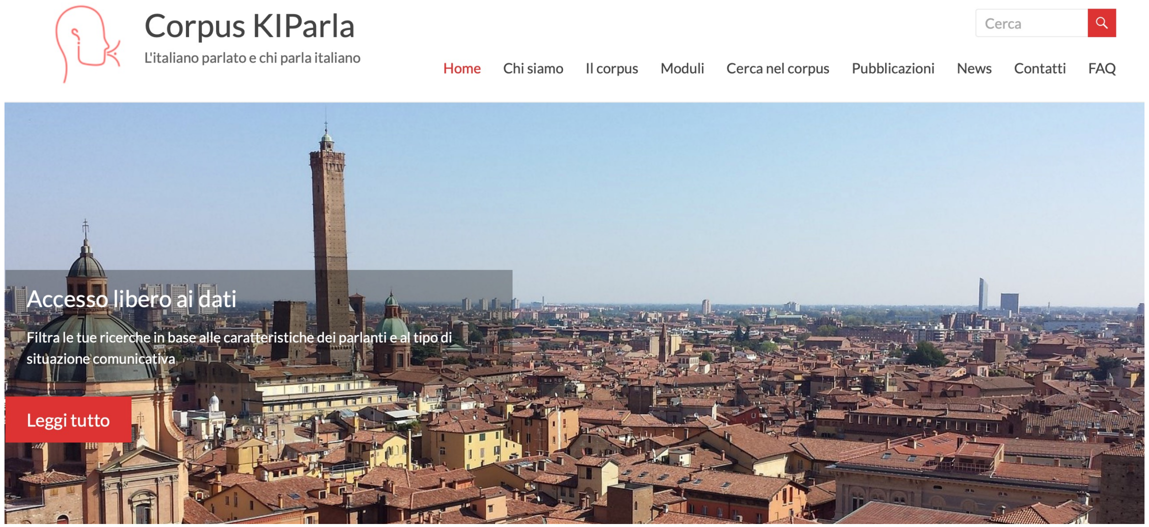
1. Corpus design and implementation

2. The modules

3. Using the corpus: perspectives and examples

4. Next steps and future challenges

Design of the KIParla ~~corpus~~ enterprise!



Corpus KIParla

L'italiano parlato e chi parla italiano

Accesso libero ai dati

Filtra le tue ricerche in base alle caratteristiche dei parlanti e al tipo di situazione comunicativa

www.kiparla.it

(Mauri et al. 2019)

The KIParla corpus aims to offer:

✓ **freely accessible** corpus

transcripts aligned with audio files

✓ **transparent metadata** system

both wrt the communicative situations and the speakers involved

✓ **search interface** based on an **international standard**

NoSketchEngine, offering advanced search functions: KWIC, frequency lists, etc.

✓ **modular, incremental, replicable infrastructure**

allowing for the expansion of the corpus over time through a modular structure

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Design of the KIParla ~~corpus~~ enterprise!

The KIParla corpus aims to offer:

✓ **freely accessible** corpus

transcripts aligned with audio files

✓ **transparent metadata** system

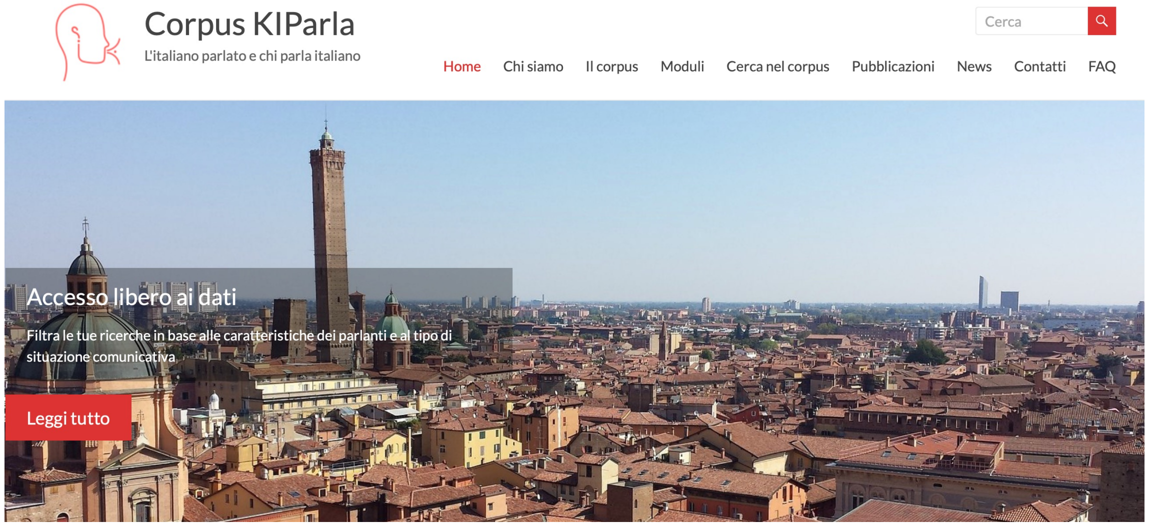
both wrt the communicative situations and the speakers involved

✓ **search interface** based on an **international standard**

NoSketchEngine, offering advanced search functions: KWIC, frequency lists, etc.

✓ **modular, incremental, replicable infrastructure**

allowing for the expansion of the corpus over time through a modular structure



Corpus KIParla

L'italiano parlato e chi parla italiano

Accesso libero ai dati

Filtra le tue ricerche in base alle caratteristiche dei parlanti e al tipo di situazione comunicativa

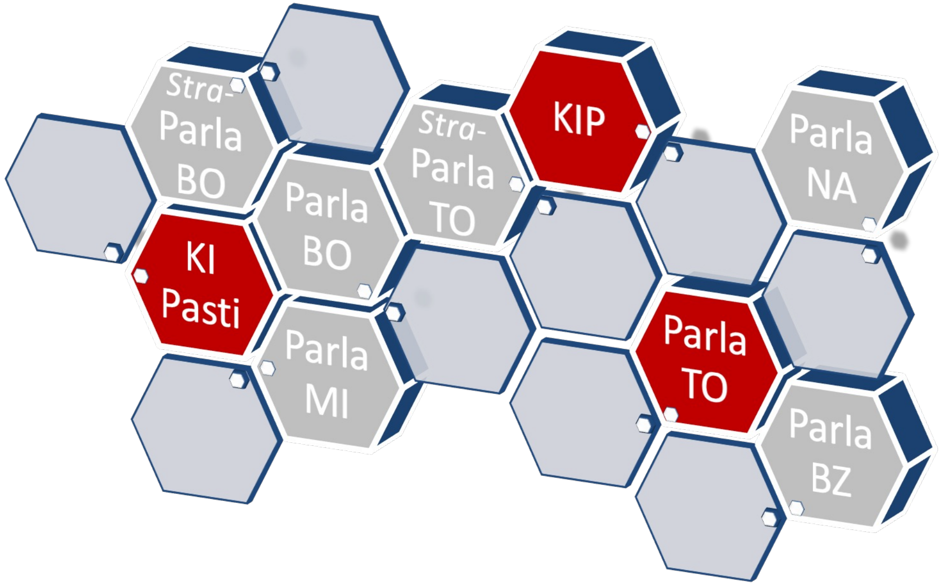
www.kiparla.it

(Mauri et al. 2019)

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

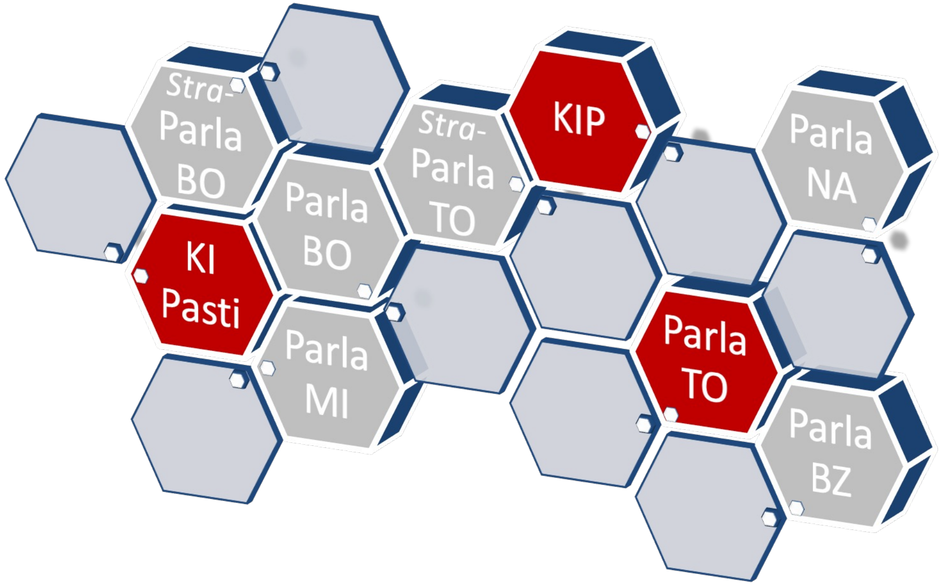
Modularity and incrementality

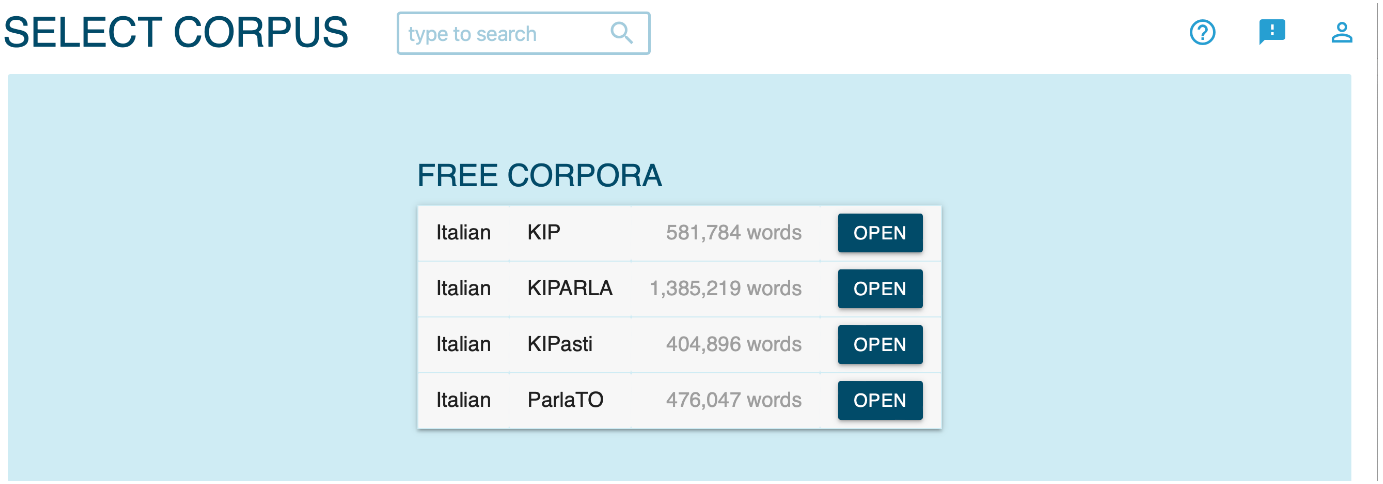
- **Modularity:** internal division into independent modules



Modularity and incrementality

- **Modularity:** internal division into independent modules
  - ✓ possibility of consulting each module separately or all modules together;




SELECT CORPUS

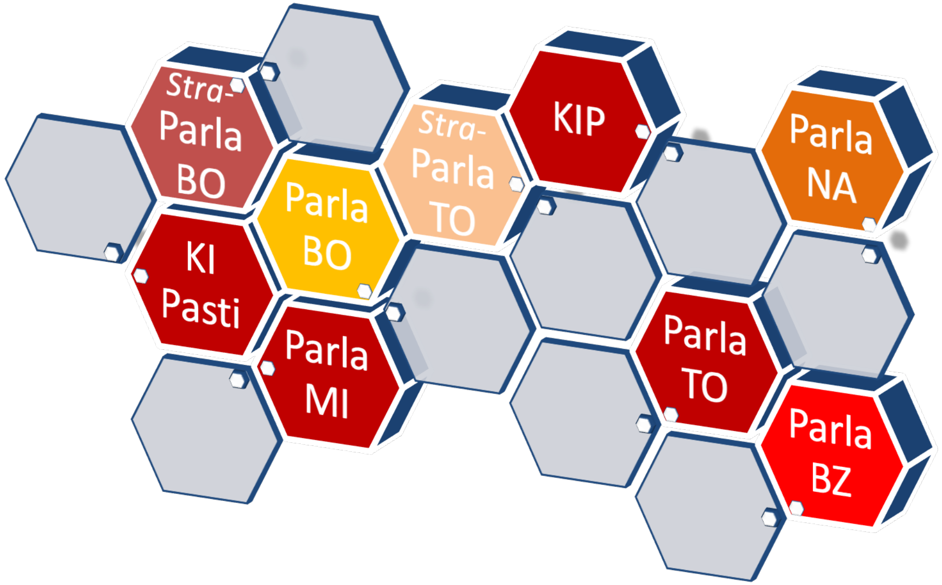
type to search

FREE CORPORA

Italian	KIP	581,784 words	OPEN
Italian	KIPARLA	1,385,219 words	OPEN
Italian	KIPasti	404,896 words	OPEN
Italian	ParlaTO	476,047 words	OPEN

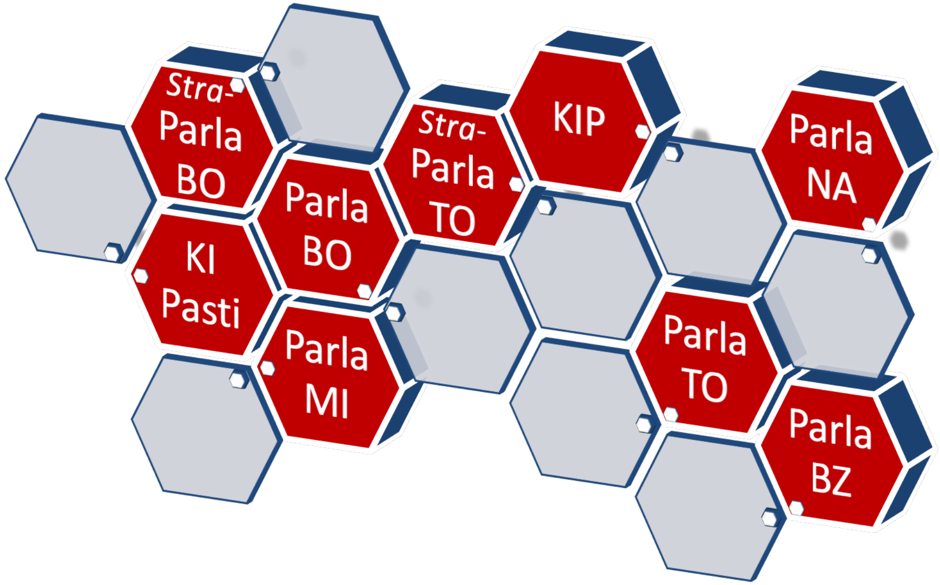
Modularity and incrementality

➢ **Incrementality**: possibility to add new modules over time



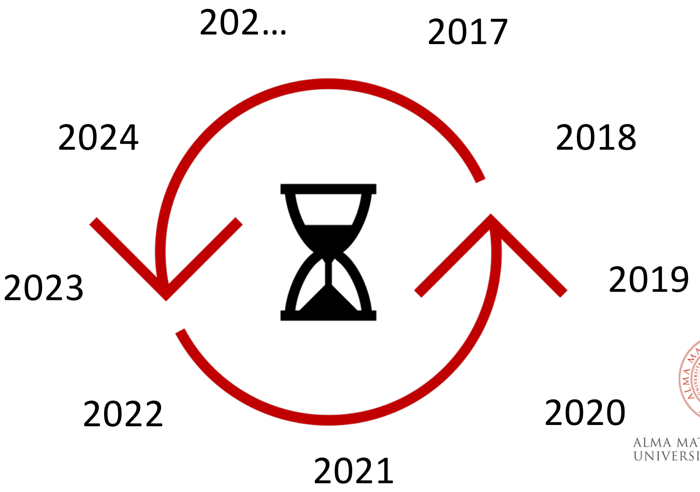
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Modularity and incrementality



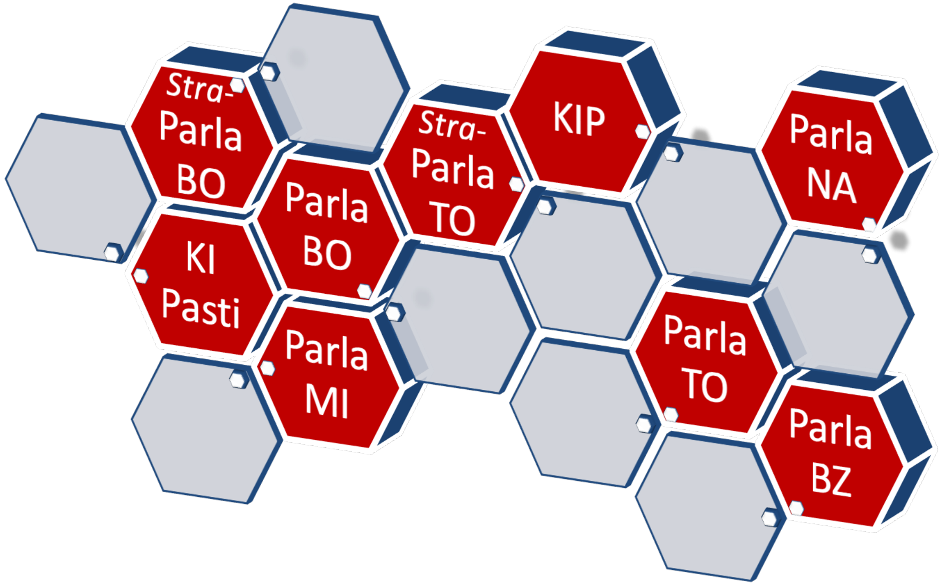
- **Incrementality**: possibility to add new modules over time

Such a dynamic nature makes the KIParla corpus suitable to document spoken language over time.



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

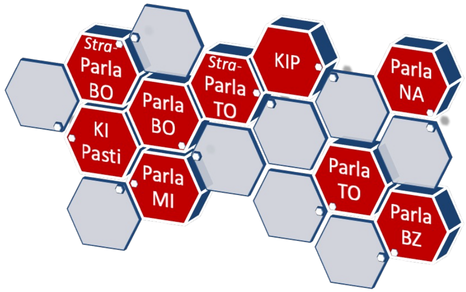
Modularity and incrementality



- **Incrementality**: possibility to add new modules over time

What are KIParla **modules**?

Modularity and incrementality



Modules are 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 vision:

the more modules are added

the more dimensions of variation can be explored

Collection and transcription: *it takes a village!*

GDPR - Regolamento 2016/679

2018 - today: more than 100 students (from the universities of Bologna and Turin) participated in the construction of the KIParla corpus. Internships, BA theses, MA theses…

Training and supervision

Frequent updates and weekly meetings

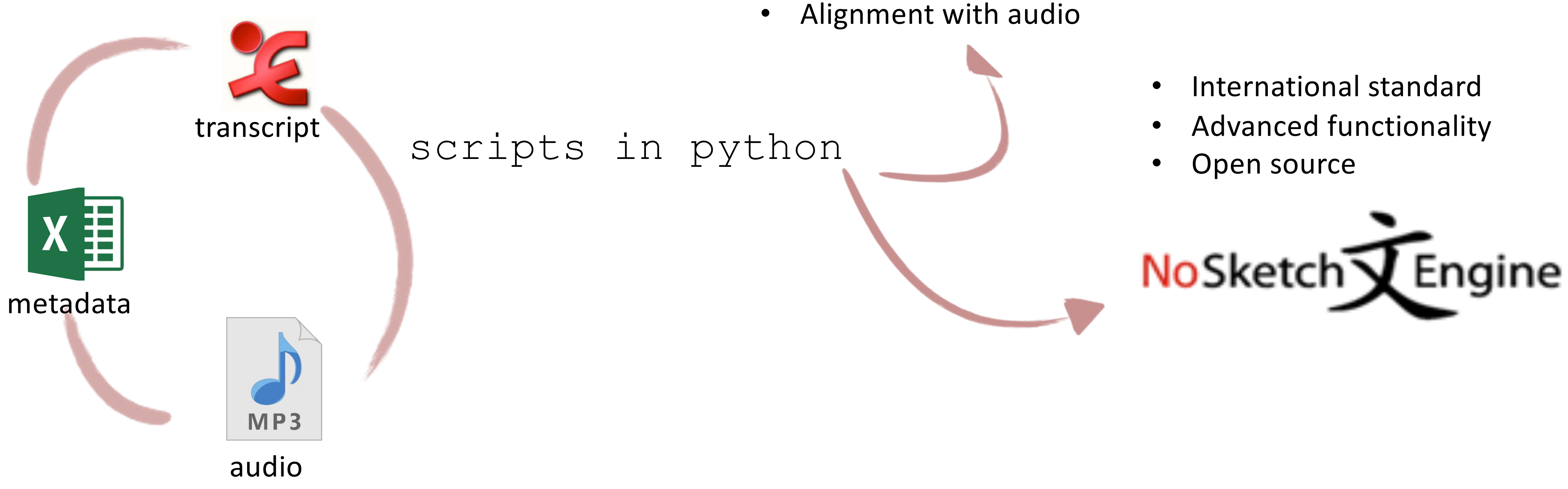
Organization and coordination

- **Transcription** by means of ELAN, simplified Jefferson (2004) for conversational aspects
- **Pseudonymization:** deletion of sensitive data from the transcript and audio track.

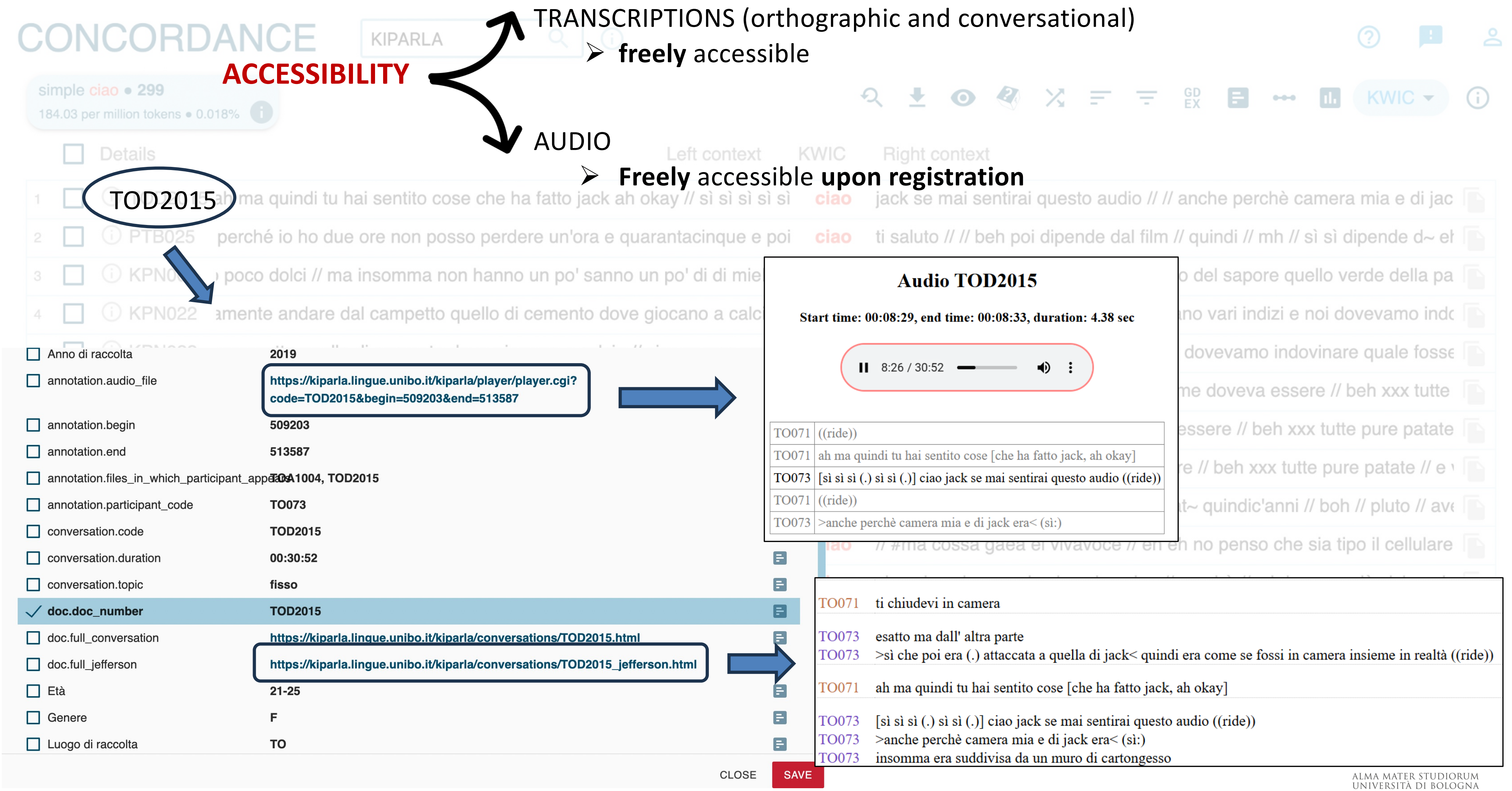
,	Rising intonation	<ciao>	Pronunciation (slower)
.	Descending intonation	[hello]	Overlaps between speakers
:	Prolonged sound	(hello)	Text difficult to understand
(.)	Short break	xxx	Unintelligible text
>ciao<	Pronunciation (faster)	((laughs))	Nonverbal behavior

Publication of data and accessibility

- XML
- Txt (orthographic version)
- Txt (conversational version)
- Occurrence/metadata association
- Alignment with audio



- International standard
- Advanced functionality
- Open source



ACCESSIBILITY

TRANSCRIPTIONS (orthographic and conversational)

- ➢ **freely** accessible

AUDIO

- ➢ **Freely** accessible **upon registration**

			Left context	KWIC	Right context
1	□	TOD2015	ah ma quindi tu hai sentito cose che ha fatto jack ah okay // sì sì sì sì sì	ciao	jack se mai sentirai questo audio // // anche perchè camera mia e di jac
2	□	PTB025	perché io ho due ore non posso perdere un'ora e quarantacinque e poi	ciao	ti saluto // // beh poi dipende dal film // quindi // mh // sì sì dipende d~ eh
3	□	KPN0	poco dolci // ma insomma non hanno un po' sanno un po' di di mie		o del sapore quello verde della pa
4	□	KPN022	amente andare dal campetto quello di cemento dove giocano a calc		ano vari indizi e noi dovevamo indo

□ Anno di raccolta	2019
□ annotation.audio_file	https://kiparla.lingue.unibo.it/kiparla/player/player.cgi?code=TOD2015&begin=509203&end=513587
□ annotation.begin	509203
□ annotation.end	513587
□ annotation.files_in_which_participant_appears	TOA1004, TOD2015
□ annotation.participant_code	TO073
□ conversation.code	TOD2015
□ conversation.duration	00:30:52
□ conversation.topic	fisso
✓ doc.doc_number	TOD2015
□ doc.full_conversation	https://kiparla.lingue.unibo.it/kiparla/conversations/TOD2015.html
□ doc.full_jefferson	https://kiparla.lingue.unibo.it/kiparla/conversations/TOD2015_jefferson.html
□ Età	21-25
□ Genere	F
□ Luogo di raccolta	TO

Audio TOD2015

Start time: 00:08:29, end time: 00:08:33, duration: 4.38 sec

TO071	((ride))
TO071	ah ma quindi tu hai sentito cose [che ha fatto jack, ah okay]
TO073	[sì sì sì (.) sì sì (.)] ciao jack se mai sentirai questo audio ((ride))
TO071	((ride))
TO073	>anche perchè camera mia e di jack era< (sì:)

TO071	ti chiudevi in camera
TO073	esatto ma dall' altra parte
TO073	>sì che poi era (.) attaccata a quella di jack< quindi era come se fossi in camera insieme in realtà ((ride))
TO071	ah ma quindi tu hai sentito cose [che ha fatto jack, ah okay]
TO073	[sì sì sì (.) sì sì (.)] ciao jack se mai sentirai questo audio ((ride))
TO073	>anche perchè camera mia e di jack era< (sì:)
TO073	insomma era suddivisa da un muro di cartongesso

Overview

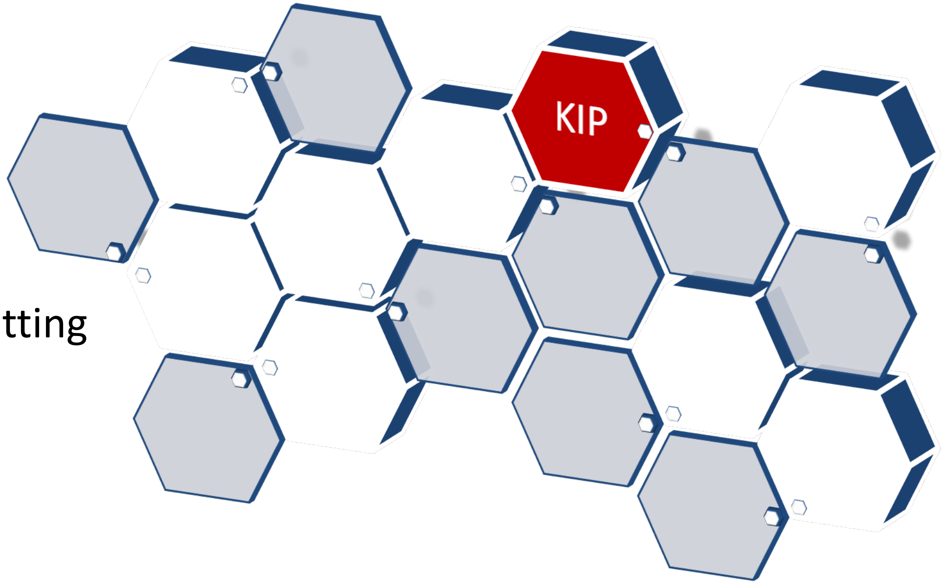
1. Corpus design and implementation

2. The modules

3. Using the corpus: some case studies

4. Next steps and future challenges

Corpus design: the KIP module



KIP module

- 273 **educated speakers** recorded in a university setting

→ Observation of register and regional variation.

✓ Data collected in Bologna and Turin (2016-2019)

✓ **Different communicative contexts**: 121 interactions

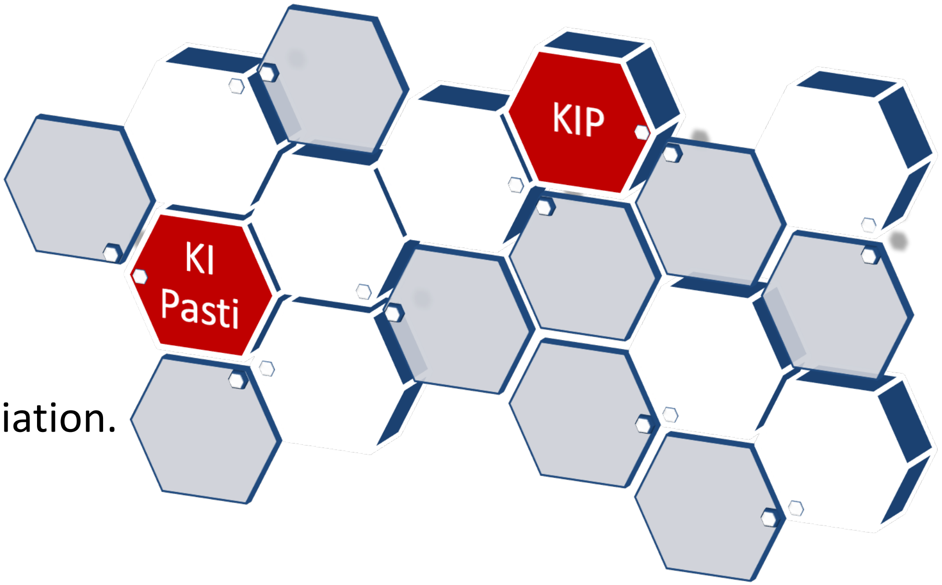
✓ **TOT: ca. 70 hrs, 660,000 tokens**

	Relationship between participants	Moderator	Topic
Free conversation	Symmetrical	Absent	Free
Semi-structured interview	Symmetrical	Present	Fixed
Office hours	Asymmetrical	Absent	Free
Exams	Asymmetrical	Absent	Free
Lessons	Asymmetrical	Present	Fixed

PUBLISHED

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Corpus design: the KIPasti module



KIPasti module

- 63 dinner-table conversations all over Italy

→ Observation of the vernacular, diastratic and regional variation.

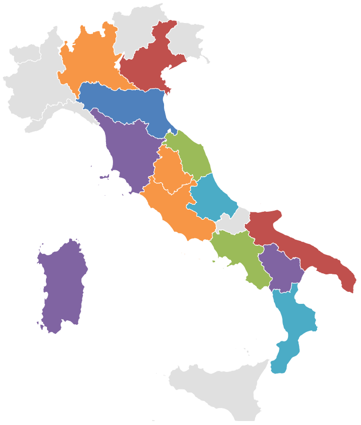
✓ Data collected in 13 regions of Italy (2020-2024)

✓ 147 Speakers **with different social characterization** (ages, education degree, occupation)

✓ One context: **dinner-table conversations**, thus shared background, informal and friendly register

✓ **TOT: ca. 42 hrs, 487,000 tokens**

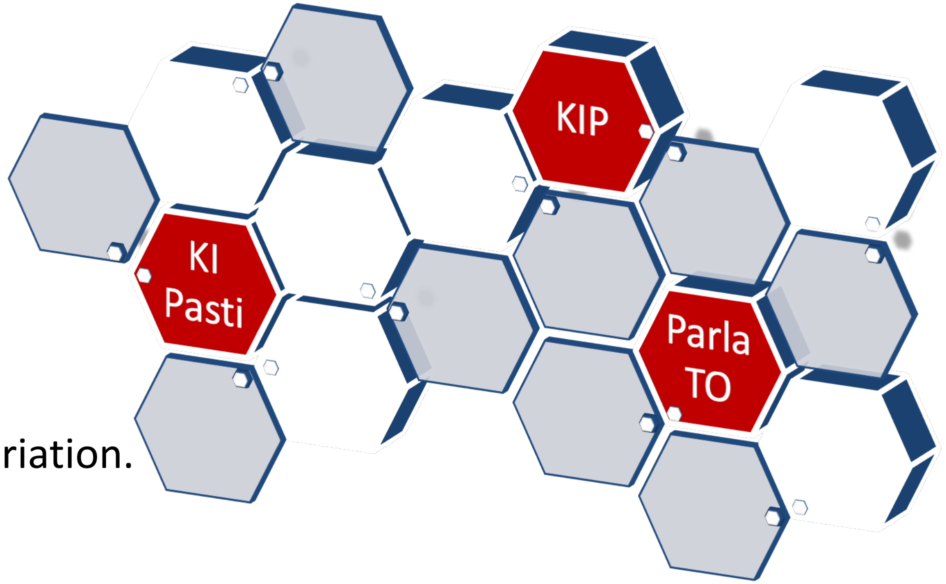
	Data collection
Northern Italy	46%
Central Italy	16%
Southern Italy	38%



PUBLISHED

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Corpus design: the ParlaTO module



ParlaTO module

- Language spoken in the metropolitan city of Turin

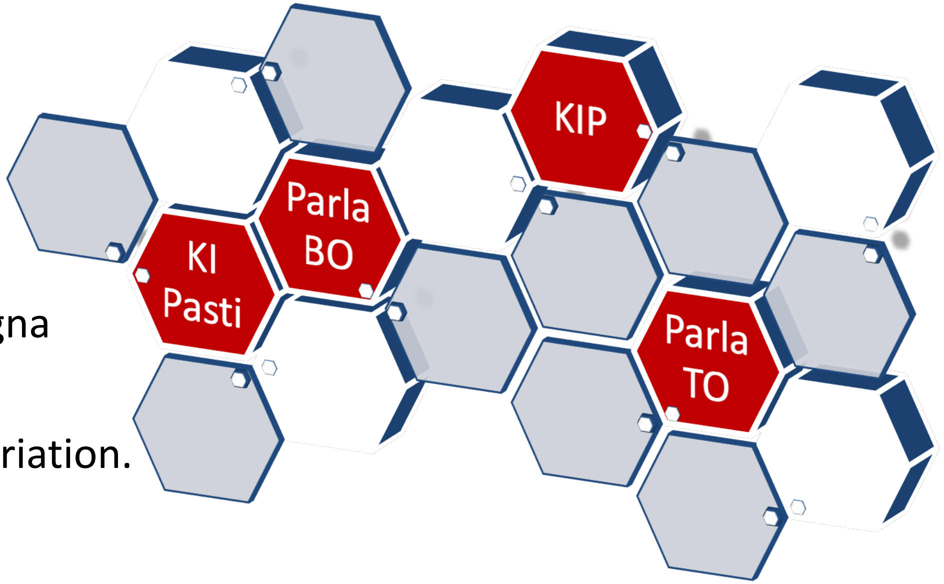
→ Observation of the vernacular, diastratic and regional variation.

- ✓ Data collected in Turin (2019)
- ✓ 88 Speakers with **different social characterization** (ages, education degree, occupation)
- ✓ One context: 67 semi-structured interviews
- ✓ **TOT: ca. 50 hrs, 560,000 tokens**

	Age groups
Young	18 ≤ x ≤ 30 years old
Adults	30 < x ≤ 60 years old
Seniors	60 < x ≤ 89 years old

PUBLISHED

Corpus design: the ParlaBO module



ParlaBO module

- Language spoken in the metropolitan city of Bologna

→ Observation of the vernacular, diastratic and regional variation.

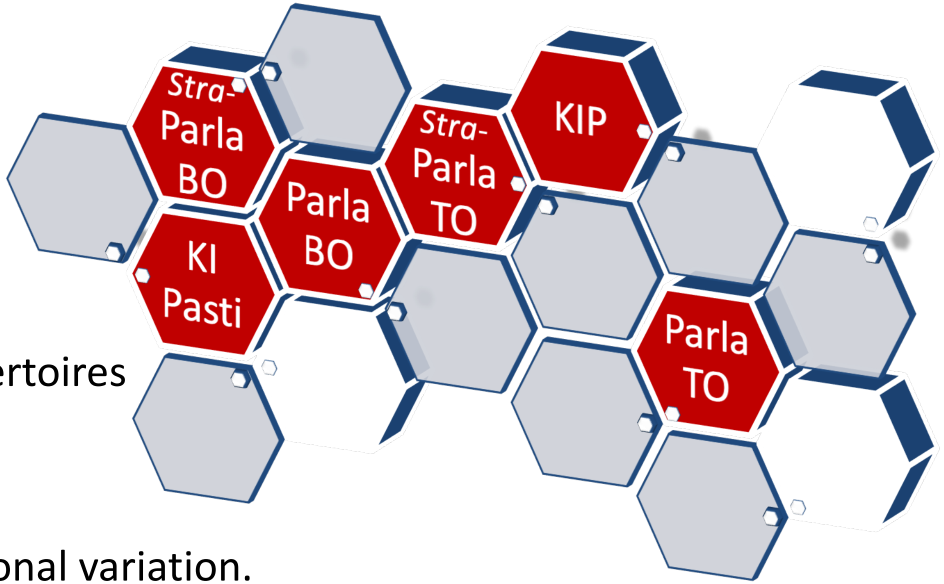
- ✓ Data collected in Bologna (2022-2024)
- ✓ 158 Speakers with **different social characterization** (ages, education degree, occupation)
- ✓ One context: 86 semi-structured interviews
- ✓ **TOT: ca. 66 hrs, 650,000 tokens**

	Data collection
18 ≤ x ≤ 30 years old	30%
30 < x ≤ 60 years old	36%
60 < x ≤ 89 years old	34%

- ✓ Data collection completed
- ➢ At the moment, we are transcribing (98%)

TO BE PUBLISHED
VERY SOON

Corpus design: Stra-Parla BO/TO modules



Next steps: Stra-ParlaBO and Stra-ParlaTO modules

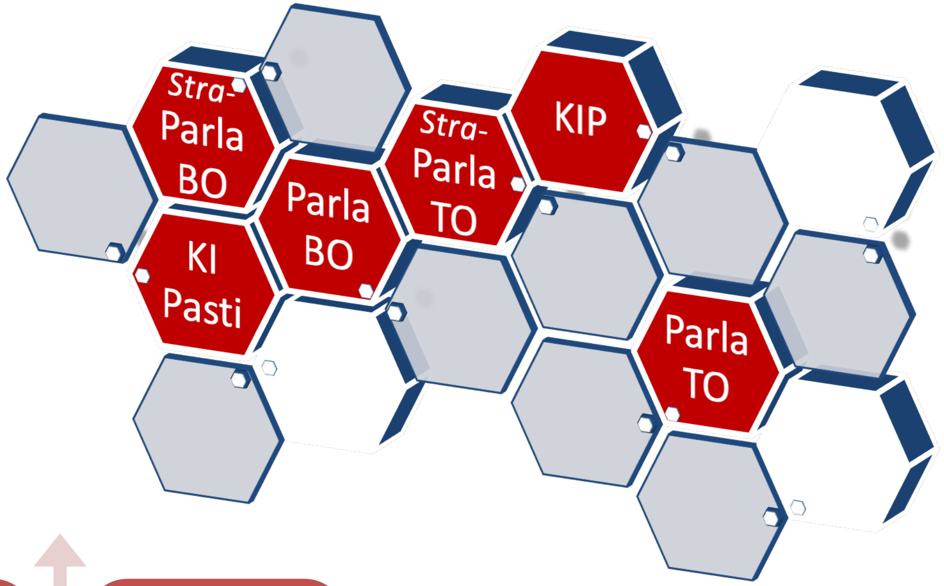
- Italian spoken within communities of **foreign-origin speakers** with complex multilingual repertoires

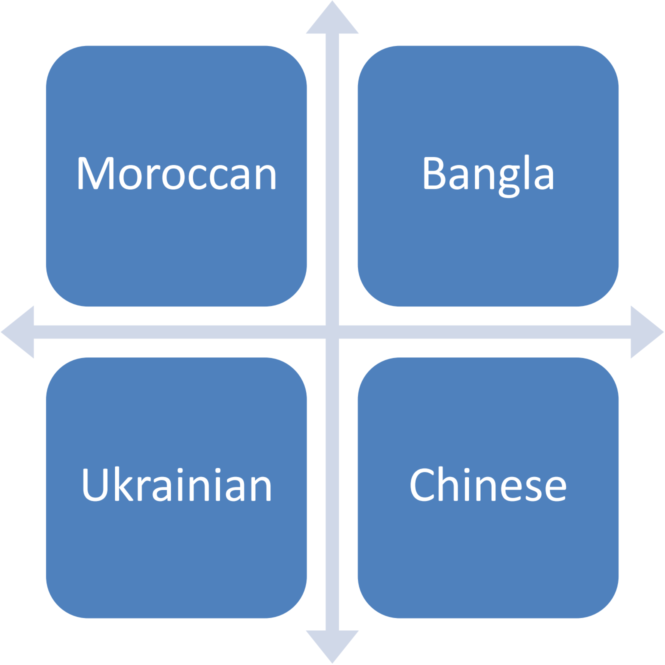
→ Observation of the learners' varieties, diastratic and regional variation.

- ✓ Data collected in Bologna and Turin (2024 - ongoing)
- ✓ Speakers with **multilingual repertoires, different social characterization and origin**
- ✓ Two contexts: semi-structured interview, free conversation
- ✓ **EXPECTED: ca. 128 hrs, 1,250,000 tokens**

Data collection in progress!

Corpus design: Stra-Parla BO/TO modules





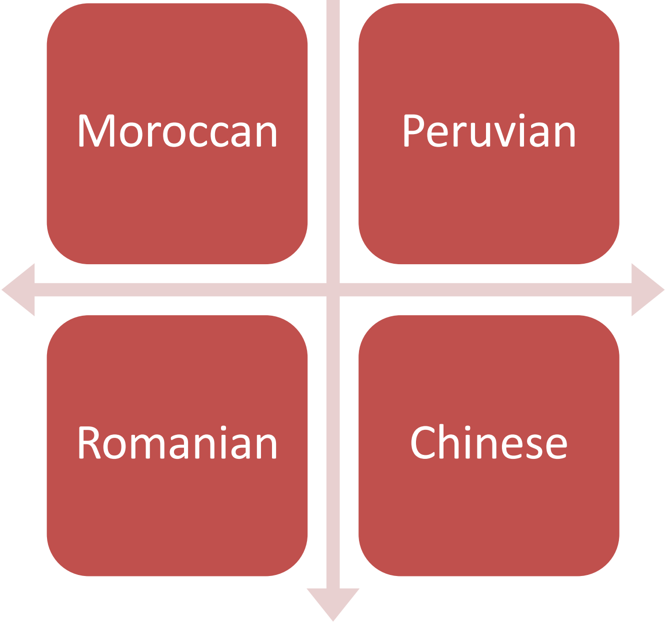
Moroccan

Bangla

Ukrainian

Chinese

Bologna



Moroccan

Peruvian

Romanian

Chinese

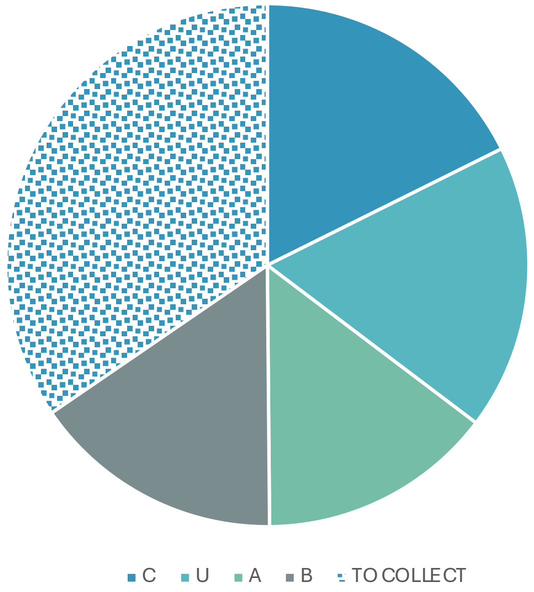
Torino

16 hours
per speech community

- 8 h Free conversation
- 8 h Interviews

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Stra-ParlaBO: where are we?



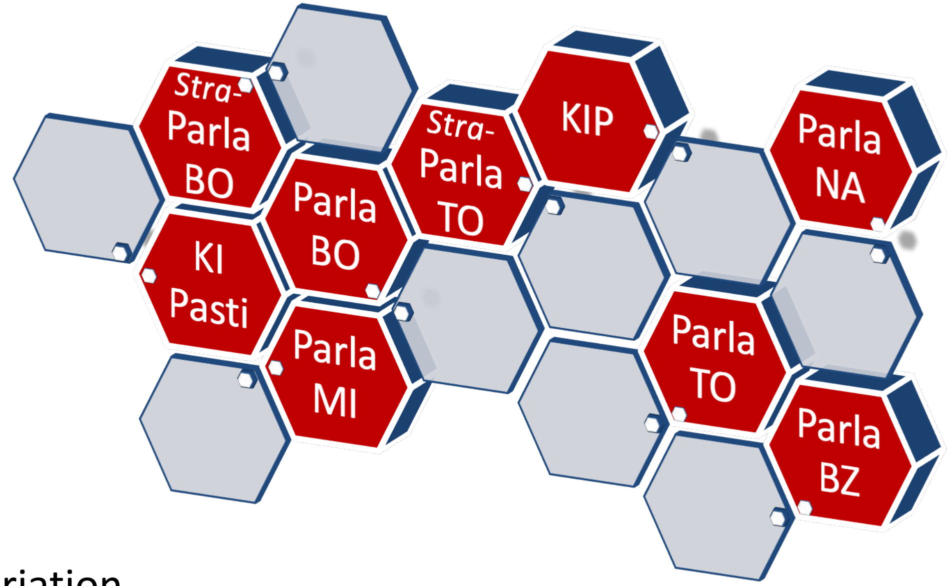
Speech community	Interviews	Free conversation	Total
Chinese	9:04:20	2:13:39	**11:17:59**
Ukranian	9:48:30	1:29:03	**11:17:22**
Moroccan Arabic	8:32:34	0:46:42	**9:19:16**
Bangla	9:58:16	0	**9:58:16**
Total	**37:23:40**	**4:29:24**	**41:53:04**

Corpus design: ongoing collaborations

Next steps: **ParlaBZ, ParlaNA, ParlaMI modules**

- Italian spoken in **Bozen** (Daniela Veronesi),
- Italian spoken in **Naples** (Margherita di Salvo),
- Italian spoken in **Milan** (Federica da Milano)

→ Observation of the vernacular, diastratic and regional variation



Work in progress!

KIParla corpus: size

KIParla Corpus (25/07/24)	
Hours	152:50:25
Informants	489
Tokens	1,439,271

- ✓ KIP
- ✓ ParlaTO
- ✓ KIPasti

KIParla Corpus (30/11/25)	
Expected Hours	343:00:00
Expected Informants	900
Expected TOT Tokens	**3,300,000**

- ✓ KIP
- ✓ ParlaTO
- ✓ KIPasti
- ✓ ParlaBO
- ✓ ParlaBZ
- ✓ ParlaNA
- ✓ Stra-ParlaBO
- ✓ Stra-ParlaTO

Overview

1. Corpus design and implementation

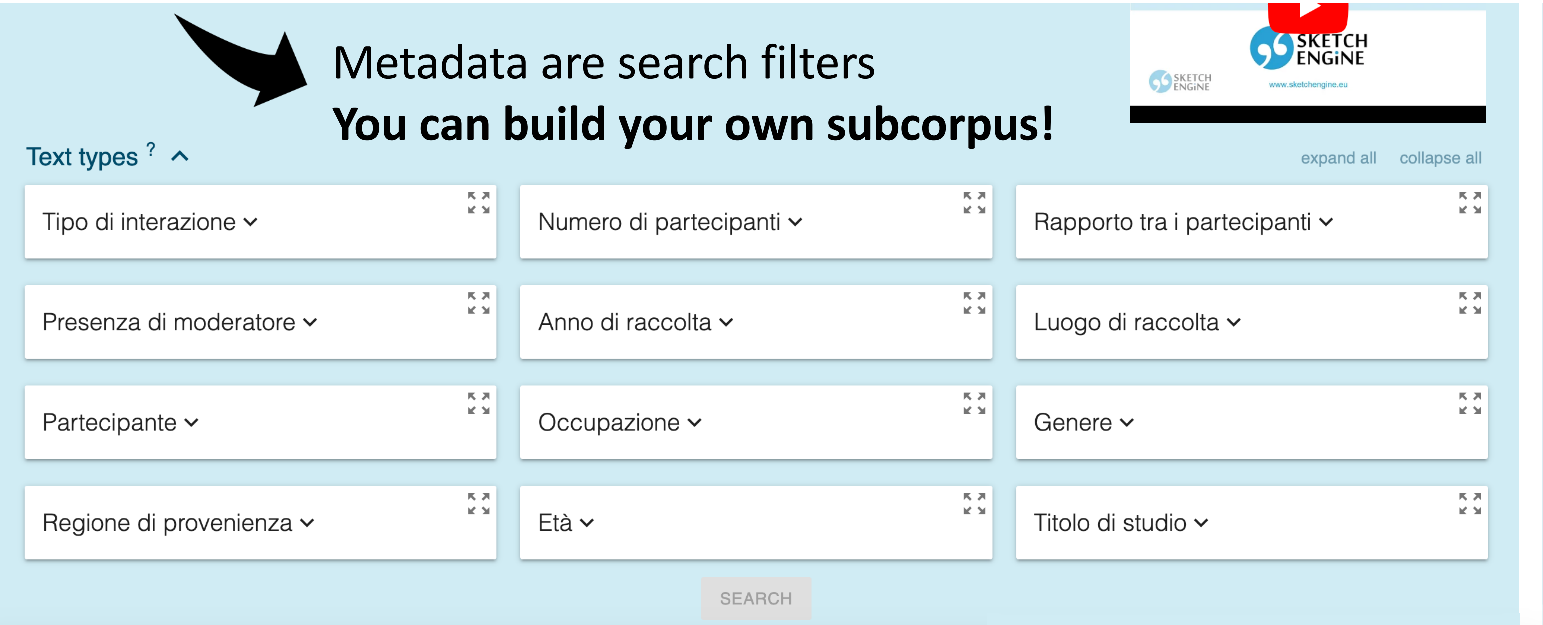
2. The modules

3. Using the corpus: perspectives and examples

4. Next steps and future challenges

Metadata are search filters

You can build your own subcorpus!



Text types ? ^

expand all collapse all

Tipo di interazione ⌄

Numero di partecipanti ⌄

Rapporto tra i partecipanti ⌄

Presenza di moderatore ⌄

Anno di raccolta ⌄

Luogo di raccolta ⌄

Partecipante ⌄

Occupazione ⌄

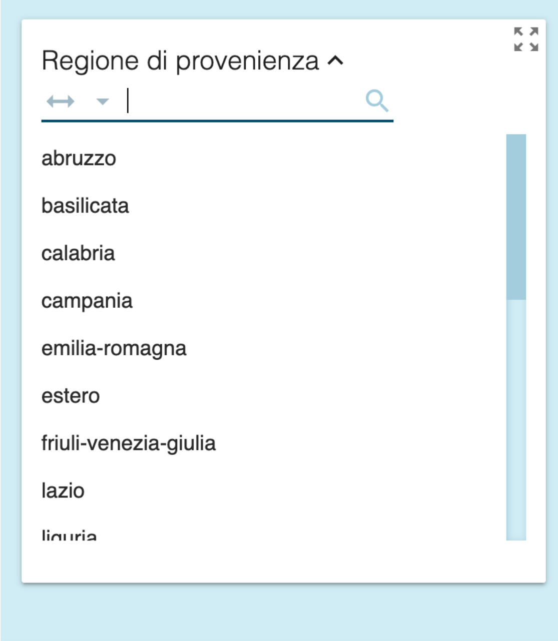
Genere ⌄

Regione di provenienza ⌄

Età ⌄

Titolo di studio ⌄

SEARCH



Regione di provenienza ^

abruzzo

basilicata

calabria

campania

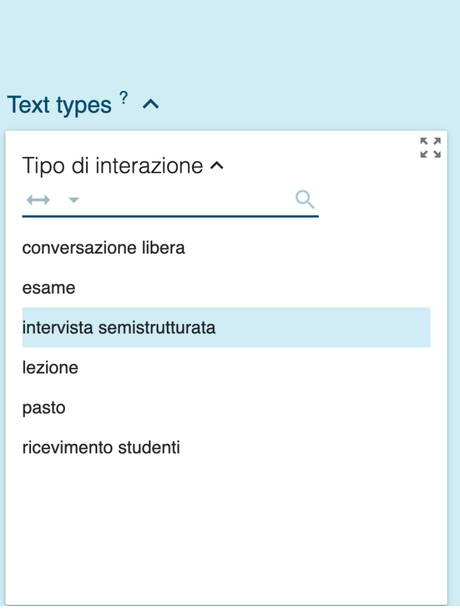
emilia-romagna

estero

friuli-venezia-giulia

lazio

liguria



Text types ? ^

Tipo di interazione ^

conversazione libera

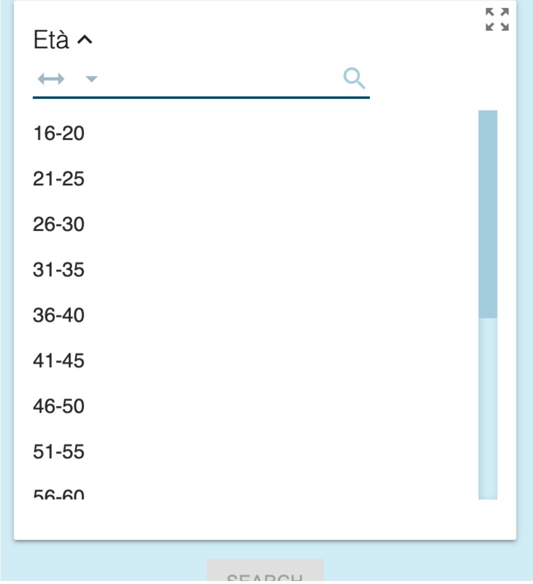
esame

intervista semistrutturata

lezione

pasto

ricevimento studenti



Età ^

16-20

21-25

26-30

31-35

36-40

41-45

46-50

51-55

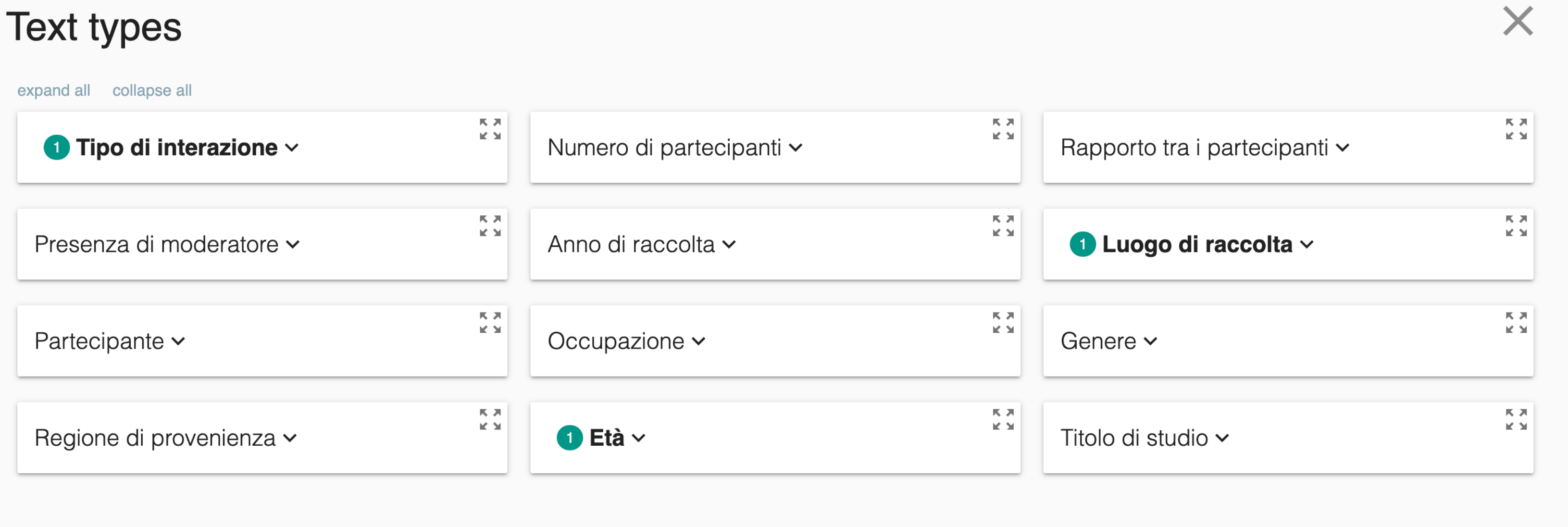
56-60

SEARCH

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

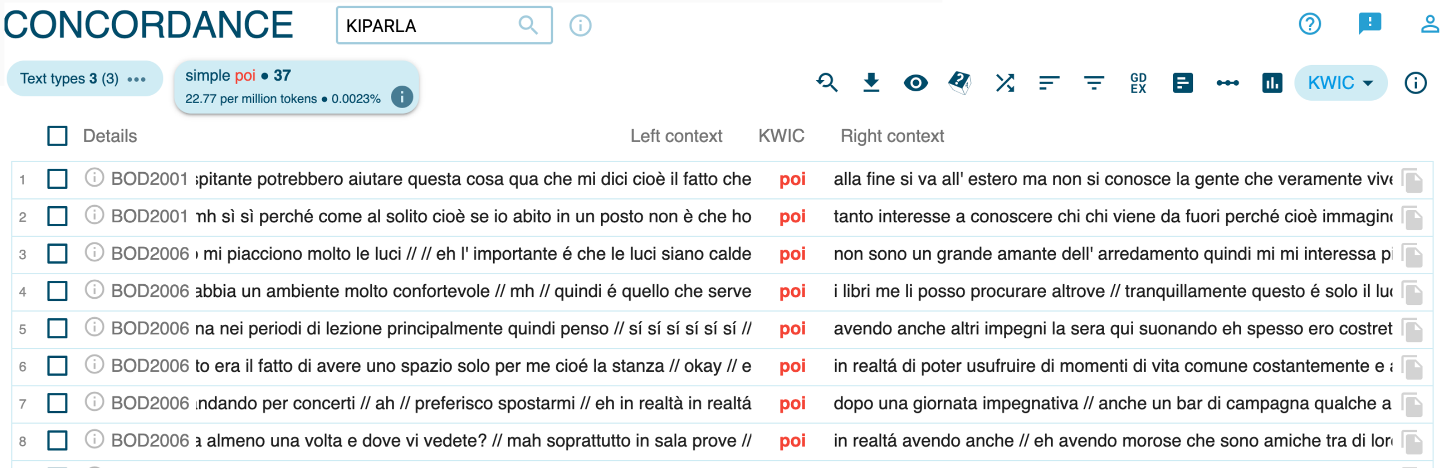
Metadata are search filters

You can build your own subcorpus!



Text types

- 1 **Tipo di interazione**
- Numero di partecipanti
- Rapporto tra i partecipanti
- Presenza di moderatore
- Anno di raccolta
- 1 **Luogo di raccolta**
- Partecipante
- Occupazione
- Genere
- Regione di provenienza
- 1 **Età**
- Titolo di studio



CONCORDANCE   KIPARLA

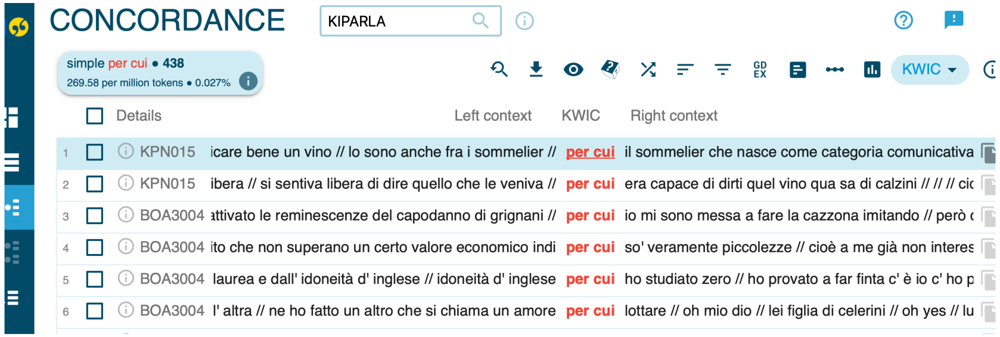
Text types **3** (3) •••

simple poi • **37**

22.77 per million tokens • 0.0023%

		Details	Left context	KWIC	Right context
1	□	BOD2001	ıpitante potrebbero aiutare questa cosa qua che mi dici cioè il fatto che	**poi**	alla fine si va all' estero ma non si conosce la gente che veramente vive
2	□	BOD2001	nh sì sì perché come al solito cioè se io abito in un posto non è che ho	**poi**	tanto interesse a conoscere chi chi viene da fuori perché cioè immagine
3	□	BOD2006	ı mi piacciono molto le luci // // eh l' importante é che le luci siano calde	**poi**	non sono un grande amante dell' arredamento quindi mi mi interessa pi
4	□	BOD2006	ıbbia un ambiente molto confortevole // mh // quindi é quello che serve	**poi**	i libri me li posso procurare altrove // tranquillamente questo é solo il luo
5	□	BOD2006	na nei periodi di lezione principalmente quindi penso // sí sí sí sí sí sí //	**poi**	avendo anche altri impegni la sera qui suonando eh spesso ero costret
6	□	BOD2006	to era il fatto di avere uno spazio solo per me cioé la stanza // okay // e	**poi**	in realtá di poter usufruire di momenti di vita comune costantemente e a
7	□	BOD2006	ındando per concerti // ah // preferisco spostarmi // eh in realtà in realtá	**poi**	dopo una giornata impegnativa // anche un bar di campagna qualche a
8	□	BOD2006	ı almeno una volta e dove vi vedete? // mah soprattutto in sala prove //	**poi**	in realtá avendo anche // eh avendo morose che sono amiche tra di lor

(1) Register variation: *per cui* lit. ‘for which’



CONCORDANCE

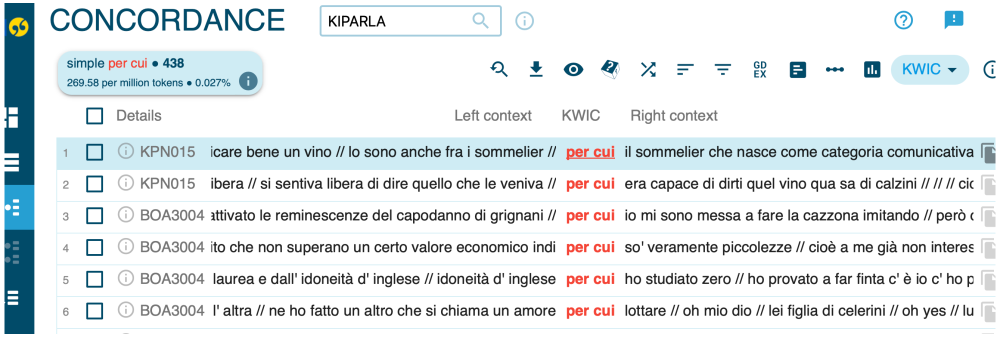
KIPARLA

simple per cui • 438
269.58 per million tokens • 0.027%

KWIC

	□	Details	Left context	KWIC	Right context
1	□	KPN015	icare bene un vino // lo sono anche fra i sommelier //	per cui	il sommelier che nasce come categoria comunicativa
2	□	KPN015	ibera // si sentiva libera di dire quello che le veniva //	per cui	era capace di dirti quel vino qua sa di calzini // // // cic
3	□	BOA3004	attivato le reminescenze del capodanno di grignani //	per cui	io mi sono messa a fare la cazzona imitando // però c
4	□	BOA3004	ito che non superano un certo valore economico indi	per cui	so' veramente piccolezze // cioè a me già non interes
5	□	BOA3004	laurea e dall' idoneità d' inglese // idoneità d' inglese	per cui	ho studiato zero // ho provato a far finta c' è io c' ho p
6	□	BOA3004	l' altra // ne ho fatto un altro che si chiama un amore	per cui	lottare // oh mio dio // lei figlia di celerini // oh yes // lu

(1) Register variation: *per cui* lit. ‘for which’



CONCORDANCE KIPARLA

simple per cui • 438
269.58 per million tokens • 0.027%

	□	Details	Left context	KWIC	Right context
1	□	KPN015	icare bene un vino // lo sono anche fra i sommelier //	per cui	il sommelier che nasce come categoria comunicativa
2	□	KPN015	ibera // si sentiva libera di dire quello che le veniva //	per cui	era capace di dirti quel vino qua sa di calzini // // // cic
3	□	BOA3004	ɪttivato le reminescenze del capodanno di grignani //	per cui	io mi sono messa a fare la cazzona imitando // però (
4	□	BOA3004	ito che non superano un certo valore economico indi	per cui	so' veramente piccolezze // cioè a me già non interes
5	□	BOA3004	laurea e dall' idoneità d' inglese // idoneità d' inglese	per cui	ho studiato zero // ho provato a far finta c' è io c' ho p
6	□	BOA3004	l' altra // ne ho fatto un altro che si chiama un amore	per cui	lottare // oh mio dio // lei figlia di celerini // oh yes // lu

✓ *Per cui* has undergone a grammaticalization process:

- **from Relative oblique construction** ‘for which’ (REL)

(1) KIP, TOD1014

TO068: quindi vi chiederei // eh vabbè mh innanzitutto
il motivo per cui avete deciso di frequentare questo corso

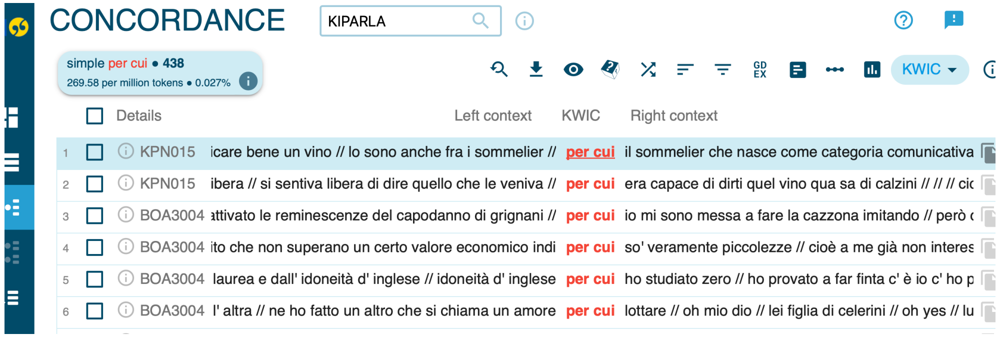
So I will ask you // first of all
The reason for which you have chosen to attend thi course

Metadata TO068
Origin: Sicily (Italy)
Age: 36-40
Degree: PhD
Occupation: Professionals

(1) Register variation: *per cui* lit. 'for which'

✓ *Per cui* has undergone a grammaticalization process:

- from Relative oblique construction 'for which' (REL)



CONCORDANCE KIPARLA

simple per cui • 438
269.58 per million tokens • 0.027%

KWIC

	Details	Left context	KWIC	Right context
1	☐ KPN015	icare bene un vino // lo sono anche fra i sommelier //	per cui	il sommelier che nasce come categoria comunicativa
2	☐ KPN015	ibera // si sentiva libera di dire quello che le veniva //	per cui	era capace di dirti quel vino qua sa di calzini // // // cic
3	☐ BOA3004	attivato le reminescenze del capodanno di grignani //	per cui	io mi sono messa a fare la cazzona imitando // però (
4	☐ BOA3004	ito che non superano un certo valore economico indi	per cui	so' veramente piccolezze // cioè a me già non interes
5	☐ BOA3004	laurea e dall' idoneità d' inglese // idoneità d' inglese	per cui	ho studiato zero // ho provato a far finta c' è io c' ho p
6	☐ BOA3004	l' altra // ne ho fatto un altro che si chiama un amore	per cui	lottare // oh mio dio // lei figlia di celerini // oh yes // lu

(1) KIP, TOD1014

TO068: quindi vi chiederei // eh vabbè mh innanzitutto
il motivo per cui avete deciso di frequentare questo corso

So I will ask you // first of all
The reason for which you have chosen to attend thi course

Metadata TO068
Origin: Sicily (Italy)
Age: 36-40
Degree: PhD
Occupation: Professionals

- to Conclusive Connective 'therefore' (CONN)

(2) ParlaTO, PTD012

TOR001: certo certo
TOI068: **per cui** diciamo che sì
boh non so dimmi che cosa vuoi che ti racconto

sure sure
***so** let's say that yes*
boh I don't know tell me what you want me to tell you

Metadata TOI068
Origin: Lazio (Italy)
Age: 46-50
Degree: University degree
Occupation: Professionals

(Ballarè & Mauri 2024)

(1) Register variation: *per cui*

Non relativizing ***per cui*** is nowadays accepted but mainly associated to informal and spoken interactions

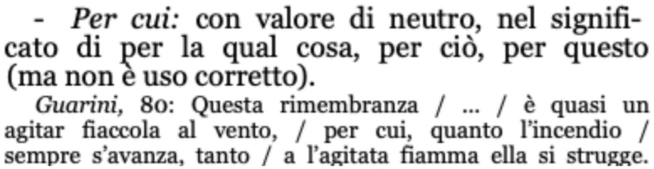
1879

Tommaseo:
It shall not be used

Dizionario *della lingua italiana*:
Per cui a modo d'avverbio, in senso di *Per la qual cosa, Dunque,* non è da dire.

1969

Battaglia:
this use is not correct



- *Per cui:* con valore di neutro, nel significato di per la qual cosa, per ciò, per questo (ma non è uso corretto).
Guarini, 80: Questa rimembranza / ... / è quasi un agitar fiaccola al vento, / per cui, quanto l'incendio / sempre s'avanza, tanto / a l'agitata fiamma ella si strugge.

2008

Treccani:
especially used in spoken language

"spesso per cui è usato, spec. nel linguaggio parlato, col valore della congiunzione conclusiva *e perciò*."

2024

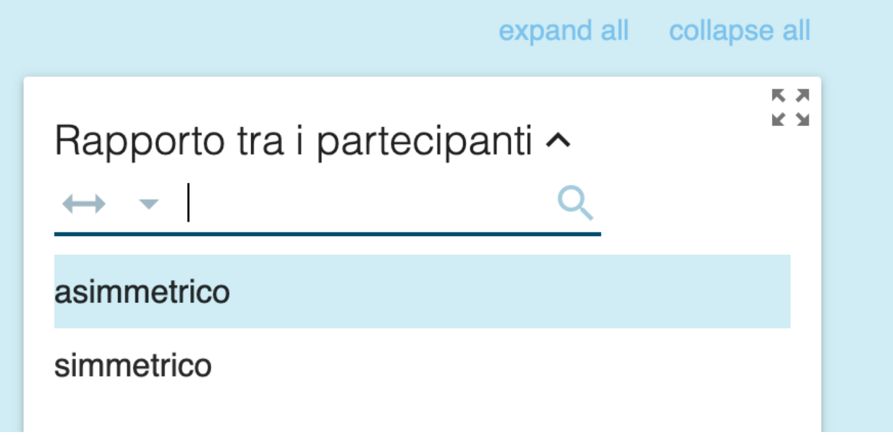
Sabatini & Coletti:
it does not suit technical-formal writings

"*Sicché, perciò*. Frequente nel parlato e nelle scritture di livello stilistico medio, non si addice alle scritture di carattere tecnico-formale"

- Sociolinguistic variation?

(1) Register variation: *per cui*

- **KIP**: **spontaneus conversations**, **semi-structured interviews**, **office hours, lessons, exams**;
- **ParlaTO**: **semi-structured interviews**;
- **KIPasti**: **dinner-table conversations;**



Rapporto tra i partecipanti

asimmetrico

simmetrico

✓ **Formal register**
TOT: 361.702 tokens

✓ **Informal register**
TOT: 1.339.021 tokens

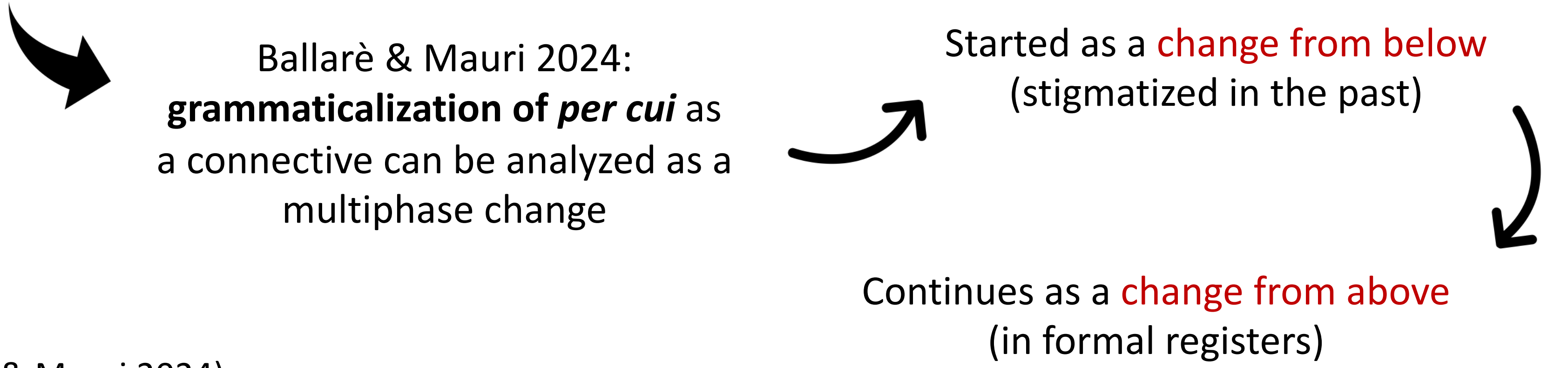
(Ballarè & Mauri 2024)

(1) Register variation: *per cui*

	REL	CONN	TOT.
Informal register	30 **(35,3%)**	55 **(64,7%)**	85
Formal register	70 **(20,2%)**	277 **(79,8%)**	347

The Fisher exact test statistic value is 0.0041. The result is **significant** at p < .01.

✓ CONNective uses of *per cui* are **significantly more frequent** in **formal interactions!**



Ballarè & Mauri 2024:
grammaticalization of *per cui* as a connective can be analyzed as a multiphase change

Started as a change from below
(stigmatized in the past)

Continues as a change from above
(in formal registers)

(Ballarè & Mauri 2024)

(2) Sociolinguistic variation: subjunctive vs. indicative in counterfactual conditionals

SUBJUNCTIVE

PROTASIS

CONDITIONAL

APODOSIS

STANDARD ITALIAN
(Berruto 1983)

(3) ParlaTO, PTA18

TOI031:	*ovviamente*	*se*	*poi*	***avesse***	***partorito***
	obviously	if	then	have:SUBJ	give birth

ci	***saremmo***	***dovuti***	***fermare***	*in*	*ambulanza*
ourselves	be:COND	must	stop	in	ambulance

'obviously if she had given birth we should have stopped in an ambulance'

Non-STANDARD ITALIAN

INDICATIVE – past imperfective

PROTASIS

APODOSIS

(4) ParlaTO, PTD006

TOI055:	*se non*	***trovavo***	*niente*	***m'***	***attaccavo***	*al*	*tram*
	if not	find:IND	nothing	myself	hang:IND	to	tram

'if I had not found anything I could have sung for it'

Mauri et al. 2023
Counterfactual conditionals:
Linguistic variation in Italian and beyond

(2) Sociolinguistic variation: subjunctive vs. indicative in counterfactual conditionals

- ➢ **SEVERAL non-STANDARD counterfactual constructions…**

Mixed cases!

(5) ParlaTO, PTA001

TOI001:	*se*	***tifavo***	*juve*	*mio*	*padre*	
	if	support:IND:PAST:1SG	Juventus	my	father	

‘if I had supported Juventus, my father’

PROTASIS

INDICATIVE

non	***sarebbe tornato***	*più*	*a*	*casa*
NEG	come.back:COND:PAST:3SG	anymore	to	home

‘never have returned home’

APODOSIS

CONDITIONAL

Mauri et al. 2023
Counterfactual conditionals:
Linguistic variation in Italian and beyond

(2) Sociolinguistic variation: subjunctive vs. indicative in counterfactual conditionals

- **SEVERAL non-STANDARD counterfactual constructions…**

What is the actual distribution of **non-standard counterfactuals?**

(5) ParlaTO, PTA001

TOI001:

se	***tifavo***	*juve*	*mio*	*padre*
if	support:IND:PAST:1SG	Juventus	my	father

‘if I had supported Juventus, my father’

PROTASIS

INDICATIVE

non	***sarebbe tornato***	*più*	*a*	*casa*
NEG	come.back:COND:PAST:3SG	anymore	to	home

‘never have returned home’

APODOSIS

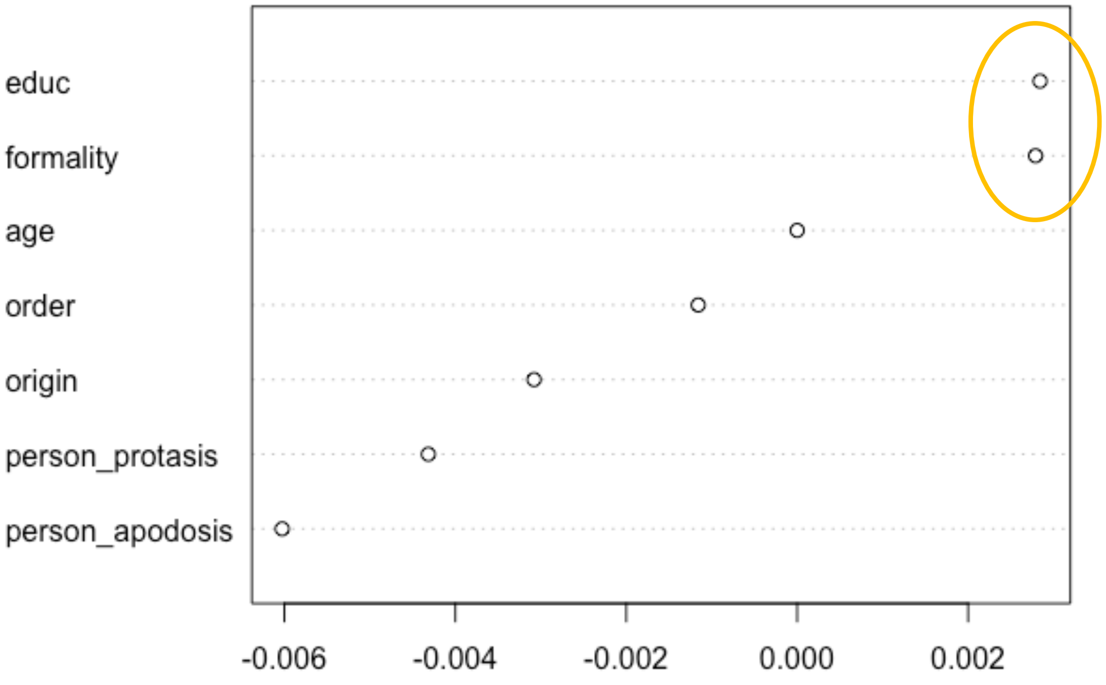
CONDITIONAL

Mauri et al. 2023
Counterfactual conditionals:
Linguistic variation in Italian and beyond

(2) Sociolinguistic variation: subjunctive vs. indicative in counterfactual conditionals

- Random forest technique

Use of indicative in counterfactuals



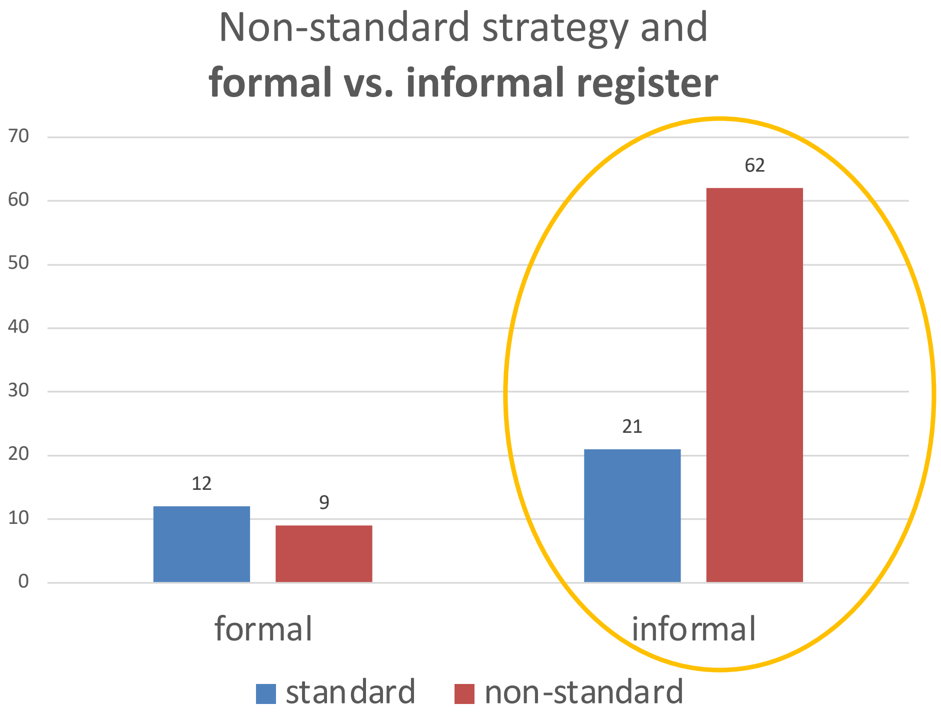
Goodness of Fit: C= 0.8014385.

Education level → Speakers metadata

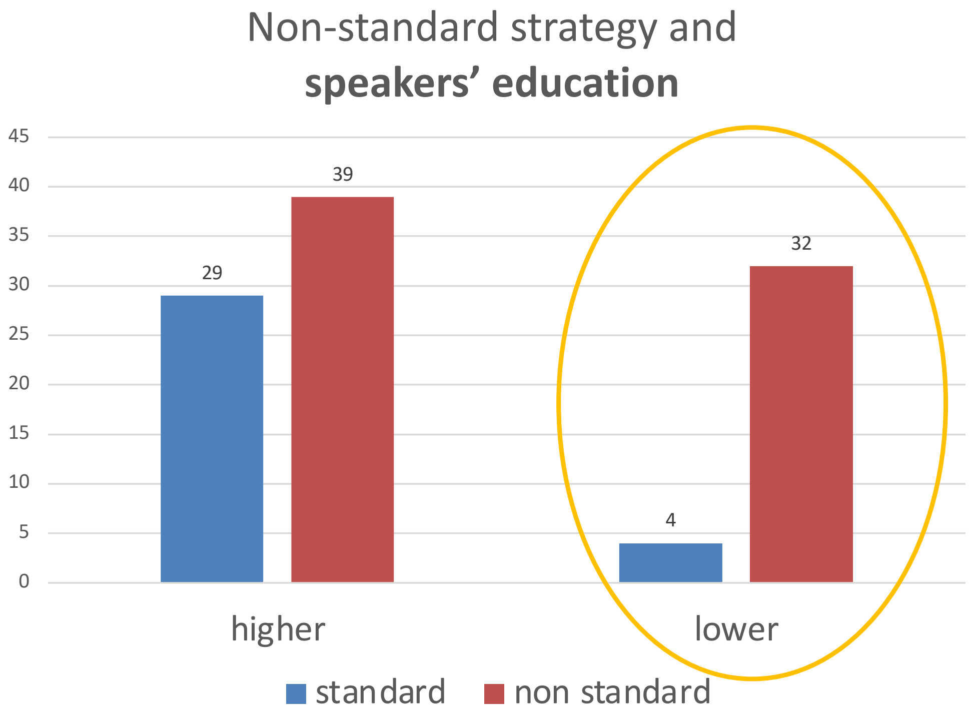
Register variation → Situation metadata

Mauri et al. 2023
Counterfactual conditionals:
Linguistic variation in Italian and beyond

(2) Sociolinguistic variation: subjunctive vs. indicative in counterfactual conditionals



According to the Fisher exact test, the result is significant at p < .01**, with p=< 0.0082, N=104.



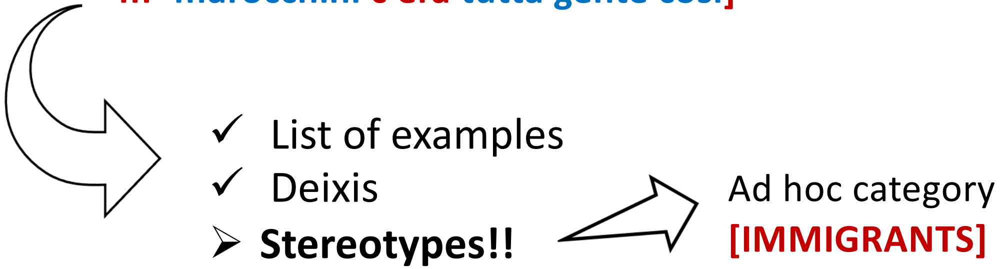
According to the Fisher exact test, the result is significant at **p < .001*****, with p=< 0.0009, N=104.

(3) Investigating speakers’ positioning and category (co-)construction

(6) ParlaBO, PBC007

Arrival of the allied troops in Romagna, during WWII

BOI019: eh sai l'inghi- l'inghilterra era una potenza anche::	*you know England was a power also*
nel nel in i:ndia dalle parti infatti lì **[c'erano degli india:ni c’era**	*in the area of India, so there were also Indians*
m- marocchini c'era tutta gente così]	*Moroccans, people like that*



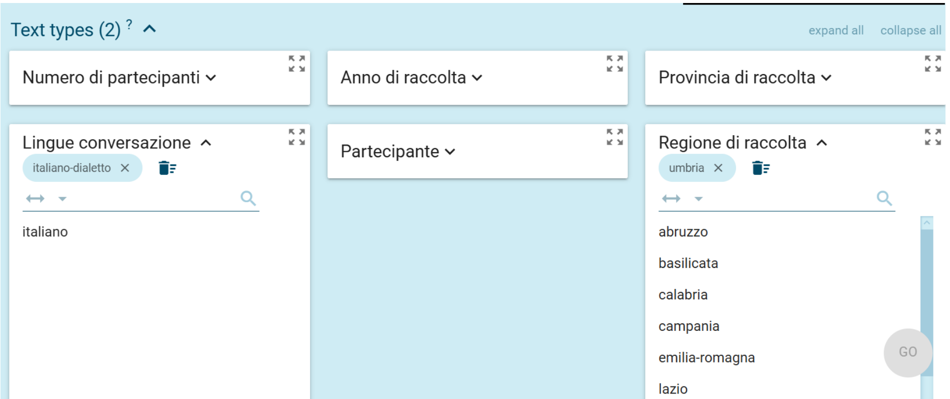
- ✓ List of examples
- ✓ Deixis
- ➢ **Stereotypes!!**

Ad hoc category
[IMMIGRANTS]

Metadata BOI019
Region of birth: Emilia-Romagna
Age: 91-95
Education: Diploma
Occupation: Retired

Mauri et al. 2022: ad hoc categorization and positioning

(4) Investigating geographical variation and dialects



Text types (2) ? expand all collapse all

Numero di partecipanti | Anno di raccolta | Provincia di raccolta

Lingue conversazione: italiano-dialetto ✕ — italiano

Partecipante

Regione di raccolta: umbria ✕ — abruzzo, basilicata, calabria, campania, emilia-romagna, lazio

(7) KIPasti, KPC009

PKP089: [°ma° diec']euro se vai a mangia' da qualche parte (.) ma **magni de brutto** eh: , (.) cioè: ((ride))
PKP088: ((ride))
PKP089: #[eh so' ita] so' ita a paghe': (.) >gli ho fatto< no: da:i pago io: perchè se no pagate sempre voi (.) m'avete fatto veni' a ce:na (.) a pranzo e:h
PKP088: [mh.]
PKP089: #so' ita a paghe' citti:. (.) ((respiro)) [QUANTO]?

with ten euros if you go eat somewhere
you eat a lot I mean ((laughs))
((laughs))
so I went to pay and I said this is on me
because you always pay, I've come for lunch I've come for dinner

so i went to pay, people ((breaths loudly) how much?

# : code-switching

✓ Use of vernacular and dialect

Metadata PKP089
Origin: Umbria (Italy)
Age: 21-25
Degree: Vocational training
Occupation: Craft and related trades worker

(5) Investigating jargon

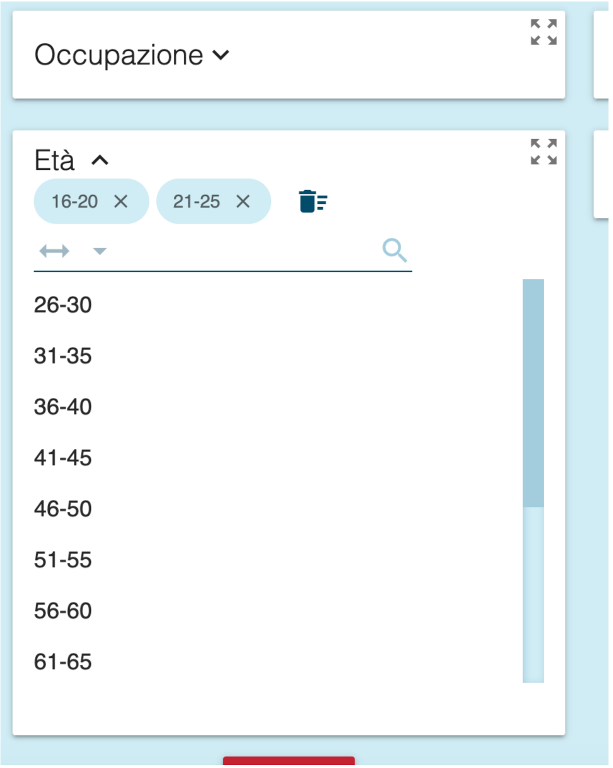
(8) KIPasti, KPN022

PKP070: beh i pulotti possono arrivare lo ste[sso] — *well the police may come anyway*

PKP076: [mh]mh quello sì — *yes that's true*

PKP075 i? — *the?*

PKP070 pulotti — *the police*



Occupazione

Età

16-20 ✕ 21-25 ✕

26-30

31-35

36-40

41-45

46-50

51-55

56-60

61-65

Metadata PKP070
Origin: Veneto (Italy)
Age: 21-25
Degree: High school
Occupation: University student

Metada PKP076
Origin: Veneto (Italy)
Age: 56-60
Degree: Middle school
Occupation: Craft workers

Metadata PKP076
Origin: Veneto (Italy)
Age: 16-20
Degree: Middle school
Occupation: School student

(6) StraParlaBO: features of learners' varieties

➢ **Comparability with other varieties of Italian!**
e.g. vernacular, non-standard

(9) SBIB002

PSB026: ehm **una** cliente **venuto** qua mio $ negozì **lasciato** una bici	***a:ART.F*** *client* ***come:PTCP.M.S*** *here my shop* ***left:PTCP.M.S*** *a bike*

$: nonexistent word

(10) SBIB009

PSB058: ah poi diverso lavorare cercare **lavorare** difficile	*then different work look for* ***work:V*** *hard*
PSB005: qui (.) o in bangladesh?	*here or in Bangladesh?*
PSB058: no: bangladesh tanti **lavorare** // **tante lavoro** c'è	*no Bangladesh a lot* ***work:V*** *//* ***many:F.P work:N*** *there is*
PSB005: qui è più difficile?	*here it is harder?*
PSB058: sì sì cercare difficile **lavorare**	*yes yes look for hard* ***work:V***
però bangladesh lavorare (.) soldi di meno	*but Bangladesh work (.) less money*

- ✓ Topic-focus organization of syntax
- ✓ No clearcut distinction in PoS
- ✓ Existential constructions
- ✓ Verbal morphology and use
- ✓ Gender agreement and marking

Metadata PSB026; PSB058
Origin: Bangladesh
Mother tongue: Bangla
Age: 36-40; 16-20
Years spent in Italy: 14; 3
Degree: University; Middle school
Occupation: Sales worker;
Vocational training student

(7) StraParlaBO: code-switching I

(11) SBIC001

PSB002: ci- cina ci sono otto ((ride)) sistemi	*China there are eight ((laughs)) systems*
PSB001: ah sì	*that's right*
PSB002: di:: >come si dice< di modo [modi] [e:] cucinare	*of how do you say that way ways of cooking*
PSB001: [mhmh] [sì]	*yes*
PSB002: #_suan leba zhong shuofa	***well, eight ways***
PSB001: #_ah, bazhong wo yiwei ni shuo caixi	***eight ways, i thought you meant eight cuisines***
PSB002: #_dui caixi wo xiang shuo	***yes, cuisines I wanted to say***
PSB001: s::ì	*yes*
PSB002 #_dan caixi wo bu zhidao zenme qu fanyi ba zhong shuofa yinggai ye buneng zheme	***but cuisines I don't know how to translate, eight ways, maybe you can't say it***
PSB001: otto filoni di	*eight styles of*
PSB005: mhmh	*mhmh*
PSB001: cioè di cucina	*I mean of cooking*
PSB005: okay	*okay*

- Participant and discourse-related code switching
- Focus on form
- Language mediation

Metadata PSB001; PSB002
Origin: China
Mother tongue: Mandarine Chinese; Xiang
Age: 26-30; 31-35
Years spent in Italy: 4; 5
Degree: PhD; Bachelor's degree
Occupation: Professionals; Student

(7) StraParlaBO: code-switching II

Metadata PSB067; PS069
Origin: Morocco
Mother tongue: Moroccan Arabic; French
Age: 41-45; 46-50
Years spent in Italy: 21; 17
Degree: diploma
Occupation: Elementary occupations; unemployed

(12) SBIA013

PSB005: qual è la più bella la città più bella del marocco	*which city is the nicest in Morocco?*
PSB069: no no agadir	*Agadir*
PSB005: okay	*okay*
PSB069: la città grande	*the big city*
PSB067: #_ zwina agadir? // no per me no	***Agadir is nice?*** *// I don't think so*
PSB069: perché mh	*because*
PSB068: #_ ma 'mri mshit liha	***I have never been there***
PSB067: #_ min casablanca	***from Casablanca***
PSB069: non è non è (.) no // non è abita (.) solo	*it is not it is not // it doesn't live, only*
PSB067: #_ wash kat 'ajbek agadir? zwina? hsen men casa?	***do you like Agadir? Is it better than Casablanca?***
PSB069: #_ batatan	***not at all***
PSB068: #_ ghadi nt'assab ((ride))	***I am getting angry ((laughs))***
PSB066: ((ride))	*((laughs))*
PSB067: no no la più bella è casablanca	*the nicest is Casablanca*
PSB005: okay	*okay*

- Participant-related code-switching

Overview

1. Corpus design and implementation

2. The modules

3. Using the corpus: some case studies

4. Next steps and future challenges

Guidelines for a Modular, Incremental, Replicable resource (MIR)

KIParla is the first Modular, Incremental, and Replicable resource of spoken Italian (**MIR resource**)

✓ All the methodological choices are being documented at every stage, in compliance with the FAIR principles (Wilkinson et al. 2016)

➢ **MIR Guidelines**: the whole **protocol will be fully accessible and replicable** also for other languages in and outside Europe (**QuiénHabla? QuiParle? WerSpricht?...).**

Next steps and future challenges

BOSCO, Cristina et al. 2020. *KIPoS @ EVALITA2020: Overview of the Task on KIParla Part of Speech Tagging* In: *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020: Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian Final Workshop* [online]. Turin: Accademia University Press, 2020. Available on the Internet: <http:// books.openedition.org/aaccademia/7743>. ISBN: 9791280136329. DOI: https://doi.org/10.4000/ books.aaccademia.7743.

Computational developments:

✓ Semi-automatic transcription (?!?)

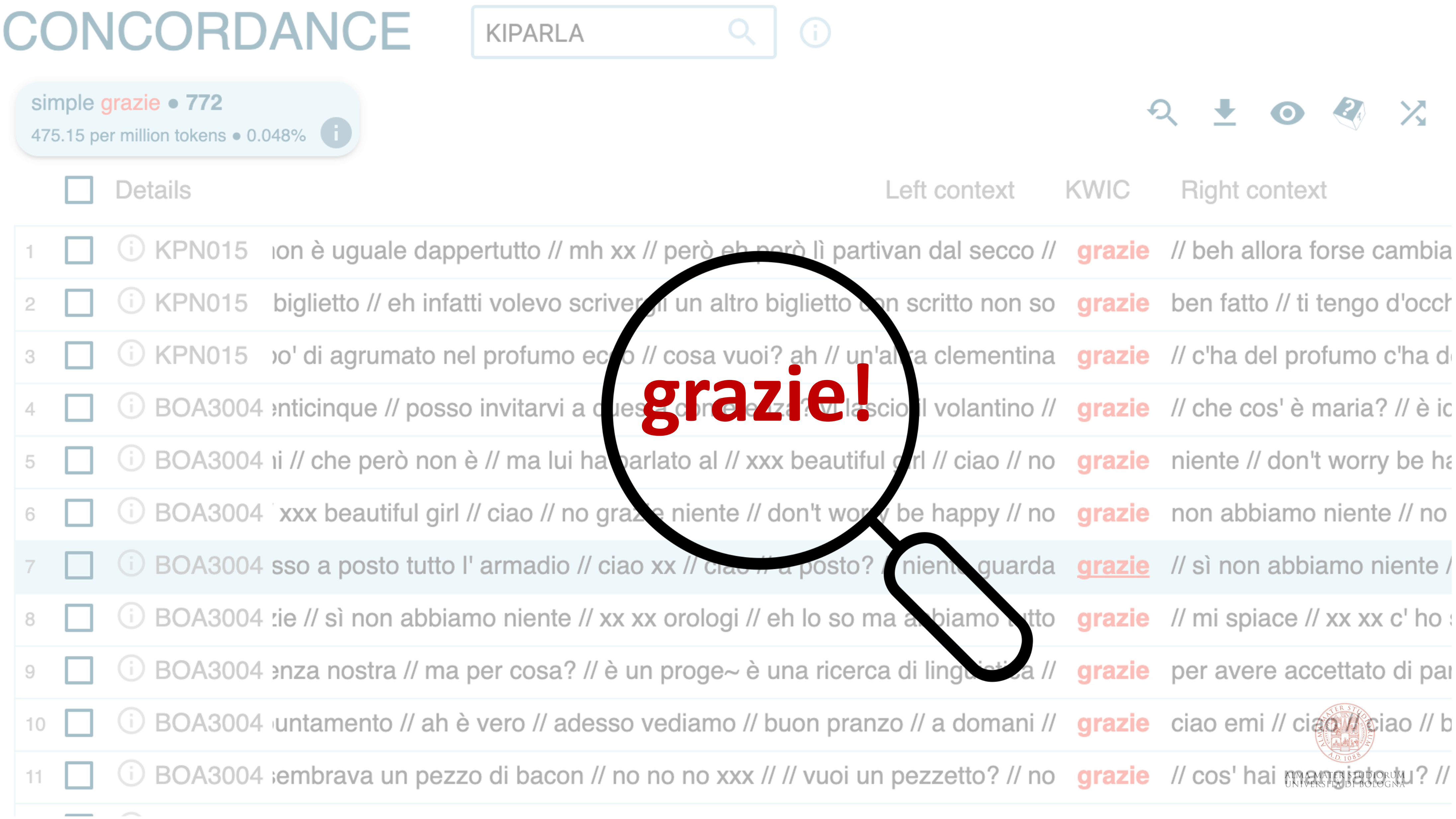
✓ Lemmatization

✓ PoS tagging

✓ UD Treebank of Spoken Italian, based on KIParla

Corpus expansion:

✓ **New modules** >> *new collaborations*



grazie!

CONCORDANCE

KIPARLA

simple grazie • 772

475.15 per million tokens • 0.048%

	□	Details	Left context	KWIC	Right context
1	□	ⓘ KPN015	on è uguale dappertutto // mh xx // però eh però lì partivan dal secco //	grazie	// beh allora forse cambia
2	□	ⓘ KPN015	biglietto // eh infatti volevo scrivergli 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	nticinque // posso invitarvi a questa conferenza? vi lascio il 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 pa
10	□	ⓘ BOA3004	untamento // ah è vero // adesso vediamo // buon pranzo // a domani //	grazie	ciao emi // ciao // ciao // b
11	□	ⓘ BOA3004	sembrava un pezzo di bacon // no no no xxx // // vuoi un pezzetto? // no	grazie	// cos' hai mangiato tu? //

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA