



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



Transcribing internal diversity: challenges and experiments from the KIParla Corpus

**Silvia Ballarè, Caterina Mauri and Eleonora
Zucchini**

**WS5: Developing models for linguistic research: transcription and
training data usage in low-resource scenarios**

Language and languages at the crossroads of disciplines

Lille, 01-03/09/2025

Outline

1. Capturing variation: the KIParla corpus

- Main features

- Methodologies and practices

2. Transcribing spoken data

- Methodologies and practices

- Challenges

- (Possible) solutions

3. Future perspectives

Outline

1. Capturing variation: the KIParla corpus

Main features

Methodologies and practices

2. Transcribing spoken data

Methodologies and practices

Challenges

(Possible) solutions

3. Future perspectives

The KIParla corpus

The building of the KIParla corpus started almost 10 years ago (!) thanks to the joint effort of **Caterina Mauri** and **Silvia Ballarè** (University of Bologna), **Eugenio Gorla** and **Massimo Cerruti** (University of Turin).



Over the years, several other researchers have joined the project, as **Eleonora Zucchini** (Masaryk University), **Ludovica Pannitto** (University of Bologna) and **Beatrice Bernasconi** (University of Turin).



Modularity and incrementality

- The KIParla corpus consists of 4 independent modules;
- It is possible to access each of them individually;
- Or have access to the whole resource (**2.328.193** tokens);
- We are currently working on new modules and others will be (hopefully!) added over time.

Modules: corpora of spoken Italian with an internal structure that provide access to a large set of metadata.

- **Main goal:** build a resource that could gather data collected in different places, at different times, with different scientific aims and different fundings.



Module 1: KIP (668.581 tokens)

- Spoken data collected in university
→ **Main focus:** register variation
- ✓ Bologna and Turin (2016-2019);
- ✓ University students and professors;
- ✓ Different interactional contexts.

Interactions	Hours
Lessons	25:45:12
Exams	06:20:22
Office hours	06:48:19
Semi-structured interviews	14:06:15
Spontaneous conversation	16:23:08
Total	69:23:08

Module 2: ParlaTO (561.388 tokens)

Module 3: ParlaBO (703.376 tokens)



→ **Main focus:** social variation

✓ **Turin** (2019)

✓ **Bologna** (2021-2024)

- ✓ Speakers with diverse social characterization;
- ✓ Only one interactional context (semi-structured interview).

Class	Hours
Young (16 < x < 29 y.o.)	16:56:47
Adults (30 < x < 59 y.o.)	15:39:01
Elderly (over 60 y.o.)	16:15:26
Total	48:51:16

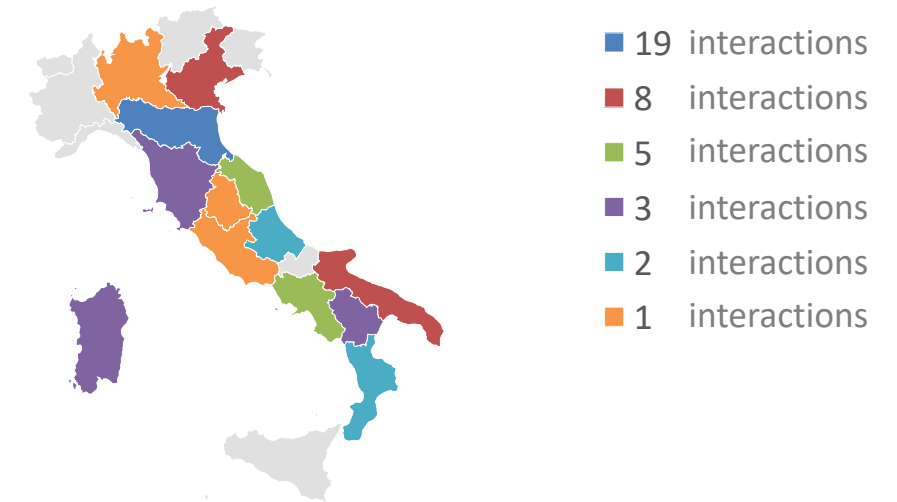
Class	Hours
Young (16 < x < 29 y.o.)	18:46:50
Adults (30 < x < 59 y.o.)	24:21:43
Elderly (over 60 y.o.)	22:34:52
Total	65:43:25



Module 4: KIPasti (482.887 tokens)

- Spoken data collected during kitchen table interactions
→ **Main focus:** geographic variation
- ✓ All over Italy (2020-2024);
- ✓ Speakers with diverse social characterization (close-knit relations);
- ✓ Only one interactional context.

Area	Hours
North	19:25:31
Center	06:50:19
South and islands	16:33:29
Total	42:49:19



Balanced sample of the population – 2020 ISTAT data

Con tecnologia Bing
© GeoNames, Microsoft, TomTom

KIParla

Tokens: 2.326.171

Metadata (core-set)

Conversations

- **Kind of interaction:** kitchen table conversation, semi-structured interviews, spontaneous conversation, exam, office hours, lesson.
- **Number of participants:** 1-6.
- **Relationship between participants:** symmetric/asymmetric.
- **Moderator:** presence/absence.
- **Year:** 2017-2024.
- **Place:** Italian regions.

Speakers

- **Gender:** M, F, N/A.
- **Age:** 16-over85.
- **Occupation:** administrative, artisan, pensioner, professional, retailer, student, technical, unemployed, unqualified.
- **Educational degree:** elementary, middle school, technical/professional school, high school, BA/MA, PhD.
- **Origin:** Italian regions.

StraParla-BO and StraParla-TO

- Italian spoken within communities of **speakers with migratory background** with complex multilingual repertoires;
- Communities involved: Bengalese, Chinese, Moldovian, Peruvian and Ukranian.

→ Observation of learners' varieties, social and geographic variation.

- Data collected in Bologna and Turin (2024 - ongoing);
- Speakers with **multilingual repertoires, different social characterization and origin**;
- Two contexts: semi-structured interview, spontaneous conversation;
- Data collection completed, transcription ongoing;
- **EXPECTED: ca. 128 hrs, 1.250.000 tokens**;
- **Release: early 2026.**



Data collection: before starting



GDPR - Regolamento 2016/679

- Who is responsible for the data?
- What metadata will be stored? Will it be aggregated?
- Where will the data be stored?
- Who has access to the data?
- Do the recordings contain sensitive information (names, addresses etc.)?



Università degli Studi di Torino
Dipartimento di Studi Umanistici



Alma Mater Studiorum - Università di Bologna
Dipartimento di Lingue, Letterature e Culture Moderne

Informazioni sul trattamento dei dati personali ai sensi dell'art. 13 del Regolamento
2016/679/UE

Versione n. 1 del ____/____/2021

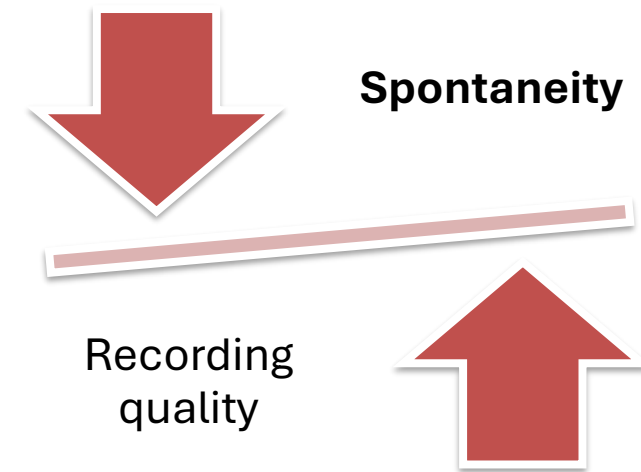
Data collection: tools

Documents:

- Privacy policy;
- Consent form.

Recording tools:

- Zoom H4n Pro recorder;
- Smartphone (for practical reasons).



Data archiving

After each recording, using a OneDrive folder, the collector has:

- Uploaded and named (with an alphanumeric code) the audio file (.mp3, .wav);
- Entered (in two separate Excel files) information about the participants and the conversation;
- Uploaded the consent forms signed by the participants;
- Started the transcription.

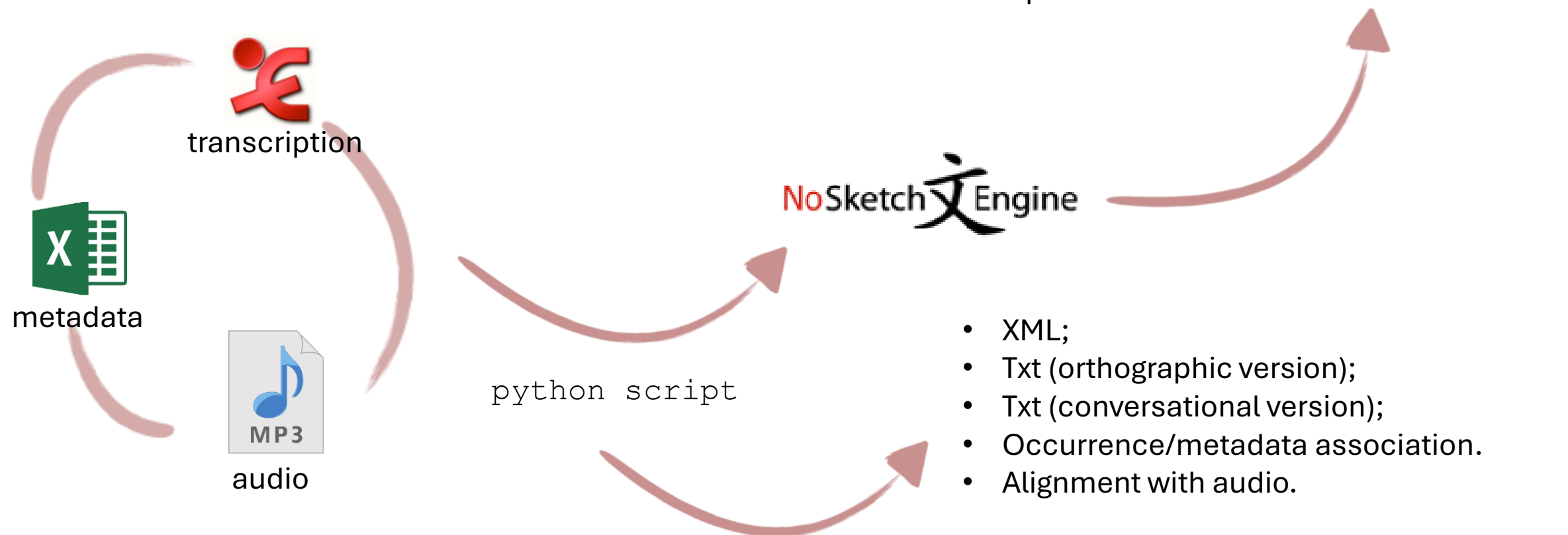
Conversations – KIP module

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	type	code	duration	icipants nur	p1	p2	p3	p4	p5	p6	p7	pants relati	moderator	topic	year	point
33	conversazione libera	BOA3017	0:30:22	4	BO139	BO145	BO146	BO147				simmetrico	no	libero	2019	BO
34	conversazione libera	BOA3018	0:27:14	2	BO139	BO145						simmetrico	no	libero	2019	BO
35	conversazione libera	BOA3019	0:23:15	4	BO149	BO150	BO151	BO152				simmetrico	no	libero	2019	BO
36	conversazione libera	BOA3020	0:22:52	1	BO152	BO153	BO154					simmetrico	no	libero	2019	BO
37	conversazione libera	BOA3021	1:11:38	4	BO155	BO156	BO157	BO158				simmetrico	no	libero	2019	BO

Participants – ParlaTO module

	A	B	C	D	E	F	G	H
1	participant code	participant occupa	participant sex	files in which participant appears	participant b	participant ag	participant degree	
2	TOR001	intell	F	PTA002	lombardia	26-30	phd	
3	TOR002	intell	M	PTA002	piemonte	26-30	phd	
4	TOR003	intell	M	PTA002	piemonte	26-30	laurea	
5	TOI002	oper	M	PTA002	piemonte	26-30	it	
6	TOI003	oper	M	PTA002	piemonte	26-30	it	

Data access



Outline

1. Capturing variation: the KIParla corpus

Main features

Methodologies and practices

2. Transcribing spoken data

Methodologies and practices

Challenges

(Possible) solutions

3. Future perspectives

Data collection and transcription: it takes a village!



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

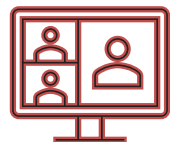
2018 – today: around 100 students (universities of Bologna and Turin) took part in the corpus building.

→ **internship**



Training

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



Supervision

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

Data transcription



Data are **manually** transcribed using elan.

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

“Nearly-globalized set of instructions for transcription” (Slembrouck 2007: 823).

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



Data transcription: a decision-making practice

Main goal: balance between accuracy and ‘searchability’ (homogeneity and processing).

Ex: Italian displays great internal variability:

We decided to systematically disregard phonetic/phonological variation and transcribe phenomena that involve morphology (at least, in principle...!).

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

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

Revision: typos, overlaps, ...

Anonymization: deletion of sensitive data from the transcription and from the audio track.

Issue n.1: language contact



- Italo-romance dialects derive, as Italian, from Latin
→ sister languages (and not variety of the same language);
- Monolingual and bilingual interactions in informal context;
- Intermediate varieties (i.e. regional Italians).
- Languages spoken by speakers with migratory background.

TOI035	#proprio una turineisa	‘a real Turinese ’
TOR001	u~ u:na delle pochissime	‘one of the few’
TOR003	°mhmh°	‘mhmh’
TOI035	[mh è vero]	‘mh it is true’
TOR001	[e:::h]	‘eh’

Issue n.1: language contact

- ITA
- DIA
- UNCERTAIN

TOI118 #vinteset a m ricordu pi nen
TOI118 #a m dan mila euro
TOI118 #ma cuntent. sun sempre dui miliun
TOI118 #poi fasu el cunt
TOI118 ma
TOI118 #a sun mac
TOI118 dieci pezzi da cento
TOI118 e venti da cinquanta
TOI118 sbaglio marco?
TOR002 no no
TOI118 #e l'è la metà ed lon che (.) eh l'aviu
TOI118 eh allora tutti i commercianti han
pensato bene di adeguarsi
TOI119 a ra[ccontare]
TOI118 #[giruma na lira] an n'eu~ an n'euro
TOR002 eh [sì] ((ride))
TOI118 [a] posto a posto (.) è stata la nostra
rovina

twentiseven, I don't remeber anymore
they give me a thousand euro
but (I was) happy. It's still 2 milions milioni
then I do the math
but
they are just
ten hundred pieces
and twenty fifty (pieces)
am I wrong, Marco?
no no
it is half of what we had
so all the traders thought it best to adapt
to narrate
let's switch from the lira to the euro
yes
alright. it was our downfall.



Issue n.1: language contact

- ITA
- DIA
- UNCERTAIN

TOI118 #vinteset a m ricordu pi nen
TOI118 #a m dan mila euro
TOI118 #ma cuntent. sun sempre dui miliun
TOI118 #poi fasu el cunt
TOI118 ma
TOI118 #a sun mac
TOI118 dieci pezzi da cento
TOI118 e venti da cinquanta
TOI118 sbaglio marco?
TOR002 no no
TOI118 #e l'è la metà ed lon che (.) eh l'aviu
TOI118 eh allora tutti i commercianti han
pensato bene di adeguarsi
TOI119 a ra[ccontare]
TOI118 #[giruma na lira] an n'eu~ an n'euro
TOR002 eh [sì] ((ride))
TOI118 [a] posto a posto (.) è stata la nostra
rovina

Pied.

a=m	ricordu	pi	nen
1SG.NOM=1SG.REFL	remember.1SG	anymore	NEG

Ita.

non	mi	ricordo	più
NEG	1SG.REFL	remember.1SG	anymore

'I do not remember anymore'



Issue n.1: language contact

- ITA
- DIA
- UNCERTAIN

TOI118 #vinteset a m ricordu pi nen
TOI118 #a m dan mila euro
TOI118 #ma cuntent. sun sempre dui miliun
TOI118 #poi fasu el cunt
TOI118 ma
TOI118 #a sun mac
TOI118 dieci pezzi da cento
TOI118 e venti da cinquanta
TOI118 sbaglio marco?
TOR002 no no
TOI118 #e l'è la metà ed lon che (.) eh l'aviu
TOI118 eh allora tutti i commercianti han
pensato bene di adeguarsi
TOI119 a ra[ccontare]
TOI118 #[giruma na lira] an n'eu~ an n'euro
TOR002 eh [sì] ((ride))
TOI118 [a] posto a posto (.) è stata la nostra
rovina

- Solution n. 1: # → Transcription unit containing at least one word in an Italo-romance dialect (unit feature)
- Solution n. 2 → Word in a different language (token feature)

Issue n.1: language contact

STIS010 – semi-structured interview with Peruvian speaker (PSB036)

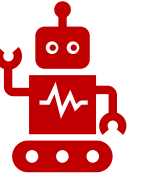
Di là non non c'è: le le classe gratuite (.) [pe]r #el \$estraniero se vuole:: \$studiare l' \$espagno[lo]

‘There are no free courses for foreigners who want to study Spanish’

SPA	ITA	?
extranjero estudiar español	straniero studiare spagnolo	estraniero estudiare espagnolo

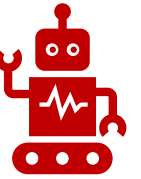
- \$ → Word that cannot be assigned to a specific language

Issue n.2: Manual transcription



- Time consuming
- Internal variability

Towards a semi-automatic transcription?



Performances of Automatic Speech Recognition (ASR) systems have recently skyrocketed thanks to AI

Introduction of ASR in the current pipeline:

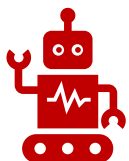
- Reduces transcription time?
- Has a positive or a negative impact on transcription consistency?
- Leads to more standardisation transcriptions?



Experiment with Whisper by OpenAI

Experiment

Whisper



ASR software by OpenAI:

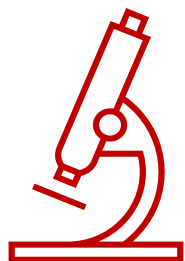
- High level of performance
- Can be used offline (no transfer of data)
- .srt output (can be imported on ELAN)

Set up:

- 11 volunteer transcribers (experts and novices)
- 2 working sessions (2 hours each)
- **Phase 1:** 10-minute recording transcribed manually
- **Phase 2:** revision of 10-minute recording transcribed by Whisper

Analysis:

- Comparison of manual and ASR-assisted transcription with manual transcription revised by expert (**Gold version**)
- To run comparison → script in python



Script: general functions

Output: **two tabular data files** and **revised version of .eaf file**

1. Vertical format

- Transcription tokenised
- Rows: tokens
- Columns: features annotated on each token (overlaps, prolongations ...); token type (linguistic; non linguistic) ...

2. Summary of errors and inconsistencies

- Rows: transcription units (TUs)
- Columns: errors or inconsistencies detected on each unit

3. New .eaf after automatic revision (errors that are not relevant for the analysis).

Script's output: vertical format

00:00:20.000	00:00:21.000	00:00:22.000	00:00:23.000	00:00:24.000	00:00:25.000	00:00:26.000	00:00:27.000	00:00:28.000	00:00:29.000
	okay.		e poi sei venuto per quanto tempo:?		[da]		di quest'anno?		okay.
(.) °l'amico°.	d[all-] dal quattro (.) ottobre (.) fino a adesso.							si.	



token_id	speaker	tu_id	unit	id	span	form	type	variation	jefferson_feats	align
10-4	PSB064	10	10	_	un	un	linguistic	none	_	_
10-5	PSB064	10	10	_	{P}	{P}	shortpause	none	_	_
10-6	PSB064	10	10	_	°l'	l'	linguistic	none	SpaceAfter=No Volume=low	_
10-7	PSB064	10	10	_	amico°.	amico	linguistic	none	Intonation=falling Volume=low	End=20.227
11-0	PSB005	11	11	_	okay.	okay	linguistic	none	Intonation=falling	Begin=20.26 End=20.86
12-0	PSB005	12	12	_	e	e	linguistic	none	_	Begin=21.5
12-1	PSB005	12	12	_	poi	poi	linguistic	none	_	_
12-2	PSB005	12	12	_	sei	sei	linguistic	none	_	_
12-3	PSB005	12	12	_	venuto	venuto	linguistic	none	_	_
12-4	PSB005	12	12	_	per	per	linguistic	none	_	_
12-5	PSB005	12	12	_	quanto	quanto	linguistic	none	_	_
12-6	PSB005	12	12	_	tempo:?	tempo	linguistic	none	Intonation=rising	End=23.589
13-0	PSB064	13	13	_	d[all-]	dall-	linguistic	none	Interrupted=Yes	Begin=23.97
13-1	PSB064	13	13	_	dal	dal	linguistic	none	_	_
13-2	PSB064	13	13	_	quattro	quattro	linguistic	none	_	_
13-3	PSB064	13	13	_	{P}	{P}	shortpause	none	_	_
13-4	PSB064	13	13	_	ottobre	ottobre	linguistic	none	_	_
13-5	PSB064	13	13	_	{P}	{P}	shortpause	none	_	_

Script's output: automatic correction and error detection

W: warning →
corrections
made
automatically

E: errors → what
you should
correct
manually

TU

Modified TU

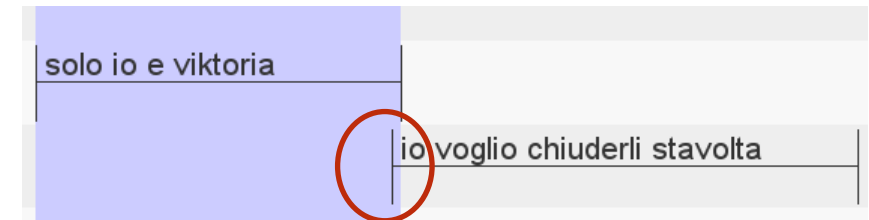
B	C	D	E	F	G	H	I	J	Q	R	S	T	U	AC	AD
speaker	start	end	duration	include	variation	W:normalized_spaces	W:numbers	W:accents	E:volume	E:pace	E:guess	E:overlap	E:overlap_mismatch	original	text
PSB086	0.0	5.022	5.022	True	—		0	0	0	False	False	False	False	vedere la scena che uno di q	vedere la scena che uno di q
PSB086	5.22	10.881	5.661	True	—		0	0	0	False	False	False	False	però stanno attaccati xxx u-	però stanno attaccati xxx u-
PSB036	10.38	10.92	0.54	True	—		0	0	0	False	False	False	False	[e dai]	[e dai]
PSB086	11.32	14.84	3.52	True	—		0	0	0	False	False	False	False	e e vedere e:h siccome lui d	e e vedere e:h siccome lui d
PSB086	16.119	20.213	4.094	True	—		0	0	0	False	False	False	False	e dovevano spostare per osp	e dovevano spostare per osp
PSB086	21.24	25.633	4.393	True	—		0	0	0	False	False	False	False	alle undici xx papÃ non riesc	alle undici xx papÃ non riesc
PSB086	26.453	31.787	5.334	True	—		0	0	0	False	False	False	False	ha fatto una faccia come far	ha fatto una faccia come fan
PSB036	31.72	32.813	1.093	True	—		0	0	0	False	False	False	False	((ride))	((ride))
PSB086	32.36	35.893	3.533	True	—		0	0	0	False	False	False	False	stessa cosa questo papÃ nc	stessa cosa questo papÃ no
PSB086	36.433	41.074	4.641	True	—		0	0	0	False	False	False	False	lo sai che figlia mag- Ã" imp	lo sai che figlia mag- Ã" impe
PSB086	41.372	43.319	1.947	True	—		0	0	0	False	False	False	False	in ospedale lo spostano no?	in ospedale lo spostano no?
PSB086	43.44	45.72	2.28	True	—		0	0	0	False	False	False	False	voglio che parli tu con medic	voglio che parli tu con medic
PSB086	45.973	49.279	3.306	True	—		0	0	0	False	False	False	False	ma papÃ ma non riesco alle	ma papÃ ma non riesco alle
PSB086	50.907	52.093	1.186	True	—		0	0	0	False	False	False	False	ci tenevo che tu ci sei	ci tenevo che tu ci sei
PSB086	52.31	54.358	2.048	True	—		0	0	0	False	False	False	False	<papÃ vengo (.) papÃ >	<papÃ vengo (.) papÃ >
PSB086	54.58	57.003	2.423	True	—		0	0	0	False	False	False	False	non so prendo la moto corro	non so prendo la moto corro
PSB086	57.37	60.971	3.601	True	—		0	0	0	False	False	False	False	una scena davvero davvero i	una scena davvero davvero i
PSB086	61.17	66.569	5.399	True	—		0	0	0	False	False	False	False	e e lui mi sempre racconta q	e e lui mi sempre racconta q
PSB005	65.906	66.546	0.64	True	—		0	0	0	False	False	False	False	[mhmhmh]	[mhmhmh]
PSB086	67.17	69.177	2.007	True	—		0	0	0	False	False	False	False	tutta la vita: le medie superi	tutta la vita: le medie superic
PSB086	69.43	73.195	3.765	True	—		0	0	0	False	False	False	False	e lui dice ah mio f- miei figli	e lui dice ah mio f- miei figli t
PSB086	73.35	75.98	2.63	True	—		0	0	0	False	False	False	False	e:h hanno questo comportar	e:h hanno questo comportar

Script's output: automatic correction and error detection

The script:

Automatically correct certain inconsistencies:

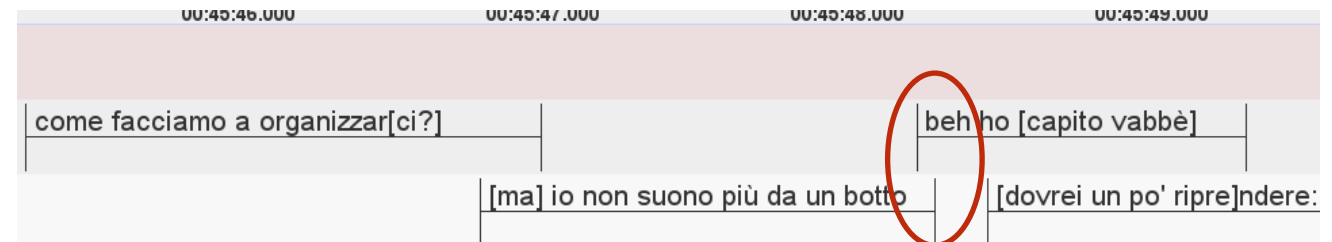
- accents
- empty annotations
- neglectable temporal overlaps



solo io e viktoria	
	io voglio chiuderli stavolta

Detect other inconsistencies and report them in a table:

- Unpaired brackets → *[quale par]te d'italia s-]*
- Missing or extra overlap annotation



00:45:46.000	00:45:47.000	00:45:48.000	00:45:49.000
come facciamo a organizzar[ci?]		beh ho [capito vabbè]	
	[ma] io non suono più da un botto		[dovrei un po' ripre]ndere:

Outline

1. Capturing variation: the KIParla corpus

Main features

Methodologies and practices

2. Transcribing spoken data

Methodologies and practices

Challenges

(Possible) solutions

3. Future perspectives

Next steps

In parallel with collection/transcription of new modules

- Check the results of the comparison between manual and semi-automatic transcription for the different data;
- Linguistic annotation: POS-tagging, lemmatization, UD.



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



Thank you! Questions?

Silvia Ballarè (University of Bologna)
Caterina Mauri (University of Bologna)
Eleonora Zucchini (Masaryk University)

Language and languages at the crossroads of disciplines

Lille, 01-03/09/2025