

Reuse by design

A pivot-based architecture for the KIParla corpus of Spoken Italian

Ludovica Pannitto & Caterina Mauri
Department of Modern Languages, Literatures and Cultures,
University of Bologna

CLARIN Café, Wed Jun 24th

The KIParla corpus



CLARIN

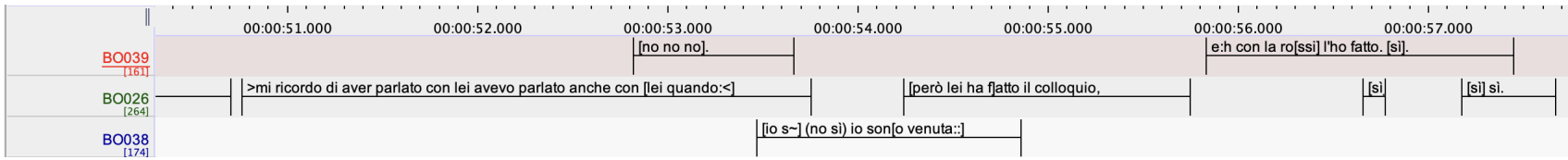
GitHub

- ✓ Large, modular corpus of **spontaneous spoken Italian**
- ✓ Initiated in 2016, collaborative project between **University of Bologna** and **University of Turin** (Mauri, Ballarè, Goria, et al., 2019)
- ✓ Currently: **6 released modules**
  KIP (Goria and Mauri, 2018), KIPasti (Mauri, Ballarè and Zucchini, 2024a), ParlaBO (Mauri, Ballarè and Zucchini, 2024b), ParlaTO (Cerruti and Ballarè, 2021), StraParlaBO & StraParlaTO (Mauri, Zucchini, et al., 2025)
- ✓ ≈ **350 hours** of recorded speech; ≈ **3.2M tokens**
- ✓ Wide range of settings: free conversations, semi-structured interviews, lectures, oral exams, office hours
- ✓ Rich set of metadata both at speaker and conversation level that enable **sociolinguistic** and **cross-context** analyses

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

What a conversation looks like

ELAN — time-aligned tiers



Jefferson linearization

```
1  BO026  >mi ricordo di aver parlato con lei avevo parlato anche con [lei quando:<]
2  BO039  [no no no].
3  BO038  [io s~] (no sì) io son[o venuta::]
4  BO026  [però lei ha f]atto il colloquio,
5  BO039  e:h con la ro[ssi] l'ho fatto. [sì].
6  BO026  [sì]
7  BO026  [sì] sì.
```

Orthographic linearization

```
1  BO026  mi ricordo di aver parlato con lei avevo parlato anche con lei quando
2  BO039  no no no
3  BO038  io s~ no sì io sono venuta
4  BO026  però lei ha fatto il colloquio
5  BO039  eh con la rossi l'ho fatto sì
6  BO026  sì
7  BO026  sì sì
```

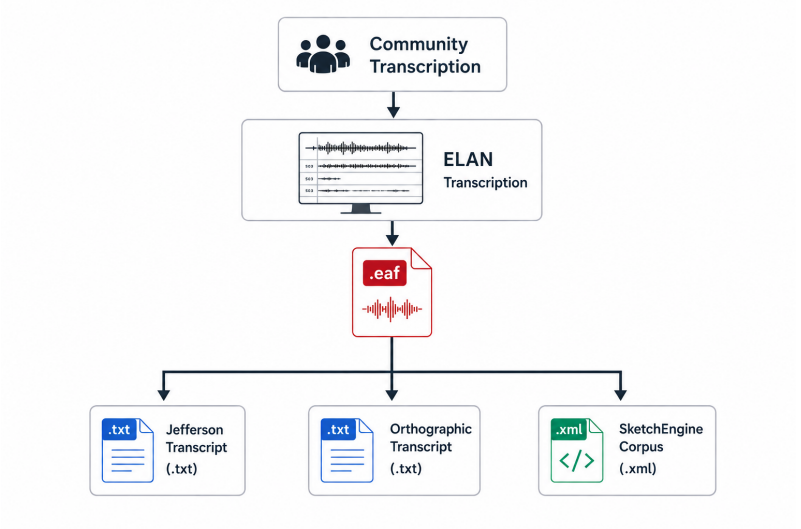
The `.eaf` output is expressive but **opaque**: overlap is implicit in time offsets, units are not well defined, and the format resists automated processing and version control.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Building a spoken corpus: metadata and transcription

Transcription: Jefferson conventions (Jefferson, 2004)

Symbol	Meaning	Type	Level
`word,`	Weakly rising intonation	Categorical	Token
`word?`	Rising intonation	Categorical	Token
`word.`	Falling intonation	Categorical	Token
`wo:rd`	Prolonged sound	Categorical	Character
`(.)`	Short pause	Categorical	Token
`word=word`	Prosodically linked units		
`°word°`	Lower volume	Categorical	Token
`WORD`	Higher volume	Categorical	Token
`wor-`	Interrupted word	Categorical	Token
`>word<`	Faster speech	Categorical	Span
`<word>`	Slower speech	Categorical	Span
`[word]`	Overlapping speech	Relational	Span
`(word)`	Uncertain transcription	Categorical	Span
`xxx`	Unintelligible sequence	Categorical	Token
`((laughs))`	Non-verbal behavior	Categorical	Token



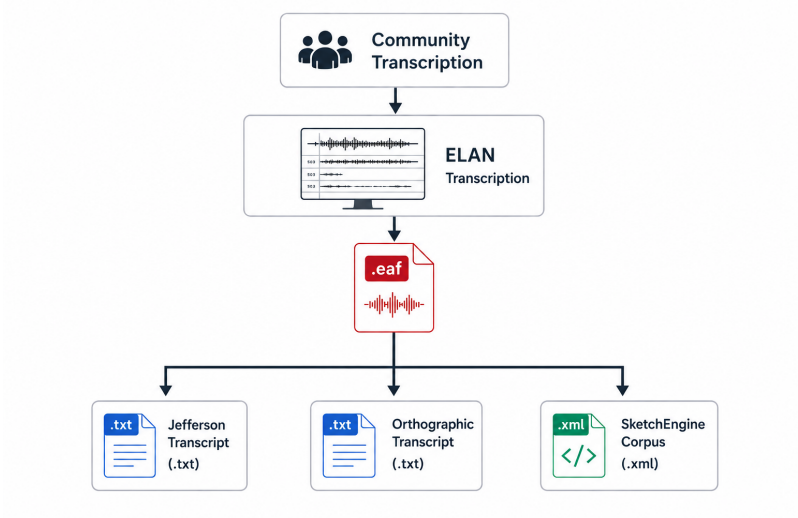
`.eaf` was the primary artefact; exports were ad hoc and unidirectional.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

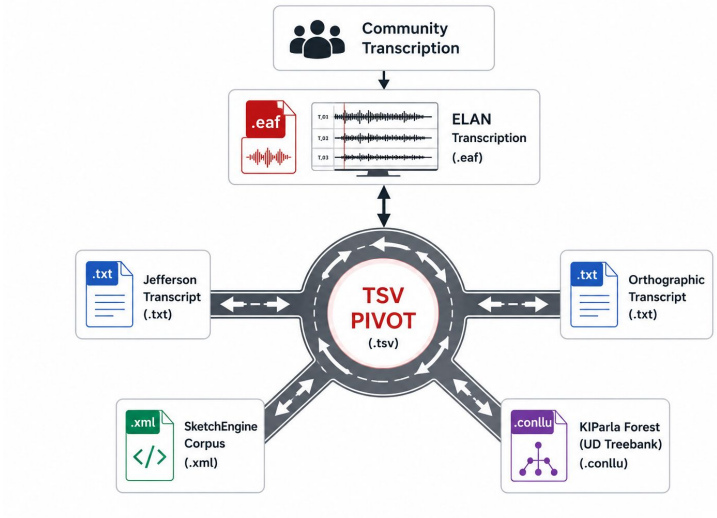
From pipeline to roundabout

→ We introduce a **pivot format** that sits at the centre: ELAN files, NLP-suitable formats, and artifacts are all derived — and **re-derived** — from it. Corrections propagate deterministically.

Before



Now



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

The pivot format

A **pseudo-tokenized, verticalized, tab-separated** format: each row = one validated token, enriched with all relevant features.

1	28-0	BO026	28	>mi	mi	linguistic	_	Begin=50.761	_	Fast=0-2	_	_
2	28-1	BO026	28	ricordo	ricordo	linguistic	_	_	_	Fast=0-7	_	_
3	28-2	BO026	28	di	di	linguistic	_	_	_	Fast=0-2	_	_
4	28-3	BO026	28	aver	aver	linguistic	_	_	_	Fast=0-4	_	_
5	28-4	BO026	28	parlato	parlato	linguistic	_	_	_	Fast=0-7	_	_
6	28-5	BO026	28	con	con	linguistic	_	_	_	Fast=0-3	_	_
7	28-6	BO026	28	lei	lei	linguistic	_	_	_	Fast=0-3	_	_
8	28-7	BO026	28	avevo	avevo	linguistic	_	_	_	Fast=0-5	_	_
9	28-8	BO026	28	parlato	parlato	linguistic	_	_	_	Fast=0-7	_	_
10	28-9	BO026	28	anche	anche	linguistic	_	_	_	Fast=0-5	_	_
11	28-10	BO026	28	con	con	linguistic	_	_	_	Fast=0-3	_	_
12	28-11	BO026	28	[lei	lei	linguistic	_	_	_	Fast=0-3	_	0-3(245)
13	28-12	BO026	28	quando:<]	quando	linguistic	_	End=53.754	5x1	Fast=0-6	_	0-6(245)
14	29-0	BO039	29	[no	no	linguistic	_	Begin=52.815	_	_	_	0-2(245)
15	29-1	BO039	29	no	no	linguistic	_	_	_	_	_	0-2(245)
16	29-2	BO039	29	no].	no	linguistic	Intonation=Falling	End=53.662	_	_	_	0-2(245)
17	30-0	BO038	30	[io	io	linguistic	_	Begin=53.465	_	_	_	0-2(245)
18	30-1	BO038	30	s~]	s~	linguistic	Interrupted=Yes	_	_	_	_	0-2(245)
19	30-2	BO038	30	(no	no	linguistic	_	_	_	_	0-2(0)	_
20	30-3	BO038	30	sì)	sì	linguistic	_	_	_	_	0-2(0)	_
21	30-4	BO038	30	io	io	linguistic	_	_	_	_	_	_
22	30-5	BO038	30	son[o	sono	linguistic	_	_	_	_	_	3-4(8)
23	30-6	BO038	30	venuta::]	venuta	linguistic	_	End=54.858	5x2	_	_	0-6(8)
24	31-0	BO026	31	[però	però	linguistic	_	Begin=54.24	_	_	_	0-4(8)
25	31-1	BO026	31	lei	lei	linguistic	_	_	_	_	_	0-3(8)
26	31-2	BO026	31	ha	ha	linguistic	_	_	_	_	_	0-2(8)
27	31-3	BO026	31	f]atto	fatto	linguistic	_	_	_	_	_	0-1(8)
28	31-4	BO026	31	il	il	linguistic	_	_	_	_	_	_
29	31-5	BO026	31	colloquio,	colloquio	linguistic	Intonation=WeaklyRising	End=55.748	_	_	_	_
30	32-0	BO039	32	e:h	eh	linguistic	_	Begin=55.833	0x1	_	_	_
31	32-1	BO039	32	con	con	linguistic	_	_	_	_	_	_
32	32-2	BO039	32	la	la	linguistic	_	_	_	_	_	_
33	32-3	BO039	32	ro[ssi]	rossi	linguistic	_	_	_	_	_	2-5(9)
34	32-4	BO039	32	l'	l'	linguistic	SpaceAfter=No	_	_	_	_	_
35	32-5	BO039	32	ho	ho	linguistic	_	_	_	_	_	_
36	32-6	BO039	32	fatto.	fatto	linguistic	Intonation=Falling	_	_	_	_	_
37	32-7	BO039	32	[sì].	sì	linguistic	Intonation=Falling	End=57.45	_	_	_	0-2(10)
38	33-0	BO026	33	[sì]	sì	linguistic	_	Begin=56.654|End=56.773	_	_	_	0-2(9)
39	34-0	BO026	34	[sì]	sì	linguistic	_	Begin=57.177	_	_	_	0-2(10)
40	34-1	BO026	34	sì.	sì	linguistic	Intonation=Falling	End=57.668	_	_	_	_
41	35-0	BO026	35	anche	anche	linguistic	_	Begin=58.027	_	_	_	_
42	35-1	BO026	35	lei?	lei	linguistic	Intonation=Rising	End=58.552	_	_	_	_
43	36-0	BO038	36	sì.	sì	linguistic	Intonation=Falling	Begin=58.64	_	_	_	_
44	36-1	BO038	36	(.)	[PAUSE]	shortpause	_	_	_	_	_	_
45	36-2	BO038	36	sì	sì	linguistic	_	_	_	_	_	_
46	36-3	BO038	36	per~	per~	linguistic	Interrupted=Yes	_	_	_	_	_
47	36-4	BO038	36	cioè	cioè	linguistic	_	_	_	_	_	_
48	36-5	BO038	36	io	io	linguistic	_	_	_	_	_	_
49	36-6	BO038	36	sono	sono	linguistic	_	_	_	_	_	_
50	36-7	BO038	36	venuta	venuta	linguistic	_	_	_	_	_	_
51	36-8	BO038	36	anche:	anche	linguistic	_	_	4x1	_	_	_
52	36-9	BO038	36	[a	a	linguistic	_	_	_	_	_	0-1(11)
53	36-10	BO038	36	ricevimento	ricevimento	linguistic	_	_	_	_	_	0-11(11)
54	36-11	BO038	36	con	con	linguistic	_	_	_	_	_	0-3(11)
55	36-12	BO038	36	lei].	lei	linguistic	Intonation=Falling	_	_	_	_	0-3(11)
56	36-13	BO038	36	sì	sì	linguistic	_	_	_	_	_	_
57	36-14	BO038	36	sì	sì	linguistic	_	_	_	_	_	_

Three-level validation

- ✓ **TU level**
- ✓ **Span level**
- ✓ **Token level**

Design goals

- ✓ **Co-readable** by humans and machines
- ✓ **Round-trip stable**: tokenize → serialize → original transcript (modulo normalization)

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Continuous integration

The pivot format enables a **DevOps-inspired** curation workflow: the corpus is treated not as a static publication but as a **versioned, continuously maintained codebase**.

Version control (git)

- ✓ Each modification tracked as a diff on specific token rows
- ✓ Corrections are **localized**, **reversible**, and **auditable**
- ✓ Corpus history is inspectable: who changed what, when, and why

Automated validation

- ✓ Well-formedness checks run at every update (character, span, TU levels)
- ✓ Structural errors surface immediately rather than propagating silently
- ✓ Corrections in the pivot propagate **deterministically** to all derived formats

Derived projects

The pivot architecture has already enabled new analytical and infrastructural projects — with more underway.

Lemmatization & POS tagging

Token sequences extracted directly from the pivot and fed to UDPipe (Straka, 2018). Syntactic sub-tokens (e.g. *del* → *di* + *il*) inserted as `type=syntactic` rows — isolatable with a single `grep`.

	token_id	speaker	tu_id	unit	id	span	form	lemma	upos	jefferson_feats
2	0-0	PKP025	0	0	1	faccio	faccio	fare	VERB	_
3	0-1	PKP025	0	0	2	il	il	il	DET	_
4	0-2	PKP025	0	0	3	tè	tè	tè	NOUN	_
5	1-0	PKP026	1	1	1	(°allora°)	allora	allora	ADV	_
6	2-0	PKP025	2	2	1	[accendo]?	accendo	accendere	VERB	_
7	3-0	PKP026	3	3	1	[va	va	andare	VERB	_
8	3-1	PKP026	3	3	2	bene]	bene	bene	ADV	_
9	4-0	PKP026	4	4	1	ci	ci	ci	PRON	_
10	4-1	PKP026	4	4	2	fermiamo	fermiamo	fermare	VERB	_
11	4-2	PKP026	4	4	3	un	un	uno	DET	_
12	4-3	PKP026	4	4	4	attimo	attimo	attimo	NOUN	_

KIParla Forest

UD treebank derived directly from the pivot (Pannitto et al., 2025): 18,050 tokens, 6 conversations (interviews, free conversations, office hours, a lecture).

- ✓ CoNLL-U from pivot; pauses in MISC field; overlaps as explicit cross-references
- ✓ Released in UD v2.17, updated in v2.18

In progress

- ✓ TEI/ISO 24624 export
- ✓ Pragmatic and discourse annotation layers
- ✓ Interactive visualizations

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Conclusions

A spoken corpus pivot format should:

- ✓ **Support version control and diff-based inspection** — corrections are localized, reversible, and auditable
- ✓ **Allow straightforward integration of additional annotation layers** — new columns without touching the token structure
- ✓ **Serve as a single maintained source** from which alternative representations can be **reproducibly derived**
- ✓ **Be co-readable** — inspectable with a text editor or `grep`, yet fully machine-processable

Interoperability is achieved not through adherence to a single format, but through **alignment of curatorial practices**: version control, layered validation, and reproducible derivation.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Thanks!

Questions?

Ludovica Pannitto & Caterina Mauri
University of Bologna
caterina.mauri@unibo.it
ellepannitto@gmail.com

KIParla on GitHub: https://www.github.com/KIParla/

References I

Cerruti, Massimo & Silvia Ballarè (2021). “ParlaTO: corpus del parlato di Torino”. In: *Bollettino dell’Atlante Linguistico Italiano (BALI)* 44, pp. 171–196.

Goria, Eugenio & Caterina Mauri (2018). “Il corpus KIParla: una nuova risorsa per lo studio dell’italiano parlato”. In: *CLUB Working Papers in Linguistics*. Ed. by Francesca Masini & Fabio Tamburini. Vol. 2. Alma Mater Studiorum Università di Bologna, pp. 96–116.

Jefferson, Gail (2004). “Glossary of transcript symbols with an introduction”. In: *Pragmatics & Beyond New Series*. Ed. by Gene H. Lerner. Amsterdam: John Benjamins Publishing Company, pp. 13–31.

Mauri, Caterina, Silvia Ballarè, Eugenio Goria, Massimo Cerruti & Francesco Suriano (2019). “KIParla Corpus: A New Resource for Spoken Italian”. In: *Proceedings of the Sixth Italian Conference on Computational Linguistics*.

Mauri, Caterina, Silvia Ballarè & Eleonora Zucchini (2024a). *Modulo KIPasti*. DOI: 10.60760/unibo/kipasti.

— (2024b). *Modulo ParlaBO*. DOI: 10.60760/unibo/parlabo.

Mauri, Caterina, Eleonora Zucchini, Eugenio Goria & Barbara Bernasconi (2025). *Il progetto DiverSIta. Documentare la diversità nell’italiano parlato*. Bologna: CLUB – Circolo Linguistico dell’Università di Bologna. DOI: 10.6092/unibo/amsacta/8647.

Pannitto, Ludovica, Eleonora Zucchini, Silvia Ballarè, Cristina Bosco, Caterina Mauri & Manuela Sanguinetti (2025). “Introducing KIParla Forest: seeds for a UD annotation of interactional syntax”. In: *Proceedings of the Eighth International Conference on Dependency Linguistics (Syntaxfest 2025)*. Ljubljana, Slovenia, pp. 54–73.

Straka, Milan (2018). “UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task”. In: *Proceedings of the CoNLL 2018 Shared Task*. Brussels, Belgium: Association for Computational Linguistics, pp. 197–207.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA