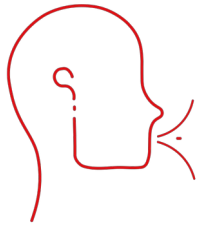


ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA





EXPERIMENTAL **LAB**

Treebanking the KIParla corpus of Spoken Italian

new formats and new perspectives for interoperability

Ludovica Pannitto, 12.05.2025 - University of Gothenburg

People involved in this work

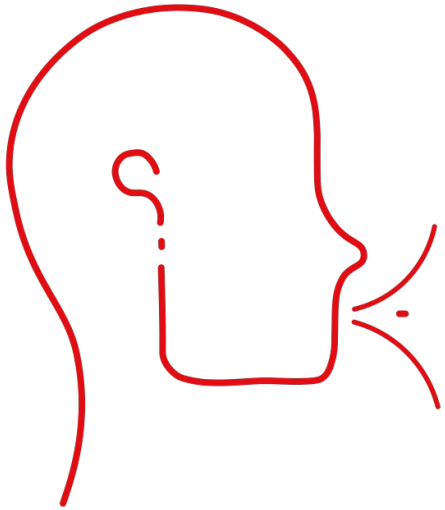
Caterina Mauri	Cristina Bosco	many
Silvia Ballarè	Manuela Sanguinetti	many
Massimo Cerruti	Eleonora Zucchini	...
Eugenio Goria	Martina Simonotti	...
		many
		student interns

In short

1. The KIParla corpus
    a. *INTERMEZZO: Data in spoken language studies*
    b. The current format of the corpus

2. KIParla Forest - the syntactic treebank
    a. *INTERMEZZO: Universal Dependencies (UD)*
    b. A new format for KIParla corpus
    c. Segmentation
    d. Morphosyntactic annotation
    e. Other choices

Enter KIParla:
a growing corpus of Spoken Italian



www.kiparla.it

> corpus of *naturally occurring* Spoken Italian

> started in 2016 as a collaboration between University of Bologna and University of Turin

> built around the concept of *modularity*: new modules are constantly added, with compatible metadata and transcription conventions

> as of today, 4 modules available, 4 to be released soon

Available modules

KIP

~70h, 600K tokens

Lectures, exams, office hours, semi-structured interviews, free conversations recorded in Bologna and Turin

KIPasti

~40h, 430K tokens

free conversations (lunch or dinner time) recorded all over Italy

ParlaBO

~65h, 640K tokens

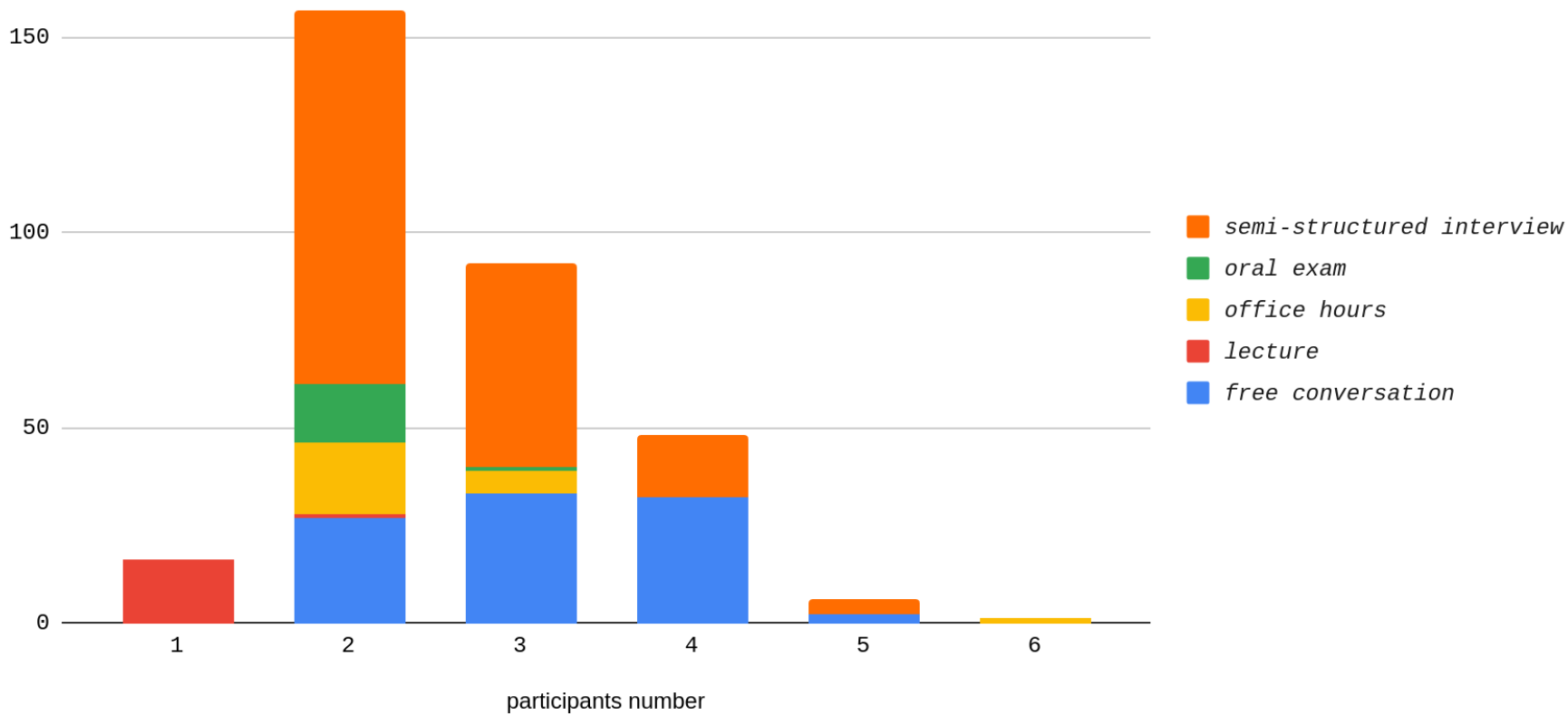
semi-structured interviews recorded in Bologna

ParlaTO

~50h, 500K tokens

semi-structured interviews recorded in Turin

Conversations per type and number of partecipants



Under Construction: StraParlaBO and StraParlaTO

> ~50h each, semi-structured interviews and free conversations among **speakers with a history of international migration** (SIMB) living in the urban area of Bologna and Turin

> It includes code switching phenomena and L2 phenomena

Metadata

Conversation metadata:

> place of collection

> type of interaction

As of today:

- **320 conversations**
- **560 speakers**

Speaker metadata:

> age range

> place of origin

> profession

> highest achieved study title

> gender

> age

\+ module-specific metadata

Participants per occupation and age-range

occupation	16-20	**21-25**	26-30	31-35	36-40	40-45	41-45	46-50	51-55	56-60	61-65	66-70	71-75	76-80	81-85	over 85	**TOT**
artisan		2	2	3					1	4		1					**13**
retailer	1	7	3	6	2				3	3							**25**
unemployed	1	6	2						3		1				1		**14**
self-employed					2												**2**
intellectual	3	8	12	8	9	1	9	8	16	10	10	2	3		2		**101**
not qualified										2							**2**
worker	1	8	5	2			2	2		2							**22**
retired										2	5	9	14	16	8	11	**65**
student	23	**218**	17	5													**263**
technical job	1	3	4	1	1				4	6	1		2	1			**24**
office job		4	2	1	2			3	9	4	1						**26**
TOT	**30**	**256**	**48**	**26**	**16**	**1**	**11**	**13**	**36**	**33**	**18**	**12**	**19**	**17**	**11**	**11**	**558**

INTERMEZZO



Data in Spoken Language Studies

The ideal scenario



(The *Mulino Bianco* repository of language data)

> we **collect** naturally occurring linguistic data in the wild

> data gets effortlessly **annotated** through shared and stable linguistic categories

> we can perform **queries** to extract datasets and formulate theories

The real scenario



(Scary house nobody would dare to enter, with a completely useless external light.)
Generated by chatGPT after several trials

> interactional data takes insane amount of **time and effort** to be collected

> the project easily becomes a layered collection of **ad-hoc categories**

> query **systems and interfaces** never really allow us to perform just the right query

Our scenario



> Human language is first and foremost **spoken language**

> There is growing interest in NLP in modelling language in interaction (i.e., besides *speech*)
("Putting Natural in Natural Language Processing" - Chrupała 2023)

> KIParla is a big (~2M tokens) and carefully constructed resource, but not
easily queryable for linguistic categories at the moment



> Full manual construction comes with a lots of inconsistencies (dis-*agreements*? in-*stability*?)

> Current standards for spoken language annotation do not easily support syntax (and vice-versa?)

> Most human-friendly annotation and query interfaces are not designed for collaborative work and versioning

Our scenario



> Human language is first and foremost **spoken language**

> There is growing interest in NLP in modelling language in interaction (i.e., besides *speech*)
("Putting Natural in Natural Language Processing" - Chrupała 2023)



> KIParla is a big (~2M tokens) and carefully constructed resource, but not
easily queryable for linguistic categories at the moment

> Full manual construction comes with a lot of inconsistencies (dis-*agreements*? in-*stability*?)

> Current standards for spoken language annotation do not easily support syntax (and vice-versa?)

> Most human-friendly annotation and query interfaces are not designed for collaborative work and versioning

FAIR data or FAIR processes?

Research object	Observable
ultimate substrate	speech event
epistemic object	utterance in the speech event
original recording	video/audio tape of the speech event
derived representation	transcription and annotation of the utterance

(Adapted from *"Data in linguistics"* - Lehmann 2004)

"Nothing is, in and of itself, a datum; instead, it is a datum for somebody in some perspective."

> In NLP there is often an unstated assumption about the **reality of the datum**, spoken data confront us with the fact that every bit of information is in fact a **stratification** of several levels of **interpretation**

"a datum for somebody in some perspective"

(1)	Noam	goes	**out**	tonight
	NOUN	VERB	ADP/PRT	ADV/NOUN
(2)	Noam	likes	**social**	media
	NOUN	VERB	ADJ/NOUN	NOUN
(3)	Noam	likes	**his**	car
	NOUN	VERB	DET/PRON	NOUN

(Example from *"Linguistically debatable or just plain wrong?"* - Plank et al. 2014)

> What we call `data` at every step of the process is in fact a form of `gold standard`, where some of the information has been normalized

> The more we depart `standard` language, the more linguistic categories become **unstable**



> We want to use annotation to **let categories emerge** from data

Issues with the traditional annotation process

"The current process of linguistic annotation retains a series of practices that evolved for pure linguistic annotation, i.e., the ***annotation of objective, often technical linguistic aspects of texts****.*

[...]

the unwritten assumption is made that there is exactly one truth, one true value for each variable defined in the annotation scheme for each instance to annotate. This assumption makes sense: ***according to any standard grammar of English****, a word in a sentence can either be a noun or a verb, or any other grammatical category, but not more than one at the same time.*

[...]

disagreement is treated as a kind of statistical noise in the data*, and removed by forcing an agreement by harmonization or automatic aggregation of the annotation"*

(From *"The Perspectivist Data Manifesto"* - pdai.info)

BACK TO KIPARLA



Current pipeline - a new module is born

1. Conversations are **recorded** in order to reach the module-specific objectives and balance

2. Each conversation is **transcribed** in Jefferson-style notation (used in Conversation Analysis)

3. An expert **pseudonymizes** files, **revises** the transcription and compiles the **metadata** files

4. Files are prepared for sharing on noSketchEngine through **ad-hoc scripts**

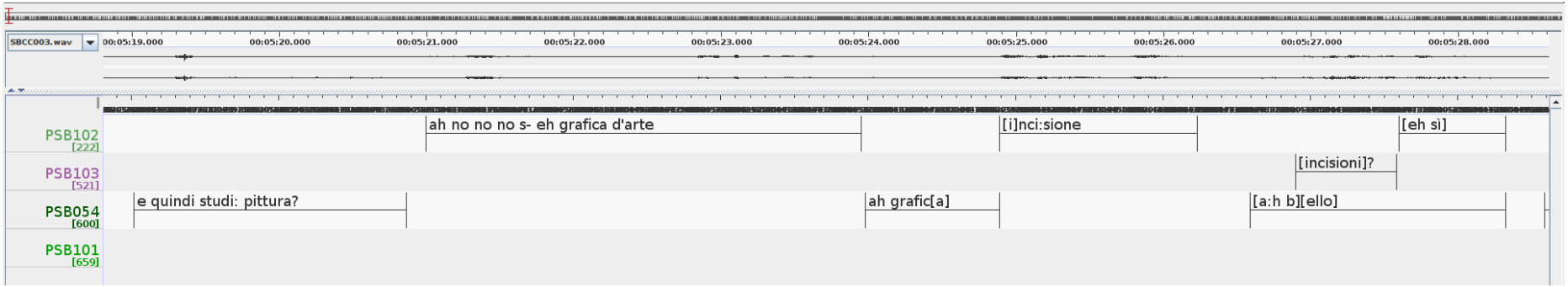
Jefferson notation

,	> weakly rising intonation
?	> rising intonation
.	> falling intonation
:	> prolonged sound
(.)	> short pause
=	> prosodic link
°blah°	> low volume
BLAH	> high volume
bl~	> interruption
>blah<	> fast pace
<blah>	> slow pace
[blah]	> overlapping speech
(blah)	> transcriber's hypothesis
xxx	> incomprehensible sequence
((laughs))	> non verbal behavior

<lo:re:m [xx do::lor> sit am(et), °conse]cte[tur adipiscing] elit° sit amet consectetur? ((sigh)) adipisc~ ELIT quisque=faucibus ex adipiscing elit >'sque fa:ucibus ex sapien vi:tae pellentesque<

ELAN

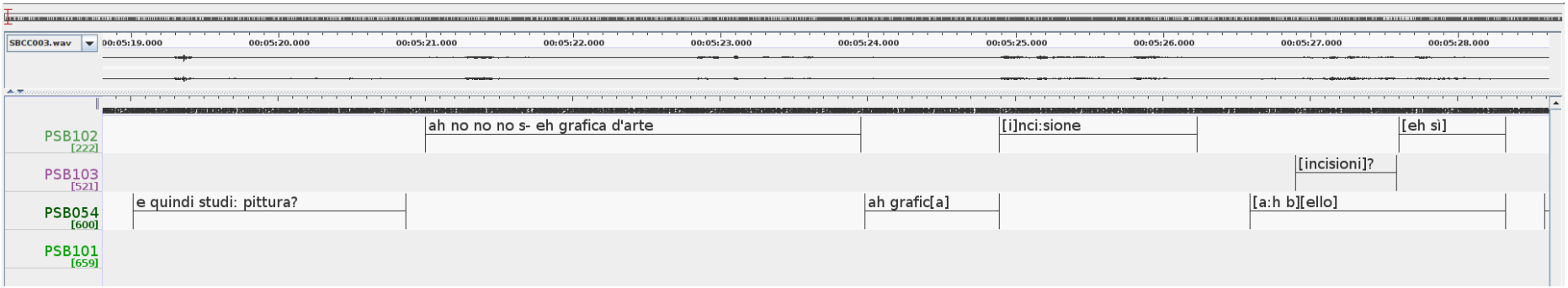
> **Transcription units** aligned with audio



ELAN

> Transcription units aligned with audio

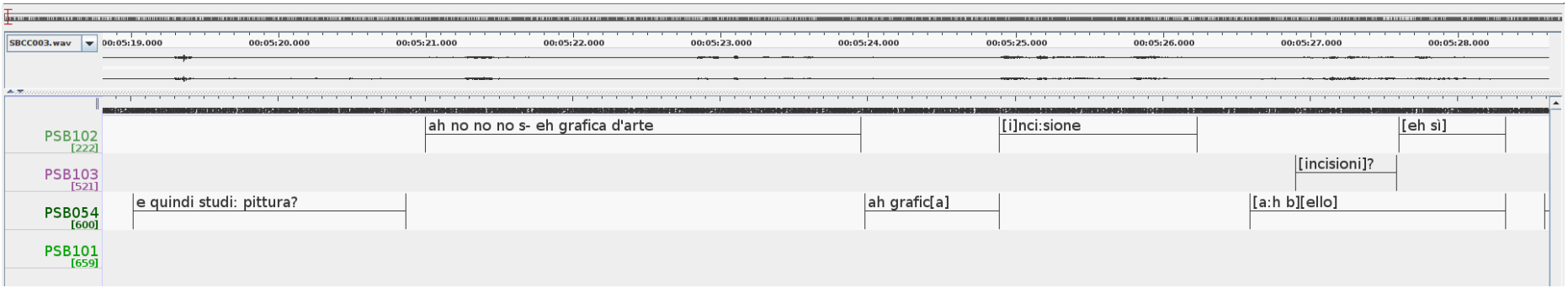
> **segment-level** annotations



<lo:re:m **[xx** do::lor**>** sit am**(**et**),** **°**conse**]**cte**[**tur adipiscing**]** elit**°** sit amet consectetur**?** **((sigh))** adipisc**~** **ELIT** quisque**=**faucibus ex adipiscing elit **>'**sque fa**:**ucibus ex sapien vi:tae pellentesque**<**

ELAN

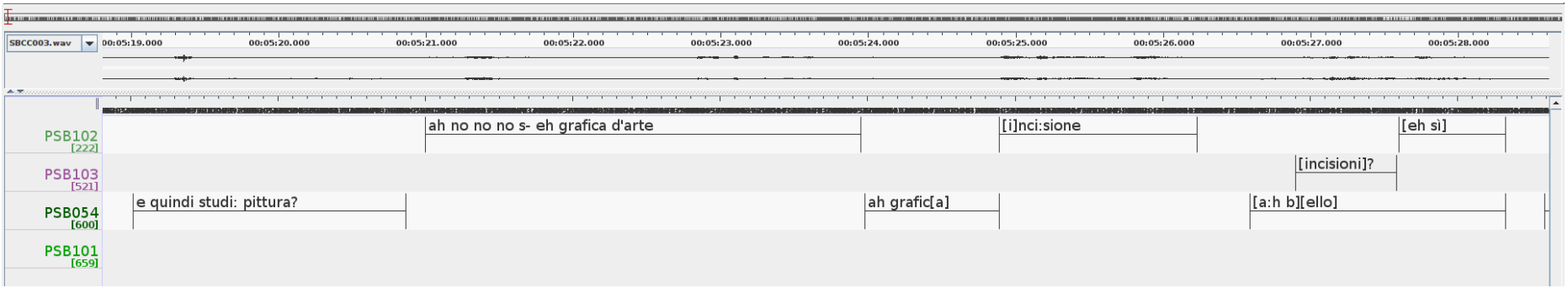
> Transcription units aligned with audio
> **segment-level** annotations



<lo:re:m **[xx** do::lor**>** sit am**(**et**),** **°**conse**]**cte**[**tur adipiscing**]** elit**°** sit amet consectetur**?** **((sigh))** adipisc**~** **ELIT** quisque**=**faucibus ex adipiscing elit **>'**sque fa**:**ucibus ex sapien vi:tae pellentesque**<**

ELAN

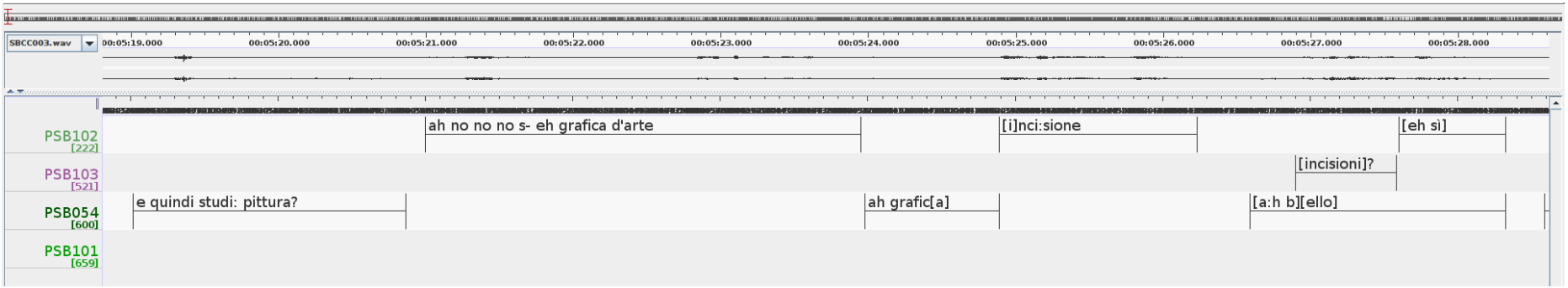
> Transcription units aligned with audio
> segment-level annotations
> **token-level** annotations



<lo:re:m **[xx** do**::**lor**>** sit am**(**et**),** °conse**]**cte**[**tur adipiscing**]** elit° sit amet consectetur**?** **((sigh))** adipisc**~** **ELIT** quisque**=**faucibus ex adipiscing elit **>'**sque fa**:**ucibus ex sapien vi:tae pellentesque**<**

ELAN

> Transcription units aligned with audio
> segment-level annotations
> token-level annotations



full jefferson script

orthographic script

noSketchEngine `vert` format

Accessibility: how does data look like today



www.kiparla.it

Why the current pipeline is sub-optimal



> the process is extremely **time consuming** and yet no queries other than form-based queries can be performed

> transcriptions are produced by hand and **not validated** against an expected format

> data preparation is **opaque** to the user *(if you were a researcher, what format would you want for this kind of data?)*

> derived data is currently **not versioned**

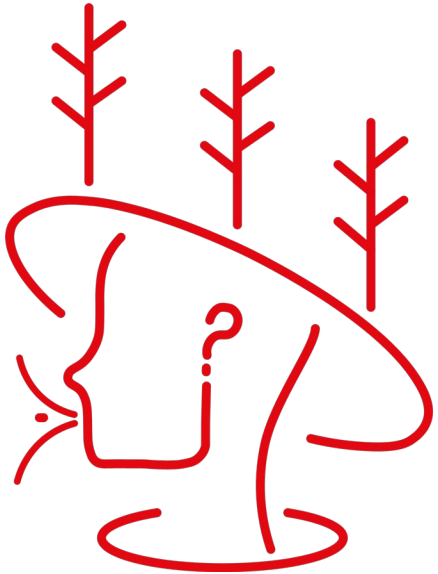
> not every annotated feature is *stable* in the same way, and adding **further levels of annotation** requires new formats

Deriving a treebank from KIParla

> Syntax of spoken language is generally under-represented in resources with respect to written language syntax

> Syntax cues play a key role in management of turn-taking by human interlocutors

> Participants use a system of dialogic affordances to converge on intentions and meaning representation



KIParla Forest

"Towards a dialogic syntax" - Du Bois 2014
"Dynamic Syntax: The Flow of Language Understanding" - Kempson et al. 2000

Co-construction and split utterances

D: Yeah I mean if you're looking at quantitative things it's really you know how much actual- How much variation happens whereas qualitative is 〈 pause 〉 you know what the actual variations

U: entails

D: entails. you know what the actual quality of the variations are.

[BNC G4V 114-117]

M: I might be a bit biased, I think they still do that but I think erm 〈 pause 〉

J: The television has 〈 pause 〉

M: the television has made a difference. I think not only just at fire stations, I think in the whole of life, hasn't it?

[BNC K69 51-54]

Examples from *"Split Utterances in Dialogue: a Corpus Study"* - Purver et al. 2009

INTERMEZZO



Universal Dependencies

Materials for the next slides: Marie-Catherine de Marneffe, Joakim Nivre, and Daniel Zeman

Not any treebank, a UD treebank

*"Universal Dependencies (UD) is a project developing **cross-linguistically consistent treebank annotation** for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a **language typology** perspective."*

Is UD good at representing variability **within** each language, besides cross-language comparability?

UD is A project — but also an **annotation scheme**, a data repository and a **community**



The UD Philosophy

1. Maximize parallelism – but don’t overdo it
    a. Don’t annotate the same thing in different ways
    b. Don’t make different things look the same
    c. Don’t annotate things that are not there

2. Universal **taxonomy** with language-specific elaboration
    a. Languages select from a universal pool of categories
    b. Allow language-specific extensions



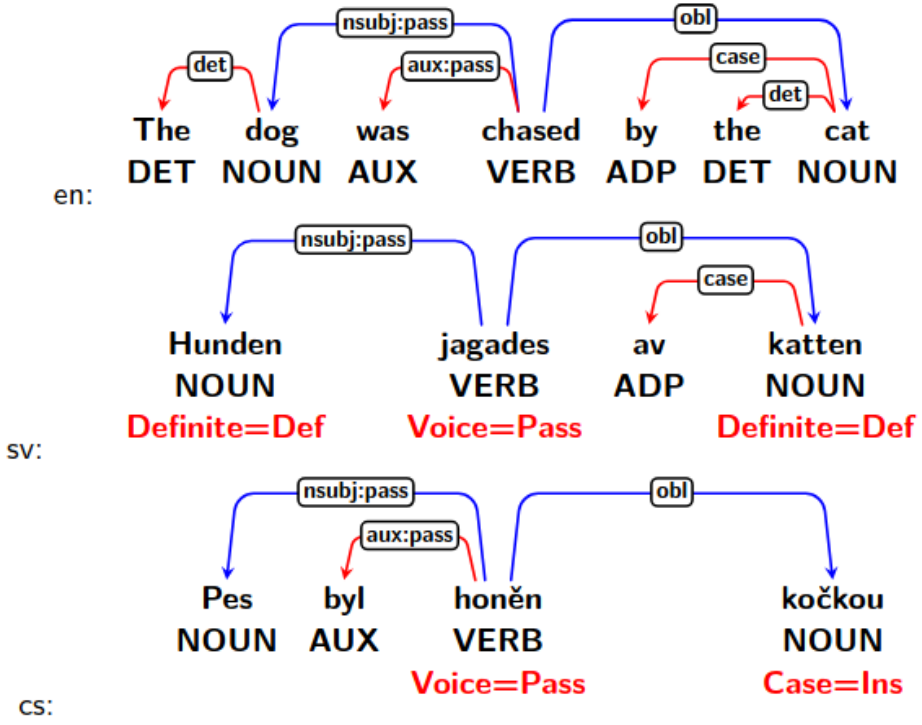
Syntactic Annotation in UD

1. Basic principles:
    a. The primacy of content words
    b. Clauses, nominals and modifier words
    c. Core arguments vs. oblique dependents

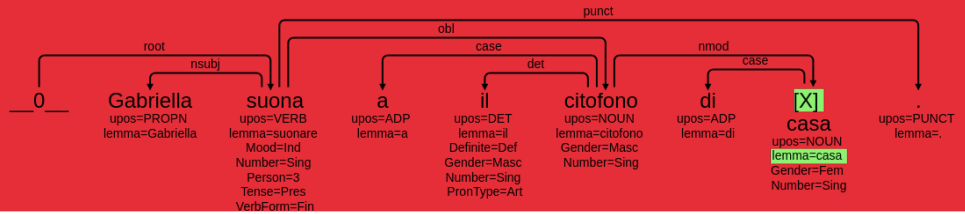
2. **Universal** and **language-specific** relations



The primacy of content words



CoNLL-U format



sent_id = isst_tanl-1371

text = Gabriella suona al citofono di casa.

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	ENHANCED	MISC
1	Gabriella	Gabriella	PROPN	SP	_	2	nsubj	_	_
2	suona	suonare	VERB	V	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	0	root	_	_
3-4	al	_	_	_	_	_	_	_	_
3	a	a	ADP	E	_	5	case	_	_
4	il	il	DET	RD	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	5	det	_	_
5	citofono	citofono	NOUN	S	Gender=Masc|Number=Sing	2	obl	_	_
6	di	di	ADP	E	_	7	case	_	_
7	casa	casa	NOUN	S	Gender=Fem|Number=Sing	5	nmod	_	SpaceAfter=No
8	.	.	PUNCT	FS	_	2	punct	_	_

	Nominals	Clauses	Modifier words	Function Words
Core arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	advmod* discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	**Headless**	**Loose**	**Special**	**Other**
conj cc	fixed flat	list parataxis	compound orphan goeswith reparandum	punct root dep

https://universaldependencies.org/u/dep/index.html

> rows correspond to **functional categories** in relation to the head

> columns correspond to **structural categories** of the dependent

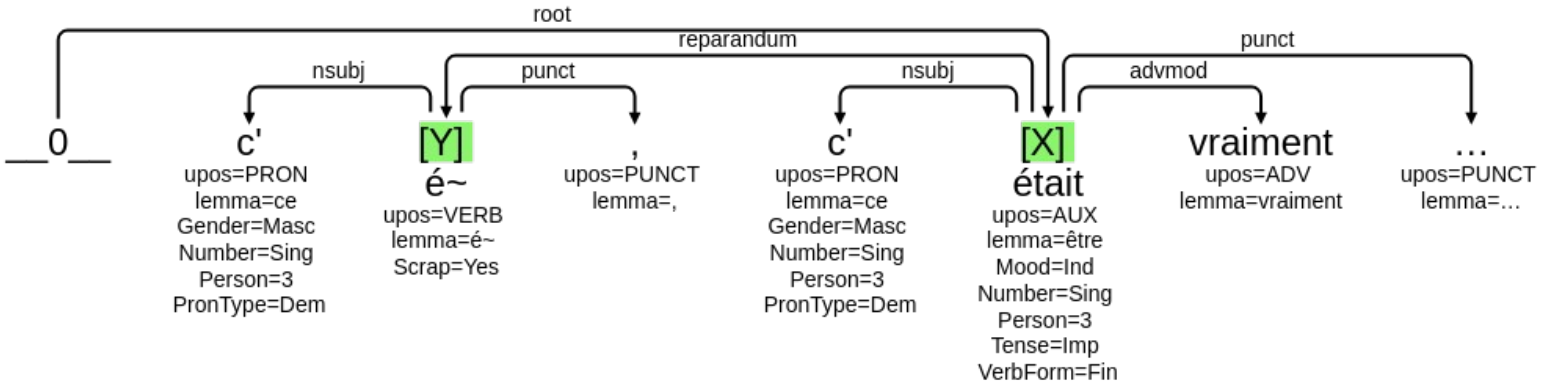


CoNLL-U format

> *discourse*: interjections and other discourse particles and elements not clearly linked to the structure of the sentence

> *reparandum*: disfluencies overridden in a speech repair

> *parataxis*: discourse relations in linking clauses, tag questions and more



SpLAn-UD - Spoken Language Annotation for Universal Dependencies

- **Event title**: SpLAn-UD - Spoken Language Annotation for Universal Dependencies, a UniDive workshop on annotation of spoken data
- **Location**: LILEC, University of Bologna, Aula Seminari 2 - Piazza San Giovanni in Monte 2 - 40124 Bologna, Italy
- **Map**: Google Maps
- **Dates**: 29-30 May 2025
- **Chairs**: Kaja Dobrovoljc, Sylvain Kahane, Caterina Mauri and Ludovica Pannitto
- **Local Organizer**: Caterina Mauri, Ludovica Pannitto and Eleonora Zucchini

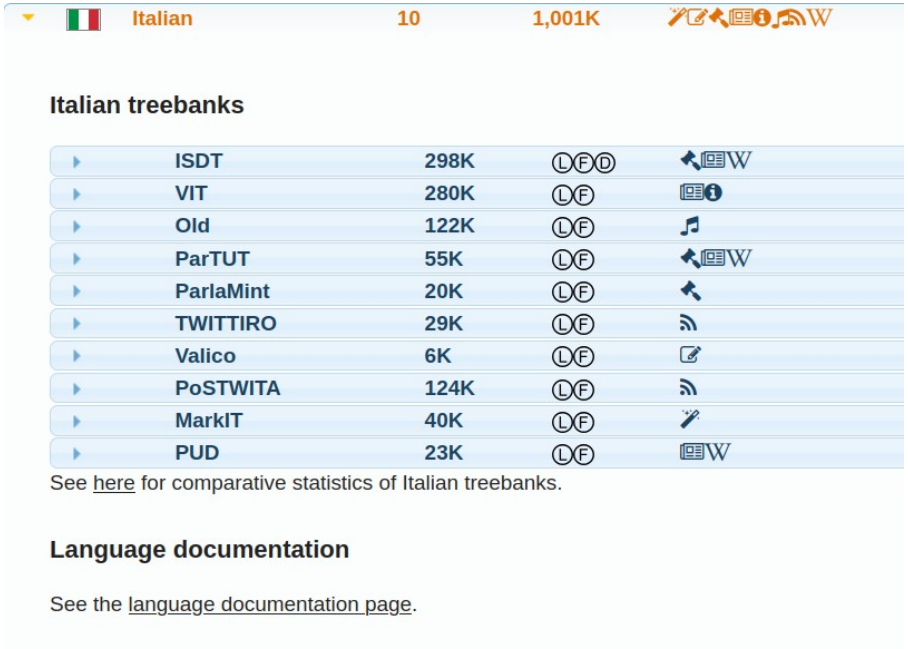
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the European Union

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

Join us in Bologna! (or virtually)

Italian resources in UD



Italian	10	1,001K	

Italian treebanks

ISDT	298K	LFD	
VIT	280K	LF	
Old	122K	LF	
ParTUT	55K	LF	
ParlaMint	20K	LF	
TWITTIRO	29K	LF	
Valico	6K	LF	
PoSTWITA	124K	LF	
MarkIT	40K	LF	
PUD	23K	LF	

See here for comparative statistics of Italian treebanks.

Language documentation

See the language documentation page.

> Italian has a very solid tradition in the UD enterprise, with resources already appearing in the first release dating back to 2015

> No treebank with **naturally occurring spoken language** available (yet)

BACK TO KIPARLA



Data selection from KIParla

CODE	TOD1005bis	BOD2018	PBB004	BOA3017
TYPE	lecture	interview	interview	free conversation
INTERACTION LEVEL	almost none	partially free	partially free	free
N. TOKENS	6788	4634	5898	4551
DURATION	00:50:44	00:28:08	00:35:54	00:30:22
PARTICIPANTS	1	2	3	4
KIPOS	yes	yes	no	yes

Table 1: Conversations selected for the first release of KIParla Forest.

> We selected four conversations based on their interactional level:
- *free interaction*
- *partially free interaction*
- *rigid interaction*
- *almost no interaction*

> total of 21833 *orthographic* tokens (not all of them also qualify as *syntactic* tokens)

Other spoken language treebanks

Name	Code	Release	Source	Tokens	Sents	Avg. length
Beja NSC	bej_nsc	v2.8	converted	1,101	56	19.7
Cantonese HK	yue_hk	v2.1	native	13,918	1,004	13.9
Chinese HK	zh_hk	v2.1	native	9,874	1,004	9.8
Chukchi HSE	ckt_hse	v2.7	native	5,389	1,004	5.4
French ParisStories	fr_parisstories	v2.9	converted	29,438	1,755	16.8
French Rhapsodie	fr_rhapsodie	v2.2	converted	34,437	2,837	16.8
Frisian-Dutch Fame	qfn_fame	v2.8	native	3,729	400	9.3
Komi-Zyrian IKDP	kpv_ikdp	v2.2	native	2,304	214	10.8
Naija NSC	pcm_nsc	v2.2	converted	140,729	9,242	15.2
Norwegian NynorskLIA	no_nynorsklia	v2.1	converted	55,410	5,250	10.6
Slovenian SST	sl_sst	v1.3	native	29,488	3,188	9.2
Turkish-German SAGT	qtd_sagt	v2.7	native	36,934	2,184	16.9

Table 1: Alphabetical list of spoken language treebanks in UD v2.9.

Table from *"Spoken Language Treebanks in Universal Dependencies: an Overview"* - Dobrovoljc 2022

Where to start from?

BO118 [io ho preso] il patentino e non ho mai preso il motorino perché avevo pau[ra]
BO140 [ah] ecco vedi? ((ride))
BO140 no io lo prendevo anche quando nevicava
BO118 ma:do' ((ride))
BO118 che po:i ((*nomepaese)) è fatta tut~ cioè tutta in sali:ta[:: son] scomode le strade
BO140 [sì eh]
BO140 sì eh cioè diciamo che il centro è tutto:: salita e discesa.
BO140 però: fuori:: tutta la periferia così:: cioè sono delle pendenze norma:li diciamo.
BO118 okay

Extract of BOD2018

> We could remove everything that is `syntactically useless` and start from there

> **ISSUE n.1**: not all information can go into UD, but we want cross-querying to be possible

> **ISSUE n.2**: the treebank has to inherit changes to the original transcriptions (also relevant for GDPR) and the corpus could inherit from the treebank (relevant for maintenance practices)

> **ISSUE n.3**: CoNLL-U and our formats (*.eaf*) are not natively compatible

A new format for KIParla

Example 1. CLAN representation of data (first three lines).

```
*MOT: look at the tree ! 2263675_2265197
%mor: v|look prep|at det|the n|tree !
%gra: 1|0|ROOT 2|1|JCT 3|4|DET 4|2|POBJ 5|1|PUNCT
```

Example 2. Representation in TEI corresponding toexample 1.

```
<timeline unit="s">
 <when absolute="0" xml:id="T0"/>
 <when interval="2263.675" since="#T0" xml:id="T1086"/>
 <when interval="2265.197" since="#T0" xml:id="T1087"/>
</timeline>
<annotationBlock end="#T1087" start="#T1086" who="MOT" xml:id="au551">
 <u>
  <seg>look at the tree !  </seg>
 </u>
 <spanGrp type="mor">
  <span>v|look prep|at det|the n|tree !  </span>
 </spanGrp>
 <spanGrp type="gra">
  <span>1|0|ROOT 2|1|JCT 3|4|DET 4|2|POBJ 5|1|PUNCT  </span>
 </spanGrp>
</annotationBlock>
```

> Many formats exist for spoken language representation, also based on TEI-xml standard

> no one is natively convertible to CoNLL-U (and back)

> Jefferson notation allows non-nested structures (e.g. (hi h[o)w are you])

> we still want data to be both human readable and machine readable

Example of data in XML-TEI from *"TEICORPO: A Conversion Tool for Spoken Language Transcription with a Pivot File in TEI"* - Parisse et al. 2020

From Jefferson to CoNLL-compatible KIParla

1. Each transcription unit is **normalized** and **validated**
    a. common errors and typos are corrected automatically (e.g., double spaces are reduced to one, number are spelled out...)
    b. sequence level annotations are checked for consistency (e.g., parentheses balance)

From Jefferson to CoNLL-compatible KIParla

2. Transcriptions units are **tokenized**

a. tokens are identified based on orthographic units
b. each token is assigned an id, a type (linguistic, unknown, non-linguistic-behavior, shortpause) and its boundaries in the transcription unit
c. linguistic tokens are checked against a regex `(['~-]?(\p{L}+:*)*\p{L}+:*[-'~]?[.,?]?)`
d. jefferson annotations are turned into token-level features

From Jefferson to CoNLL-compatible KIParla

3. **Overlaps** among transcription units are computed and checked

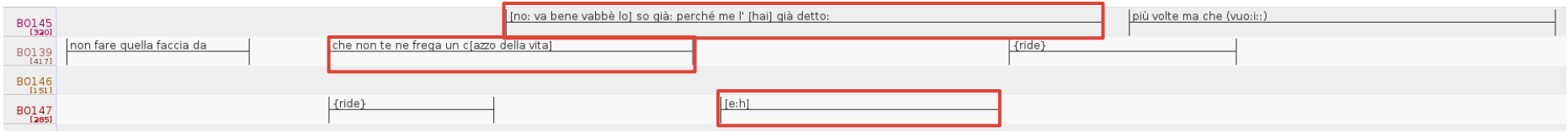
a. Non-linguistic-behaviours are not considered as overlapping items

b. Overlapping events can involve more than two speakers

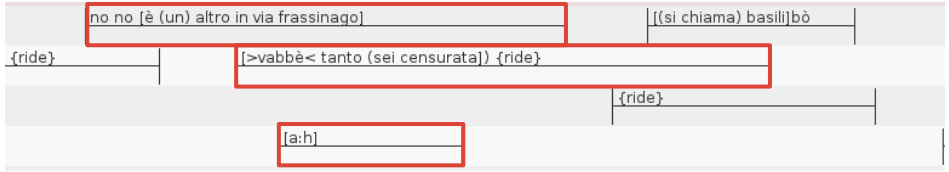
c. A transcription unit can participate in more than one overlapping event

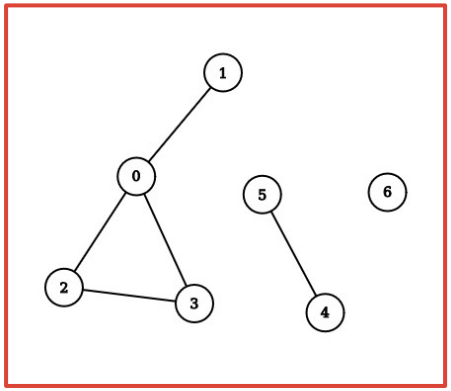
Overlapping speech

Case 1: unit participating in more than one overlapping event



Case 2: overlap among three speakers





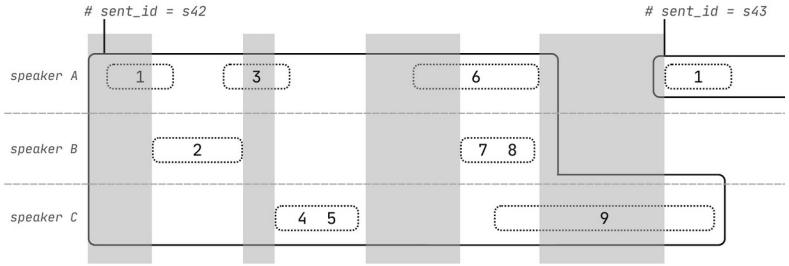
(near-)Algorithm

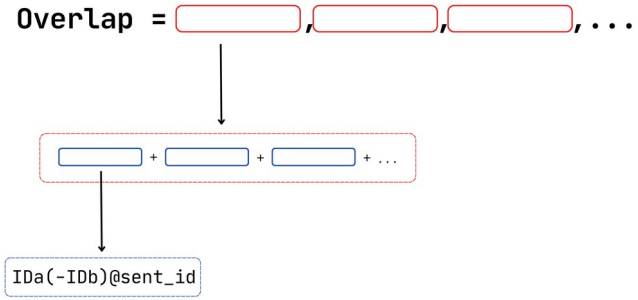
> Each unit is **linked** in a graph data structure to all units to which it overlaps

> Each transcription units also has a number of **character-spans** annotated as overlapping

> We identify **cliques** (i.e., strongly connected components) in the graph as overlapping events

> Overlapping events and overlapping spans are **matched**





CoNLL-compatible KIParla: the fields

> Speaker ID

> Token ID (transcription unit ID + progressive ID)

> type (linguistic, metalinguistic, unknown, shortpause)

> original span

> CoNLL-U `sent_id`

> CoNLL-U fields

> alignment features

> jefferson-derived features

	token_id	speaker	tu_id	unit	id	span	form	...	type	jefferson_feats	align	prolong	pace	guesses	overlaps
1	38-0	BO147	38	_	_	(xx	x	_	unknown		Begin=49.327		_	0-2(0)	_
2	38-1	BO147	38	_	_	ma	ma	_	linguistic		_		_	0-2(0)	_
3	38-2	BO147	38	_	_	io	io	_	linguistic		_		_	0-2(0)	_
4	38-3	BO147	38	_	_	da	da	_	linguistic		_		_	0-2(0)	_
5	38-4	BO147	38	_	_	pallotti	pallotti	_	linguistic		_		_	0-8(0)	_
6	38-5	BO147	38	_	_	ci	ci	_	linguistic		_		_	0-2(0)	_
7	38-6	BO147	38	_	_	piglio	piglio	_	linguistic		_		_	0-6(0)	_
8	38-7	BO147	38	_	_	le	le	_	linguistic		_		_	0-2(0)	_
9	38-8	BO147	38	_	_	paste	paste	_	linguistic		_		_	0-5(0)	_
10	38-9	BO147	38	_	_	>cioè:<	cioè	_	linguistic		_	3x1	Fast=0-4(0)	0-4(0)	_
11	38-10	BO147	38	_	_	ci	ci	_	linguistic		_		_	0-2(0)	_
12	38-11	BO147	38	_	_	prend[o:	prendo	_	linguistic	ProsodicLink=Yes	_	5x1	_	0-6(0)	5-6(39)
13	38-12	BO147	38	_	_	mh	mh	_	linguistic		_		_	0-2(0)	0-2(39)
14	38-13	BO147	38	_	_	le	le	_	linguistic		_		_	0-2(0)	0-2(39)
15	38-14	BO147	38	_	_	brio~])	brio~	_	linguistic	Interrupted=Yes	End=52.997		_	0-5(0)	0-5(39)
16	39-0	BO146	39	_	_	[le	le	_	linguistic		Begin=51.927		_	_	0-2(38)
17	39-1	BO146	39	_	_	lasagne]	lasagne	_	linguistic		End=52.887		_	_	0-7(38)

And eventually, the treebank!

> Syntactic relations can **only** be observed **within the sentence boundaries**: we don't want segmentation to obscure relations and we want to **natively represent syntactic co-construction among speakers**

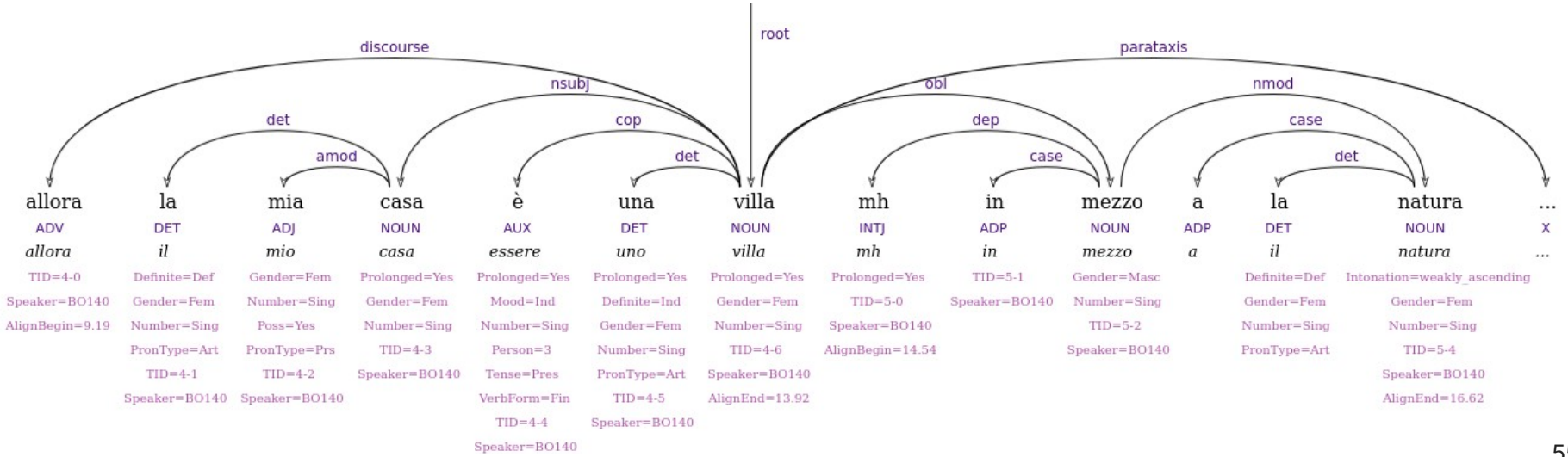
> Transcription unit boundaries in the KIParla cannot be treated as reliable basis for segmentation: there are many examples where core grammatical relations hold between elements of different transcription units

4, BO140 allora la mia ca:sa:: è::: una: villa::
so DEF my house is INDEF villa

‘so my house is a villa’

5, BO140 mh: in mezzo alla natura,
mh in middle of.DEF nature

‘mh, immersed in nature’



And eventually, the treebank!

> Other treebanks use the relation `AttachTo` in the MISC field for cases of co-construction, but this is not currently supported by tools

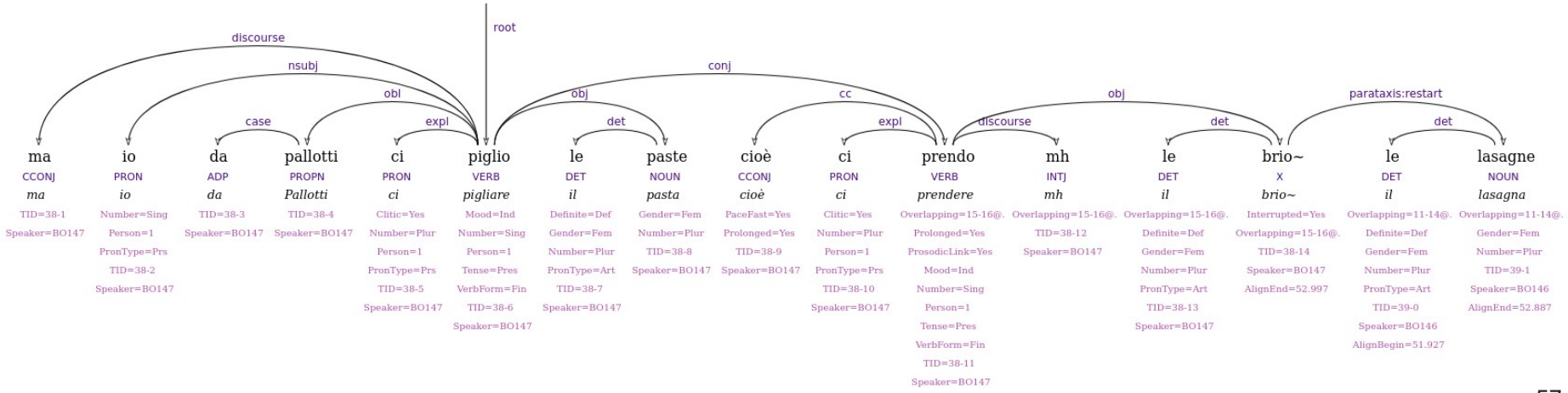
> we begin with annotating dependencies between words, establishing a unit boundary when no dependency link can be found. Such anapproach allows boundaries to emerge **bottom-up**

38, BO147 (xx ma io da pallotti ci piglio le paste >cioè:< ci prend[o: =mh le
UNK but I by Pallotti LOC take:PRS.1SG DEF pastries that.is LOC take:PRS.1SG mh DEF
brio~])
crois(sants)

'but from Pallotti I buy pastries I mean I buy croissants'

39, BO146 [le lasagne]
DEF lasagna

'lasagna'



Criteria for segmentation

> rely on prosody: syntactic annotation has to be performed while listening to the recording

> **ISSUE n.1**: connectives that often develop discourse functions:
- identifiability of a clear head to which the connective should be attached. If the connective is linked to a larger portion of conversation and/or it is not possibile to identify a clear preceding head, then we set a maximal unit boundary before the connective itself.

73, BOI013 se perdo un autobus poi
if lose:1SG INDEF bus after
devo s~ (a)spettare un'altra ora
must:1SG w~ wait another hour
'if I miss a bus then I have to wait fot another hour'

74, BOR005 >ah okay **quindi**< la mobilità è
ah okay **so** DEF mobility is
molto ridotta
very reduced
'ah okay **so** mobility is very reduced'

Criteria for segmentation

> **ISSUE n.2**: the grammatical relationship is not overtly marked

- it may be difficult to decide whether we are dealing with listing or implicit reformulation occurring within the same unit, or with independent segments.

925, BO139 la prima c'e~ {ride}
DEF first there.wa~ laughs
'the first one there wa~ {laughter}'

926, BO139 una ragazza che: raccontava di ex coinquiline cristiane che tipo
INDEF girl that talked of ex roommates christian who like
'a girl that was talking about ex christian roommates that like'

927, BO139 le avevano rubato un pr~ un =eh non mh un dildo che lei aveva nel =eh
to.her had:3PL stolen INDEF pr~ INDEF eh not mh INDEF dildo that 3SG.F had in.DEF eh
{ride}
laughs
'they stole her a co~ mh a dildo that she had in {laughter}'

928, BO147 {ride}
laughs
'{laughter}'

929, BO139 nel comodino
in.INDEF bedside.table
'in her bedside table'

930, BO139 perché col sospetto che lei lei stava studiando delle cose di magia nel
because with.DEF suspect that 3SG.F 3SG.F was studying INDED.PL things of magic in.DEF
sen[so >cioè< {P} di antropologia]
sense I.mean PAUSE of anthropology
'because suspecting she was studying something about magic, I mean, anthropology'

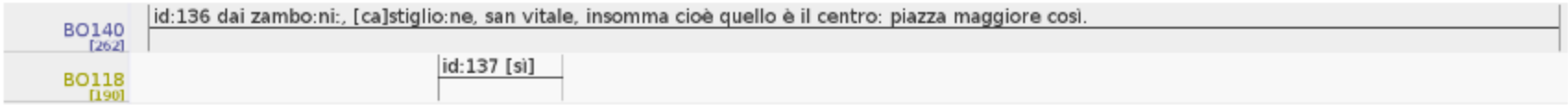
931, BO145 [a:h me l' hai raccontato]
ah to.me it have.2SG told
'oh you told me'

932, BO139 eh
eh
'eh'

933, BO139 uno studio sui riti magici nella sicilia tipo dell' ottoce~ metti una roba
INDEF study on.tDEF rituals magic in.DEF sicily like of.DEF eight.hundr~ take:2SG a thing
del genere
of.DEF genre
'a study on magic rites in the Sicily of ninet~ century, for example, something like that'

Criteria for segmentation

> **ISSUE n.3**: feedback phenomena
- we consider as internal to the maximal units all feedback phenomena that do not interfere with the main syntactic flow , regardless of the speaker who is uttering them;
- in such cases, we link the feedback expression to the unit head through the `dep:feedback` label;
- in case of feedback phenomena that interrupt the syntactic flow, which may or may not cause replanning, we have to proceed being aware that we are dealing with a continuum: we set a maximal unit boundary if the portion following feedback has no clear syntactic relation to the portion preceding it.



136, BO140 dai zambo:ni:, [ca]stiglio:ne, san vitale, insomma cioè quello è il centro: piazza
come.on Zamboni Castiglione san Vitale to.sum.up I.mean that is the center square
maggiore così.
major so

'come on, Zamboni, Castiglione, San Vitale I mean. That is DEF city centre, Piazza Maggiore, like that'

137, BO118 sì
yes

'yes'

Tagging morphology

EVALITA
Evaluation of NLP and Speech Tools for Italian

> 2020: shared task in the Italian Evaluation campaign for PoS tagging the KIParla corpus, but guidelines were not fully compliant with UD v.2

> PoS are annotated based on **morphological features**, sticking to the main category of each word, even though they perform a different function in the syntactic context:

- the word *basta* (en. 'stop') is an inflected form of the verb *bastare* (en. 'to stop') but works as discourse marker meaning 'that's it'.
- we took a restricted approach at determiners, using the DET label only for articles, demonstratives and quantifiers, while considering any other elements of the noun phrase, both preceding and following the head, as modifiers of the noun. This allows for a consistent annotation of diatopical variation

UPOS for what words are, DEPREL for what they do

Joint discussion with Arianna Masciolini

> What do we want from the treebank? To **maximize the amount of information** that we can *encode* in (and therefore *query* from) the resource

> Generally (but not explicitly), UD allows the idea of functionally defined universal **constructions** (DEPRELS) and morphosyntactic **strategies** (UPOS), but actually the guidelines (hence, the validator) prevent some of this to be annotated

> There are cases, especially in `non-standard` language, where the two layers clearly diverge

Example (1): validator is happy

quindi è una roba che dovrebbe essere anche piuttosto personale no

il contatto con le persone e com'è nata l'idea della ricerca eccetera però ho **tipo** un libro di metodologia che è un manuale

allora mi sono presa male che magari avrei dovuto scrivere qualcosa **tipo** sai come come ha fatto tal dei tali **oppure** in accordo col metodo di sticazzi

so it's something that should even be quite personal right

getting in touch with people and how the research idea was born but I **like** had a methodology book which is a manual

then I felt upset* that I should have written something **like** you know how how did that guy go **or** according to who-fucking-ever's methodology

Example (2-3): validator is angry

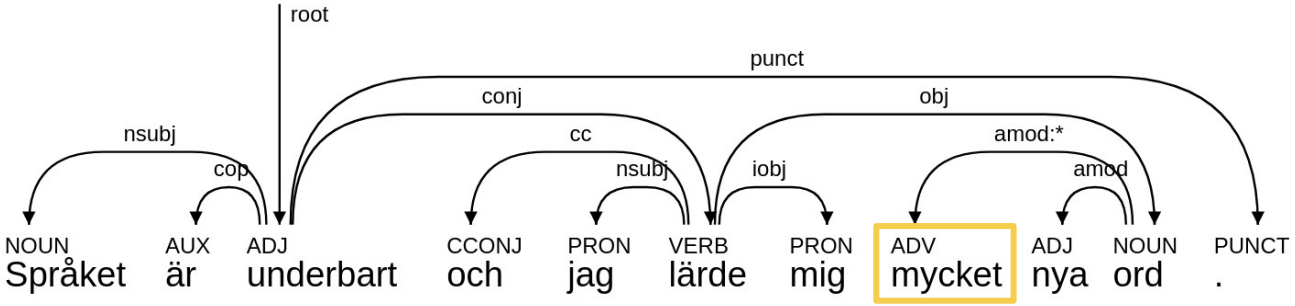
e: eravamo più contente di dormire:: tra donne che avere un uomo in casa anche perché **magari** cioè nel senso si possono creare anche diversi imbarazzi

and we were happier to sleep among women than having a man in the house because **maybe** you know it could be very awkward

magari

ma|gà|ri

inter., cong., avv.



She language is wonderful and I learned **very* (many)** new words.

Miscellanea

> some orthographic tokens do not make it into the treebank, for instance non-linguistic behaviours that do not have any syntactic relevance

> some non-linguistic-haviours (e.g., shortpauses or ways of speaking) end up in MISC

> multiword tokens generate new tokens: these inherit token id from the multiword token and jefferson-derived features remain attached to the multiword token

> code switching

> interrupted tokens

To sum up

> We want to treebank `non-standard` language because there is where many interesting syntactic phenomena occur

> We want to use UD to analyze internal language variation besides cross-language variation

> Our resources need to be interoperable with each other but also as informative as possible

> We need to be able to go back and forth between levels of representations

Thank you! Questions?