Introducing KIParla Forest: *seeds for a UD annotation of interactional syntax*

Ludovica Pannitto[1], Eleonora Zucchini[1], Silvia Ballarè[1], Cristina Bosco[2], Caterina Mauri[1], Manuela Sanguinetti[3]

[1]Alma Mater Studiorum - University of Bologna,[2]University of Turin,[3]University of Cagliari

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

SYNTAX FEST Ljubljana August 26-29 2025

EXPERIMENTAL LAB lilec.lab@unibo.it

UniDive

www.kiparla.it

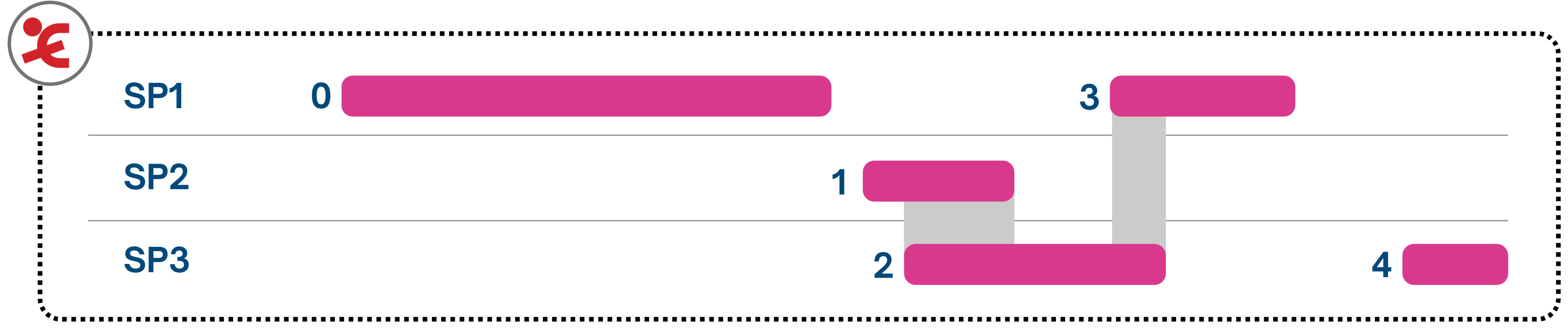
From KIParla to *KIParla Forest*

The KIParla resource [1,2,3] is a corpus of spoken italian manually transcribed following Jeffersonian notation for Conversation Analysis (CA) [4]. All summed up, the KIParla counts ca. **228 hours** of recordings and approximately **2M transcribed tokens**.

```
0 bene (.) allora e::h (.) >non siete
tutti e due di< bologna quindi
1 [solo] io
2 [n~] (.) no s[olo lui]
3 [di bolo]gna,
4 napoli
```

```
0 well (.) then e::h (.) >you're not
both from< bologna then
1 [just] me
2 [n~] (.) no j[ust him]
3 [from bolo]gna,
4 napoli
```

The corpus is currently segmented into **transcription units (TUs)**, based on pseudo-prosodic hints. TUs may **overlap.**



SP = Speaker. Representation of overlaps.

Speech-specific metadata

```
# sent_id
# text
# jefferson_text
# audio_url
```

At sentence level, metadata include the **sentence id** (aligned to conversation), original Jefferson transcription and url to audio (available for research purposes). Further metadata is retained in a separate *json* file available in the repository.

Each token retains (in MISC):

- **speaker ID**
- **boundary-related** features,
- **language variation** features (foreign languages + dialects)
- and **pseudonymization** information

```
AlignBegin=xxx(ms) and  AlignEnd=xxx(ms)
UnitBoundary=Yes
Lang=(NO_ISO_CODE|iso-code)
Anonymized=Yes
```

- Information derived from CA annotation:
  - **intonation** pattern
  - **prolonged sounds**
  - **volume** and **pace**
  - **short pauses** and **prosodic links**

```
Intonation=(Rising|WeaklyRising|Descending)
Prolongation=Yes
PauseAfter=Yes
Volume=(High|Low)
Pace=(Fast|Slow)
ProsodicLink=Yes
```

- information about **overlapping speech** (reference to overlapping tokens)

Morphosyntactic information

Tokenization, lemmatization

Pauses are removed and transformed into a feature in MISC.

text	form	upos	lemma	MISC
casa (.)	casa	NOUN	casa	PauseAfter= Yes

Multi-word tokens keep CA features.

Interrupted words are lemmatized as their complete version when context informative enough + specific feature

form	upos	lemma	MISC	*gloss*
c'	PRON	ci		there
era	VERB	essere		was
so~	ADV	solo	Interrupted=Yes	on~
c'	PRON	ci		there
era	VERB	essere		was
solo	ADV	solo		only
casa	NOUN	casa		house
mia	ADJ	mio		my

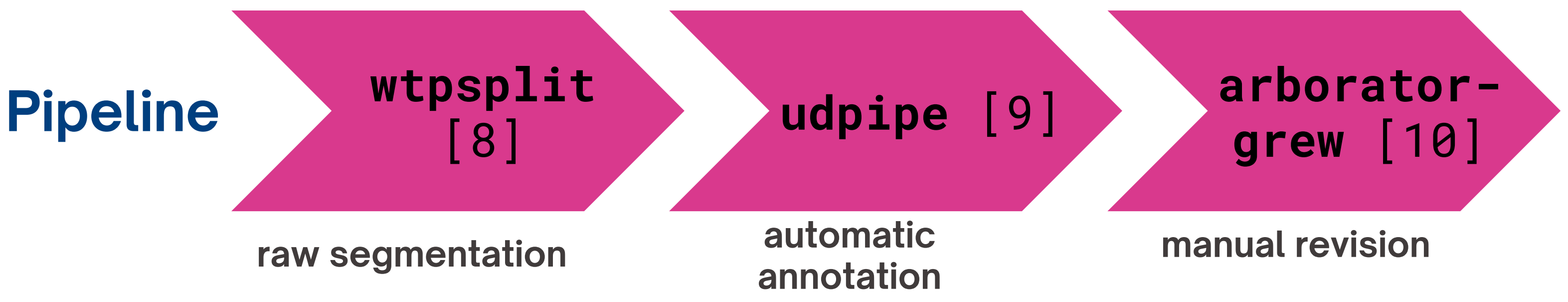
PoSTagging

Conservative approach: tag main category of each word

> ***basta*** *lit. 3sg of* ***bastare****, to suffice* tagged as VERB - en. 'stop'
>
> ***tipo*** *lit.* ***type*** tagged as NOUN - en. 'for example', 'like'

Inter-annotator agreement

- Cohen's κ > 0.87
- most disagreement on CCONJ and ADV



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

SpLAn-UD

Increased attention to the syntactic annotation of spoken varieties within the Universal Dependencies framework is attested by the fact that the number of treebanks including or completely dedicated to spoken language is on the rise.

Treebank curators took different directions in the creation of their resources, which could impact on derived measures or performance on downstream tasks [5,6]

Join!

Task 1.5

UniDive

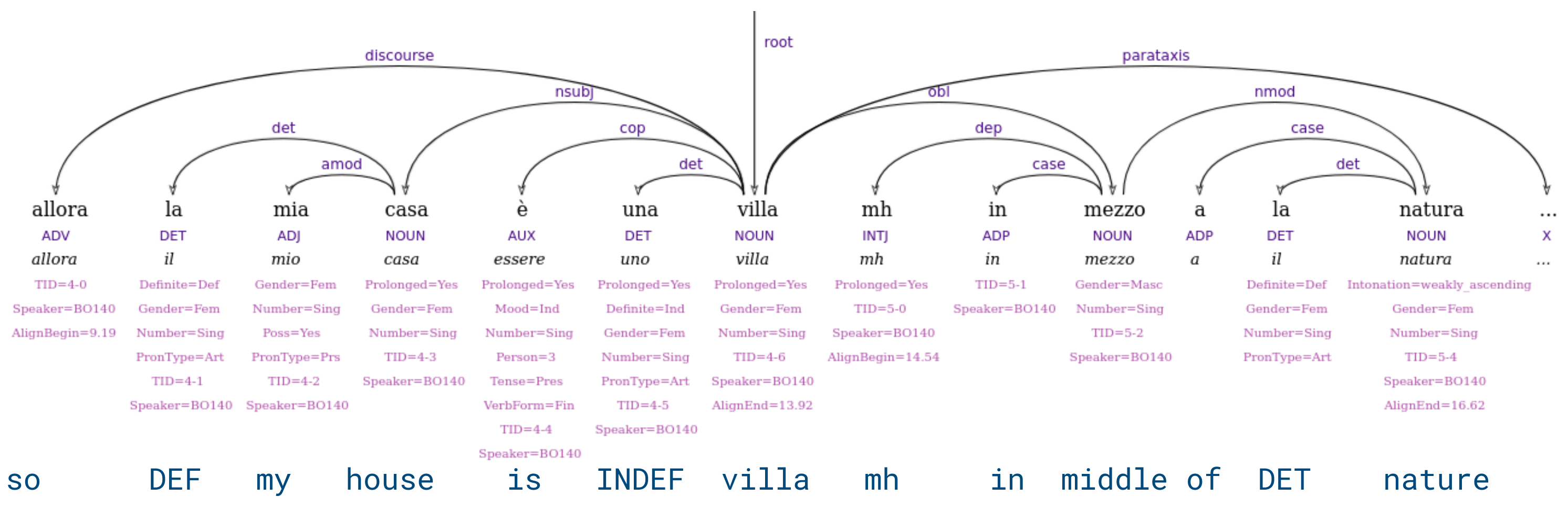
Data sample

First release (Nov 2025) aims at ~22K tokens selected based on **type of interaction** (free turn-taking, partially free turn-taking, rigid turn-taking, close to no interaction).

		conversation id	info
free	table conversations + free interactions	BOA3017	4551 tokens, 4 participants
partial	semi-structured interviews	PBB004	5898 tokens, 3 participants
		BOD2018	4634 tokens, 2 participants
close to none	lessons	TOD1005bis	6788 tokens, 1 participant

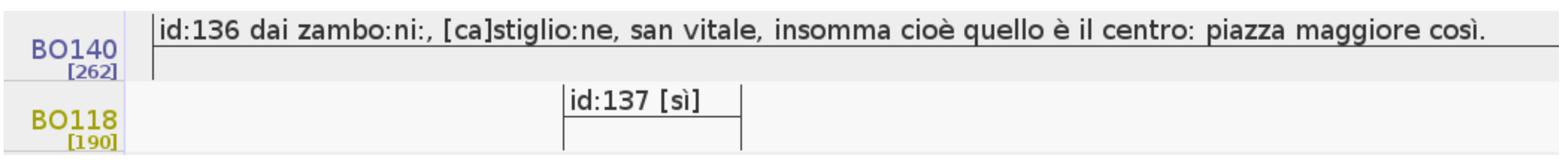
Segmentation and syntax

TU boundaries are not reliable as sentence boundaries: we identify maximal-units **based on syntactic dependencies**, in order to focus on **interactional syntax** and **cross-speaker syntactic affordances**


root
discourse
nsubj
det
amod
cop
det
obl
parataxis
dep
case
nmod
case
det
allora ADV allora TID=4-0 Speaker=BO140 AlignBegin=9.19
la DET il Definite=Def Gender=Fem Number=Sing PronType=Art TID=4-1 Speaker=BO140
mia ADJ mio Gender=Fem Number=Sing Poss=Yes PronType=Prs TID=4-2 Speaker=BO140
casa NOUN casa Prolonged=Yes Gender=Fem Number=Sing TID=4-3 Speaker=BO140
è AUX essere Prolonged=Yes Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin TID=4-4 Speaker=BO140
una DET uno Prolonged=Yes Definite=Ind Gender=Fem Number=Sing PronType=Art TID=4-5 Speaker=BO140
villa NOUN villa Prolonged=Yes Gender=Fem Number=Sing TID=4-6 Speaker=BO140 AlignEnd=13.92
mh INTJ mh Prolonged=Yes TID=5-0 Speaker=BO140 AlignBegin=14.54
in ADP in TID=5-1 Speaker=BO140
mezzo NOUN mezzo Gender=Masc Number=Sing TID=5-2 Speaker=BO140
a ADP a
la DET il Definite=Def Gender=Fem Number=Sing PronType=Art
natura NOUN natura Intonation=weakly_ascending Gender=Fem Number=Sing TID=5-4 Speaker=BO140 AlignEnd=16.62
... X ...
so DEF my house is INDEF villa mh in middle of DET nature

Special cases:

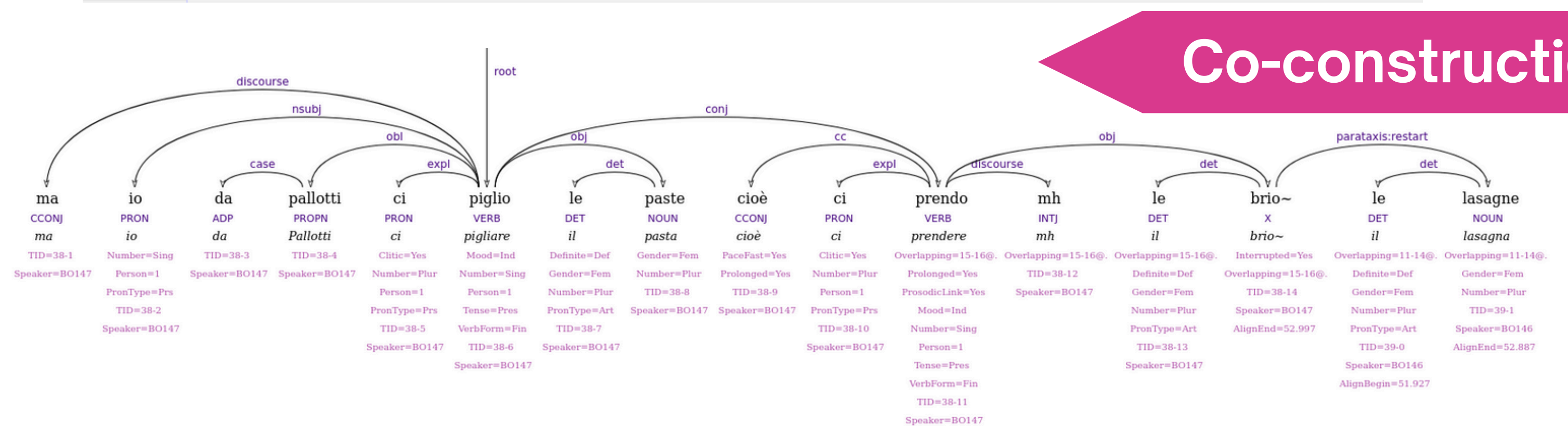
- **connectives with discourse functions:**
  - unit boundary is postulated if no relation with previous portion exists
- **feedback phenomena:**
  - no unit boundary is postulated if speech flow is uninterrupted



BO140 [262] id:136 dai zambo:ni:, [ca]stiglio:ne, san vitale, insomma cioè quello è il centro: piazza maggiore così.

BO118 [190] id:137 [sì]

Co-construction


root
discourse
nsubj
obl
case
expl
conj
obj
det
cc
expl
discourse
obj
det
parataxis:restart
det
ma CCONJ ma TID=38-1 Speaker=BO147
io PRON io Number=Sing Person=1 PronType=Prs TID=38-2 Speaker=BO147
da ADP da TID=38-3 Speaker=BO147
pallotti PROPN Pallotti TID=38-4 Speaker=BO147
ci PRON ci Clitic=Yes Number=Plur Person=1 PronType=Prs TID=38-5 Speaker=BO147
piglio VERB pigliare Mood=Ind Number=Sing Person=1 Tense=Pres VerbForm=Fin TID=38-6 Speaker=BO147
le DET il Definite=Def Gender=Fem Number=Plur PronType=Art TID=38-7 Speaker=BO147
paste NOUN pasta Gender=Fem Number=Plur TID=38-8 Speaker=BO147
cioè CCONJ cioè PaceFast=Yes Prolonged=Yes TID=38-9 Speaker=BO147
ci PRON ci Clitic=Yes Number=Plur Person=1 PronType=Prs TID=38-10 Speaker=BO147
prendo VERB prendere Overlapping=15-16@. Prolonged=Yes ProsodicLink=Yes Mood=Ind Number=Sing Person=1 Tense=Pres VerbForm=Fin TID=38-11 Speaker=BO147
mh INTJ mh Overlapping=15-16@. TID=38-12 Speaker=BO147
le DET il Overlapping=15-16@. Definite=Def Gender=Fem Number=Plur PronType=Art TID=38-13 Speaker=BO147
brio~ X brio~ Interrupted=Yes Overlapping=15-16@. TID=38-14 Speaker=BO147 AlignEnd=52.997
le DET il Overlapping=11-14@. Definite=Def Gender=Fem Number=Plur PronType=Art TID=39-0 Speaker=BO146 AlignBegin=51.927
lasagne NOUN lasagna Overlapping=11-14@. Gender=Fem Number=Plur TID=39-1 Speaker=BO146 AlignEnd=52.887

REFERENCES

[1] https://kiparla.it
[2] C. Mauri, et al., Kiparla corpus: a new resource for spoken italian, in *Proceedings of the 6th Italian Conference on Computational Linguistics CLiC-it* (2019).
[3] S. Ballarè, C. Mauri, et al.. La creazione del corpus KIParla: criteri metodologici e prospettive future. *Rivista italiana di dialettologia* (2020)
[4] G. Jefferson, et al., Glossary of transcript symbols with an introduction, *Conversation analysis* (2004)
[5] K. Dobrovoljc. Spoken language treebanks in universal dependencies: An overview. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (2022)
[6] S. Kahane, et al. Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories* (2021)
[7] C. Bosco et al. The Evalita 2014 Dependency Parsing task. in *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it* (2014)
[8] B. Minixhofer et al. Where's the point? self-supervised multilingual punctuation-agnostic sentence segmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (2023)
[9] M. Straka. 2024. Universal dependencies 2.15 models for UDPipe 2 (2024-11-21). *LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics* (2024).
[10] G. Guibon et al., When collaborative treebank curation meets graph grammars. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (2020)