The “KIPARLA Forest” treebank of spoken Italian

an overview of initial design choices

Ludovica Pannitto, Caterina Mauri

Alma Mater Studiorum - University of Bologna

UniDive

EXPERIMENTAL **LAB**

lilec.lab@unibo.it

Funded by the European Union

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

From KIParla to *KIParla Forest*

The KIParla resource [1,2,3] is a corpus of spoken italian manually transcribed following Jeffersonian notation [4]. All summed up, the KIParla counts ca. **228 hours** of recordings and approximately **2M transcribed tokens**.
Among existing Italian UD treebanks, none is explicitly addressing spoken varieties.

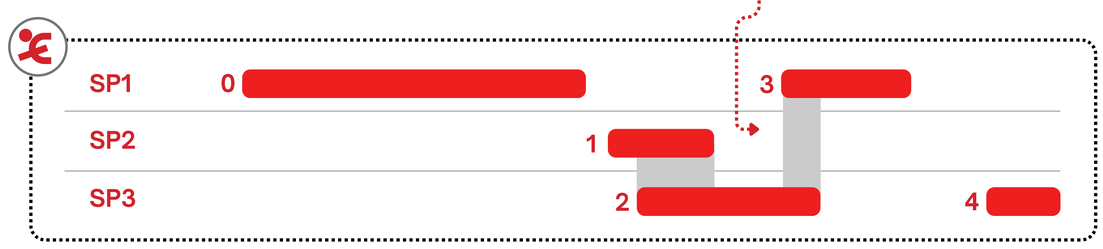
WG1

WG4

```
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
```

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

TUs may **overlap.**



SP = Speaker. Representation of overlaps.

An intermediate CoNLL format is built, formally equivalent to the jeffersonian notation. The *KIParla Forest* CoNLL-U is thus derived ensuring **full recoverability** of the original transcription.

Segmenting in maximal units

Currently the corpus is segmented into TUs. This has several issues:

- choices taken by transcribers are not stable across the corpus
- larger spans allow for broader syntactic investigations, including the analysis of strategies for **co-construction of discourse**

```
1. bottom-up identification of government relations, both within each
   unit and among different units
2. identify nuclei with autonomous illocutionary force [5]
3. merge TUs into maximal units based on the annotated dependency
   relations

AttachTo=sent_id@tok_id
Rel=deprel
```

Prosody

Each token retains information about its **intonation pattern**, the presence of a **prolonged sound**, its **volume**, its **pace**. Presence of **short pauses** and **prosodic links** are also annotated.

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

Data sample

First release aims at 110K tokens selected based on **type of interaction** (free turn-taking, partially free turn-taking, rigid turn-taking, close to no interaction).

		conversations	tokens
free	table conversations + free interactions	4	31K
partial	semi-structured interviews	9	40K
rigid	exams + office hours meetings	6	25K
close to none	lessons	4	30K

Other token-level features

Each token retains information about its **intonation pattern**, the presence of a **prolonged sound**, its **volume**, its **pace**. Presence of **short pauses** and **prosodic links** are also annotated.

```
AlignBegin=xxx(ms) and AlignEnd=xxx(ms)
UnitBoundary=Yes
OrigLang=(dialect|iso-code|uncertain)
Anonymized=Yes
```

Metadata

```
# sent_id
# text
# jefferson_text
# audio_id
```

```
(with reference to external json)
# speaker_id
# conversation_id
# notes
```

Non-lexical tokens

		UPOS	deprel
Fillers	ehm, beh, mh, ...	INTJ	discourse:filler
Cut-off words	que~,	X or same as repair	reparandum or parataxis:restart
Metalinguistic annotations	{ride}, {borbotta}	X	dep:extra

These are also translated into **modality of expression**, **events outside of the interaction** and **notes from the transcriber.**

Workplan

Data preparation
- Semi-automatic **error correction** on current transcription

Lemmatization and PoS tagging
- Automatic lemmatization and PoS tagging
- Manual revision for lemmatization and PoS tagging on sample + coder agreement
- Manual lemmatization and PoS tagging revision on full treebank

Segmentation and UD Parsing
- Automatic dependency parsing
- Manual revision of dependency annotation on sample + coder agreement
- Manual dependency revision on full treebank

Release
- Release for SyntaxFest

We would like to thank the **KIParla research group**, and in particular dr. Silvia Ballarè and dr. Eleonora Zucchini for the fruitful discussions during the design phases. Many thanks also to prof. Cristina Bosco and dr. Manuela Sanguinetti who are also actively involved in the *KIParla Forest* project.
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"

REFERENCES

[1] https://kiparla.it
[2] C. Mauri, S. Ballarè, E. Goria, M. Cerruti, F. Suriano, Kiparla corpus: a new resource for spoken italian, n Bernardi, R., R. Navigli & G. Semeraro (eds.), in: Proceedings of the 6th Italian Conference on Computational Linguistics CLiC-it (2019).
[3] Silvia Ballarè, Caterina Mauri, et al. 2020. La creazione del corpus kiparla: criteri metodologici e prospettive future. RID, RIVISTA ITALIANA DI DIALETTOLOGIA, 44:53–69.
[4] G. Jefferson, et al., Glossary of transcript symbols with an introduction, Conversation analysis (2004) 13–31
[5] Paola Pietrandrea, Sylvain Kahane, Anne Lacheret-Dujour, and Frédéric Sabio. 2014. The notion of sentence and other discourse units in corpus annotation. Spoken corpora and linguistic studies, pages 331–364.

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA