Towards the first UD Treebank of Spoken Italian: the *KIParla forest*

Ludovica Pannitto, Experimental Laboratory - University of Bologna - Italy

April 2024

The present project endeavors to enrich the linguistic resources available for Italian by constructing a treebank for the KIParla corpus [Mauri et al., 2019, Ballarè et al., 2020], an existing and well known resource for spoken Italian. At the moment of writing, in fact, none of the available italian UD treebanks is explicitly addressing spoken varieties. In particular, the Venice Italian Treebanks, VIT [Delmonte et al., 2007], contains approximately 60k tokens of spoken dialogues from different Italian varieties: they were however collected already 20 years ago now and as the result of a Map Task, therefore not in ecological, conversational situation. Other treebanks such as ParTUT [Sanguinetti and Bosco, 2015] contain portions of controlled transcribed speech, such as transcribed Ted Talks [Cettolo et al., 2012]. Transcripts of formal speeches also appear in the ParlaMINT treebank [Agnoloni et al., 2022]. Differently from the aforementioned resources, the genesis of the KIParla corpus stems from the imperative to capture the linguistic diversity inherent in spoken language, representing variation with respect to the properties of the interaction and the properties of the speakers involved. Geographical differentiation is represented in the corpus both by the influence and occasional use of local dialects and by regional traits, perceptible even in the speech of educated individuals. More generally, spoken data are rare in the broader context of Universal Dependency treebanks [Dobrovoljc, 2022] but offer unique insights into linguistic phenomena and support various research fields, ranging from typology to NLP. A greater availability of spoken treebanks would open the path to large-scale studies on phenomena such as conversational patterns, discourse markers, and syntactic variation, which are hard to scale above the lexical level with available resources. From the NLP perspective, currently available pipelines often drastically drop their accuracy rate when running on spoken language varieties and no spoken resource is currently available to train accurate annotation pipelines tailored to speech data.

The KIParla corpus, initially a collaborative effort between the University of Bologna and the University of Turin, encompasses a diverse range of Italian spoken varieties, manually transcribed following Jefferson guidelines[1] and aligned with audio files. The main novelty of KIParla in the landscape of spoken data collection is that it is structured in an incremental and modular fashion, which allows for the addition of new corpus modules over time, focusing on different dimensions of linguistic variation and geographical areas. Thanks to thoughtful guidelines on formats, transcriptions and metadata, the integration of new modules into the overall resource is easily achieved. The entire corpus is freely available for consultation[2], and an orthographic transcription is already provided upon request. Each module within the overarching KIParla project is balanced to ensure its coherence and self-sufficiency as a linguistic resource (Table 1).

Module	Description
KIP[3]	data from university lectures, examinations, student consultations, semi-structured interviews, and spontaneous speech. Duration: 70 hours, 661,175 tokens. Status: Available.
ParlaTO[4]	individual interviews and group discussions. Duration: 50 hours. Status: Available.
ParlaBO	semi-structured interviews. Duration: 50 hours. Status: Close to publication.
ParlaBZ	semi-structured interviews and dinner table conversations. Duration: Approximately 5 hours. Status: Close to publication.
KIPasti	spontaneous speech during family or friendly gatherings. Duration: 40 hours, 487,607 tokens. Status: Available
Stra-ParlaBO	semi-structured interviews and spontaneous speech from speakers with international migration backgrounds. Duration: Approximately 50 hours. Status: Under construction.
Stra-ParlaTO	semi-structured interviews and spontaneous speech from speakers with international migration backgrounds. Duration: Approximately 50 hours. Status: Under construction.
ParlaMI	semi-structured interviews and dinner table conversations. Duration: Approximately 60 hours. Status: Planned.
ParlaNA	various communicative contexts. Status: Planned.

Table 1: KIParla Project Modules

Preliminary linguistic annotation efforts on

---

the KIParla corpus were initiated during the EVALITA[5] evaluation campaign in 2020. The KIPoS task [Bosco et al., 2020], focusing on Part-of-Speech tagging of KIParla data, received attention within this campaign. The dataset provided for the task comprised approximately 200,000 tokens, sourced from the KIP module, with around 30,000 tokens undergoing manual review and correction after automatic processing through UDPipe [Straka, 2018].

Participating teams obtained satisfactory results on the task, highlighting the challenges in morphologically and syntactically tagging speech, especially in unconstrained speech contexts. Those will especially be the focus of my participation in the training school and could be addressed during the brainstorming hackathon, namely: the identification of mode-specific phenomena, tracing the greater variety of non-standard and creative alternatives to be assigned back to the same linguistic phenomenon (e.g., forms to be assigned to the same lemma), and dealing with different types of interaction and registers. The newly collected modules, more specifically the KIPasti one and those targeting international migration, are bound to introduce **further** challenges, especially concerning code-switching phenomena both towards local dialects and towards other languages present in the speakers' linguistic background. In the KIPoS task, for instance, two Part-of-Speech labels were introduced to tag tokens coming from *Italo-Romance dialects* and from *Languages other than Italian* (namely, DIA and LIA respectively). Revisions may be necessary to accommodate structural code-switching phenomena, aligning with approaches in other multilingual UD treebanks [Çetinoğlu and Çöltekin, 2019, Braggaar and Goot, 2021]. As pointed out in Dobrovoljc [2022], many other aspects need specific attention in the case of spoken treebanks: in the case of KIParla, some of them have already been addressed in the creation of the resource. For instance, as far as **orthography and punctuation** are concerned, the corpus is already provided with a standard orthography transcription, but no sentence-initial capitalization is performed and standard punctuation is inserted in cases of questions and exclamations. **Segmentation** into sentence-like units is notoriously a controversial topic when it comes to speech. In the KIPoS task the original intonation-based segmentation was taken to

define sentence boundaries: the Rhapsodie treebank [Kahane et al., 2021] considers instead illocutory units (i.e., a speech segment that corresponds to a single speech act) as basic units for segmentation, thus differentiating between microsyntactic (government) units and macrosyntactic maximal units [Pietrandrea et al., 2014]. This approach seem to present numerous advantages, namely the fact that it proceeds bottom-up building dependency relations among tokens and that it can be more easily operationalized in a series of tests to guide annotators. A connected aspect is that of **speakers' overlap**, which is also taken track of in the Rhapsodie approach. Table 2 summarizes the different approaches taken in available UD spoken treebanks with respect to the mentioned aspects: the last column reports features that a treebank based on KIParla could provide.

	Beja	Cantonese	Chinese	Chukchi	ParisStories	Rhapsodie	Frisian-Dutch	Komi-Zyrian	Naija	Norwegian	Slovenian	Turkish-German	KIParla forest
Sound file ID	yes	no	no	yes	yes	no	no	no	yes	no	no	no	yes
Text-sound alignment	yes	no	no	yes	yes	no	no	no	yes	no	no	no	yes
Speaker ID	no	no	no	no	yes	yes	yes	no	yes	yes	no	no	yes
Language variety	no	no	no	no	no	yes	yes	no	yes	no	no	no	yes
Standard ortography	no	no	yes	yes	yes	yes	yes	no	no	yes	yes	yes	yes
Capitalization	no	no	no	no	no	no	yes	no	yes	no	no	no	no
Pronunciation	yes	no	no	yes	no	no	no	no	no	no	yes	no	no
Speaker overlap	no	no	no	no	no	yes	no	no	no	no	no	no	yes
Final punctuation	yes	yes	yes	yes	yes	yes	no	yes	yes	yes	no	yes	no
Other punctuation	yes	yes	yes	yes	yes	no	yes	yes	yes	yes	no	yes	no
Incomplete words	no	no	no	yes	yes	yes	no	no	yes	yes	yes	yes	yes
Fillers	no	no	no	no	yes	yes	yes	no	yes	yes	yes	yes	yes
Silent pauses	yes	yes	no	no	no	no	no	no	yes	yes	yes	no	yes
Incidents	no	no	no	no	no	no	no	no	no	no	yes	no	yes

Table 2: Features to be found in spoken UD treebanks, from Dobrovoljc [2022]

The proposal is therefore to build on the already existing portion of annotated data for the KIPoS task, enlarging it with 30,000 tokens coming from the new conversational contexts from the KIPasti module, therefore planning a first release consisting of approximately 60,000 tokens. Participation in the training school will be crucial for the project, because it will allow us to focus on the challenges just highlighted, discussing the possible solutions with expert scholars. Given its size and the variety of conversational contexts recorded in the KIParla, the resulting treebank would represent an invaluable resource to investigate syntax-level phenomena in relation to regional variation, standard vs. substandard and formal vs. informal varieties, sociolectal variation and multilingualism, providing us with the first large scale picture of the linguistic richness and complexity of variation in spoken Italian and its usage across different contexts and communities.

2

References

T. Agnoloni, R. Bartolini, F. Frontini, S. Montemagni, C. Marchetti, V. Quochi, M. Ruisi, and G. Venturi. Making italian parliamentary records machine-actionable: The construction of the parlamint-it corpus. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 117–124, 2022.

S. Ballarè, C. Mauri, et al. La creazione del corpus kiparla: criteri metodologici e prospettive future. *RID, RIVISTA ITALIANA DI DIALETTOLOGIA*, 44:53–69, 2020.

C. Bosco, S. Ballarè, M. Cerruti, E. Goria, and C. Mauri. Kipos @ evalita2020: Overview of the task on kiparla part of speech tagging, 2020. URL www.kiparla.it,.

A. Braggaar and R. V. D. Goot. Challenges in annotating and parsing spoken, code-switched, frisian-dutch data, 2021. URL https://fy.wikipedia.org/wiki/Fryske.

M. Cerruti and S. Ballarè. Parlato: corpus del parlato di torino. *BOLLETTINO DELL'ATLANTE LINGUISTICO ITALIANO*, 44:171–196, 2020.

Ö. Çetinoğlu and Ç. Çöltekin. Challenges of annotating a code-switching treebank. In *Proceedings of the 18th international workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 82–90, 2019.

M. Cettolo, C. Girardi, and M. Federico. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the Conference of European Association for Machine Translation (EAMT)*, pages 261–268, 2012.

R. Delmonte, A. Bristot, and S. Tonelli. Vit-venice italian treebank: Syntactic and quantitative features. In *Sixth International Workshop on Treebanks and Linguistic Theories*, volume 1, pages 43–54. Northern European Association for Language Technol, 2007.

K. Dobrovoljc. Spoken language treebanks in universal dependencies: an overview, 2022.

E. Goria and C. Mauri. Il corpus kiparla: una nuova risorsa per lo studio dell'italiano parlato. In *Club Working Papers in Linguistics. Vol. 2*, volume 2, pages 96–116. CLUB-Circolo linguistico dell'Università di Bologna, 2018.

S. Kahane, B. Caron, E. Strickland, and K. Gerdes. Annotation guidelines of ud and sud treebanks for spoken corpora. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages pp–35. Association for Computational Linguistics, 2021.

C. Mauri, S. Ballarè, E. Goria, M. Cerruti, and F. Suriano. Kiparla corpus: A new resource for spoken italian. In *Proceedings of the 6th Italian Conference on Computational Linguistics CLiC-it.*, 2019.

P. Pietrandrea, S. Kahane, A. Lacheret-Dujour, and F. Sabio. The notion of sentence and other discourse units in corpus annotation. *Spoken corpora and linguistic studies*, pages 331–364, 2014.

M. Sanguinetti and C. Bosco. Parttut: The turin university parallel treebank. *Harmonization and development of resources and tools for italian natural language processing within the parli project*, pages 51–69, 2015.

M. Straka. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In D. Zeman and J. Hajič, editors, *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. doi: 10.18653/v1/K18-2020. URL https://aclanthology.org/K18-2020.