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

Ludovica Pannitto
University of Bologna
ludovica.pannitto@unibo.it

Eleonora Zucchini
University of Bologna
eleonora.zucchini2@unibo.it

Silvia Ballarè
University of Bologna
silvia.ballare@unibo.it

Cristina Bosco
University of Torino
cristina.bosco@unito.it

Caterina Mauri
University of Bologna
caterina.mauri@unibo.it

Manuela Sanguinetti
University of Cagliari
manuela.sanguinetti@unica.it

Abstract

The present project endeavors to enrich the linguistic resources available for Italian by introducing *KIParla Forest*, a treebank for the KIParla corpus - an existing and well-known resource for spoken Italian. This article contextualizes the project, describes the treebank creation process and design choices, and highlights future plans for next improvements.

1 Introduction

Today, the Universal Dependencies (henceforth, UD, de Marneffe et al. 2021) body of resources[1] counts 296 treebanks for 168 languages. While many different genres are represented among the corpora, ranging from news to fiction to legal texts, spoken language is surely underrepresented. This aspect strikes as counterintuitive if one thinks that language resources should mirror language use; however, it is actually in line with a tendency in the Natural Language Processing (NLP) community to rely on what is, or was, easily accessible for processing rather than truly representative. Spoken language, in fact, poses unique challenges when it comes to its representation for processing, some of which derive from the long-standing but unstated assumption that NLP is primarily Written Language Processing (Linell, 2019; Chrupała, 2023). As a result, while there is a shared consensus on the primate of spoken over written language, only approximately 20 out of the 168 UD languages have a dedicated spoken treebank (Dobrovoljc, 2022), and Italian is not among those. A greater availability of spoken treebanks would open the path to large-scale studies on phenomena typical of interactional data, such as conversational patterns, discourse markers, and syntactic variation, which are hard to scale above the lexical level with available resources. The NLP community has only recently begun to focus on spoken languages, taking into account not only institutional languages but also dialects and endangered languages (Bird and Yibarbuk, 2024). The great diversity of these languages and their wide distribution make starting to study them particularly urgent. From the NLP perspective, accuracy rates of currently available pipelines drastically drop when running on spoken language varieties, and no spoken resource is currently available to train accurate annotation pipelines tailored to speech data (see, among others, Liu and Prud'hommeaux 2023).

We therefore introduce *KIParla Forest*, the first Universal Dependency treebank of Spoken Italian, derived from the KIParla corpus project (Mauri et al., 2019a; Ballarè et al., 2020). In this paper, we examine the motivations and major design choices taken in the first phases of the creation of the resource, focusing in particular on the pipeline from segmentation to syntactic annotation. KIParla Forest is planned for release in UD 2.17 in November 2025. Because of their complexity and the need for linguistic glosses, most examples are reported in Appendix A.

2 Universal Dependencies for Spoken Language

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. UDv2.0 already included UD_Slovenian-SST (Dobrovoljc and Nivre, 2016), a treebank composed entirely of spoken data, and some spoken data in mixed-genre treebanks. Despite the fact that UDv2.16 sees now 48 treebanks counting both spoken-specific and mixed-genre treebanks that contain spoken data, a full set of guidelines dedicated to spoken-specific phenomena is yet to be released. Currently, a dedi-

[1] www.https://universaldependencies.org/

Proceedings of the Eighth International Conference on Dependency Linguistics (Depling, SyntaxFest 2025), pages 54–73
August 27-28, 2025 ©2025 Association for Computational Linguistics

cated taskforce within the UniDive COST Action[2] is dedicated to analyzing and harmonizing current practices for morphosyntactic annotation of speech-specific phenomena. Currently, in fact, treebank curators took different directions in the creation of their resources, which could impact on derived measures or performance on downstream tasks (see Table 1 for an overview). Most spoken treebanks include information about alignment and metadata about speakers and language variety. As far as capitalization and punctuation are concerned, some take a written-derived approach, normalizing the transcription with added capitalization and written-like punctuations, while others (for instance, UD_Beja-Autogramm Kahane et al. 2022) employ it to represent prosodic traits. Fillers and filled pauses are reported in most treebanks, mostly with conventionalized transcriptions (e.g., *euh* in French, *e* in Norwegian or *ähm* in Turkish-German), either marked as `X` or `INTJ` (we choose the latter) and generally labeled as `discourse` or `discourse:fillers`, attaching to the root of the sentence. Discourse markers are generally marked according to their syntactic category (they could be verbal, adverbial, interjections, etc). They are generally labeled as `discourse`, while Naija NSC (Caron et al., 2019), Slovenian SST (Dobrovoljc and Nivre, 2016) and Turkish-German SAGT (Çetinoğlu and Çöltekin, 2019) use `parataxis:discourse` for distinguishing clausal markers.

	Beja	Cantonese	Chinese	Chukchi	ParisStories	Rhapsodie	Frisian-Dutch	Komi-Zyrian	Naija	Norwegian	Slovenian	Turkish-German	**KIParlaForest**
Sound file ID	yes	no	no	yes	yes	no	no	no	yes	no	no	no	no*
Text-sound alignment	yes	no	no	yes	no	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	no	yes	yes	no	yes	no	yes	yes
Standard ortography	no	no	yes	yes	yes	yes	yes	no	no	yes	yes	yes	yes
Capitalization	no	no	no	yes	no	no	no	yes	no	no	no	yes	yes*
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	yes	no	yes
Final punctuation	yes	yes	yes	yes	yes	yes	no	yes	yes	yes	no	yes	no
Other punctuation	yes	yes	yes	no	yes	yes	no	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	no	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 1: The table, adapted from (Dobrovoljc, 2022), compares features to be found in spoken UD treebanks and features that will be available in KIParlaForest (rightmost column). The table is not exhaustive as, since Dobrovoljc's paper in 2022, new treebanks of spoken data have appeared within the UD family of resources.

3 Universal Dependencies for Italian

Italian has a very solid tradition in the UD enterprise, with resources already appearing in the first release dating back to 2015 (Nivre et al., 2015). The first UD-based treebank ever released for Italian is ISDT (Italian Stanford Dependency Treebank), originally developed for the dependency parsing shared task of EVALITA-2014 (Bosco et al., 2014). In the current release, ISDT contains approximately 298K tokens and includes texts pertaining to the legal domain, or harvested from news and Wikipedia. Another Italian treebank, UD-VIT (Alfieri et al., 2016) was obtained by semi-automatically converting the Venice Italian Treebank (Delmonte et al., 2007), which included approximately 60K words of spoken data in its original version. However, to the best of our knowledge, this portion was not ported into the UD resource.

Spoken data, or as we could better define it 'conceptually-written' (Koch and Oesterreicher, 2012) or 'spoken-written' (Nencioni, 1976) language, is also collected in the ParlaMint corpus (Agnoloni et al., 2022; Alzetta et al., 2024), built from stenographic verbatim records of parliamentary speeches. Whereas, as the authors say, 'debates of the COVID-19 period are mostly characterised by traits specific to the spontaneous speech', no detailed description of such features is provided and no measures are described to adapt UD guidelines to such genre. Similarly, a section of the parallel treebank ParTUT (Sanguinetti and Bosco, 2014, 2015) features annotated data from the Europarl corpus (Koehn, 2005), a collection of texts from the proceedings of the European Parliament.

Two resources that are not specific for spoken language but are still relevant for our work are PoSTWITA-UD (Sanguinetti et al., 2018) and TWITTIRÒ-UD (Cignarella et al., 2019), which contain collections of tweets: in these cases, explicit choices were made to extend UD guidelines to non-standard productions, in particular extending the `parataxis` relation to systematically cover a class of juxtaposition phenomena. Many of these guidelines are collected in Sanguinetti et al. (2023), that describes annotation choices for user-generated content. Lastly, among the resources of interest to our domain, is MarkIT (Paccosi et al., 2023), which contains around 800 sentences, extracted from students' essays, covering seven types of marked constructions, many of which are also typical of spoken data, such as for instance hang-

[2] CA21167, https://www.cost.eu/actions/CA21167/

ing topic sentences or sentences with presentative *there*. In this scenario, KIParla Forest would thus represent the first attempt to develop a fully-spoken treebank for Italian. The following section will outline the corpus from which this treebank originates.

4 Data

The KIParla corpus[3] (Mauri et al., 2019a; Ballarè et al., 2020) is a resource for the study of spoken Italian and is a product of a collaborative effort between the Universities of Bologna and Turin. It is structured in an incremental and modular fashion that allows the addition of new corpus modules over time. To date, KIParla encompasses a diverse range of Italian spoken varieties and involves participants of various age, genders and backgrounds and with different professional and educational achievements. As a whole, the KIParla counts ca. 228 hours of recordings and approximately 2M transcribed tokens. At the time of writing, the corpus is freely available for consultation through a custom noSketchEngine service[4], that provides transcriptions, carried out manually following Jefferson guidelines (Jefferson, 2004), aligned with audio files; access to full transcripts is also provided. Preliminary linguistic annotation efforts on the KIParla corpus were initiated during the EVALITA[5] evaluation campaign in 2020. The KIPoS task[6] (Bosco et al., 2020) precisely focused on Part-of-Speech tagging of KIParla data, comprising approximately 200K tokens automatically annotated with UDPipe and partially manually revised. KIParla contains recordings collected in different conversational settings. To create the core of KIParla Forest, a balanced sample of such data was selected to showcase syntactic annotation of conversations presenting different degrees of interactional freedom, and including various number of speakers. Then, the chosen conversations were organised based on interactional levels identified in the KIParla corpus, ranging from *free interaction* (free conversations), to *partially free interaction* (semi-structured interviews), *rigid interaction* (university exams and office hours) and situations with *almost no interaction* (lectures).

When selecting conversations, we made sure we included those that had been already manually annotated during the KIPoS task, in order to capitalize on the gold part of speech annotations already in place[7]. The final selection is reported in Table 2. All summed up, the treebank counts 21.871 tokens.

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 2: Conversations selected for the first release of KIParla Forest.

4.1 Data preparation

KIParla conversations are manually transcribed through ELAN[8] (Max Planck Institute for Psycholinguistics, The Language Archive, 2024) and stored in `.eaf` format. The native transcription format includes a Jefferson-inspired set of conventions to represent features of spoken language (intonation, pace, pauses, overlaps, repair...). The first step towards the construction of the treebank consisted, therefore, in fully transforming the current notation into a columnar format, therefore isolating orthographic tokens from prosodic features annotated in Jefferson notation. Since not all Jefferson features will be included in the UD treebank, we made sure that each orthographic token bears a unique token identifier (`TID`) in order to retrieve, in combination with `sent_id`, more specific features and to ensure backward compatibility with the KIParla resource. As a result of our normalization process, each conversation is represented in a `conll`-like format. The conversation is divided into Transcription Units (TUs), manually identified by transcribers and aligned with audio. TUs are then split into orthographic tokens, each annotated with Jefferson-derived features.

4.2 Speech-specific metadata

Most spoken treebanks include speech-specific metadata such as links to audio files, information about the speaker and on language variety. As audio access is restricted to registered users, for privacy reasons, an explicit link to the audio file cannot be provided as of today in KIParla Forest. All audio files and speaker-specific metadata are available upon request, only for research purposes.

[3] www.kiparka.it

[4] https://search.corpuskiparla.it/corpus/crystal/#open

[5] https://www.evalita.it/

[6] http://www.di.unito.it/~tutreeb/kipos-evalita2020/index.html

[7] see Section 7 regarding the modifications that were implemented.

[8] https://archive.mpi.nl/tla/elan

Two attributes (AlignBegin and AlignEnd, expressed in milliseconds), typically attributed to the first and last token of each TU, are provided at token level in the MISC field of the CoNLLU files. Differently from other spoken corpora, that provide speaker information at the maximal unit level, in our treebank each token bears a special SpeakerID feature that contains the id of the speaker as a value. Each speaker is then described through its metadata (including data such as gender, age, origin, education level, profession) in a separate json file available in the treebank repository. The same applies to conversation-specific metadata (i.e., number of participants, place and date of recording, type of interaction). The resource also contains information about *overlaps*: these represent a particularly challenging feature both to annotate and to parse, as single tokens can participate in more than one overlapping span, and overlaps can happen among two or multiple speakers. We have adopted a special Overlap feature in the MISC column, attributed to all tokens that participate in overlapping spans. The feature value is composed as a comma-separated list of ranges, where each range has format: idX-idY@sent_id-n+...+idT-idS@sent_id-m (see Figure 1). Figure 2 shows how the feature is rendered in the different overlapping scenarios. Figures 5 and 6 in Appendix A demonstrate the annotation.

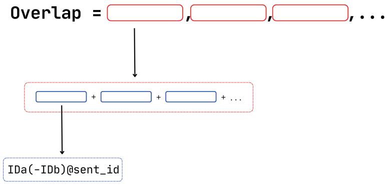
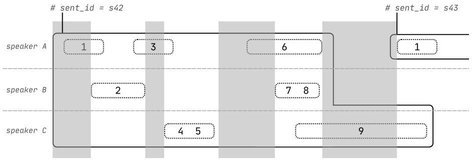


Figure 1: The figure shows the composition of the Overlap feature. The feature value is composed as a comma-separated list of pointers (top tier of the scheme), where each range is in turn composed of a + separated list of ranges. Each range is then a reference to a specific token or sequence of tokens, identified by their CoNLLU IDs and sentence identifier.

The next sections describe the design choices we made to transform such preliminary data collection into proper UD-compatible data and to operationalize certain annotation decisions, starting from the basic segmentation steps up to the syntactic level.



ID	Overlap	ID	Overlap
1	2@s42	6	7-8@s42+9@s42
2	2@s42,3@s42	7	6@s42+9@s42
3	2@s42,4-5@s42	8	6@s42+9@s42
4	3@s42	9	6@s42+7-8@s42,1@s43
5	3@s42		

Figure 2: The figure shows some cases of overlaps. The example shows an extract of a conversation where three speakers (A, B and C) are involved. Dotted boxes represent TUs, while numbers represent orthographic tokens as they would be numbered in CoNLLU. TUs are in fact grouped in two maximal units, namely s42, composed by 9 tokens, and s43, where only token 1 is visible. Token 2 in sentence s42, for instance, overlaps partially with token 1 and partially with token 3: its Overlap feature would therefore be constituted by two different span references. Token 6, on the other hand, participates to a complex span, where all three speakers overlap. This translates into a + separated sequence of references, accounting for the fact that overlap with tokens 7-8 and with token 9 is simultaneous. Token 9 shows a combination of the two overlapping situations.

5 Segmentation into maximal units

Segmentation of spoken data into maximal units has already been tackled by existing spoken language trebanks in UD; however, the documentation mostly lacks information on the formal criteria that were adopted (Dobrovoljc, 2022): a popular choice is that of illocutionary units (IUs, Cresti et al. 1995; Pietrandrea et al. 2014), defined as speech segments that correspond to a single speech act, or linguistic units serving to express a single primary idea. On the other hand, it is well known that, especially in conversational settings, syntactic affordances are exploited across turns by different speakers, to co-construct the structure of discourse (Du Bois, 2014). As a consequence, given that syntactic relations can be observed only within the sentence boundaries, their identification should be carried out in a way that does not obscure relations within a broader linguistic context and allows to natively represent syntactic co-construction among speakers (see Figure 7 in Appendix A). The need for a more careful definition of the maximal unit, i.e., the domain in which syntactic relations hold, is also demonstrated by the introduction of the feature AttachTo in treebanks of spoken language (Kahane et al., 2021b), for cases of co-construction.

When dealing with the transcriptions of the KIParla corpus, it is necessary to consider that TU boundaries cannot be treated as reliable basis for segmentation since they do not represent sentence-like units or utterances in any meaningful way: their boundaries remain highly subjective and there are many examples where core grammatical relations hold between elements of different TUs (see Figure 8 in Appendix A). As a result, we decide for the KIParla Forest to explore a different perspective: we begin with annotating dependencies between words, establishing a unit boundary whenever no dependency link can be found. Such an approach allows boundaries to emerge bottom-up, purely based on the existence of grammatical relations, rather than the reverse; moreover, it enables us to distribute syntax along the speech stream and represent the cooperative - potentially inter-speaker and inter-turn - organization of syntax. While in fact we acknowledge the centrality of IUs in speech, we do not believe they are useful tools to identify syntactic maximal units, as their speaker-bound nature can obscure this cooperative nature of spoken syntax.

This said, for technical reasons we needed a preliminary sentence-like segmentation in order to run an automatic pipeline to pre-annotate parts of speech and also to visualize text in a reasonable way on the annotation tool of our choice (i.e., ArboratorGrew, Guibon et al. 2020). Therefore, we decided to employ `wtpsplit` (Minixhofer et al., 2023), an unsupervised multilingual sentence segmentation system that does not rely on punctuation marks to segment a textual stream. We applied the model to the entire conversation, ordering tokens based on their time of utterance, regardless of speaker. The output of `wtpsplit` constitutes the basis for further steps of analysis and annotation: as it was done in KIPoS, automatic PoS tagging and syntactic parsing was performed through UDPipe (Straka, 2018), based on the model trained on PoSTWITA-UD data from release 2.15 (Straka, 2024). The model was chosen in continuity with what was done during KIPoS task, under the assumption that social media data would share some of the features of spoken language, such as for instance the extended use of discourse markers and increased use of parataxis relations. This constitutes the basis on which annotators will then be free to merge or split such automatically-identified units in the syntactic annotation phase, based on the identification of local grammatical relations. The segmentation phase is thus divided into an automatically-driven one, initial and purely operational, and a manually revised one, which emerges out of the identification and annotation of locally relevant grammatical relations. As a consequence, the identification of maximal units is not anterior to the annotation stage in our approach, but rather emerges from it (see Section 8 for more details).

6 Tokenization and lemmatization

Aside of the Jefferson notation, the KIParla project uses standard orthography and spelling, which, in the case of Italian is not particularly problematic. The only difference between UD-like tokenization and the one natively available through the transcription is the case of multiword tokens, which are used in Italian treebanks for article-preposition contractions and cases of clitics attached to verbs. These were split by our pre-annotation pipeline, which introduced multiword tokens. Metadata such as `SpeakerID`, `TID`, and Jefferson-derived features remain on the multiword token, while new syntactic tokens receive distinct `TID`s for backward compatibility with KIParla (these are created during the parsing step and kept in an intermediate pivot file that allows to cross-reference the corpus and the treebank).

The KIParla resource includes both code-switching and dialectal variation, which is currently identified at TU level by the introduction of a # symbol in the Jefferson transcription, at the beginning of each TU. The information about variation remains therefore available among the metadata of each maximal unit (`# contains_variation`). As new KIParla modules will involve L2 speakers and showcase examples of code-switching, we are experimenting the ‘Code-switched analysis‘ currently proposed by UD guidelines[9]: we therefore differentiate between cases where the foreign material is borrowed and incorporated in Italian, by fully considering them same as Italian material, and cases where we apply the analysis (either morphological or syntactic) of the target language. Such cases are marked by the feature `Lang=CODE` in `MISC`. Ambiguous cases will be annotated as foreign language only when considering them Italian is impossible; unknown cases will be coded with `Lang=UNKNOWN`$_I$`SO`. Dialects, for the moment,

[9] https://universaldependencies.org/foreign.html#option-1-code-switched-analysis

have been coded with `Lang=NO_ISO_CODE` for the lack of a dedicated ISO-639 code (see Figure 3). Furthermore, KIParla contains special tokens to represent non-linguistic behaviour and instances of anonymization (home addresses; work places and the like). Non-linguistic behaviour includes short pauses, tags expressing actual non linguistc behaviour (NLB, e.g., *((laughter))*), annotations expressing modality of utterance (e.g., *((reading))*), events happening outside of the interaction (OOI, e.g., *((phone ringing))*) and notes (e.g., *((recording interrupted))*). These cases are treated differently when imported into the treebank. More specifically, short pauses are transformed into a `PauseAfter=Yes` feature in `MISC`. Cases of true non linguistic behaviour are only kept when relevant to the syntactic construction of the units, with their forms and lemmas uppercased and a feature `Type=NLB` is added to the `MISC` column (see Figure 9 in Appendix A). Modalities are not included in the treebank as tokens, but a feature `Type=reading|singing|...` is added in the `MISC` column on the relevant linguistic tokens (see Figure 10 in Appendix A). OOI events and annotations are kept as metadata at the maximal unit level. Finally, as far as anonymized tokens are concerned, as done in PoSTWITA-UD, instances of anonymized tokens are prefixed by `@`: examples include cases such as '@nomepaese' (en. '@villagename'). It is worth mentioning that all personal first names (except for recognizable names e.g., celebrities) are pseudonymised: they are replaced with a different name of approximately the same length; therefore, such instances are considered as normal tokens.

961, BOI012 >lo so che bologna< è basket
3SG.OBJ know.1SG that Bologna is basket
⋆

city ma::
city but
⋆

'I know that Bologna is "basket city" but'

(a)

#pa se vuoi fazzu eu
da(d) if want:2SG do:1SG 1SG
⋆ ⋆

'dad if you want I can do it'

(b)

Figure 3: Both examples show code-switching phenomena, example 3a includes English elements while 3b includes elements from an Italo-Romance dialect. Tokens marked with ⋆ have features `Foreign=Yes` and `Lang=eng` in 3a, and feature `Lang=NO_ISO_CODE` in 3b. (from conversations PBB004 and KPS001)

Concerning lemmatization, a few choices need to be discussed. While the original transcription contains no capital letters at all, all proper nouns' lemmas have been capitalized in order to facilitate downstream tasks that might require named entity recognition. Words interrupted during speech (i.e., false starts) are lemmatized as their complete version whenever the context is informative enough, either because there is a repetition surrounding the interrupted word (see Example 1) or because there is compatible syntactic context preceding or following the token, as in Example 2. We did not trust semantic predictability to be informative enough, so we did not lemmatize cases as the one in Example 3. In this case semantics would suggest 'persone' (en. 'people') as the interrupted lemma, but we excluded these cases as a clear repetition was missing. A feature `Interrupted=Yes` is reported in `MISC` in all cases.

(1) vabbè scusa è **sta**∼ è **stato**
well sorry AUX.3SG be∼ AUX.3SG been
più...
more
'sorry it's be∼ it's been more...' (conversation BOA3017)

(2) e non è una città **vic**∼ che è vicino a
and NEG is INDEF city cl∼ that is close to
tante possibilità
many possibilities
'and it is not a city cl∼ that is close to many possibilities' (conversation BOD2018)

(3) generalmente: [non] conosco person∼
generally NEG know:1SG peop∼
famiglie: che bolognesi che abitano in
families that from.bologna that live:3PL in
centro
centre
'I generally don't know peop∼ families that from Bologna that live in the city centre' (conversation BOD2018)

Acronyms are transcribed through their phonetic realization, at the form level, and they are lemmatized as their dictionary entries ('RSA', en. 'nursing home', lemmatized as such but transcribed as *erreessea*, its phonetic realization). Interjections and ideophones are transcribed but are normalized, at lemma level, to the lexical entry that can be found in Italian dictionaries (for instance, 'okay' is kept as such at the form level but lemmatized as 'ok').

7 PoS tagging and morphology

The KIPoS task, the first attempt at PoS annotation on the KIParla corpus, was carried out using a tagset only inspired by UD tagset, that included also PoS labels introduced on purpose. Specifically, NEG was employed for sentence negation, PARA for particles pertaining to paraverbal communication, DIA and LIN as subtypes of any UPOS to mark Italo-Romance dialectal variation and languages other than Italian[10]. Therefore, in our process, after automatic annotation we aligned our data with KIPoS gold datasets, having restored the UPOS labels, in order to retrieve as much gold annotation as possible. This was then the basis for manual correction. Annotation (both for morphosyntax and syntax) was performed collaboratively by the authors through ArboratorGrew.

We operated by the following criteria. Fillers and filled pauses, which include cases such as *beh*, *eh*, *ehm* and *mh*, are marked as INTJ. Interrupted words are tagged either with the PoS of their repair (see Section 6) or with X. We align with French Rhapsodie, ParisStories (Kahane et al., 2021a) and Naija NSC (Caron et al., 2019) marking them with the $\sim$ symbol at the end of the form in order to avoid any possible overlap with Italian words that contain an hyphen. We adopt a rather conservative approach when assigning PoS labels, sticking to the main category of each word, even though they perform a different function in the syntactic context. An example may be the word *basta* (en. 'stop'), which is an inflected form of the verb *bastare* (en. 'to stop') but also works as discourse marker meaning 'that's it'. In line with choices taken by other spoken language treebanks, all discourse markers are marked according to their morphological category (e.g. verbs, adverbs, interjections, etc.). We specifically questioned the annotation of determiners: we restricted the use of 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 (be they *adjectival*, *numeral* or *possessive*). This allows for a consistent annotation of diatopical variation concerning, for instance, the position of elements such as *mio, tuo, suo...* (en. *my, your, his/her...*), which may precede or follow the nominal head (e.g, the use of 'il mio libro' over 'il libro mio' can depend on a number of factors, which also include simple diatopical variation with no implications on the linguistic relation between 'mio' - *my* - and 'libro' - *book*). However, in cases where modifiers such as possessives exclude the presence of a properly defined determiner (e.g., 'mia mamma', en. *my mom*), these are tagged as DET. We manually revise morphological features while we do not annotate XPOS.

We computed Cohen's κ to evaluate inter-annotator agreement (Artstein and Poesio, 2008) on UPOS labeling, obtaining almost perfect agreement (above 0.87[11]) in all our scenarios. For the agreement task, an external annotator was asked to annotate UPOS labels on approximately 1500 tokens from each of our four conversations. We provided the annotator files pre-annotated with the PoSTWITA UDPipe model and set up a dedicated project on ArboratorGrew. We also instructed the annotators with the criteria described in this section. As expected, most disagreement is registered between CCONJ and ADV, which are the ones more prone to develop discourse functions.

8 Syntax

Dependency trees in the UD formalism are directed acyclic graphs that have tokens as nodes and *grammatical relation*s as edges, with no notion of constituency or bracketing allowed. However, not all edges allowed in UD represent syntactic relation in the strict sense (Mel'cuk, 1988; Tesnière, 2015): there exist relations like *flat* or *goeswith* that aim at representing exocentric constructions or at allowing to treat phenomena that are more pertinent to the form in which data presents itself, rather than its actual linguistic structure (de Marneffe and Nivre, 2019). UD is also already equipped with a set of relations that seem to have been introduced with speech in mind: in particular *discourse* that is used for interjections and other discourse particles and elements not clearly linked to the structure of the sentence, *reparandum* for disfluencies overridden in a speech repair and *parataxis*, whose cases of applications are manifold and include discourse relations in linking clauses and tag questions. In designing our treebank, we tried to took advantage of these already defined relations, while questioning the need for more fine-grained analysis of spoken-language specific phenomena. More specif-

[10] Moreover, because of technical issues, the data employed for the KIPoS task was not entirely identical to the current version of the corpus.

[11] More precisely, $\kappa = 0.87$ for BOD2018, $\kappa = 0.88$ for BOA3017, $\kappa = 0.88$ for TOD1005bis, $\kappa = 0.91$ for PBB004.

73, BOI013 se perdo un autobus poi devo
if lose:1SG INDEF bus after must:1SG
s∼ (a)spettare un'altra ora
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 molto
ah okay **so** DEF mobility is very
ridotta
reduced
'ah okay **so** mobility is very reduced'

Figure 4: A case where 'quindi' (e. 'so'), usually used as a connective, develops a discourse function as the antecedent of the connective is missing. (from conversation PBB004)

ically, we label interjections and filled pauses as *discourse*, attaching them to the closest projective head. The relation `reparandum` is to be used for disfluencies and self-repairs, that concern both individual words or longer chunks. In this case, the false start or interrupted token is linked to its repair. The biggest issues arise when dealing with clause-linking criteria and, therefore, in relation to segmentation. As described in Section 5, we choose not to segment conversations based on a priori criteria, but we rather start from the annotation of local, purely syntactic dependencies and establish a boundary whenever no further dependency can be found. Such an approach to segmentation seems to be more adherent to how speakers construct syntactic relations, that is, incrementally, by progressive, unplanned expansions (see Figure 11 in Appendix A) that exploit syntactic connections to keep discourse tightly interwoven and cohesive (e.g. through relative clauses, conjunctions, lists - see Masini et al. 2018; Mauri et al. 2019b -, etc.). Such an approach, however, comes with consequences, both for segmentation and relation labeling. The first problem regards connectives, because they frequently develop discourse functions (see Example 4, the connective is rendered in bold) and it may be difficult to tell cases in which they create a local syntactic link from cases in which they indicate some general anaphoric discourse relation; however, choosing among the two strongly affects segmentation.

We are still working on a precise set of criteria to deal with such cases. Two parameters that we currently rely on are prosody (i.e., the syntactic annotation task has to be performed while listening to the recording) and the 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. In this latter case, the connective is identified as having a *discourse marker* function. Discourse markers are indeed typical of spoken language and are connected to their head through the `discourse` relation. Interestingly, discourse markers frequently involve more than one word. Following what is done with complex prepositions, we chose to identify well established cases through the `fixed` relation. A further problem in unit boundaries identification occurs when the grammatical relationship is not overtly marked, as in parataxis. In such cases (Figure 11), in particular in the case of 'eh uno studio sui 'riti magici' (en. 'well, a study on magic rites'), it may be difficult to decide whether we are dealing with listing or implicit reformulation occurring within the same unit, or with independent segments. Prosody here plays a crucial role, because independent segments typically correlate to separate prosodical contours (Mithun, 1988), while there is rich evidence in the literature for clearly identifiable intonational patterns associated to lists (Masini et al., 2025). Moreover, spoken data require to find a specific way to treat feedback phenomena; since the extent of such phenomena varies greatly, we have hypothesized the following solutions. We consider as internal to the maximal units all feedback phenomena that do not interfere with the main syntactic flow (see Figure 12), regardless of the speaker who is uttering them; in such cases, we propose to link the expression providing feedback to the unit head through the *ad hoc* label `discourse:feedback`; no maximal unit boundary is thus identified. 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. Obviously, there may be cases that are more difficult to assign to one of these two types; as with the other pending issues, we are working on testing the validity of our hypothesis on larger sets of data.

9 Conclusion and Future work

One might wonder: why taking the trouble to create a treebank for spoken language, if most of the categories seem to be ill-defined when applied to spoken data? We asked ourselves the same question while working on this project, and we came to the following conclusion. On the one hand, we hope to lay the foundations of a future discussion on the categories themselves, giving a contribution from the perspective of ecological spoken data, i.e. naturally-occurring spoken data, collected in real-life communicative contexts, rather than in artificial or experimental settings. This perspective emphasizes the importance of capturing language as it is actually used by speakers in their everyday social interactions, preserving the features of spontaneity, interaction, variation, and context-dependence that characterize real-world speech. On the other hand, the creation of a spoken treebank is, in our case, also aimed to offer an additional level for accessing and querying the resource: when it comes to spoken data, in fact, interfaces are typically limited to form-based queries, highly restricting the range of possible data explorations. For these reasons, the choices we made about what kind of relations are considered *grammatical relations* are tailored to represent the interactional architecture of ecological spoken data. In our work, our design choices try to follow the competing criteria taken as design choices for the UD formalism (de Marneffe et al., 2021), without favoring one perspective over the others (we comment here on the relevant ones):

- *UD needs to be reasonably satisfactory on linguistic analysis of individual languages*: Italian shows great internal variation, not only wrt. to written vs. oral modalities, but also in terms of regional and register variation. The development of KIParla Forest aims to move a step forward in the process of representing intra-linguistic diversity;

- *UD needs to be good for linguistic typology*: a treebank of spoken language avoiding *a priori* segmentation, based instead on local (and incremental) syntactic relations, allows to represent and extract phenomena of syntactic chains along speech. This allows for better typological comparability with (often purely oral) languages showing, for instance, clause-chaining phenomena (Mansfield and Barth, 2021);

- *UD must be suitable for rapid and consistent annotation*: we kept the modifications to the UD annotation procedure to the minimum, to favor consistency and rapidity and limit the need to learn new rules;

- *UD must support well downstream tasks*: while the role of resources in the LLM era is a challenging discussion topic, we believe that our choices could be fit to support tasks that require rich semantic representations, based on larger discourse context, as well as open the path to benchmarks dedicated to interactional fluency.

While our morphosyntactic choices have been satisfactorily validated through inter-coder agreement, future work is needed on the validation of syntactic criteria. Moreover, as more and more spoken treebanks are being released, we foresee a broader discussion within the community to agree on common annotation guidelines for spoken-specific phenomena.

10 Limitations

The paper describes initial steps towards the release of a new resource. We therefore acknowledge that many of our statements are to be considered preliminary and are likely to be rediscussed and updated as new data are integrated in the resource. Moreover, UD is still lacking stable and shared guidelines on the annotation of spoken data. We will participate in the community debate to develop shared guidelines and update our choices accordingly.

11 Ethical considerations

The KIParla corpus is compliant with current European data protection regulations (Data protection - European Commission[12]); all data are recorded with overt microphones and speakers provide a written consent to the collection and usage of the data for research purposes. Before upload, audio tracks and transcriptions are pseudonymized, by removing all sensitive information. Metadata regarding speakers and conversations are stored and shared in an aggregated format that prevents speakers' recognition. The treebank is automatically linked to the original data, and the choices taken ensure the possibility of removing data, should speakers revoke their consent.

[12] https://commission.europa.eu/law/law-topic/data-protection_en?prefLang=it

Acknowledgments

The research leading to these results has received funding from Project ”DiverSIta-Diversity in spoken Italian”, prot. P2022RFR8T, CUP J53D23017320001, funded by EU in NextGenerationEU plan through the Italian "Bando Prin 2022 - D.D. 1409 del 14-09-2022". This work also received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology). We want to specifically thank student that helped and will help up annotating the treebank.

References

Tommaso Agnoloni, Roberto Bartolini, Francesca Frontini, Simonetta Montemagni, Carlo Marchetti, Valeria Quochi, Manuela Ruisi, and Giulia Venturi. 2022. 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.

Linda Alfieri, Fabio Tamburini, and 1 others. 2016. (almost) automatic conversion of the venice italian treebank into the merged italian dependency treebank format. In *CEUR WORKSHOP PROCEEDINGS*, volume 1749, pages 19–23. AAccademia University Press.

Chiara Alzetta, Simonetta Montemagni, Marta Sartor, and Giulia Venturi. 2024. Parlamint-it: an 18-karat UD treebank of Italian parliamentary speeches. *Language Resources and Evaluation*.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Silvia Ballarè, Caterina Mauri, and 1 others. 2020. La creazione del corpus kiparla: criteri metodologici e prospettive future. *RID, RIVISTA ITALIANA DI DIALETTOLOGIA*, 44:53–69.

Steven Bird and Dean Yibarbuk. 2024. Centering the speech community. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics - Volume 1: Long Papers*, page 826–839. ACL.

Cristina Bosco, Silvia Ballare, Massimo Cerruti, Eugenio Goria, and Caterina Mauri. 2020. Kipos@ evalita2020: Overview of the task on kiparla part of speech tagging. *EVALITA Evaluation of NLP and Speech Tools for Italian-December 17th, 2020.*

Cristina Bosco, Felice Dell’Orletta, Simonetta Montemagni, Manuela Sanguinetti, and Maria Simi. 2014. The Evalita 2014 Dependency Parsing task. *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014 : 9-11 December 2014, Pisa*, pages 1–8. Publisher: Pisa University Press.

Bernard Caron, Marine Courtin, Kim Gerdes, and Sylvain Kahane. 2019. A surface-syntactic ud treebank for naija. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 13–24. Association for Computational Linguistics.

Özlem Çetinoğlu and Çağrı Çöltekin. 2019. 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.

Grzegorz Chrupała. 2023. Putting Natural in Natural Language Processing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7820–7827, Toronto, Canada. Association for Computational Linguistics.

Alessandra Teresa Cignarella, Cristina Bosco, and Paolo Rosso. 2019. Presenting twittirò-ud: An italian twitter treebank in universal dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 190–197.

Emanuela Cresti and 1 others. 1995. Speech act units and informational units. In *E. Fava (haz.), Speech Acts and Linguistic Research, Proceedings of the Workshop, Center for Cognitive Science, State University of New York at Buffalo, Nemo, Padova*, pages 89–107.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Marie-Catherine de Marneffe and Joakim Nivre. 2019. Dependency grammar. *Annual Review of Linguistics*, 5(Volume 5, 2019):197–218.

Rodolfo Delmonte, Antonella Bristot, and Sara Tonelli. 2007. 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 Technologies.

Kaja Dobrovoljc. 2022. Spoken language treebanks in universal dependencies: An overview. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1798–1806.

Kaja Dobrovoljc and Joakim Nivre. 2016. The universal dependencies treebank of spoken slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1566–1573.

John W. Du Bois. 2014. Towards a dialogic syntax. *Cognitive Linguistics*, 25(3):359–410.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *LREC 2020-12th Language Resources and Evaluation Conference*.

Gail Jefferson. 2004. Glossary of transcript symbols with an introduction. In Gene H. Lerner, editor, *Pragmatics & Beyond New Series*, volume 125, pages 13–31. John Benjamins Publishing Company, Amsterdam.

Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes. 2021a. 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.

Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes. 2021b. Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 35–47, Sofia, Bulgaria. Association for Computational Linguistics.

Sylvain Kahane, Martine Vanhove, Rayan Ziane, and Bruno Guillaume. 2022. A morph-based and a word-based treebank for beja. In *TLT 2021-20th International Workshop on Treebanks and Linguistic Theories. 21-25 March 2021, Sofia, Bulgaria*, pages 48–60.

Peter Koch and Wulf Oesterreicher. 2012. Language of immediacy-language of distance: Orality and literacy from the perspective of language theory and linguistic history.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.

Per Linell. 2019. The Written Language Bias (WLB) in linguistics 40 years after. *Language Sciences*, 76. Publisher: Elsevier Ltd.

Zoey Liu and Emily Prud'hommeaux. 2023. Data-driven parsing evaluation for child-parent interactions. *Transactions of the Association for Computational Linguistics*, 11:1734–1753.

John Mansfield and Danielle Barth. 2021. Clause chaining and the utterance phrase: Syntax–prosody mapping in matukar panau. *Open Linguistics*, 7(1):423–447.

Francesca Masini, Claudia Roberta Combei, and Roberta Cicchirillo. 2025. The prosody of list constructions. In Kiki Nikiforidou and Mirjam Fried, editors, *Multimodal Communication from a Construction Grammar Perspective*, Constructional Approaches to Language, pages 116–151. John Benjamins Publishing Company.

Francesca Masini, Caterina Mauri, Paola Pietrandrea, and 1 others. 2018. List constructions: Towards a unified account. *Italian Journal of Linguistics*, 30(1):49–94.

Caterina Mauri, Silvia Ballarè, Eugenio Goria, Massimo Cerruti, and Francesco Suriano. 2019a. Kiparla corpus: A new resource for spoken italian. In *Proceedings of the 6th Italian Conference on Computational Linguistics CLiC-it.*

Caterina Mauri, Eugenio Goria, and Ilaria Fiorentini. 2019b. Non-exhaustive lists in spoken language: A construction grammatical perspective. *Constructions and frames*, 11(2):290–316.

Max Planck Institute for Psycholinguistics, The Language Archive. 2024. ELAN (version 6.9) [computer software]. `https://archive.mpi.nl/tla/elan`. Retrieved from https://archive.mpi.nl/tla/elan.

Igor Aleksandrovic Mel'cuk. 1988. *Dependency Syntax: Theory and Practice*. SUNY Press.

Benjamin Minixhofer, Jonas Pfeiffer, and Ivan Vulić. 2023. Where's the point? self-supervised multilingual punctuation-agnostic sentence segmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7215–7235, Toronto, Canada. Association for Computational Linguistics.

Marianne Mithun. 1988. The grammaticization of coordination. In John Haiman and Sandra Thompson, editors, *Clause combining in grammar and discourse*, pages 331–60. Benjamins, Amsterdam; Philadelphia.

Giovanni Nencioni. 1976. Parlato-parlato, parlato-scritto, parlato-recitato.

Joakim Nivre, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richárd Farkas, Jennifer Foster, Filip Ginter, Yoav Goldberg, Jan Hajič, Jenna Kanerva, Veronika Laippala, Alessandro Lenci, Teresa Lynn, Christopher Manning, Ryan McDonald, Anna Missilä, Simonetta Montemagni, Slav Petrov, and 7 others. 2015. Universal dependencies 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Teresa Paccosi, Alessio Palmero Aprosio, and Sara Tonelli. 2023. Adding a Novel Italian Treebank of Marked Constructions to Universal Dependencies. *IJCoL. Italian Journal of Computational Linguistics*, 9(1). Number: 1 Publisher: Accademia University Press.

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.

Manuela Sanguinetti and Cristina Bosco. 2014. Converting the parallel treebank partut in universal stanford dependencies. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014: 9-11 December 2014, Pisa*, pages 316–321. Pisa University Press.

Manuela Sanguinetti and Cristina Bosco. 2015. 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.

Manuela Sanguinetti, Cristina Bosco, Lauren Cassidy, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2023. Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations. *Language Resources and Evaluation*, 57(2):493–544.

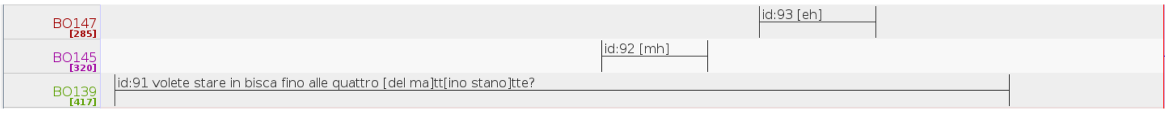
Manuela Sanguinetti, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, Oronzo Antonelli, and Fabio Tamburini. 2018. Postwita-ud: an italian twitter treebank in universal dependencies. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Milan 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 (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Lucien Tesnière. 2015. *Elements of Structural Syntax*. John Benjamins Publishing Company, Amsterdam.

A Examples



BO147 [285] id:93 [eh]

BO145 [320] id:92 [mh]

BO139 [417] id:91 volete stare in bisca fino alle quattro [del ma]tt[ino stano]tte?

91, BO139	volete	stare	in	bisca	fino	alle	quattro	[del	ma]tt[ino	stano]tte?
	want:2PL	stay	in	casino	until	to.DEF	four	of.DEF	morning	tonight

‘do you want to stay up until four in the morning tonight?’

92, BO145	[mh]
	mh

‘mh’

93, BO147	[eh]
	eh

‘eh’

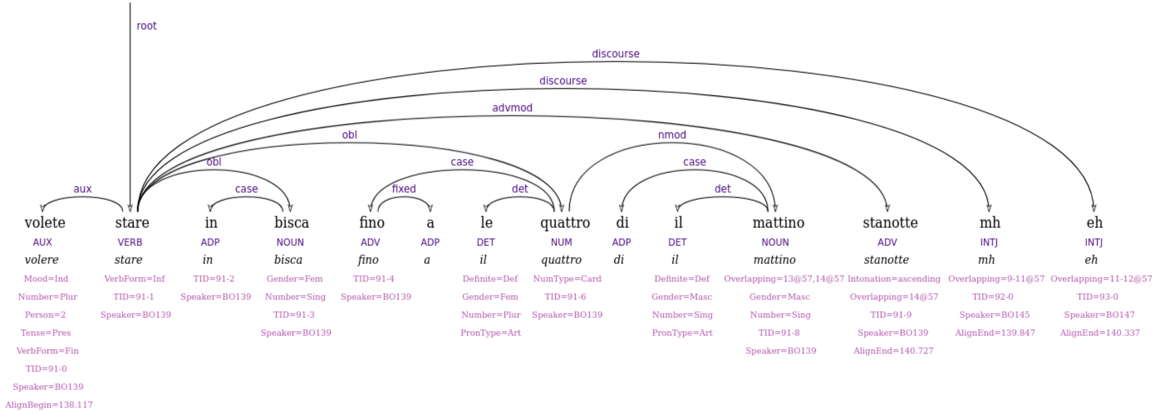
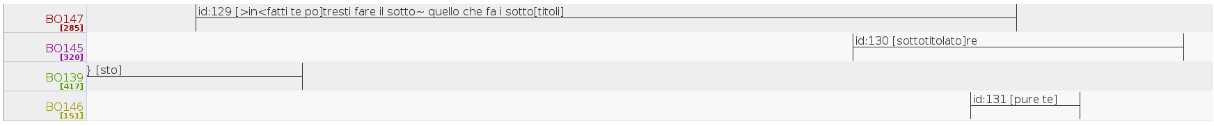

root
discourse
discourse
advmod
obl
nmod
obl
case
case
aux
case
fixed
det
det
volete AUX volere Mood=Ind Number=Plur Person=2 Tense=Pres VerbForm=Fin TID=91-0 Speaker=BO139 AlignBegin=138.117
stare VERB stare VerbForm=Inf TID=91-1 Speaker=BO139
in ADP in TID=91-2 Speaker=BO139
bisca NOUN bisca Gender=Fem Number=Sing TID=91-3 Speaker=BO139
fino ADV fino TID=91-4 Speaker=BO139
a ADP a
le DET il Definite=Def Gender=Fem Number=Plur PronType=Art
quattro NUM quattro NumType=Card TID=91-6 Speaker=BO139
di ADP di
il DET il Definite=Def Gender=Masc Number=Sing PronType=Art
mattino NOUN mattino Overlapping=13@57,14@57 Gender=Masc Number=Sing TID=91-8 Speaker=BO139
stanotte ADV stanotte Intonation=ascending Overlapping=14@57 TID=91-9 Speaker=BO139 AlignEnd=140.727
mh INTJ mh Overlapping=9-11@57 TID=92-0 Speaker=BO145 AlignEnd=139.847
eh INTJ eh Overlapping=11-12@57 TID=93-0 Speaker=BO147 AlignEnd=140.337

Figure 5: A case where a token (i.e., ‘mattino’, en. ‘morning’) is participating in two distinct overlapping spans. Its Overlapping feature is in fact composed by two distinct references, separated by a comma. (from conversation BOA3017)



BO147 [285] id:129 [>in<fatti te po]tresti fare il sotto∼ quello che fa i sotto[titoli]

BO145 [320] id:130 [sottotitolato]re

BO139 [417] } [sto]

BO146 [151] id:131 [pure te]

129, BO147

[>in<fatti	te	po]tresti	fare	il	sotto∼	quello	che	fa	i	sotto[titoli]
indeed	you.SG	could:2SG	do	DEF	sub∼	the.one	that	does	DEF	subtitles

‘indeed you could be a subtit∼ the person that makes subtitles’

130, BO145

[sottotitolato]re
subtitler

‘subtitler’

131, BO146

[pure	te]
also	you.SG

‘you too’

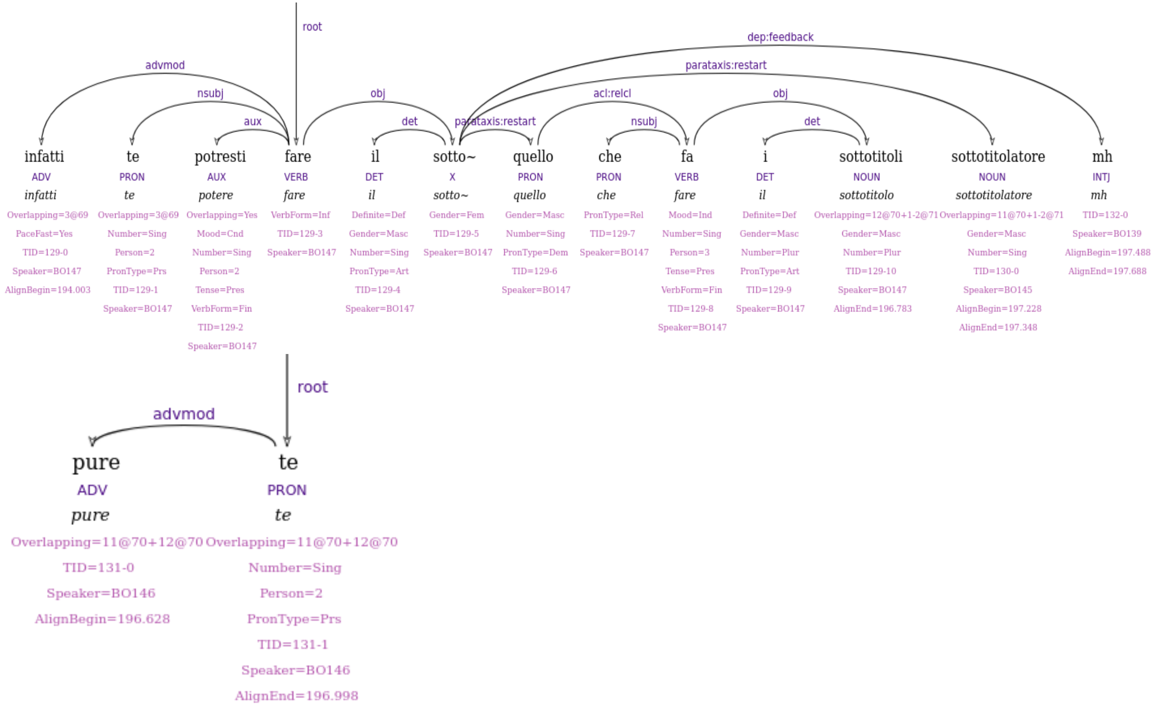

root
dep:feedback
advmod
parataxis:restart
nsubj
obj
acl:relcl
obj
aux
det
parataxis:restart
nsubj
det
infatti ADV infatti Overlapping=3@69 PaceFast=Yes TID=129-0 Speaker=BO147 AlignBegin=194.003
te PRON te Overlapping=3@69 Number=Sing Person=2 PronType=Prs TID=129-1 Speaker=BO147
potresti AUX potere Overlapping=Yes Mood=Cnd Number=Sing Person=2 Tense=Pres VerbForm=Fin TID=129-2 Speaker=BO147
fare VERB fare VerbForm=Inf TID=129-3 Speaker=BO147
il DET il Definite=Def Gender=Masc Number=Sing PronType=Art TID=129-4 Speaker=BO147
sotto~ X sotto~ Gender=Fem TID=129-5 Speaker=BO147
quello PRON quello Gender=Masc Number=Sing PronType=Dem TID=129-6 Speaker=BO147
che PRON che PronType=Rel TID=129-7 Speaker=BO147
fa VERB fare Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin TID=129-8 Speaker=BO147
i DET il Definite=Def Gender=Masc Number=Plur PronType=Art TID=129-9 Speaker=BO147
sottotitoli NOUN sottotitolo Overlapping=12@70+1-2@71 Gender=Masc Number=Plur TID=129-10 Speaker=BO147 AlignEnd=196.783
sottotitolatore NOUN sottotitolatore Overlapping=11@70+1-2@71 Gender=Masc Number=Sing TID=130-0 Speaker=BO145 AlignBegin=197.228 AlignEnd=197.348
mh INTJ mh TID=132-0 Speaker=BO139 AlignBegin=197.488 AlignEnd=197.688
root
advmod
pure ADV pure Overlapping=11@70+12@70 TID=131-0 Speaker=BO146 AlignBegin=196.628
te PRON te Overlapping=11@70+12@70 Number=Sing Person=2 PronType=Prs TID=131-1 Speaker=BO146 AlignEnd=196.998

Figure 6: A case of overlap among multiple (i.e., more than two) speakers. The token ‘sottotitoli’ (en. ‘subtitles’), with `TID=129-10`, overlaps with token ‘sottotitolatore’ (en. ‘subtitler’, `TID=130-0` in the first sentence) and ‘pure te’ (en. ‘you too’, in the second one). This is expressed by means of the + symbol in the `Overlap` feature. (from conversation BOA3017)

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 lasagne
‘lasagna’

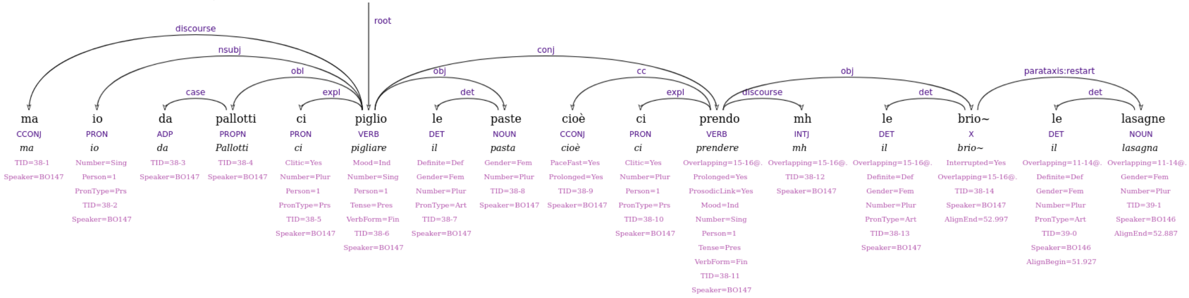

root
discourse
nsubj
obl
case
expl
conj
obj
det
cc
expl
discourse
obj
det
parataxis:restart
det
ma CCONJ ma TID=38-1 Speaker=BO147
io PRON io Number=Sing Person=1 PronType=Prs TID=38-2 Speaker=BO147
da ADP da TID=38-3 Speaker=BO147
pallotti PROPN Pallotti TID=38-4 Speaker=BO147
ci PRON ci Clitic=Yes 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

Figure 7: A case of co-construction, where the second speaker provides material (i.e., ‘le lasagne’) that is syntactically dependent on the verb uttered by the first speaker (i.e., ‘ci prendo’). (from conversation BOA3017)

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’

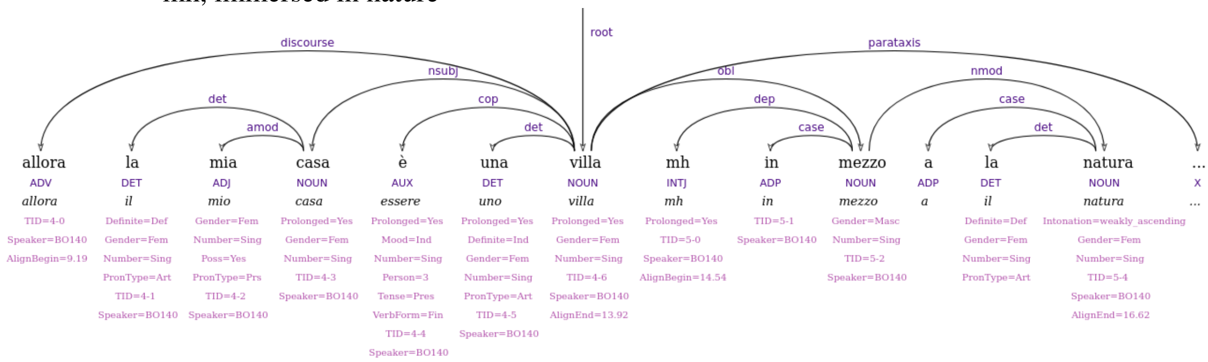

root
discourse
nsubj
det
amod
cop
det
parataxis
obl
nmod
dep
case
case
det
allora ADV allora TID=4-0 Speaker=BO140 AlignBegin=9.19
la DET il Definite=Def Gender=Fem Number=Sing PronType=Art TID=4-1 Speaker=BO140
mia ADJ mio Gender=Fem Number=Sing Poss=Yes PronType=Prs TID=4-2 Speaker=BO140
casa NOUN 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 ...

Figure 8: A case where, during transcription, a TU boundary was introduced breaking a intra-phrase relation: the nominal modifier ‘in mezzo alla natura’ would be separated from its head ‘villa’ if segmentation was performed based on TU boundaries. For space reasons, only the first part of the full unit is shown. (from conversation BOD2018)

342, BO139 ho scoperto chi è perché era
have.PRS.1SG found.out who is why was
'I found out who that was because they was'

343, BO139 un un
INDEF INDEF
'a a'

344, BO147 {tossisce}
coughs*
'{coughs}'

345, BO139 x un' in[tervista in cui x=ava]
UNK INDEF interview in which
'in an interview where'

346, BO145 [(ti) strozzi]
you.OBJ choke:2SG
'you're choking'

347, BO139 salute
bless.you
'bless you'

348, BO147 {tossisce}
coughs*
'{coughs}'

349, BO139 era tipo (un) ri[cevimento]
it.was like INDEF meeting
'it was like a meeting'

350, BO147 [trascrivi] {ride} {tossisce}
transcribe laughs$^-$ coughs*
'transcribe {laughter} {coughs}'

351, BO139 {ride}
laughs$^-$
'{laughter}'

352, BO145 cough cough
cough cough
'cough cough'

353, BO147 cou[gh cough] {ride}
cough cough laughs$^-$
'cough cough {laughter}'

354, BO139 [tossisce] {ride}
coughs laughs$^-$
'coughs'

355, BO139 chiusa parentesi
closed parenthesis
'parenthesis closed'

Figure 9: Curly brackets mark non-linguistic behavior in Jefferson notation. In the glosses, elements marked with $\star$ indicate NLBs that are annotated with feature Type=NLB in the treebank, while elements marked by $-$ have not been ported as tokens to the treebank. (from conversation BOA3017)

435, BO147	che	cosa	vuoi	da	me?	{cantando}	{ride}
	what	thing	want.2SG	from	me	singing	laughs

‘what do you want from me’

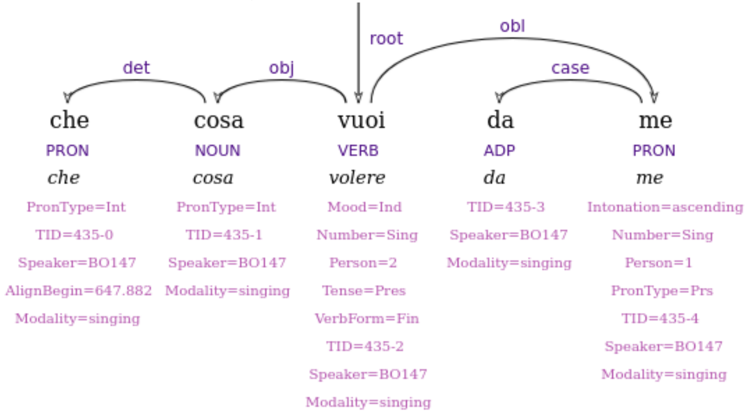


Figure 10: A case where a token originally present in the resource (i.e., {cantando}, en. ‘singing’, with `TID=435-5`) has been transformed into the `Modality=singing` feature on the tokens that were uttered while singing. Consequently, the token is not included in the treebank as a syntactic token. The example also shows a case where an NLB token (i.e., {ride}, en. ‘laughs’) is removed as not syntatically relevant. (from conversation BOA3017)

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

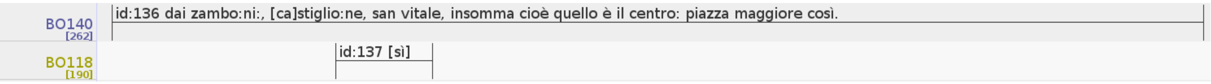
INDEF study on.tDEF rituals magic in.DEF sicily like of.DEF eight.hundr∼ take:2SG a thing

del genere

of.DEF genre

'a study on magic rites in the Sicily of ninet∼ century, for example, something like that'

Figure 11: The example shows how syntax develops incrementally and not necessarily planned in advance by the speaker. The syntactic cohesion of the discourse portion is also marked by the final 'una roba del genere' (en. 'something like that') that can be identified as the closing element of a listing. Figure 13a shows the parsed tree. (from conversation BOA3017)



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]

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

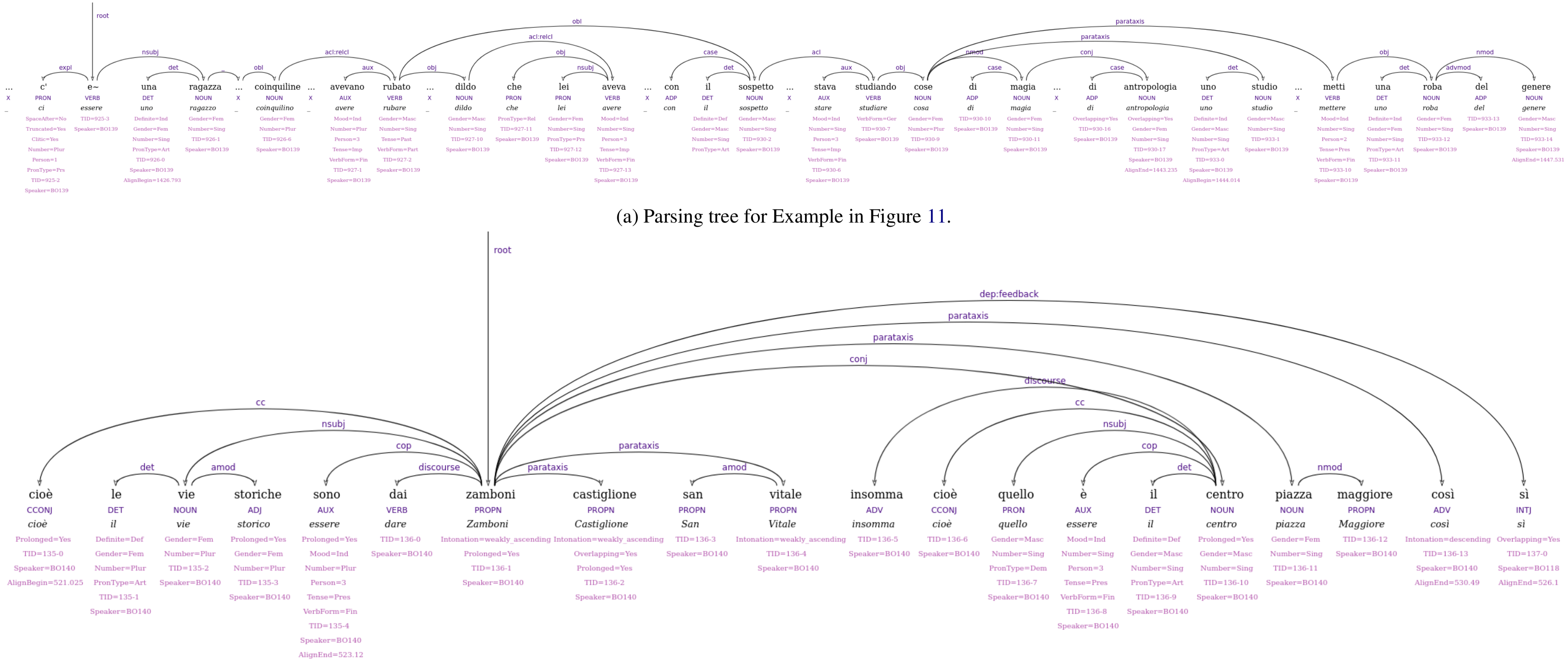
‘come on, Zamboni, Castiglione, San Vitale I mean. That is DEF city centre, Piazza Maggiore, like that’

137, BO118

sì
yes

‘yes’

Figure 12: A case where feedback from speaker BO118 (‘si’, en. ‘yes’) does not interrupt the syntactic construction of speaker BO140. Figure 13b shows the parsed tree. (from conversation BOD2018)


root
expl
nsubj
det
obl
acl:relcl
aux
obj
acl:relcl
obj
nsubj
obl
case
det
acl
aux
obj
nmod
case
parataxis
parataxis
conj
case
det
obj
det
advmod
nmod
... c' e~ una ragazza ... coinquiline ... avevano rubato ... dildo che lei aveva ... con il sospetto ... stava studiando cose di magia ... di antropologia uno studio ... metti una roba del genere
X PRON VERB DET NOUN X NOUN X AUX VERB X NOUN PRON PRON VERB X ADP DET NOUN X AUX VERB NOUN ADP NOUN X ADP NOUN DET NOUN X VERB DET NOUN ADP NOUN
_ ci essere uno ragazzo _ coinquilino _ avere rubare _ dildo che lei avere _ con il sospetto _ stare studiare cosa di magia _ di antropologia uno studio _ mettere uno roba del genere
root
dep:feedback
parataxis
parataxis
conj
cc
nsubj
cop
det
amod
discourse
parataxis
parataxis
amod
discourse
cc
nsubj
cop
det
nmod
cioè le vie storiche sono dai zamboni castiglione san vitale insomma cioè quello è il centro piazza maggiore così sì
CCONJ DET NOUN ADJ AUX VERB PROPN PROPN PROPN PROPN ADV CCONJ PRON AUX DET NOUN NOUN PROPN ADV INTJ
cioè il vie storico essere dare Zamboni Castiglione San Vitale insomma cioè quello essere il centro piazza Maggiore così sì

(a) Parsing tree for Example in Figure 11.

(b) Parsing tree for Example in Figure 12.

Figure 13