Coconstructions in spoken Universal Dependencies: guidelines and first results

Ludovica Pannitto[1], **Sylvain Kahane**[2], **Kaja Dobrovoljc**[3], **Elena Battaglia**[1], **Bruno Guillaume**[4], **Caterina Mauri**[1], **Eleonora Zucchini**[5]

[1] University of Bologna, Bologna - Italy;
[2] Université Paris Nanterre, Modyco, Paris - France;
[3] University of Ljubljana and Jozef Stefan Institute, Ljubljana - Slovenia;
[4] Université de Lorraine, CNRS, Inria, LORIA, Nancy - France;
[5] Masaryk University, Brno - Czech Republic

Relevant UniDive working groups: WG1

1 Introduction

Unlike written text, spoken language unfolds in time and is jointly produced by multiple participants, who continuously cooperate in the real-time production and reception of syntactic structure. Linearized transcriptions, however, necessarily flatten the temporal and multi-party dimension of spoken data. Universal Dependencies (UD, de Marneffe et al. 2021) inherits a broadly shared assumption in syntactic theory: that if syntax operates within the boundaries of a sentence. While this assumption is largely unproblematic for written language (but see Redaelli and Sprugnoli 2024; Read et al. 2012; Ait Azzi et al. 2019), it raises difficulties for spoken interaction, as syntactic dependencies may extend across speaker turns, interruptions, and overlapping contributions. Existing spoken UD treebanks do not adopt a uniform segmentation strategy, and this makes coconstructed spoken syntax difficult to annotate consistently. We introduce a novel proposal for annotation guidelines that can enable resource creators and users to account for syntactic dependencies holding among speech produced by different speakers.

2 Coconstructions in spoken syntax

No clear agreement emerges in the literature about what a *sentence* (as in, maximal unit of segmentation) in spoken language is (Sacks et al., 1974; Pietrandrea et al., 2014; Mettouchi and Vanhove, 2020) and, despite sentences being the main building blocks of UD resources, no explicit definition of what should count as a sentence is ever given in the current annotation guidelines. Our proposal distinguishes two complementary units of analysis: *speaker-dependent maximal units*, which preserve the organization of turns and anchor segmentation to speech production, and *structurally autonomous rectional units* (Kahane and Pietrandrea, 2012), which capture syntactic governance independently of turn boundaries. Currently, spoken treebanks in UD are organized in speaker-dependent units, and only Rhapsodie and KIParla Forest provide special features for cases of coconstructions. On this basis, we distinguish three broad configurations: *coconstruction proper* (Examples 1 and 2), *wh-question coconstructions* (Example 3), and *backchannelling* (Example 4). These configurations are distinct from ellipsis as in ellipsis the missing material is recoverable from a linguistic antecedent or from world knowledge, without needing to be produced by anyone; in coconstruction, the completing speaker actively supplies the dependent, performing syntactic work in real time.

(1) English (Ono and Thompson 1996, p. 72)
L: his position is **pretty uh**
A: . . . **stable**.

(2) French (Rhapsodie, D2001[1])
L2 “eh bien” je crois que je ne me suis pas
well I think I haven’t
conduit d’une façon conforme à ce qu’on
behaved as one would
attend “euh” { { **d’une jeune fille d’abord**
expect from a young girl first
| ^ et d’une femme ensuite } |} //+
and from a woman afterwards
L1 {| **d’une jeune bourgeoise** |} ?//+
from a young bourgeois girl?

[1] Curly brackets indicate conjuncts in a stacking construction. See Section 3.

L2 {| **“disons” d’une jeune bourgeoise** } //
let’s say from a young bourgeois girl

(3) Italian (KIParla, PTA007)
TOR001: **dove** vai ad arrampicare?
where do you climb?
TOI007: **al bi side**[2] vicino alla colletta
at the bi side close to colletta

(4) Italian (KIParla, BOA3017)
BO145: ma perché mamma c’ha dei
but because my mom has
pregiudizi nei miei confronti
prejudices against me
BO139: **mhmh**
mhmh
BO145: e poi daniela non devi avere
and then daniela you shouldn’t have
pregiudizi su di me
prejudice against me

3 Proposed guidelines

Building on Kahane et al. 2021 and to encode these phenomena in a UD-compatible way, we introduce the `MISC` features `Coconstruct` and `Backchannel`, to provide a lightweight mechanism to signal linkage between speaker-based maximal units involved in coconstruction phenomena. As a general principle, coconstruction is always marked on the element that comes second in time, both in the case of `Coconstruct` and `Backchannel` features: the feature is assigned to the responding token (i.e. the token that occurs later in the discourse and is syntactically governed by a prior contribution) and its value identifies the eliciting or projecting token in the earlier utterance (i.e. the syntactic or interactional anchor to which the responding token must be linked). This also enables multiple successive tokens to point to the same earlier anchor, reflecting shared dependency on the same projected structure. Formally, `Coconstruct` encodes a backward pointer of the form ⟨`deprel`⟩`::`⟨`sent_id`⟩`::`⟨`tok_id`⟩, while `Backchannel`, which uniquely identifies a dependency relation `discourse:backchannel`, encodes a backward pointer of the form ⟨`sent_id`⟩`::`⟨`tok_id`⟩. The `Coconstruct` feature is used in two main configurations, which differ in the syntactic status of the dependency slot targeted by the responding token: (i) **Completion** of an open dependency slot, when the responding token realizes a dependency that was projected but not realized in the preceding unit and (ii) **Stacking** on an already realized dependency slot, when the responding token provides an alternative realization of a dependency that is already present in the preceding unit.

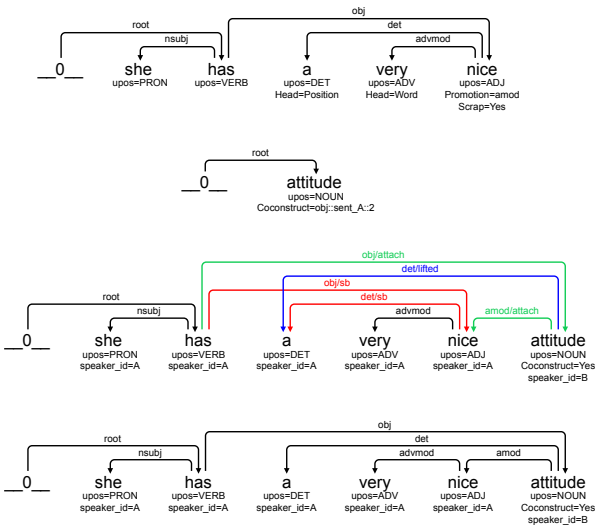

Figure 1: Example of automatic conversion on English data.

4 Automatic conversion between speaker-based and dependency-based

The speaker-based view, properly enriched with the features introduced in Section 3, can be converted into the dependency-based view, which relies on the definition of ‘sentence’ as ‘rectional unit’, enabling coconstructions to be realised as regular syntactic dependencies. In Figure 1, the first line is the speaker-based view and the last line is the dependency-based view. An intermediate richer structure is introduced in which: new dependency relations are identified with a suffix `/attach` and they are drawn in green. The richer structure helps to understand the conversion process and adds the possibility to query these modified dependencies specifically. There may be other dependency rela-

[1] L1’s contribution *d’une jeune bourgeoise* might appear to be a case of nominal ellipsis, but UD’s ellipsis mechanisms (promotion and `orphan`) presuppose that the elided head is recoverable within the same syntactic unit. Here the governing verb *attend* belongs to L2’s preceding turn and lies outside L1’s sentence in the treebank. The gap is not, in fact, empty: rather, it is filled in real time by another speaker.

[2] ‘bi side’ is the quasi-phonetic rendering of the name of a climbing gym, ‘Bside’.

tions needing to be restored in dependency-based view. The new features `Promotion` and `Head` are used to reconstruct the expected dependency structure.

These guidelines were partially applied to three UD v2.18 corpora: French-Rhapsodie (81 coconstructions, 315 backchannels), Italian-KIParlaForest (Pannitto et al., 2025) (123 coconstructions, 309 backchannels) and Slovenian-SST (42 coconstructions, 386 backchannels).

5 Conclusion

While UD has been first developed on the basis of written corpora, the collection now includes several treebanks of spoken data. Due to the lack of clear guidelines for spoken data, important discrepancies exist among treebanks for spoken data (Dobrovoljc, 2022). This work is part of a common effort to propose guidelines and unify the annotation of different treebanks. The first step in the syntactic analysis of spoken data is segmentation into minimal units (words or morphemes) and maximal units. The latter question is the main topic of this paper and the question of coconstructions is one of the very first problems to be solved. Concretely, the annotated treebanks are currently released in the speaker-based view, and a conversion script is provided to derive the dependency-based view on demand. We do not yet distribute parallel treebanks in both formats, but we hope this work will open a broader discussion on what the appropriate release format for spoken UD treebanks should be, in order to adequately account for the double representation inherent to coconstructed spoken syntax.

References

Abderrahim Ait Azzi, Houda Bouamor, and Sira Ferradans. 2019. The FinSBD-2019 Shared Task: Sentence boundary detection in PDF Noisy text in the Financial Domain. In *Proceedings of the first workshop on financial technology and natural language processing*, pages 74–80.

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

Kaja Dobrovoljc. 2022. Spoken Language Treebanks in Universal Dependencies: an Overview. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1798–1806, Marseille, France. European Language Resources Association.

Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes. 2021. 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 and Paola Pietrandrea. 2012. La typologie des entassements en français. *SHS Web of Conferences*, 1:1809–1828.

Amina Mettouchi and Martine Vanhove. 2020. Prosodic segmentation and grammatical analysis in cross-linguistic corpora. In *42nd Annual Conference of the German Linguistic Society (DGfS), Workshop Corpus-based typology: spoken language from a cross-linguistic perspective*.

Tsuyoshi Ono and Sandra A. Thompson. 1996. Interaction and Syntax in the Structure of Conversational Discourse: Collaboration, Overlap, and Syntactic Dissociation. In *Computational and Conversational Discourse*, pages 67–96, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ludovica Pannitto, Eleonora Zucchini, Silvia Ballarè, Cristina Bosco, Caterina Mauri, and Manuela Sanguinetti. 2025. Introducing KIParla forest: seeds for a UD annotation of interactional syntax. In *Proceedings of the Eighth International Conference on Dependency Linguistics (depling, Syntaxfest 2025)*, pages 54–73, Ljubljana, Slovenia. Association for Computational Linguistics.

Paola Pietrandrea, Sylvain Kahane, Anne Lacheret-Dujour, and Frédéric Sabio. 2014. The notion of sentence and other discourse units in corpus annotation. In Tommaso Raso and Heliana Mello, editors, *Spoken Corpora and Linguistic Studies*, pages 331–364.

Jonathon Read, Rebecca Dridan, Stephan Oepen, and Lars Jørgen Solberg. 2012. Sentence boundary detection: A long solved problem? In *Proceedings of COLING 2012: Posters*, pages 985–994, Mumbai, India. The COLING 2012 Organizing Committee.

Arianna Redaelli and Rachele Sprugnoli. 2024. Is Sentence Splitting a Solved Task? Experiments to the Intersection between NLP and Italian Linguistics. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 813–820, Pisa, Italy. CEUR Workshop Proceedings.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.