Spoken Language Treebanks in Universal Dependencies: an Overview

Kaja Dobrovoljc
Faculty of Arts, University of Ljubljana
Jozef Stefan Institute, Ljubljana, Slovenia

Introduction

- Syntactically annotated collections of transcribed speech are one of the fundamental language resources for spoken language research in NLP and linguistics alike.
- To enable cross-resource explorations of this limited and costly domain-specific data, there has been a growing number of spoken language treebanks adopting the Universal Dependencies annotation scheme (UD) aimed at cross-linguistically consistent treebank annotation. The scheme also proposes some basic categories and guidelines pertaining to speech-specific phenomena, such as disfluencies.
- To date, the scheme has been applied to nearly 200 treebanks in over 100 languages. Among the 26 UD treebanks containing some amount of spoken data, 12 treebanks consist of speech transcriptions only.
- To support cross-treebank data explorations on the one hand, and encourage further treebank harmonization on the other, this poster presents a comparative overview of the current treatment of speech-specific phenomena in the spoken language treebanks adopting the UD scheme.

Spoken Language Treebanks in UD

Treebank name	Release	Tokens	Sentences
Beja NSC	2.8	1,101	56
Cantonese HK	2.1	13,918	1,004
Chinese HK	2.1	9,874	1,004
Chukchi HSE	2.7	5,389	1,004
French ParisStories	2.9	29,438	1,755
French Rhapsodie	2.2	34,437	2,837
Frisian-Dutch Fame	2.8	3,729	400
Komi Zyrian IKDP	2.2	2,304	214
Naija NSC	2.2	140,729	9,242
Norwegian NynorskLIA	2.1	55,410	5,250
Slovenian SST	1.3	29,488	3,188
Turkish German SAGT	2.7	36,934	2,184

Table 1. Alphabetical list of spoken language treebanks in UD v2.9.

Comparison of Speech Transcriptions

Some examples of speech-specific data in CONLL-U:

Sentence-level comments
`# sound_url`
`# timestamp  # speaker`
`# speaker_id  # dialect`
`lang=  # phonetic_text`
`# text[phon]`

Token-level MISC
`Overlap=`
`AttachTo=, Rel=`
`AlignBegin=, AlignEnd=`
`lang=, Lang=, OrigLang=`
`word=`

Specific tokens
`? ! . , / // ?// &//`
`wor- wor~ wor—`
`eee, euh, ähm, e`
`/ #  ## [pause]`
`[laughter] […]`

	Beja NSC	Cantonese HK	Chinese HK	Chukchi HSE	French ParisStories	French Rhapsodie	Frisian-Dutch Fame	Komi Zyrian IKDP	Naija NSC	Nor. NynorskLIA	Slovenian SST	Tur.-German SAGT
Sound file ID	yes	no	no	yes	yes	no	no	no	yes	no	no	no
Text-sound alignment	yes	no	no	yes	no	no	no	no	yes	no	no	no
Speaker ID	no	no	no	no	yes	yes	yes	no	yes	yes	no	no
Language variety	no	no	no	no	no	no	yes	yes	no	yes	no	yes
Standard orthography	no	no	yes	yes	yes	yes	yes	no	no	yes	yes	yes
Capitalization	no	no	no	yes	no	no	no	yes	no	no	no	yes
Pronunciation	yes	no	no	yes	no	no	no	no	no	no	yes	no
Speaker overlap	no	no	no	no	no	yes	no	no	no	no	yes	no
Final punctuation	yes	yes	yes	yes	yes	yes	no	yes	yes	yes	no	yes
Other punctuation	yes	yes	yes	no	yes	yes	no	yes	yes	yes	no	yes
Incomplete words	no	no	no	yes	yes	yes	no	no	yes	yes	yes	yes
Fillers	no	no	no	no	yes	yes	yes	no	yes	yes	yes	yes
Silent pauses	yes	no	no	no	no	no	no	no	yes	yes	yes	no
Incidents	no	no	no	no	no	no	no	no	no	no	yes	no

Table 2. Overview of transcription principles in spoken language UD treebanks.

The treebanks vary considerably with respect to what aspects of speech are transcribed and in what way. Our results suggest that future consolidation could be achieved through:

- Adding **rich metadata** if available by following existing solutions
- Faithful transcriptions of **all speaker-uttered phenomena** (but not other sounds)
- Transcriptions in **lowercase spelling** and **standard orthography**
- Inclusion of both sentence-medial and sentence-final **punctuation** by using **written-like symbols**
- Moving the treebank-specific markup to **MISC/comments** sections in CONLL-U
- Detailed documentation of **sentence segmentation principles**

Comparison of UD Annotations

Some examples of divergent annotations:

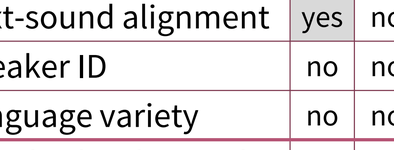
Punctuation marks
punct vs. *dep*

Filler words
INTJ vs. X
discourse vs. *discourse:filler*

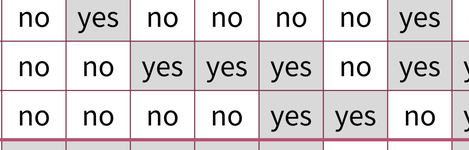
Clausal discourse markers
discourse vs.
parataxis:discourse

Repaired words
reparandum vs. *discourse:filler*

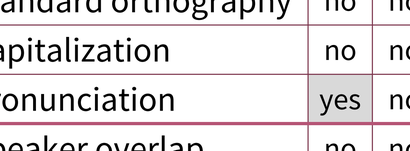
Repaired clauses
reparandum vs.
parataxis:restart vs.
parataxis:deletion



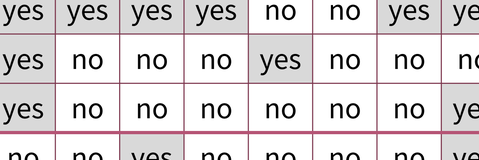
(Yes, it was, er, cumbersome.)



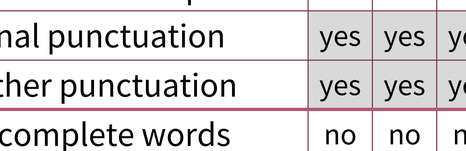
(And, er, I do have Frisian roots, my, er ...)



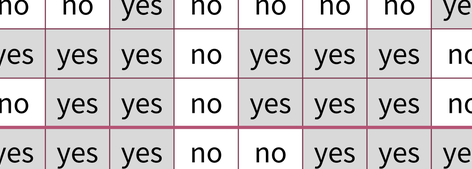
(True, they always ordered pizza.)



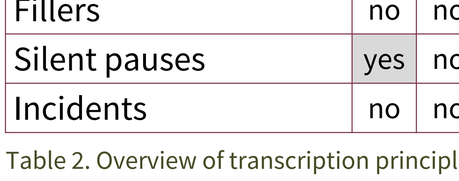
(Well, I hurt my ankle, you see.)



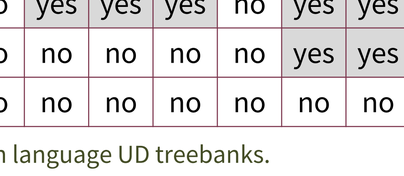
(He was warmed by the wolves.)



(The image shows how they ... This man is sitting.)



(They drove, well, was the truck a lot and ...)



(The girl didn't want to be ... marry him.)

The treebanks vary with respect to UD annotation of speech-specific phenomena. Our results suggest that future consolidation could be achieved through:

- Adhering to the **general annotation guidelines** for phenomena which are not unique to speech alone (e.g. *parataxis*).
- Following the prevailing solutions for **closed class phenomena**, such as punctuation marks and discourse fillers, with preference for core labels over extensions.
- Reconsidering the distinction between non-clausal and **clausal discourse markers**.
- Ensuring **head-attachment consistency** for loose-joining syntactic phenomena.
- Adherence to the general guidelines on the **right-to-left *reparandum*** attachment.
- Further development of the **UD guidelines on the use of *reparandum* label** and the treatment of speech-repairs in general.

Jožef Stefan Institute
University of Ljubljana
arrs JAVNA AGENCIJA ZA RAZISKOVALNO DEJAVNOST REPUBLIKE SLOVENIJE