Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1798–1806
Marseille, 20-25 June 2022

Spoken Language Treebanks in Universal Dependencies: an Overview

Kaja Dobrovoljc
Faculty of Arts, University of Ljubljana
Jožef Stefan Institute, Ljubljana, Slovenia
kaja.dobrovoljc@ff.uni-lj.si

Abstract

Given the benefits of syntactically annotated collections of transcribed speech in spoken language research and applications, many spoken language treebanks have been developed in the last decades, with divergent annotation schemes posing important limitations to cross-resource explorations, such as comparing data across languages, grammatical frameworks, and language domains. As a consequence, there has been a growing number of spoken language treebanks adopting the Universal Dependencies (UD) annotation scheme, aimed at cross-linguistically consistent morphosyntactic annotation. In view of the non-central role of spoken language data within the scheme and with little in-domain consolidation to date, this paper presents a comparative overview of spoken language treebanks in UD to support cross-treebank data explorations on the one hand, and encourage further treebank harmonization on the other. Our results show that the spoken language treebanks differ considerably with respect to the inventory and the format of transcribed phenomena, as well as the principles adopted in their morphosyntactic annotation. This is particularly true for the dependency annotation of speech disfluencies, where conflicting data annotations suggest an underspecification of the guidelines pertaining to speech repairs in general and the *reparandum* dependency relation in particular.

Keywords: Universal Dependencies, treebanks, speech transcriptions, morphosyntactic annotation, dependency syntax, disfluencies, speech repairs

1. Introduction

Spoken language treebanks, i.e. syntactically annotated collections of transcribed speech, represent one of the fundamental language resources for spoken language processing tasks, such as syntactic parsing (Caines et al., 2017; Dobrovoljc and Martinc, 2018; Braggaar and van der Goot, 2021; Liu and Prud'hommeaux, 2021) and information retrieval (Davidson et al., 2019; Liu et al., 2021). Spoken language treebanks are equally important in linguistic research in general, not only due to speech being the primary and prevailing mode of human communication that exhibits several idiosyncrasies in comparison to writing (Biber et al., 2010; Carter and McCarthy, 2015), but also due to the fact that many of the world languages have no written form at all.

Many spoken language treebanks have been created since the pivotal work on the Switchboard section of the Penn Treebank (Marcus et al., 1993), such as the multilingual Tübingen (Hinrichs et al., 2000) and CHILDES (Sagae et al., 2004; MacWhinney, 2014) treebank collections, not to mention the number of treebanks developed for individual languages. For most, customized treebank-specific syntactic annotation schemes were developed, adopting divergent approaches to annotating syntactic phenomena in general and speech-specific phenomena in particular, posing important limitations to various kinds of cross-resource explorations, such as data comparisons across languages, grammatical frameworks or language varieties.

In line with the growing need for spoken data standardization and consolidation, there has also been an increasing number of spoken language treebanks adopting the Universal Dependencies annotation scheme aimed at cross-linguistically consistent treebank annotation for many human languages (de Marneffe et al., 2021). In essence, the UD scheme provides a universal inventory of grammatical categories (parts of speech, morphological features and syntactic dependencies) and guidelines for their application, which also include some broad recommendations pertaining to speech-specific phenomena, such as various kinds of disfluencies.

While recently Kahane et al. (2021a) proposed more detailed recommendations on the treatment of speech-related phenomena based on their experience in developing the Beja, Naija and French UD treebanks, there has been no systematic and exhaustive analysis of the current state of spoken language treebanks in UD to establish the differences and similarities between them. Such a review is not only a prerequisite for an adequate interpretation of empirical results arising from cross-treebank explorations, but also represents an essential first step for further harmonization work on this limited and costly domain-specific data.

To bridge this gap, this paper gives a comparative overview of the current treatment of speech-specific phenomena in spoken language treebanks adopting the Universal Dependencies annotation scheme, based on evidence from data and treebank-related documentation. After a short presentation of the treebanks under

Name	Code	Release	Source	Tokens	Sents	Avg. length
Beja NSC	bej_nsc	v2.8	converted	1,101	56	19.7
Cantonese HK	yue_hk	v2.1	native	13,918	1,004	13.9
Chinese HK	zh_hk	v2.1	native	9,874	1,004	9.8
Chukchi HSE	ckt_hse	v2.7	native	5,389	1,004	5.4
French ParisStories	fr_parisstories	v2.9	converted	29,438	1,755	16.8
French Rhapsodie	fr_rhapsodie	v2.2	converted	34,437	2,837	16.8
Frisian-Dutch Fame	qfn_fame	v2.8	native	3,729	400	9.3
Komi-Zyrian IKDP	kpv_ikdp	v2.2	native	2,304	214	10.8
Naija NSC	pcm_nsc	v2.2	converted	140,729	9,242	15.2
Norwegian NynorskLIA	no_nynorsklia	v2.1	converted	55,410	5,250	10.6
Slovenian SST	sl_sst	v1.3	native	29,488	3,188	9.2
Turkish-German SAGT	qtd_sagt	v2.7	native	36,934	2,184	16.9

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

investigation in Section 2, we discuss the differences and similarities in the treatment of specific speech-related phenomena, both with respect to spoken language transcriptions (Section 3) and UD morphosyntactic annotations (Section 4), and conclude by some preliminary recommendations on the possible points of convergence in the future (Section 5).

2. Spoken Language Treebanks in UD

To date, the UD annotation scheme has been applied to nearly 200 treebanks in over 100 languages. Among 26 UD treebanks containing some amount of spoken data as of UD release v2.9 (Zeman and others, 2021), 12 treebanks consist of spoken language transcriptions only.[1] Listed chronologically by first UD release, these include the Slovenian SST treebank (Dobrovoljc and Nivre, 2016), Norwegian NynorskLIA (Øvrelid et al., 2018), Chinese HK (Leung et al., 2016), Cantonese HK (Wong et al., 2017), Komi-Zyrian IKDP (Partanen et al., 2018), Naija NSC (Caron et al., 2019), French Rhapsodie and French ParisStories (Kahane et al., 2021a), Chukchi HSE (Tyers and Mishchenkova, 2020), the code switching Turkish-German SAGT (Çetinoğlu and Çagri Çöltekin, 2019) and Frisian-Dutch Fame (Braggaar and van der Goot, 2021) treebanks, as well as the recently added Beja NSC treebank (Kahane et al., 2021b). For low-resource languages, such as Beja, Cantonese, Chukchi, Frisian and Naija, these are also the only UD treebanks available.

As can be seen in Table 1, UD treebanks for spoken language vary in size,[2] with the majority being much smaller than the average (text-based) UD treebank, which is expected given the costly nature of their creation. In terms of dependency relation annotation, all spoken language treebanks were annotated manually, either in UD (native annotation) or an alternative annotation scheme, such as SUD (Gerdes et al., 2018), from which the Beja NSC, Naija NSC and both French spoken language treebanks have been converted.

3. Comparison of Speech Transcriptions

Given that the representation of speech in written form depends on a multitude of factors, there is no standardized convention on which aspects of spoken communication should be transcribed and in what way (Dittmar, 2012). This is also evident when comparing spoken language treebanks in UD, as the inventory of transcribed phenomena and their formal representation in the standardized CONLL-U format varies considerably:[3] from extensive coverage of all audible phenomena with minimum additional interventions, to various kinds of transcription editing to make speech look more like writing.

[1] The 14 UD v2.9 treebanks with mixed written and spoken data include Danish DDT, English LinES, English GUM, Greek GDT, Khunsari AHA, Latvian LVTB, Nayini AHA, Persian Seraji, Polish LFG, Scottish Gaelich ARCOSG, Skolt Sami Giellagas, Soi AHA, South Levantine Arabic MADAR and Swedish LinES. However, due to limited documentation on the integration and annotation of speech-specific phenomena in these treebanks, we limit our analysis on spoken language treebanks only. In addition to the treebanks distributed within the official UD data release, English treebanks with modified versions of the UD scheme have also been developed in related work on spoken language parsing (Liu and Prud'hommeaux, 2021; Davidson et al., 2019).

[2] Token and sentence counts in Table 1 follow the official UD 2.9 statistics and nomenclature, according to which tokens (very roughly) correspond to orthographic tokens including punctuation. For treebanks that deviate from the general UD tokenization principles, such as the Beja NSC morph-based treebank (Kahane et al., 2021b), specific calculations should be made.

[3] In CONLL-U, UD annotations are encoded as tab-separated text files with predetermined columns for specific annotation levels (`https://universaldependencies.org/format.html`). However, the format allows treebank creators to add unrestricted additional annotations both on sentence level (as part of the comment lines starting with #) and token level (as part of the final MISC column).

	Beja NSC	Cantonese HK	Chinese HK	Chukchi HSE	French ParisStories	French Rhapsodie	Frisian-Dutch Fame	Komi-Zyrian IKDP	Naija NSC	Norwegian NynorskLIA	Slovenian SST	Turksih 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 characteristics in spoken UD treebanks. The *no* mark denotes both 'absent' and 'not applicable'.

We give a summarized overview of our findings in Table 2 and discuss the differences and similarities in specific aspects of speech transcription below. While some pertain to spoken data representation only and can thus be considered less relevant for direct grammatical exploration, others have implications for subsequent morphosyntactic analysis (discussed in Section 4) as well.

3.1. Speech-Specific Metadata

In addition to the mandatory CONLL-U sentence-level information, such as the unique sentence identifier and the plain surface text, some treebanks include additional speech-related metadata. Information on the location of the original **soundfile** is provided in the Beja NSC, French ParisStories and Naija NSC treebanks (as `# sound_url`), with Beja and Naija treebanks also including information on **text-sound alignment** in the form of `AlignBegin` and `AlignEnd` markers in the MISC column. In contrast, the Chukchi HSE treebank marks the sentence offset in the sound file as part of the `# timestamp` comment line.

Speaker information is included in the French, Frisian-Dutch (all as `# speaker`), Naija (as `# speaker_id`) and Norwegian (as part of the `# dialect` comment line) treebanks. In addition to the dialects in the Norwegian NynorskLIA treebank, token-level (MISC) information on **language variety** is also included in the code-switching Frisian-Dutch Fame and Turkish-German SAGT treebanks (`Lang`), as well as the Komi-Zyrian IKDP treebank (`OrigLang`) in which Russian words are marked.

3.2. Orthography

With the exception of treebanks for languages with no codified (Beja, Naija) or no standardized (Cantonese) orthography and the Komi-Zyrian IKDP treebank, which follows speech-specific orthographic conventions, most treebanks use standard (written-like) **orthography** and spelling. However, with languages using alphabetic writing systems, there is no uniform approach to **sentence-initial capitalization**: while the Chukchi HSE, Komi-Zyrian IKDP and Turkish-German SAGT use written-like capitalization, the French, Frisian-Dutch Fame, Naija NSC, Norwegian NynorskLIA and Slovenian SST treebanks opt for lowercase transcription of all tokens except for proper names (see, for example, a Frisian-Dutch example in Figure 1).

Most treebanks omit any kind of **pronunciation-based transcription**, with the exception of phonetic transcriptions in the Beja NSC treebank (`# phonetic_text`), phonemic transcriptions in the Chukchi HSE (`# text[phon]`) treebank, and the pronunciation-based word spelling in the Slovenian SST treebank (`word` attribute in the MISC column).

3.3. Segmentation

In contrast to written text, where capitalization, punctuation and white spaces act as relatively reliable indicators of sentence boundaries, **segmenting** speech into sentence-like units (utterances) presents a much more

challenging task given the complex interaction of syntactic, semantic and prosodic features (Degand and Simon, 2009). Given the documentation of most spoken language treebanks in UD lacks information on the formal criteria adopted in sentence segmentation, it is difficult to make any direct comparisons. However, the average length of sentences with respect to the number of tokens given in Table 1 suggests that potentially different approaches were adopted by individual treebanks, especially if we compare treebanks with similar tokenization, punctuation and genre distribution.

A related phenomena of **overlapped speech** is explicitly addressed only in the French Rhapsodie treebank (by using the `Overlap` mark in the MISC column and proposing the `AttachTo` and `Rel` features to mark co-constructed trees) and in the Slovenian SST treebank, which uses the [...] token to mark the point where the overlapping speech begins.

3.4. Punctuation

With the exception of the Frisian-Dutch Fame treebank, which does not use any **punctuation** at all, and the Slovenian SST treebank, which only uses marks for non-declarative sentence intonation, all treebanks include some sort of punctuation symbols, but not in a uniform way. When it comes to **sentence-final punctuation**, most treebanks use standard punctuation symbols, such as full stops, exclamation (!) and question (?) marks, with the Beja NSC and Naija NSC treebanks employing a more detailed markup for different types of segment-final breaks (e.g. /, //, ?//, &//).

There are even more differing solutions when it comes to **sentence-medial punctuation**, from treebanks employing a wide range of written-like punctuation, including commas, colons, quotations marks and similar (e.g. Cantonese HK, Chinese HK, Komi-Zyrian IKDP, Turkish-German SAGT) to treebanks with no punctuation inside sentences at all (e.g. Chukchi HSE, Frisian-Dutch Fame, Slovenian SST). Beja NSC, Naija NSC and Norwegian NynorskLIA treebanks also use special punctuation characters, such as /, # or ##, to mark sentence-medial **speech breaks** (silent pauses).

3.5. Non-Lexical Tokens

While Cantonese HK, Chinese HK and Komi-Zyrian IKDP spoken language treebanks only include word and punctuation tokens (i.e. written-like phenomena), most other treebanks include additional tokens marking other types of uttered phenomena, such as disfluencies in the form of cut-off, **incomplete words**, usually marked by a special token-final character, such as ~ (French Rhapsodie and ParisStories, Naija NSC), - (Slovenian SST, Norwegian NynorskLIA) or -- (Turkish-German SAGT), and **fillers** (filled pauses), such as *euh* in French, *e* in Norwegian or *ähm* in Turkish-German. Slovenian SST treebank also features tokens denoting other types of audible **incidents**, such as laughter and applause, although these seem less relevant for subsequent morphosyntactic analysis.

4. Comparison of UD Annotations

To facilitate consistent annotation of similar constructions across languages, Universal Dependencies annotation scheme provides a universal inventory of 17 part-of-speech categories, 24 morphological features and 37 dependency relations, while allowing language-specific extensions when necessary. This flexible broad-coverage design makes the scheme easily applicable to a wide range of language domains, including speech, especially given that the universal inventory already includes relations pertaining to the speech-frequent clause-peripheral relations, such as *vocative*, *parataxis*, *discourse* or *reparandum*.

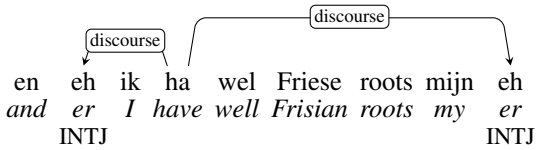
Nevertheless, the official UD guidelines on how to treat language-specific phenomena have been relatively scarce, leading to various interpretations when applying the scheme to actual data. This is also confirmed by our comparison of the differences and similarities in UD annotations for selected speech-related phenomena, presented below. Specifically, we focus on the comparison of dependency labels and part-of-speech tags, with other dimensions of comparison, such as a detailed analysis of head-attachment principles and related (non-)projectivity, left for future work.

4.1. Punctuation-like tokens

Annotation of punctuation-like tokens, such as punctuation marks and markers of prosodic breaks (Section 3.4), is quite consistent across treebanks (tokens tagged as *PUNCT* and labeled as *punct*, as illustrated in Figures 2 and 3), with the exception of the Naija NSC treebank that treats punctuation-like tokens as *dep* unspecified dependencies (Figure 9).

4.2. Fillers

In treebanks that include filler words (Section 3.5), these are mostly labeled as *discourse* (Figure 1), with the exception of the Norwegian NynorskLIA and Slovenian SST treebanks that use a special *discourse:filler* extension to distinguish them from other types of discourse particles (Figure 2). The Norwegian NynorskLIA treebank is also the only treebank to tag filler words as *X* (other) rather than interjections (*INTJ*).

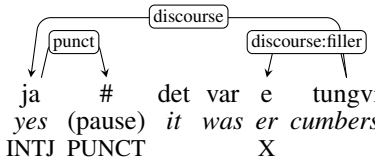


(And, er, I do have Frisian roots, my, er ...)

Figure 1: Example annotation of filler words in the Frisian-Dutch Fame treebank.

4.3. Discourse Particles

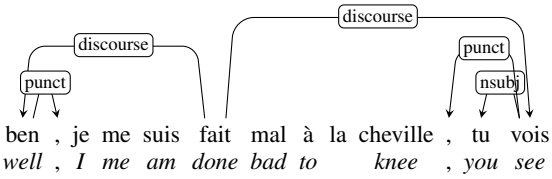
Most treebanks comply with the general UD guidelines on labeling interjections and other discourse particles



(Yes, it was, er, cumbersome.)

Figure 2: Example annotation of filler words and speech breaks in the Norwegian NynorskLIA treebank.

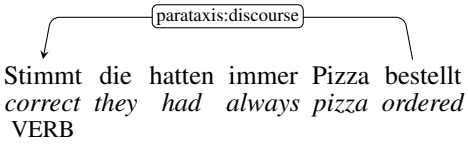
as *discourse*. However, given that words occurring as discourse particles (e.g. English *so*, *well*, *like*) take on a variety of different syntactic positions and functions, a much more detailed analysis would be required to establish the cross-treebank consistency in their potential delimitation with competing interpretations, such as as conjunctions (*cc*) or adverbial modifiers (*advmod*).



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

Figure 3: Example annotation of verbal and non-verbal discourse markers in the French ParisStories treebank.

For discourse particles deriving from clauses, in particular, such as *you know* in English, annotation principles differ–while most treebanks do not make any category-based distinctions for this type of multi-word expressions (see Figure 3 for French ParisStories), the Naija NSC, Slovenian SST and Turkish-German SAGT treebanks introduce a dedicated *parataxis:discourse* extension for distinct annotation of clausal discourse markers (Figure 4).



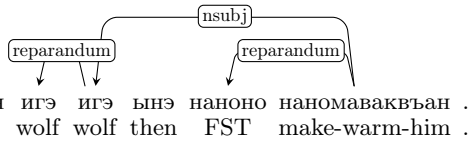
(True, they always ordered pizza.)

Figure 4: Example annotation of verbal discourse markers in the Turkish-German SAGT treebank.

4.4. Speech Repairs

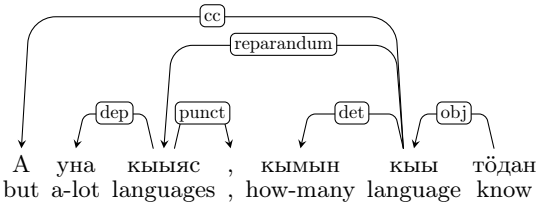
For annotation of disfluencies, in which the speaker corrects part of the intended verbalization (speech repairs),[4] the UD scheme introduces the *reparandum* dependency relation with the repaired unit (reparandum) being dependent on its repair. This relation occurs in all spoken UD treebanks, but in various ways.

Adherence to general UD guidelines on speech repairs can be observed in the Chinese HK, Frisian-Dutch Fame, Turkish-German SAGT and Chukchi HSE treebanks, which use the *reparandum* relation to label repairs of single words, word fragments (Figure 5) and longer sequences of words, with the Cantonese HK and Komi-Zyrian IKDP treebanks also marking the interruption point explicitly (see comma punctuation in Figure 6).



(He was warmed by the wolves.)

Figure 5: Example annotation of repaired words and word fragments in the Chukchi HSE treebank (FST = false start).



(But, a lot of languages ... how many languages do you know?)

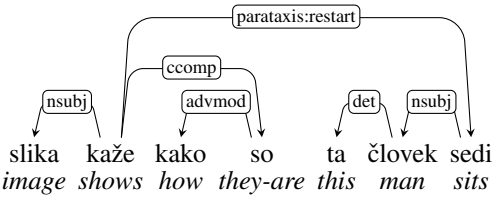
Figure 6: Example annotation of repaired phrases in the Komi-Zyrian IKDP treebank.

However, given that speech repairs often involve complex syntactic units, such as abandoned phrases or clauses with possibly shared dependants between the reparandum and repair, some treebanks introduce additional annotation principles related to self-repairing disfluencies.

For example, the Slovenian SST treebank introduces a category-based distinction, according to which repaired units involving predicates are considered as heads of restarted paratactical sentences (see the *parataxis:restart* label in Figure 7) rather than reparandums.

On the other hand, the Norwegian NynorskLIA treebank keeps the right-to-left attachment for all types of repairs, including clauses, however, it introduces a semantic-based distinction between repairs which are clearly related to the preceding reparandum on the one hand, and repairs that bear no relation to preceding context. For the latter, the *parataxis:deletion* label is used

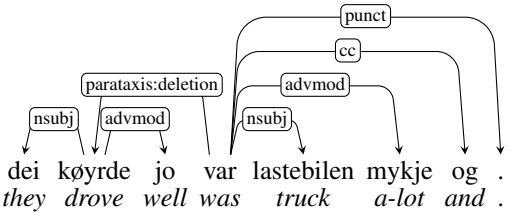
[4]We use *speech repair* as a cover term for various related concepts with sometimes overlapping, theory-dependent definitions and delimitations, such as repetition, substitution, insertion, deletion, speech error, reformulation, restart, self-repair, etc. See, for example, the typology proposed by Shriberg (1996).



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

Figure 7: Example annotation of abandoned clauses in the Slovenian SST treebank.

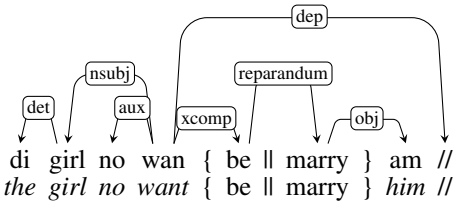
(Figure 8). A similar distinction is made for incomplete Norwegian words, labeled as either *reparandum* if replaced with the same word or *discourse:filler* if replaced with a new word.



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

Figure 8: Example annotation of deleted clauses in the Norwegian NynorskLIA treebank.

Finally, the SUD-derived treebanks for Beja, French and Naija adopt an opposite approach, in which the repair depends on the reparandum (left-to-right attachment), as illustrated in Figure 9. This solution contradicts the current UD guidelines on the treatment of repairs, and also makes the repaired trees more difficult to filter out.[5]



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

Figure 9: Example annotation of speech repairs in the Naija NSC treebank.

Although a much more detailed analysis could be performed on the treatment of speech repairs in individual treebanks, the brief comparison above suggests the need for more detailed guidelines pertaining to speech repairs in general and the *reparandum* relation in particular, especially when repaired or abandoned units involve incomplete syntactic trees. On the other hand, reformulations involving full words and phrases, open an equally challenging question of their delimitation on the spectrum from coordination to actual speech repairs (Kahane et al., 2021a).

[5]As with other types of disfluencies, such as discourse particles and fillers, there is a general preference for considering such constructions as terminal nodes in the tree, which also allows their straightforward removal, if necessary.

4.5. Other Constructions

In addition to the most prominent speech-related syntactic phenomena discussed above, some additional types of frequent constructions in speech have been discussed in documentation related to individual spoken language UD treebanks (see Section 2 for references), such as parenthetical clauses, asyndetic coordination, reported speech, ellipsis, general extenders, dislocated elements, and atypical word order. However, despite these phenomena being much more frequent in speech than in writing, they are not unique to spoken language alone. Most spoken language treebanks thus follow the general UD annotation guidelines on such constructions, without the need for new domain-specific solutions. The use of some extensions inspired by work on spoken language data, such as *parataxis:parenth* (Kahane et al., 2021a), *parataxis:dislocated* (Caron et al., 2019) and *conj:extend* (Dobrovoljc and Nivre, 2016), thus remains limited to individual treebanks only.

5. Recommendations

A comparative overview of spoken language UD treebanks presented above shows the treebanks differ considerably with respect to the inventory and the format of transcribed phenomena (Section 3), and the principles adopted in their morphosyntactic annotation (Section 4). Although this is partially understandable in view of the diverse original resources the treebanks derive from, there is also room for further homogenization, especially given the fact that most speech-related phenomena are not language-specific, but universal to human communication in general.

Given the current heterogeneity of spoken language treebanks in UD, further community-driven discussions should be encouraged before any definitive solutions are proposed. Nevertheless, some preliminary recommendations for future treebank consolidation could also be given based on the observations reported in this analysis.

5.1. Consolidating Speech Transcriptions

To enable and support various kinds of potential explorations of spoken language data in UD, **adding rich metadata** information to the treebanks should be encouraged, especially in view of the fact that a wide variety of metadata is typically already available in the original spoken corpora from which the treebanks derive.

Kahane et al. (2021a) propose a good starting point for encoding information on the speakers, the sound file

and its alignment with the text (including speaker overlap). For other types of typical speech-related metadata, such as information on the sentence- and token-level language variety, or the different types of transcriptions available, we recommend following the **existing solutions** in the individual treebanks that use them (Section 3.1) until standardized encoding is proposed.

In terms of the textual representation of speech, existing spoken language treebanks in UD show a preference for faithful transcription of **all speaker-uttered phenomena** (including fillers, incomplete words and false starts in general), which are mostly transcribed in **lowercase spelling** and follow the **standard orthography** where applicable.

In line with the UD approach to segmentation in general, the theory-dependent segmentation of speech should remain at the discretion of treebank developers, however, a more detailed **sentence segmentation documentation** should be encouraged. On the other hand, there is room for treebank consolidation with respect to the treatment of punctuation, with most treebanks under investigation **including punctuation** both in sentence-medial and sentence-final positions.

However, in contrast to Kahane et al. (2021a) who propose treebank-specific symbols for punctuation and boundary marking, our analysis shows there is a general preference for the common, **written-like punctuation symbols**, such as commas, full stops, question and exclamation marks, with the exception of speech break markers, for which various solutions have been proposed (Section 3.4). Naturally, treebank- and theory-specific markup can always be preserved as part of the MISC CONLL-U column.

5.2. Consolidating UD Annotations

In terms of annotations, the treebanks already display a relatively high degree of convergence in labeling speech-specific closed-class phenomena, such as punctuation marks (***PUNCT + punct***) and filled pauses (***INTJ + discourse***), with preference for core labels over extensions. For discourse particles in particular, our analysis confirms the previously suggested unnecessary distinction between clausal and other discourse markers (Kahane et al., 2021a), as most treebanks opt to label both as ***discourse***.

Nevertheless, further analysis is needed to establish the level of **head-attachment consistency** both within and across treebanks for these particular relations, given their loose-joining syntactic nature on the one hand and their ubiquitous sentence distribution on the other. In general, the *punct* and *discourse* attachment should follow the general UD principles on attaching these nodes to the highest possible node **preserving projectivity** and preventing the nodes to have **any dependents**.

Perhaps most strikingly, our results show the need for further discussions on the **under-specified *reparandum* relation** and the dependency-based formalization of speech repairs in general, as existing UD treebanks adopt inconsistent and often conflicting annotations of this frequent phenomenon in speech. Contrary to Kahane et al. (2021a), we argue for the original **right-to-left *reparandum* attachment** principle, which has also been adopted by most of the treebanks under investigation, according to which the repaired unit is considered the dependent of the repair that follows it.

Nonetheless, there is less agreement on what types of false starts should actually be considered as *reparandum* on the spectrum from incomplete word fragments to incomplete sentences. Rather than proposing a definitive solution, we therefore encourage further theory- and data-driven discussions on the issue with the goal of proposing systematic and exhaustive UD guidelines pertaining to speech repairs and disfluencies in general.

For the prominent spoken language phenomena which are not unique to speech alone, however, our analysis shows that **adhering to the universal annotation guidelines** should be encouraged before any domain-specific solutions are proposed.

6. Conclusion

This paper presented a comparative overview of spoken language treebanks adopting the Universal Dependencies annotation scheme, which shows the treebanks differ with respect to the principles adopted in spoken language transcription and morphosyntactic annotation. Although the established differences do not constrain future exploitation of this data, they should be taken into account when cross-treebank analyses are performed.

In parallel, our findings also highlight the need for further cross-treebank harmonization. This seems an easily achievable goal for transcription and annotation principles on which most treebank developers agree (but with differing formal implementations), such as those pertaining to metadata formalization, speech transcription and closed-class phenomena annotation.

On the other hand, the more compelling discrepancies, such as the inconsistent dependency annotation of speech repairs, showcase the need for further development of UD annotation guidelines in general, where important insights on speech-related phenomena could also be gained from communities with longer tradition in spoken language annotation (MacWhinney, 2014; Marcus et al., 1993).

We hope the overview presented in this paper presents a helpful reference point for such endeavours, and facilitates research on existing and emerging spoken language treebanks adopting the UD annotation scheme.

7. Acknowledgements

The work described in this paper was supported by the research program "Language Resources and Technologies for Slovene" (P6-0411) funded by the Slovene Research Agency. A special thanks goes to Lilja Øvrelid,

Francis M. Tyers and Niko Partanen who provided the English translations for the Norwegian, Chukchi and Komi examples, respectively.

8. Bibliographical References

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (2010). The Longman Grammar of Spoken and Written English. De Gruyter Mouton.

Braggaar, A. and van der Goot, R. (2021). Challenges in annotating and parsing spoken, code-switched, Frisian-Dutch data. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 50–58, Kyiv, Ukraine, April. Association for Computational Linguistics.

Caines, A., McCarthy, M., and Buttery, P. (2017). Parsing transcripts of speech. In *Proceedings of the Workshop on Speech-Centric Natural Language Processing*, pages 27–36. Association for Computational Linguistics.

Caron, B., Courtin, M., Gerdes, K., and Kahane, S. (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, Paris, France, August. Association for Computational Linguistics.

Carter, R. and McCarthy, M. (2015). Spoken grammar: Where are we and where are we going? *Applied Linguistics*, 38(1):1–20, 01.

Davidson, S., Yu, D., and Yu, Z. (2019). Dependency parsing for spoken dialog systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1513–1519, Hong Kong, China, November. Association for Computational Linguistics.

de Marneffe, M. C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Degand, L. and Simon, A. C. (2009). On identifying basic discourse units in speech: theoretical and empirical issues. *Discours*, 4.

Dittmar, N. (2012). Transcription. *The Encyclopedia of Applied Linguistics*.

Dobrovoljc, K. and Martinc, M. (2018). Er ... well, it matters, right? On the role of data representations in spoken language dependency parsing. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 37–46, Brussels, Belgium, November. Association for Computational Linguistics.

Dobrovoljc, K. and Nivre, J. (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, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Gerdes, K., Guillaume, B., Kahane, S., and Perrier, G. (2018). SUD or surface-syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74, Brussels, Belgium, November. Association for Computational Linguistics.

Hinrichs, E. W., Bartels, J., Kawata, Y., Kordoni, V., and Telljohann, H. (2000). The Tübingen treebanks for spoken German, English, and Japanese. In Wolfgang Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, Artificial Intelligence, pages 550–574. Springer Berlin Heidelberg.

Kahane, S., Caron, B., Strickland, E., and Gerdes, K. (2021a). 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. Association for Computational Linguistics.

Kahane, S., Vanhove, M., Ziane, R., and Guillaume, B. (2021b). A morph-based and a word-based treebank for Beja. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, Syntaxfest 2021)*, pages 48–60. Association for Computational Linguistics.

Leung, H., Poiret, R., Wong, T.-s., Chen, X., Gerdes, K., and Lee, J. (2016). Developing Universal Dependencies for Mandarin Chinese. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 20–29, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Liu, Z. and Prud'hommeaux, E. (2021). Dependency parsing evaluation for low-resource spontaneous speech. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 156–165, Kyiv, Ukraine, April. Association for Computational Linguistics.

Liu, P., Ning, Y., Wu, K. K., Li, K., and Meng, H. (2021). Open intent discovery through unsupervised semantic clustering and dependency parsing.

MacWhinney, B. (2014). *The Childes Project*. Psychology Press, January.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Øvrelid, L., Kåsen, A., Hagen, K., Nøklestad, A., Solberg, P. E., and Johannessen, J. B. (2018). The LIA treebank of spoken Norwegian dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 4482–4488, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Partanen, N., Blokland, R., Lim, K., Poibeau, T., and Rießler, M. (2018). The first Komi-Zyrian Universal Dependencies treebanks. In *Proceedings of*

the Second Workshop on Universal Dependencies (UDW 2018), pages 126–132, Brussels, Belgium, November. Association for Computational Linguistics.

Sagae, K., MacWhinney, B., and Lavie, A. (2004). Adding syntactic annotations to transcripts of parent-child dialogs. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 1815–1818, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Shriberg, E. (1996). Disfluencies in Switchboard. In *Proceedings of International Conference on Spoken Language Processing*, volume 96, pages 11–14.

Tyers, F. and Mishchenkova, K. (2020). Dependency annotation of noun incorporation in polysynthetic languages. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 195–204, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Wong, T., Gerdes, K., Leung, H., and Lee, J. (2017). Quantitative comparative syntax on the Cantonese-Mandarin parallel dependency treebank. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 266–275, Pisa,Italy, September. Linköping University Electronic Press.

Zeman, D. et al. (2021). Universal dependencies 2.9. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Çetinoğlu, Ö. and Çagri Çöltekin. (2019). Challenges of annotating a code-switching treebank. *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*.