PDTSC 2.0 - Spoken Corpus with Rich Multi-layer Structural Annotation

Marie Mikulová(✉), Jiří Mírovský, Anja Nedoluzhko, Petr Pajas, Jan Štěpánek, and Jan Hajič

Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic
{mikulova,mirovsky,nedoluzhko,pajas,hajic}@ufal.mff.cuni.cz

Abstract. We present a richly annotated spoken language resource, the Prague Dependency Treebank of Spoken Czech 2.0, the primary purpose of which is to serve for speech-related NLP tasks. The treebank features several novel annotation schemas close to the audio and transcript, and the morphological, syntactic and semantic annotation corresponds to the family of Prague Dependency Treebanks; it could thus be used also for linguistic studies, including comparative studies regarding text and speech. The most unique and novel feature is our approach to syntactic annotation, which differs from other similar corpora such as Treebank-3 [8] in that it does not attempt to impose syntactic structure over input, but it includes one more layer which edits the literal transcript to fluent Czech while keeping the original transcript explicitly aligned with the edited version. This allows the morphological, syntactic and semantic annotation to be deterministically and fully mapped back to the transcript and audio. It brings new possibilities for modeling morphology, syntax and semantics in spoken language – either at the original transcript with mapped annotation, or at the new layer after (automatic) editing. The corpus is publicly and freely available.

Keywords: Speech · Spoken corpus · Syntax · Semantics · Coreference · Treebank · Annotation

1 Introduction

Spontaneous speech breaks many rules by which written texts are constituted. Despite most spontaneous oral communications not meeting the basic written-text standards, the mutual understanding among humans does usually not get harmed. Posing no problem for humans, spontaneous speech is yet very difficult to handle for machines. POS taggers, parsers and semantic analyzers trained on written texts cannot cope with the morphological and syntactic irregularities typical of spontaneous speech. In this paper, we describe the (manually built) Prague Dependency Treebank of Spoken Czech 2.0 aimed at automatic recognition of spontaneous speech and its "understanding". We present our annotation scheme – which includes a speech "reconstruction" layer – above a corpus of

K. Ekštein and V. Matoušek (Eds.): TSD 2017, LNAI 10415, pp. 129–137, 2017.
DOI: 10.1007/978-3-319-64206-2_15

spontaneous dialogs. The reconstruction layer enables standard structural annotation, while linking the original transcript to syntax and semantics as well. The overall scheme conforms to the complex PDT-style annotation scenario that spans from linear text to dependency based syntax and semantics. The annotation scheme and the internal linking allows for future machine learning experiments using either the reconstruction layer or directly the combined links across layers.

2 Related Work

There is a wide range of corpora with disfluency annotation and subsequent syntax annotation, e.g., Switchboard corpus in Treebank-3 [8], Childes Database [18], the treebank of English, German, and Japanese created within the Verbmobil project [7], Corpus Gesproken Nederlands [19], or Treebank of Spoken French [2]. All these projects aim at identifying and labeling segments of the original audio (and transcript) for the chosen disfluencies. However, this style of disfluency annotation (consisting only in identifying and labeling spoken phenomena) cannot, in general, arrive at grammatical, fluent and understandable text readable for the human readers as well as appropriate for subsequent manual syntactic annotation or automatic processing. The development of a robust speech understanding pipeline requires not only a knowledge of what is a disfluency and where the disfluencies occur in an annotated spoken language corpus, but also how to understand them (cf. Sect. 4.1).

3 Prague Dependency Treebank of Spoken Czech

PDTSC 2.0[1] is a new release of Prague Dependency Treebank of Spoken Czech. It is a corpus of spoken language, consisting of 742,257 tokens and 73,835 sentences, representing 6,174 min (over 100 h) of spontaneous dialogs. The dialogs have been recorded, transcribed and edited in several interlinked layers: audio recordings, automatic and manual transcripts and manually reconstructed text. These layers along with morphological annotation were part of the first version of the corpus (PDTSC 1.0[2]; [3]). Version 2.0 is extended by annotation at the dependency syntax layer and the "deep" syntax layer, which contains semantic roles and relations as well as annotation of coreference. PDTSC 2.0 is freely and publicly available. Table 1 shows the inclusion and status of layers of annotation in both versions of the corpus.

3.1 The Data

PDTSC recordings consist of two parts covering two types of dialogs; both parts contain mostly colloquial Czech, even though some people spoke close to the

[1] http://ufal.mff.cuni.cz/pdtsc2.0.
[2] http://ufal.mff.cuni.cz/pdtsc1.0/en/index.html.

Table 1. Annotation in PDTSC 1.0 and PDTSC 2.0

PDTSC 2.0		t-layer	**Coreference**	manually
			Deep Syntax Annotation	manually
		a-layer	Dependency Syntax Parsing	automatically
	PDTSC 1.0	m-layer	Tagging and Lemmatization	automatically
			Speech Reconstruction	manually
		w-layer	**Transcript**	manually
		z-layer	Speech Recognition	automatically
			Audio	

standard. First, it contains a part of the Czech portion of the Malach project corpus, i.e., lightly moderated interviews (testimonies) with Holocaust survivors, originally recorded by the Shoa Visual History Foundation.[3] The second part of the corpus consists of dialogs recorded for the Companions project.[4] The domain is also personal memories, but in a Wizard-of-Oz setting where the two dialog participants chat over a collection of personal photographs. The goal of this project was to create virtual companions that would be able to have a natural conversation with humans. Domain-identical dialogs were created also in English (corpus PDTSE 1.0[5]), allowing comparison with the Czech data, even if the English data have not yet been upgraded to version 2.0.

The markup used in PDTSC 2.0 is the language-independent Prague Markup Language (PML), which is an XML subset customized for multi-layered linguistic annotation [16].

4 Layers of Annotation

PDTSC 2.0 is a treebank from the family of PDT-style corpora developed in Prague (for more information, see [4]). The main features of this annotation style are:

- based on a well-developed dependency syntax theory which is known as the Functional Generative Description [20],
- interlinked hierarchical layers of standoff annotation,
- "deep" syntax layer.

PDTSC differs from other PDT-style corpora mainly in the "spoken" part of the corpus. The layers stack starting at the external base layer with audio files (in the Vorbis format). The bottom layer of the corpus (**z-layer**) contains automatic speech recognition output synchronized to audio. The next layer, **w-layer**, contains manual transcript of the audio, i.e. everything the speaker has said including all slips of the tongue as well as non-speech events like coughing, laugh, etc.

[3] http://sfi.usc.edu/collections/holocaust.
[4] http://companions-project.org.
[5] http://ufal.mff.cuni.cz/pdtse1.0/en/index.html.

W-layer is synchronized to the automatic transcript and through it thus to the original audio. The subsequent **m-layer** contains a manually "reconstructed", i.e. edited, grammatically corrected version of the transcript, including punctuation and assumed sentence boundaries. The reconstructed tokens are automatically morphologically tagged and lemmatized. From this point on, annotation on the upper layers is the same as in the other PDT-style corpora. The dependency syntax layer (**a-layer**) is parsed automatically, while the "deep" syntax layer (**t-layer**) is annotated manually. There is a one-to-one correspondence between the tokens at the m-layer and the nodes at the a-layer. The syntactic dependencies are provided with dependency relations (e.g., Subject or Adverbial). The t-layer, which is also a tree-shaped graph (with content words only), is the highest and most complex linguistic representation that combines syntax and semantics in the form of semantic labeling, coreference annotation and argument structure description based on a valency lexicon.

In order not to lose any piece of the original information, tokens (nodes) on a lower layer are explicitly referenced from the corresponding closest (immediately higher) layer. These links allow for tracing every unit of annotation all the way down to the original audio and transcript, with the exception of reconstructed

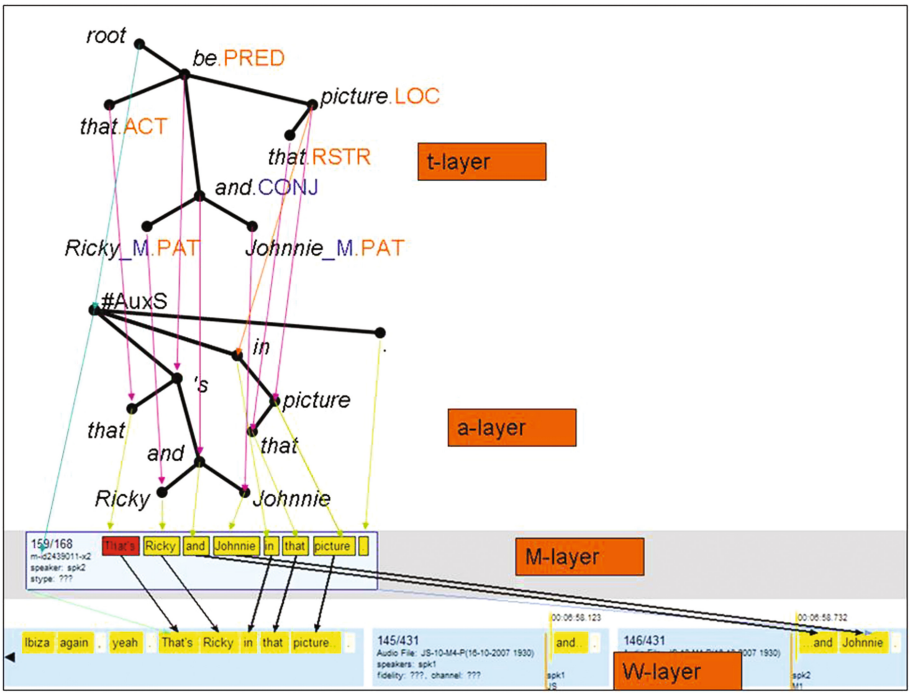

Fig. 1. Layers of annotation in PDTSC 2.0 (demonstrated on a English sentence *That's Ricky and Johnnie in that picture.*; audio not shown.)

ellipsis, which might only point in between audio segments. Figure 1 shows the relations between the layers as annotated and represented in the data.

In the following subsections, the manual annotation of the most important corpus parts (i.e., speech reconstruction and deep syntax annotation) is shortly described.

4.1 Spontaneous Speech Reconstruction

Spontaneous speech is "ungrammatical", full of a class of phenomena called disfluencies, such as false starts, repetitions, fillers, ellipses, etc. These phenomena cause problems for any subsequent processing. The purpose of speech reconstruction as defined in the present work is to "translate" the input spontaneous speech to a written text, before it is tagged and parsed. The transcript is segmented into sentence-like segments and these segments are edited to meet written-text standards, which means cleansing the text from the discourse-irrelevant and contentless material (superfluous connectives and deictic words, false starts, repetitions, etc. are removed) and re-chunking and re-building the original segments into grammatical sentences with acceptable word order and proper morpho-syntactic relations between words. The annotators are thus simulating the work of, e.g., magazine editors when preparing recorded interviews to appear in printed form. There are two basic annotation principles they have to follow:

A. The Content-Preservation Principle: the modifications of the original transcript may not affect the content.

B. The Minimal Modification Principle: modifications are only performed when it is necessary to follow written-text standards.

The annotators are also required to correctly **link the reconstructed text tokens to the original transcription** (which is, of course, then linked implicitly by using the synchronization marks to both the automatically recognized audio (z-layer) and to the audio itself). Even though the rules are relatively simple, certain conventions had to be introduced:

- source deletions: not linked (implicit links only based on order),
- word and punctuation insertions: not linked (implicit links as above),
- word substitution changes: linked to the source tokens that are the ones edited (and most similar in case of ambiguity),
- no change (identity between source and annotation): links to the source token,
- the reconstructed sentence (segment) boundaries (begin, end) are mapped onto the raw-transcript segments. These two links indicate the span of transcript that was used as the input for the given reconstructed sentence.
- word order changes are not labeled since they are deterministically extractable from the (crossing) links.

An example of linking the reconstructed text to original transcript is depicted in Fig. 1 (links between the M-layer and W-layer).

Manual annotation of speech reconstruction was the crucial part of the first version of the corpus. The annotation is described in more detail in [3] and the

guidelines are also specified in the annotation manual [9]. PDTSC annotation scheme of speech reconstruction has been developed in parallel (and often in cooperation) with Fitzgerald and Jelinek [1].

4.2 Deep Syntax and Coreference Annotation

One of the important distinctive features of the PDT-style annotation is the fact that in addition to the morphological and syntactic (dependency) layer, it includes complex semantically based annotation on the highest annotation layer (t-layer).

On the t-layer, every sentence is represented as a rooted tree with labeled nodes and edges. The tree reflects the **underlying dependency structure** of the sentence. The nodes stand for content words only. Unlike at the a-layer, not all the original tokens from the edited transcript (or, in the case of text, all the word tokens) are represented at the t-layer as nodes. Function words (prepositions, auxiliary verbs, etc.) do not have nodes of their own, but their contribution to the meaning of the sentence is not lost – several attributes are attached to the t-nodes the values of which represent such a contribution (e.g. tense for verbs). Some of the t-nodes do not correspond to any morphological token; they are added in case of surface deletions (ellipses). The types of the (semantic) dependency relations are represented by the "functor" attribute attached to all t-nodes.

The core ingredient in the annotation of the t-layer is **valency** (the theoretical description of the valency theory as developed in the framework of Functional Generative Description is summarized mainly in [17]). The valency criterion divides functors into the argument functors and adjunct functors. There are five arguments: Actor (`ACT`), Patient (`PAT`), Addressee (`ADDR`), Origin (`ORIG`) and Effect (`EFF`). In addition, we distinguish about 50 types of adjuncts (temporal, local, casual, etc.). The valency lexicon that all the PDT-family corpora use, PDT-Vallex [6,21], was built in parallel with the annotation of sentences and it has been used for consistent annotation of valency modifications in the annotated sentences. The t-layer annotation of PDTSC extended PDT-Vallex with approximately 1,500 new lemmas and 2,500 new valency frames [14].

The PDTSC 2.0 also captures grammatical and textual **coreference** relations. Grammatical coreference is based on language-specific grammatical rules, whereas to resolve textual coreference, the context knowledge is needed. Textual coreference annotation is based on the "chain principle", the anaphoric entity always referring to the last preceding coreferential antecedent. Coreference relations are technically part of the t-layer.

Annotation principles used at the t-layer and the annotation guidelines are described in the annotation manuals [10,11]. Compared to the anchoring original project of Prague Dependency Treebank[6] [5], the t-layer annotation in PDTSC 2.0 is slightly simplified; e.g., it does not contain information structure annotation (topic-focus).

[6] http://ufal.mff.cuni.cz/prague-dependency-treebank.

5 Annotation Quality Checking (Inter-annotator Agreement)

There are many ways to produce correct written text from a literal transcript. To capture this fact, we provide multiple parallel annotations for each transcript, but we do not unify the individual annotation streams. We believe that it will lead to more possibilities of training and evaluation of any tools that might be developed using such data, in a similar vein to the way multiple reference translations are used for automatic machine translation evaluation (more about speech reconstruction quality checking see in [3]).

A multiple parallel annotation of the same data becomes impossible (with regard to time and work) if the treebank is large and the annotated information is complex. For measuring an inter-annotator agreement (IAA) of deep syntax annotation, only a subset of the data was annotated in parallel. Since there is no "golden" annotation, we measure the agreement of all the pairs of annotators. A system of automatic quality checking of the annotated data was developed as well (see [12]). For more detailed account how the IAA for deep syntax annotation and for coreference relations are measured, see [13] and [15]. Table 2 shows average values of IAA measurements for deep syntax annotation and for textual coreference relations. Problems of low inter-annotator agreement and ambiguity in annotation of coreference relations are also described in [15].

Table 2. IAA in deep syntax annotation and coreference relations

Syntax	Annot1	Annot2	Annot3	Annot4
Annot1	-	98.7	98.4	98.6
Annot2	98.7	-	98.4	98.5
Annot3	98.4	98.4	-	98.4
Annot4	98.6	98.5	98.4	-
Coreference	Annot1	Annot2	Annot3	Annot4
Annot1	-	87.5	-	87.0
Annot2	87.5	-	86.8	-
Annot3	-	86.8	-	89.5
Annot4	87.0	-	89.5	-

6 Conclusion: What Is the Data Good For?

With the release of PDTSC 2.0, we have to a large extent closed the gap between the full annotation of the Prague Dependency Treebank (which is a written text-based corpus) and the Prague spoken dialog corpus, the PDTSC. We are not aware of any other spoken language corpus that would have both the "disfluencies" marked and a full annotation of syntax and semantics. In addition, we

have kept the unique "reconstruction" layer of annotation, which allows different views of and annotation mapping onto the original data: either the annotation can be mapped all the way to audio (or its automatic or manual transcripts), getting the usual style of speech corpora annotation with syntax built over the original transcript, or one might attempt to use the reconstruction layer - for example, one can perform the reconstruction step directly, using the upper layer annotation possibly only as a "hidden" layer (or not at all). Either way, we hope that this resource can help build automatic speech understanding and dialog systems.

As with similar projects, this release is a step towards bigger corpora, with more manual annotation. The PDTSC 2.0 will be also extended in the future, most notably by manual annotation on the m- and a-layers, and will become part of a consolidated Prague Dependency Treebanks release in 2018, which will contain four different treebanks of Czech, uniformly annotated using the scheme described in part here, with data coming from text, speech and internet sources.

Acknowledgments. The research reported in the paper was supported by the Czech Science Foundation under the projects GA16-05394S and GA17-12624S. This work has also been supported by the LINDAT/CLARIN project of Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

References

1. Fitzgerald, E., Jelinek, F.: Linguistic resources for reconstructing spontaneous speech text. In: Proceedings of the 6th LREC, Marrakech, Moroco (2008)
2. Gerdes, K., Kahane, S., Lacheret, A., Truong, A., Pietrandrea, P.: Intonosyntactic data structures: the Rhapsodie Treebank of spoken French. In: Proceedings of the 6th Linguistic Annotation Workshop, Jeju, Korea, pp. 85–94. ACL (2012)
3. Hajič, J., Cinková, S., Mikulová, M., Pajas, P., Ptáček, J., Toman, J., Urešová, Z.: PDTSL: an annotated resource for speech reconstruction. In: Proceedings of the 2008 IEEE Workshop on Spoken Language Technology, Goa, India, pp. 93–96 (2008)
4. Hajič, J., Hajičová, E., Mikulová, M., Mírovský, J.: Prague dependency treebank. In: Handbook on Linguistic Annotation, Volume II, pp. 555–594. Springer, Dordrecht (2017)
5. Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., Ševčíková-Razímová, M., Urešová, Z.: Prague Dependency Treebank 2.0 (LDC2006T01) (2006)
6. Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., Pajas, P.: PDT-VALLEX: creating a large-coverage valency lexicon for treebank annotation. In: Proceedings of the 2nd Treebanks and Linguistic Theories Workshop, pp. 57–68. Vaxjo University Press, Vaxjo (2003)
7. Hinrichs, E.W., Bartels, J., Kawata, Y., Kordoni, V., Telljohann, H.: The verbmobil treebanks. In: KONVENS, pp. 107–112 (2000)
8. Marcus, M., Santorini, B., Marcinkiewicz, M.A., Taylor, A.: Penn Treebank-3. Linguistic Data Consortium, LDC99T42, University of Pennsylvania (1999)
9. Mikulová, M.: Rekonstrukce standardizovaného textu z mluvené řeči v Pražském závislostním korpusu mluvené češtiny. Manuál pro anotátory. Technical report ÚFAL TR-2008-38 (2008)

10. Mikulová, M.: Annotation on the tectogrammatical level. Additions to annotation manual (with respect to PDTSC and PCEDT). Technical report ÚFAL TR-2013-52 (2014)
11. Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Kučová, L., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K., Žabokrtský, Z.: Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Technical report 30, Prague, Czech Republic (2006)
12. Mikulová, M., Štěpánek, J.: Annotation quality checking and its implications for design of Treebank (in Building the Prague Czech-English Dependency Treebank). In: Proceedings of 8th Treebanks and Linguistic Theories Workshop, Milano, Italy, pp. 137–148 (2009)
13. Mikulová, M., Štěpánek, J.: Ways of evaluation of the annotators in building the Prague Czech-English Dependency Treebank. In: Proceedings of the 7th LREC, Valletta, Malta, pp. 1836–1839 (2010)
14. Mikulová, M., Štěpánek, J., Urešová, Z.: Liší se mluvené a psané texty ve valenci? Korpus "gramatika" axiologie **8**, 36–46 (2013)
15. Nedoluzhko, A., Mírovský, J.: Annotators' certainty and disagreements in coreference and bridging annotation in Prague Dependency Treebank. In: Proceedings of the 2nd International Conference on Dependency Linguistics, Prague, Czech Republic, pp. 236–243 (2013)
16. Pajas, P., Štěpánek, J.: Recent advances in a feature-rich framework for treebank annotation. In: Proceedings of the 22nd International Conference on Computational Linguistics, Manchester, UK, vol. 2, pp. 673–680 (2008)
17. Panevová, J.: On verbal frames in functional generative description. Prague Bull. Math. Linguist. **22**, 3–40 (1974)
18. Sagae, K., MacWhinney, B., Lavie, A.: Adding syntactic annotations to transcripts of parent-child dialogs. In: Proceedings of the 4th LREC, Lisbon, Portugal (2004)
19. Schuurman, I., Goedertier, W., Hoekstra, H., Oostdijk, N., Piepenbrock, R., Schouppe, M.: Linguistic annotation of the spoken Dutch corpus: if we had to do it all over again. In: Proceedings of the 4th LREC, Lisbon, Portugal (2004)
20. Sgall, P., Hajičová, E., Panevová, J.: The Meaning of the Sentence and Its Semantic and Pragmatic Aspects. Academia/Reidel Publishing Company, Prague/Dordrecht (1986)
21. Urešová, Z.: Building the PDT-VALLEX valency lexicon. In: Proceedings of the 5th Corpus Linguistics Conference, pp. 1–18. University of Liverpool, Liverpool (2012)