

# Prague Dependency Treebank - Consolidated 2.0: Enriching a Complex Annotation Scheme

Marie Mikulová, Jiří Mírovský, Milan Straka, Pavlína Synková, Jan Štěpánek,  
Barbora Štěpánková, Jan Hajič

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics  
Malostranské náměstí 25, 118 00 Prague 1, Czech Republic  
{mikulova,mirovsky,straka,synkova,stepanek,stepenakova,hajic}@ufal.mff.cuni.cz

## Abstract

The Prague Dependency Treebank framework is unique in its attempt to systematically include and link different layers of language, including a meaning representation with several types of inter-sentential phenomena, especially coreference and discourse relations. We present its second consolidated version (PDT-C 2.0), which concludes almost 30-years long project of sustained development of the resource to a uniformly and coherently annotated, genre-diversified, almost 4 million token language resource of Czech language, with accompanying fully compatible lexicons. In addition to continuous linguistic research, the richly linguistically annotated corpus is also widely used in international comparisons of the development of traditional and novel NLP tools as well as in conversions into other formalisms. The corpus and the trained parsers are available under the CC BY-NC-SA licence.

**Keywords:** treebank, morphology, syntax, semantics, coreference, discourse, tagger, parser, lexicon

## 1. Introduction

We present the **Prague Dependency Treebank - Consolidated 2.0** (PDT-C 2.0; Hajič et al., 2024a) a second consolidated release of the existing original PDT-corpora of Czech data published in one package to allow for easier data handling. Compared to the previous 1.0 version, the data are now fully manually annotated. Included is a morphological dictionary MorfFlex (Hajič et al., 2024b), and a valency lexicon PDT-Vallex (Urešová et al., 2024). Both external resources are fully compatible with PDT-C annotation. In the paper, we summarize in a balanced manner all the aspects of this large language resource: data genres, multi-layer annotation scheme, various types of rich linguistic annotation, the linked lexical resources and their relation to the annotation. The presented language resource is unique in the following aspects:

- **multi-layer architecture:** the complex structure of language is captured through inter-linked hierarchical layers of annotation as illustrated in Fig. 1. This enables researchers to process specific linguistic aspects independently, allowing for more detailed and precise analyses. At the same time, the interconnectivity of the layers makes it possible to study how meaning is linked to text.
- **rich linguistic annotation** spanning from morphology and syntax to semantics, including **inter-sentential phenomena** (especially coreference, discourse relations). See Tab. 2 for overview.

- **genre-diversified datasets:** written, translated, spoken, and user-generated. See Tab. 1 for the datasets overview.
- **large volume of data:** more than 4 million tokens; see Tab. 1 for the volume overview.
- all annotations were performed **manually**.

PDT-C 2.0 is provided as a digital open resource accessible to all users via LINDAT/CLARIAH-CZ.<sup>1</sup>

The paper is organized as follows: The PDT multi-layer annotation scheme is described in Sect. 2; the genre diversity of the data is presented in Sect. 3; the voluminous manual annotation is emphasized in Sect. 4; the richness of the linguistic annotation is outlined in Sect. 5; external, fully compatible language resources are mentioned in Sect. 6; the application of the treebank in the field of NLP (parser development and conversion into other frameworks) is presented in Sect. 7. We conclude in Sect. 9, describing also the future work and ongoing annotation efforts (Sect. 8).

### 1.1. From 1.0 to PDT-C 2.0

Compared to the previous 1.0 version (Hajič et al., 2020), the data are now fully manually annotated. The novelty lies in the following:

- Manual annotation at the **surface syntax** layer (Sect. 5.3) is now performed in those parts of the corpus that were previously annotated only by automatic tools. The goal of the annotation

<sup>1</sup><http://hdl.handle.net/11234/1-5813>

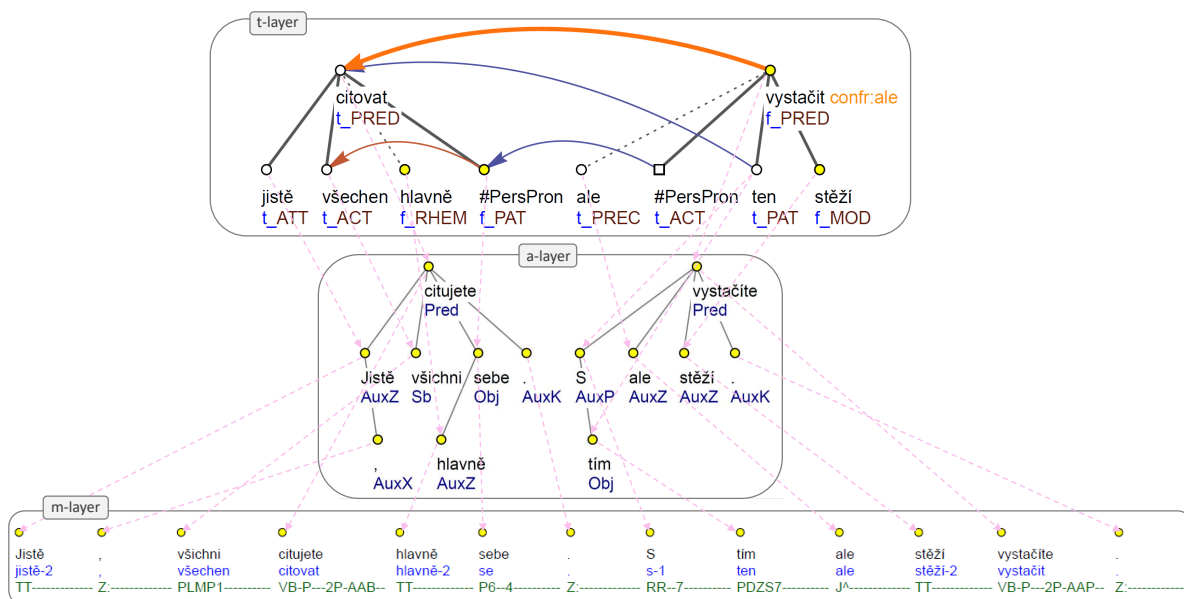


Figure 1: Multi-layer annotation scheme in the PDT-C treebank, illustrated on the text:  
*Jistě, všichni citujete hlavně sebe. S tím ale stěží vystačíte.*  
 Of-course, you-all cite mainly yourself. With that but hardly suffice-you.  
 ‘Of course, you all mainly cite yourself. But that’s hardly enough.’

work was also to consolidate the manual annotation across all layers, including previously manually annotated parts. Annotators follow all annotation layers during the annotation process. This resulted in many modifications and corrections to the original annotation.

- Manual annotation of **discourse** relations (Sect. 5.7) is now provided for all datasets.
- Manual annotation of **coreference** (Sect. 5.6) is now provided for all datasets.

## 2. Multi-layer Architecture

The PDT annotation scheme is based on the well-developed theory of language description, Functional Generative Description (Sgall et al., 1986) and was reflected in several annotation manuals available from the project website.<sup>2</sup>

The multi-layer architecture (linked from meaning to text) allows a comprehensive description of the relations between morphological properties, syntactic function and expressed meaning, and thus contributes to greater accuracy in the description of the language and to the overall consistency of the annotated data (cf. Hajičová et al., 2022; Mikulová et al., 2025). The multi-layer architecture is schematically illustrated in Fig. 1: each annotation layer is indicated by a separate box. The links

between the layers are indicated by the light dotted arrows. There are the three layers of annotation:

- **morphology** annotation (*m-layer* box in Fig. 1): all tokens get a lemma and morphological tag (see Sect. 5.2).
- **surface syntax** (*a-layer*): a dependency tree capturing syntactic relations such as subject, object, adverbial, etc. (see Sect. 5.3),
- **deep syntax** and other **semantics** annotations (*t-layer*), capturing deep syntactic structure (Sect. 5.4), valency (5.5), coreference (Sect. 5.6), discourse (Sect. 5.7), etc.

In addition to the above-mentioned three annotation layers in the PDT scheme, there is also the **raw text layer** (it is not shown in Fig. 1), where the text is segmented into documents and paragraphs and individual tokens are assigned unique identifiers. There is additional audio signal and speech recognition layer in the spoken data (Sect. 3.3). In the spoken data part, the raw text layer is in fact also an “annotated” layer, namely the manually provided transcription of the audio signal.

**Linking the layers.** To avoid losing any of the original information, tokens (nodes) at a lower layer are explicitly referenced from the closest (immediately higher) layer. These links enable every unit of annotation to be traced all the way down to the original text or transcript and audio (in spoken data).

<sup>2</sup><https://ufal.mff.cuni.cz/pdt-c>

	Written	Translated	Spoken	User-generated	Total
Morphological layer	1,957,150	1,152,289	742,316	33,836	3,885,591
Surface syntactic layer	1,503,637	1,152,289	742,316	33,836	3,432,078
Deep syntactic layer	833,180	1,152,289	742,316	33,836	2,761,621

Table 1: Volume of the datasets in PDT-C 2.0 (number of tokens)

### 3. Genre Diversified Data

PDT-C 2.0 consists of four different datasets: written texts (Sect. 3.1), translated texts (Sect. 3.2), spoken texts (Sect. 3.3), and of user-generated texts (Sect. 3.4).

The datasets are uniformly published in three **formats**: pml, mrp, and treex. The Prague Markup Language format (PML, Pajas and Štěpánek, 2008) is a language-independent, XML-based format customized for multi-layer linguistic annotation. Treex is technically also a PML format, used in the NLP system Treex (all annotation layers are in a single file; Žabokrtský, 2011). MRP is a JSON-based format used in the CoNLL 2019 and 2020 shared tasks on meaning representation parsing (Oepen et al., 2019, 2020); unlike the PML and Treex formats, the conversion to the MRP format, described in detail in Zeman and Hajič (2020), is lossy because it extracts only part of the annotation.

**Quality and consistency** of the annotations were monitored, measured, and ensured using various tools (such as multiple annotations and automated checks; cf. Mikulová and Štěpánek, 2010; Mikulová et al., 2022, 2025).

#### 3.1. Written Data

The dataset of written texts coming from the **Prague Dependency Treebank**, the first PDT corpus, in development since the 1990s (Hajič, 1998). The data consist of Czech newspaper and journal texts from three domains: daily news, business, and science. Compared to other datasets, the annotation in the written dataset is the richest one, some special annotations are added; see Tab. 2.

#### 3.2. Translated Data

The dataset of translated texts comes from the **Prague Czech-English Dependency Treebank** (PCEDT, originally published in 2012, Hajič et al., 2012). PCEDT is a (partially) manually annotated Czech-English parallel corpus. The English part consists of the Wall Street Journal sections of the Penn Treebank (Marcus et al., 1993). The Czech part, used in the PDT-C consolidated edition, has been manually (and professionally, with multiple quality control passes) translated from the English original, sentence to sentence.

#### 3.3. Spoken Data

The dataset of spoken texts is taken from the **Prague Dependency Treebank of Spoken Czech** (originally published in 2017; Mikulová et al., 2017). It contains slightly moderated testimonies of Holocaust survivors from the Shoa Foundation Visual History Archive and dialogues in which two participants chat over a collection of photographs.

The spoken data differs from the other included PDT-corpora mainly in the “spoken” part of the corpus. In addition to the three annotation layers described in Sect. 2, the corpus also contains audio signal, transcript produced by an automatic speech recognition engine, and manual transcription of the recorded speech. The process starts at the “audio” layer, which contains the audio signal. The next layer contains the transcript as produced by an automatic speech recognition engine. The word layer contains manual transcription of the recorded speech, and the morphological layer contains the reconstructed, i.e. grammatically corrected version of the sentences (see Sect. 5.1). From this point on, annotation on the upper layers is standard.

#### 3.4. User-generated Data

The dataset of user-generated texts comes from the **PDT-Faust** corpus, which is a small treebank containing short segments (very often with non-standard as well as expressive, obscene, or vulgar content) typed in by various users on the [reverso.net](http://reverso.net) web page for translation. The Czech data includes manual annotations of Czech reference translations of English source texts. This texts were translated independently by three translators and all three reference translations were annotated.

## 4. Volume of Data

The data volume is given in Tab. 1. Altogether, the consolidated treebank contains of almost 4 million tokens with manual morphological annotation (Sect. 5.2) and 3.5 million tokens with manual surface syntactic annotation (Sect. 5.3) and 2.7 million with manual deep syntactic and other semantic annotations (Sect. 5.4). The different number of tokens in the case of the written data is due to the fact that some annotations are only available for the morphological and/or surface syntactic layer.

Dataset/Type of annotation	Written	Translated	Spoken	User-generated
Audio	non-applicable	non-applicable	provided	non-applicable
ASR transcript	non-applicable	non-applicable	provided	non-applicable
Transcript	non-applicable	non-applicable	manually	non-applicable
Translation	non-applicable	manually	non-applicable	manually
Morphological layer				
Speech reconstruction	non-applicable	non-applicable	manually	non-applicable
Lemmatization	manually	manually	manually	manually
Tagging	manually	manually	manually	manually
Surface syntactic layer				
<b>Dependency structure</b>	manually	<b>manually</b>	<b>manually</b>	<b>manually</b>
<b>Surface syntactic functions</b>	manually	<b>manually</b>	<b>manually</b>	<b>manually</b>
Clause segmentation	manually	not annotated	not annotated	not annotated
Deep syntactic layer				
Deep syntactic structure	manually	manually	manually	manually
Deep syntactic functions	manually	manually	manually	manually
Valency	manually	manually	manually	manually
<b>Coreference</b>	manually	manually	manually	<b>manually</b>
<b>Discourse</b>	manually	<b>manually</b>	<b>manually</b>	<b>manually</b>
Grammatemes	manually	not annotated	not annotated	not annotated
Topic-focus articulation	manually	not annotated	not annotated	not annotated
Bridging relations	manually	not annotated	not annotated	not annotated
Genre specification	manually	not annotated	not annotated	not annotated
Quotation	manually	not annotated	not annotated	not annotated
Multiword expressions	manually	not annotated	not annotated	not annotated

Table 2: Overview of various types of annotation and their realization in the datasets (new manual annotation made to PDT-C 2.0 is indicated in bold)

## 5. Rich Linguistic Annotation

The long-run Prague Dependency Treebank project is unique in its attempt to systematically cover and link different layers of language description including a rich semantic annotation. Tab. 2 provides an overview of the different types of annotation across the three annotation layers (see Sect. 2) for each dataset (see Sect. 3), along with information on how the annotations were carried out. Newly added manual annotations in the PDT-C 2.0 version are highlighted in bold. The table shows that all datasets include manual annotations for lemmatization, tagging, dependency structure, deep syntactic structure, valency, coreference, and discourse. In the spoken and written datasets, there are also additional specialized annotations. In the following subsections, the annotations of the most important phenomena are shortly described.

### 5.1. Speech Reconstruction

Spontaneous speech reconstruction is a special type of manual annotation at the morphological layer that only belongs to the spoken data. The purpose of speech reconstruction is to “translate” the “ungrammatical” spontaneous speech to a writ-

ten text, before it is tagged and parsed. The transcript is divided into sentence-like segments and the segments are edited to meet written-text standards, which means cleansing the text from the discourse-irrelevant and content-less material (e.g., superfluous words, false starts, repetitions, etc. are removed) and re-building the original segments into grammatical sentences with acceptable word order and proper morphosyntactic relations between words. See more in Hajič et al. (2008) and Mikulová et al. (2017).

### 5.2. Lemmatization and Tagging

At the morphological layer, a lemma and a tag is assigned to each token. Czech is a highly inflectional language. A 15-character tag describes the inflectional forms of (declined) nouns and adjectives and (conjugated) verbs. All tokens of a sentence are traditionally also assigned a POS category within the tag. The annotation contains no syntactic structure, no attempt is made to put together analytical verb forms or other types of multiword expressions. The annotation is described in Mikulová et al. (2020); Hajič et al. (2020).

COREFERENCE	count
grammatical coreference	66,168
textual coreference	331,783
segment of text	5,903
situational context	20,753

Table 3: Volume of coreference annotations

### 5.3. Surface Syntactic Annotation

A surface dependency structure is captured by a rooted tree with the specification of the head for each node and the assignment of a syntactic function such as subject (*Sb*), object (*Obj*), adverbial (*Adv*), or attribute (*Attr*). Every token of the raw text (including punctuation marks; cf. nodes for comma (*AuxX*) and terminal symbol of the sentence (*AuxK*) in Fig. 1) is represented by a node and no additional nodes are allowed. The annotation guidelines are described in Mikulová et al. (2026a).

### 5.4. Deep Syntactic Structure

At the deep syntactic layer, every sentence is represented as a tree-like graph. Unlike the lower layers, not all of the original tokens are represented as nodes, but the nodes only stand for content words (e.g., there is only one node for the preposition phrase *s tím* ‘with that’ at the t-layer in Fig. 1). Function words (prepositions, auxiliaries, etc.) do not have nodes of their own, their contribution to the meaning of the sentence is captured by several attributes attached to the nodes, the values of which represent this contribution (e.g., tense for verbs; see Sect. 5.9). In case of surface deletions, extra nodes are added (in Fig. 1, the restoration of a deletion is illustrated by the *#PersPron* (personal pronoun) node for the Actor (*ACT*) of the second sentence’s predicate). The types of the (semantic) dependency relations are represented by the *functor* attribute attached to all nodes. Annotation principles are described in several manuals (Mikulová et al., 2006; Mikulová, 2014).

### 5.5. Valency

The core ingredient in the annotation of deep structure is valency (predicate-argument structure annotation). The valency criterion divides functors into argument and adjunct functors. There are five arguments: Actor (*ACT*), Patient (*PAT*), Addressee (*ADDR*), Origin (*ORIG*), and Effect (*EFF*). In addition, about 50 types of adjuncts (temporal, spatial, manner, causal, regard, etc.) are used. For a particular verb (or more precisely, verb sense), a subset of the functors is obligatory, while others are either not present at all or are optional. Each

DISCOURSE	Count
Discourse relations: COMPARISON	
concession	2,736
confrontation	2,432
correction	1,390
gradation	1,108
opposition	10,348
pragmatic contrast	231
restrictive opposition	1,019
Discourse relations: CONTINGENCY	
condition	4,801
explication	373
pragmatic condition	372
pragmatic reason-result	493
purpose	2,710
reason-result	10,927
Discourse relations: EXPANSION	
conjunction	29,509
conjunctive alternative	854
disjunctive alternative	623
equivalence	544
generalization	522
instantiation	687
specification	1,428
Discourse relations: TEMPORAL	
precedence-succession	6,600
synchrony	2,799
DISCOURSE RELATIONS: TOTAL	82,506

Table 4: Volume of discourse annotations

occurrence of a verb in all corpora is linked to the appropriate valency frame in the valency lexicon (see Sect. 6.2).

### 5.6. Coreference

Coreference annotation (Hajičová et al., 2000) captures a referential identity relation between entities (nodes; ex. *Einstein - he - the famous scientist*). Several types are distinguished. Grammatical pronominal coreference is based on language-specific grammatical rules, whereas resolving textual coreference (both pronominal and nominal) requires contextual knowledge. Textual coreference annotation follows the “chain principle”, where the anaphoric entity always refers to the last preceding antecedent. Coreference can also be cataphoric (pointing to a subsequent part of the text). Two special cases of coreference are further annotated: reference to situational context and reference to a segment of text. In Fig. 1, coreference relations are represented by the brown (grammatical) and blue (textual) arrows. In PDT-C 2.0, there is now manual coreference annotation in all four datasets; the volume of the coreference annotations is in Tab. 3.

## 5.7. Discourse

Annotation of discourse relations covers local relations that hold between two spans of text (usually clauses and sentences) marked by primary or secondary discourse connectives. Primary connectives are grammaticalized, mostly one-word expressions (such as *a* ‘and’, *ale* ‘but’, *protože* ‘because’), secondary connectives are more loose expressions (*z toho důvodu* ‘for that reason’, *na druhou stranu* ‘on the other hand’, etc.). Each relation is accompanied by a discourse type (such as *reason-result*, *condition*, *purpose*, *equivalence*, etc.; see Tab. 4 for a complete list). See more in Poláková et al. (2012); Mírovský et al. (2024). Discourse relations are annotated between roots of the relevant subtrees (sentences, clauses, phrases) in the trees. In Fig. 1, a discourse relation (of *confrontation*) is represented by the orange arrow between the predicates of the sentences. In PDT-C 2.0, there is now manual discourse annotation in all four datasets.

## 5.8. Topic-Focus Articulation

A basic aspect of the deep structure is also the topic-focus articulation (for arguments on its semantic relevance see Sgall et al., 1986; Hajičová et al., 1998), indicated by the blue values  $\text{t}$  and  $\text{f}$  (in front of the functor values) in Fig. 1:  $\text{t}$  is for contextually bound and  $\text{f}$  for contextually non-bound nodes. The ordering of nodes corresponds to the information structure of a sentence (cf. different position of particle *stěží* ‘hardly’ and *ale* ‘but’ at the a-layer and t-layer in Fig. 1).<sup>3</sup> In PDT-C 2.0, topic-focus articulation is captured only in the written dataset (cf. Tab. 2).

## 5.9. Grammatemes

So called grammatemes (Razímová and Žabokrtský, 2006; Panevová and Ševčíková, 2010) are attached to some nodes; they provide information about the node that cannot be derived from the deep syntactic structure, the functor and other attributes. Grammatemes are counterparts of those morphological categories which bear relevant semantic information (e.g., tense of predicate, number of entities, modality of the sentence). They are annotated only in the written dataset (cf. Tab. 2).

## 5.10. Bridging

In the written dataset, apart from the coreference relations, non-coreferential association relations

are annotated as bridging relations if they are related in one of specific types of semantic or conceptual ways to their antecedents. Several types are distinguished, e.g., metonymical relation between a part and a whole (*part-of*; ex. *room – ceiling*); relation between a set and its subsets (*set-subset*, ex. *students – some students – a student*), the relation between an entity and a singular function on this entity (*function*; ex. *prime minister – government*). See more in Nedoluzhko and Mírovský (2011).

## 5.11. Other Annotation

In the written dataset, noun valency, and other semantic properties of the sentence such as genre specification, multiword expressions, quotation are also annotated. More information of these special annotations can be found in Mikulová et al. (2013).

## 6. External Resources

An important part of annotation also involves various dictionaries. They can be used to distinguish the different meanings of words and also to maintain or monitor the consistency of annotations. The PDT-C annotation is associated with the morphological (Sect. 6.1) and valency (6.2) dictionaries.

### 6.1. MorfFlex

MorfFlex (the latest version is **MorfFlex CZ 2.1**, Hajič et al., 2024b, Hlaváčová et al., 2026) is the Czech morphological dictionary. For each word form, full inflectional information is coded in a positional tag. Word forms are organized into paradigms according to their morphological behaviour. The paradigm is identified by a unique lemma. The description also contains some semantic, stylistic and derivational information. MorfFlex is distributed as a flat list of *form - lemma - tag* triplets. MorfFlex CZ 2.1 contains 126,906,921 such triplets. It is fully compatible with the PDT-C 2.0 morphological annotation.

### 6.2. PDT-Vallex

The valency dictionary PDT-Vallex (Hajič et al., 2003; Urešová, 2012) was developed in parallel with the annotation and contains almost exclusively verbs and their meanings that occurred in the annotated data, whose valency the annotators needed to know in order to correctly annotate valency. The latest version, **PDT-Vallex 4.5** (Urešová et al., 2024), includes nearly 8,500 verbal lemmas and 14,500 valency frames. It is part of PDT-C 2.0 release, and the valency frames are directly linked to the data through explicit references.

<sup>3</sup>The nodes at the lower layers are naturally ordered based on the surface word order.

Tags	Tagging Accuracy					Lemmatization Accuracy					Performance	
	Predicted	Faust	PCEDT	PDT	PDTSC MacroAvg	Faust	PCEDT	PDT	PDTSC MacroAvg	Speed	Size	
Full tags	95.08	97.24	95.95	97.63	96.47	97.69	98.98	98.29	98.83	98.45	26k $\frac{\text{toks}}{\text{s}}$	25MB
First 2 pos	98.24	99.30	98.61	99.10	98.81	97.31	98.79	98.10	98.69	98.22	328k $\frac{\text{toks}}{\text{s}}$	10MB

Table 5: Performance of morphological tagging and lemmatization of two MorphoDiTa tagger models, one predicting full tags (15 positions) and the other predicting only first 2 tag positions.

Test Data	Morpho Train		Syntax Train		Raw Predictions		With MorphoDiTa			UAS	LAS
	PDT	PDT-C	PDT	PDT-C	Tags	Lemmas	Tags	Lemmas	LemmasEM		
PDT-C: Faust	✓	×	✓	×	95.64	96.36	97.02	98.22	97.85	83.72	78.94
	×	✓	✓	×	97.13	98.27	97.77	98.91	98.46	84.20	79.97
	×	✓	×	✓	97.18	98.24	97.82	99.15	98.70	91.44	88.40
PDT-C: PCEDT	✓	×	✓	×	98.11	98.46	98.44	98.94	98.40	95.57	93.63
	×	✓	✓	×	98.84	99.25	99.00	99.51	98.86	95.74	93.86
	×	✓	×	✓	98.89	99.27	99.06	99.55	98.90	96.52	94.97
PDT-C: PDT	✓	×	✓	×	98.18	98.39	98.49	98.92	98.73	95.17	92.63
	×	✓	✓	×	98.26	98.50	98.49	98.93	98.73	95.25	92.76
	×	✓	×	✓	98.20	98.39	98.45	98.86	98.67	95.24	92.80
PDT-C: PDTSC	✓	×	✓	×	97.30	97.67	97.86	98.42	98.36	93.60	89.97
	×	✓	✓	×	98.68	99.08	98.91	99.34	99.27	93.69	90.08
	×	✓	×	✓	98.71	99.08	98.96	99.35	99.29	94.96	92.33
PDT-C: MacroAvg	✓	×	✓	×	97.31	97.72	97.95	98.62	98.34	92.01	88.79
	×	✓	✓	×	98.23	98.77	98.54	99.17	98.83	92.22	89.17
	×	✓	×	✓	98.24	98.74	98.57	99.23	98.89	94.54	92.12

Table 6: Comparison of UDPipe morphosyntactic performance in percents using either whole PDT-C or just PDT as morphological/syntactic annotation training data, evaluated on its four subsets, with or without MorphoDiTa dictionary during inference. Predicted lemma is considered correct when a raw lemma plus an optional lemma sense match; LemmasEM compares full lemmas including all additional information.

## 7. Related Data and Tools

Throughout their development, the PDT corpora have served as an invaluable resource for linguistic research, for enriching the description of the Czech language system, and for developing general methods of language description. The richly linguistically annotated PDT corpora are also widely used in international comparisons in the NLP field. In what follows, we briefly introduce the latest morphological, syntactic, and semantic analysers (Sect. 7.1), as well as the most recent conversions of PDT-C into other formalisms (Sect. 7.2).

### 7.1. Tools

We now describe and evaluate morphological, morphosyntactic and deep syntactic analyzers trained on PDT-C 2.0; where appropriate, we compare them to models trained only on PDT syntactic data, the only available syntactic data in PDT-C 1.0. All models are released under the CC BY-NC-SA license.

**MorphoDiTa** We train a fast CPU-based tagger and lemmatizer using MorphoDiTa (Straková et al., 2014). The performance of the resulting model is presented in Table 5 and the model is available at <https://hdl.handle.net/11234/1-5985>.

**UDPipe** We also train morphosyntactic parser model using a GPU-based UDPipe (Straka, 2018; Straka et al., 2019). Notably, we utilize the variant described by Straka and Straková (2024) which improves performance by consulting the morphological dictionary provided by MorphoDiTa during inference.

Table 6 shows the performance of three UDPipe models: trained only on the PDT data, trained on full PDT-C morphological data but only on PDT syntactic data (best configuration trainable on PDT-C 1.0; Straka and Straková, 2024), and trained on the full PDT-C 2.0 data. Using the latter model reduces the macro-averaged error rate in syntactic parsing accuracy by more than 27%.

Compared to the MorphoDiTa tagger, UDPipe achieves 60% error reduction in tagging accuracy and 50% in lemmatization; however, MorphoDiTa

Test Data	Syntax Train		MRP Train		Labels	Properties	Anchors	Edges	Attributes	Tops	All
	PDT	PDT-C	PDT	PDT-C							
PDT-C: Faust	✓	✗	✓	✗	88.74	51.72	90.77	56.30	54.28	92.00	66.28
	✗	✓	✓	✗	92.30	85.32	94.04	77.79	75.24	99.00	85.68
	✗	✓	✗	✓	92.48	85.88	95.01	77.67	76.25	96.67	86.01
PDT-C: PCEDT	✓	✗	✓	✗	94.99	53.17	96.62	67.45	59.04	99.83	72.97
	✗	✓	✓	✗	96.97	92.95	97.91	87.19	80.27	100.00	92.79
	✗	✓	✗	✓	96.96	92.18	98.02	87.39	81.30	99.83	92.71
PDT-C: PDT	✓	✗	✓	✗	96.55	92.87	97.66	82.41	74.76	99.09	91.36
	✗	✓	✓	✗	96.62	88.78	97.79	81.87	75.57	99.93	89.56
	✗	✓	✗	✓	96.39	89.20	97.85	82.68	77.42	99.56	89.93
PDT-C: PDTSC	✓	✗	✓	✗	93.21	52.72	97.05	67.50	72.82	99.60	72.48
	✗	✓	✓	✗	96.22	92.64	98.30	86.87	83.87	99.95	92.59
	✗	✓	✗	✓	96.24	92.78	98.30	87.54	84.72	99.60	92.85
PDT-C: MacroAvg	✓	✗	✓	✗	93.37	62.62	95.52	68.41	65.22	97.63	75.77
	✗	✓	✓	✗	95.53	89.92	97.01	83.43	78.74	99.72	90.16
	✗	✓	✗	✓	95.52	90.01	97.30	83.82	79.92	98.91	90.38

Table 7: Comparison of PERIN semantic parsing MRP score in percents on the four PDT-C subsets, evaluated using either whole PDT-C or just PDT as syntactic/semantic annotation training data.

model is by a decimal order of magnitude faster on a single CPU.

The model trained on full PDT-C 2.0 data is available at <http://hdl.handle.net/11234/1-6045>.

**PERIN** Finally, we train a meaning representation parser producing semantic graphs (Zeman and Hajič, 2020) using PERIN (Samuel and Straka, 2020), the winning system of the 2020 CoNLL shared task on Cross-Framework Meaning Representation Parsing (MRP 2020; Oepen et al., 2020).

To produce a semantic graph, the PERIN parser processes not just an input sentence but also its syntactic tree; therefore, it builds on top of a UD-Pipe parser. We therefore train three parser configurations: first training both UDPipe and PERIN on PDT data (the setting of MRP 2020), then training UDPipe on PDT data and PERIN on PDT-C data (the best setting in PDT-C 1.0), and finally using the complete PDT-C data for both systems.

The results of all configurations is presented in Table 7. When using just PDT to train the semantic parser, the results on the other subsets are considerably worse; on the other hand, using all PDT-C training data decreases performance on PDT slightly (indicating either limited parser capacity or mild annotation differences).

Maybe surprisingly, when using all PDT-C data to train the semantic parser, the quality of the syntactic trees has only minor influence on the final results, i.e., even with slightly lower-quality trees on non-PDT subsets, PERIN successfully learns to produce high-quality semantic graphs from them.

The model trained on full PDT-C 2.0 data is available at <http://hdl.handle.net/11234/1-6118>.

## 7.2. Conversion to Other Frameworks

An important application (and promotion) of an annotated corpus is its conversion into another formalisms. By being incorporated into different frameworks, the robustness and universality of the chosen format become evident. The PDT corpora are used for conversions into various frameworks.

First, it is important to mention the popular **Universal Dependencies (UD)** formalism (de Marneffe et al., 2021). After the conversion of PDT-C 2.0 (in UD version 2.16),<sup>4</sup> Czech represents the largest language coverage (Mikulová et al., 2026b).

The PDT corpora are part of other expanding UD projects: **DeepUD**<sup>5</sup> (Droganova and Zeman, 2019) enriches the basic UD annotation with deep syntactic annotations; the **CorefUD** project<sup>6</sup> (Nedoluzhko et al., 2022) is an initiative for harmonizing coreference corpora into a unified format.

The discourse annotation in the PDT-C 2.0 is also converted into the **Penn Discourse Treebank format** (Prasad et al., 2008; Mírovský et al., 2023) within the **Prague Discourse Treebank 4.0** (PDiT 4.0)<sup>7</sup> release (Mírovský and Synková, 2026).

The rich annotation at the t-layer serves as a source for conversion into various semantic and knowledge representations. Among earlier conversions, let us mention the transformation into the formal-logical format **Minimal Recursion Semantics** (Copestake et al., 2005; Jakob et al., 2010).

<sup>4</sup><http://hdl.handle.net/11234/1-5901>

<sup>5</sup><http://hdl.handle.net/11234/1-3720>

<sup>6</sup><http://hdl.handle.net/11234/1-5896>

<sup>7</sup><http://hdl.handle.net/11234/1-5680>

Most recently, the PDT-C 2.0 data have been converted into the **Uniform Meaning Representation format** (Van Gysel et al., 2021; Lopatková et al., 2024).<sup>8</sup>

## 8. Future Work

The description of language is far from complete. Despite the remarkable success of large language models (LLMs), we are still far from achieving systems that truly understand natural language, and fundamental linguistic research remains essential. In this respect, we aim to continue our efforts in systematically describing language from form to meaning. For the next version, PDT-C 3.0, the following annotation efforts have already been initiated: (i) event-type annotation (Urešová et al., 2025; Straková et al., 2026), (ii) implicit discourse relations (Zikánová et al., 2019), and (iii) fine-grained classification of (circumstantial) semantic roles, such as temporal, spatial, manner, and causal roles (Mikulová, 2024). These extensions will further enrich the corpus, enabling deeper linguistic analysis and providing a robust foundation for future NLP research and applications.

## 9. Conclusion

In the contribution, we present the Prague Dependency Treebank – Consolidated 2.0, a comprehensive, multi-layer linguistic resource that integrates semantic, syntactic, and morphological information, including inter-sentential phenomena such as coreference and discourse relations. The long-term development of the PDT framework has resulted in a uniformly annotated, genre-diversified corpus of almost 4 million Czech tokens, supported by fully compatible lexicons (morphological and valency). Its rich annotation makes it an invaluable tool for both linguistic research and the development of traditional and advanced NLP applications, as well as for conversion into many other formats across all levels of linguistic description. The corpus is freely available for use.

## 10. Limitations

While we present a large, genre-diversified, and richly annotated language resource, there are several limitations we are aware of. Some types of annotation are currently available only for a subset of the treebank — this applies in particular to grammemes (Sect. 5.9), as well as to bridging relations (Sect. 5.10) and multiword expressions (Sect. 5.11); cf. Tab. 2. Moreover, the treebank is monolingual. Although an English counterpart

exists, it is released separately as part of a parallel dataset (cf. Sect. 3.2). The new release of the Czech-English parallel treebank (PCEDT 3.0) is planned for late 2026.

The evaluation of predicted semantic graphs is currently based on MRP score (Oepen et al., 2020), which considers only the subset of annotation converted to the MRP format (Zeman and Hajič, 2020). Evaluating the full annotations will require to devise a new evaluation metric.

## 11. Acknowledgements

The research reported here has been supported by the Czech Science Foundation under the project 22-03269S. The work described herein has also been supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ.<sup>9</sup>

## 12. Bibliographical References

- Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2005. Minimal Recursion Semantics: An Introduction. *Research on Language and Computation*, 3(4):281–332.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. **Universal Dependencies**. *Computational Linguistics*, 47(2):255–308.
- Kira Drohanova and Daniel Zeman. 2019. **Towards Deep Universal Dependencies**. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 144–152, Paris, France. Association for Computational Linguistics.
- Jan Hajič, Eduard Bejček, Jaroslava Hlavačová, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. **Prague Dependency Treebank - Consolidated 1.0**. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5208–5218, Marseille, France. European Language Resources Association.
- Jan Hajič, Silvie Cinková, Marie Mikulová, Petr Pajas, Jan Ptáček, Josef Toman, and Zdeňka Urešová. 2008. PDTSL: An Annotated Resource For Speech Reconstruction. In *Proceedings of the 2008 IEEE Workshop on Spoken Language Technology*, pages 93–96, Goa, India.

<sup>8</sup><http://hdl.handle.net/11234/1-5951>

<sup>9</sup><https://lindat.cz>

- Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. [Announcing Prague Czech-English Dependency Treebank 2.0](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 3153–3160, Istanbul, Turkey. European Language Resources Association.
- Jan Hajič, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. 2003. PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In *Proceedings of The Second Treebanks and Linguistic Theories Workshop*, pages 57–68, Vaxjo, Sweden. Vaxjo University Press.
- Eva Hajičová, Jarmila Panevová, and Petr Sgall. 2000. [Coreference in Annotating a Large Corpus](#). In *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece. European Language Resources Association.
- Eva Hajičová, Petr Sgall, and Barbara Partee. 1998. *Topic-focus articulation, tripartite structures, and semantic content*. Kluwer Academic Publishers, Dordrecht, Boston.
- Jan Hajič. 1998. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In Eva Hajičová, editor, *Issue of Valency and Meaning*, pages 106–132. Karolinum, Prague.
- Eva Hajičová, Marie Mikulová, Barbora Štěpánková, and Jiří Mírovský. 2022. [Advantages of a Complex Multilayer Annotation Scheme: The Case of the Prague Dependency Treebank](#). In *Proceedings of the 16th Linguistic Annotation Workshop within LREC2022*, pages 70–78, Marseille, France. European Language Resources Association.
- Jaroslava Hlaváčová, Marie Mikulová, Barbora Štěpánková, Milan Straka, and Jan Hajič. 2026. MorfFlex: Handling Rich Morphology. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, Palma de Mallorca, Spain. European Language Resources Association.
- Max Jakob, Markéta Lopatková, and Valia Kordoni. 2010. [Mapping between Dependency Structures and Compositional Semantic Representations](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Valletta, Malta. European Language Resources Association.
- Markéta Lopatková, Eva Fučíková, Federica Gamba, Jan Štěpánek, Daniel Zeman, and Šárka Zikánová. 2024. [Towards a Conversion of the Prague Dependency Treebank Data to the Uniform Meaning Representation](#). In *Proceedings of the 24th Conference Information Technologies – Applications and Theory*, pages 62–76, Košice, Slovakia. CEUR-WS.org.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. [Building a Large Annotated Corpus of English: The Penn Treebank](#). *Computational Linguistics*, 19(2):313–330.
- Marie Mikulová. 2014. [Annotation on the Tectogrammatical Level. Additions to Annotation Manual \(with respect to PDTSC and PCEDT\)](#). Technical Report TR-2013-52, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic.
- Marie Mikulová. 2024. [Fine-grained Classification of Circumstantial Meanings within the Prague Dependency Treebank Annotation Scheme](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 7314–7323, Torino, Italia. ELRA and ICCL.
- Marie Mikulová, Eduard Bejček, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Pavel Straňák, Magda Ševčíková, and Zdeněk Žabokrtský. 2013. [From PDT 2.0 to PDT 3.0 \(Modifications and Complements\)](#). Technical Report TR-2013-54, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic.
- Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský. 2006. [Annotation on the Tectogrammatical Level in the Prague Dependency Treebank. Annotation Manual](#). Technical report, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic.
- Marie Mikulová, Jan Hajič, Alevtina Bémová, Eva Buráňová, Jiří Kárník, Petr Pajas, Jarmila Panevová, Barbora Štěpánková, Jan Štěpánek, and Zdeňka Urešová. 2026a. Manual for Annotation at Analytical Layer, Revision for the Prague Dependency Treebank — Consolidated 2024 release. Technical report, Institute of Formal and Applied Linguistics, Charles University, Prague.
- Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Petr Pajas, Jan Štěpánek, and Jan Hajič. 2017. PDTSC 2.0 - Spoken corpus with rich multi-layer

- structural annotation. In *Text, Speech, and Dialogue, 20th International Conference*, Lecture Notes in Computer Science, pages 129–137, Cham / Heidelberg / New York / Dordrecht / London. Springer.
- Marie Mikulová and Jan Štěpánek. 2010. [Ways of Evaluation of the Annotators in Building the Prague Czech-English Dependency Treebank](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Valletta, Malta. European Language Resources Association.
- Marie Mikulová, Barbora Štěpánková, and Jan Štěpánek. 2025. [From Form to Meaning: The Case of Particles within the Prague Dependency Treebank Annotation Scheme](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2163–2175, Abu Dhabi, UAE. Association for Computational Linguistics.
- Marie Mikulová, Barbora Štěpánková, Daniel Zeman, Jan Štěpánek, Milan Straka, and Jan Hajič. 2026b. [Meet UD\\_Czech-PDTC: A Large and Genre-Rich Treebank in Universal Dependencies](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, Palma de Mallorca, Spain. European Language Resources Association.
- Marie Mikulová, Milan Straka, Jan Štěpánek, Barbora Štěpánková, and Jan Hajič. 2022. [Quality and Efficiency of Manual Annotation: Pre-annotation Bias](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2909–2918, Marseille, France. European Language Resources Association.
- Marie Mikulová, Jaroslava Hlaváčová, Jan Hajič, Jiří Hana, Hana Hanová, Barbora Hladká, Barbora Štěpánková, and Daniel Zeman. 2020. [Manual for Morphological Annotation, Revision for the Prague Dependency Treebank - Consolidated 1.0](#). Technical Report TR-2020-64, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic.
- Jiří Mírovský, Pavlína Synková, Lucie Poláková, and Marie Paclíková. 2024. [Cost-Effective Discourse Annotation in the Prague Czech-English Dependency Treebank](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 4067–4077, Torino, Italia. ELRA and ICCL.
- Jiří Mírovský, Magdalena Rysová, Pavlína Synková, and Lucie Poláková. 2023. [Prague to Penn Discourse Transformation](#). *Prague Bulletin of Mathematical Linguistics*, 120:5–30.
- Jiří Mírovský and Pavlína Synková. 2026. [Presenting the Prague Discourse Treebank 4.0](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, Palma de Mallorca, Spain. European Language Resources Association.
- Anna Nedoluzhko and Jiří Mírovský. 2011. [Annotating Extended Textual Coreference and Bridging Relations in the Prague Dependency Treebank](#). Technical Report 44, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic.
- Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. [CorefUD 1.0: Coreference Meets Universal Dependencies](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872, Marseille, France. European Language Resources Association.
- Stephan Oepen, Omri Abend, Lasha Abzianidze, Johan Bos, Jan Hajič, Daniel Hershcovich, Bin Li, Tim O’Gorman, Nianwen Xue, and Daniel Zeman. 2020. [MRP 2020: The Second Shared Task on Cross-Framework and Cross-Lingual Meaning Representation Parsing](#). In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 1–22, Online. Association for Computational Linguistics.
- Stephan Oepen, Omri Abend, Jan Hajič, Daniel Hershcovich, Marco Kuhlmann, Tim O’Gorman, Nianwen Xue, Jayeol Chun, Milan Straka, and Zdeňka Urešová. 2019. [MRP 2019: Cross-Framework Meaning Representation Parsing](#). In *Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning*, pages 1–27, Hong Kong. Association for Computational Linguistics.
- Petr Pajas and Jan Štěpánek. 2008. [Recent Advances in a Feature-Rich Framework for Treebank Annotation](#). In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 673–680, Manchester, UK. Coling 2008 Organizing Committee.
- Jarmila Panevová and Magda Ševčíková. 2010. [Annotation of Morphological Meanings of Verbs Revisited](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Valletta, Malta. European Language Resources Association.
- Lucie Poláková, Pavlína Jínová, Šárka Zikánová, Zuzanna Bedřichová, Jiří Mírovský, Magdaléna

- Rysová, Jana Zdeňková, Veronika Pavlíková, and Eva Hajičová. 2012. [Manual for Annotation of Discourse Relations in Prague Dependency Treebank](#). Technical Report 47, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Milt-sakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. [The Penn Discourse TreeBank 2.0](#). In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, Marrakech, Morocco. European Language Resources Association.
- Magda Razímová and Zdeněk Žabokrtský. 2006. Annotation of Grammatemes in the Prague Dependency Treebank 2.0. In *Proceedings of the LREC Workshop on Annotation Science*, pages 12–19, Genova, Italy. European Language Resource Association.
- David Samuel and Milan Straka. 2020. [ÚFAL at MRP 2020: Permutation-invariant Semantic Parsing in PERIN](#). In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 53–64, Online. Association for Computational Linguistics.
- Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague/Dordrecht.
- 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 and Jana Straková. 2024. Open-Source Web Service with Morphological Dictionary–Supplemented Deep Learning for Morphosyntactic Analysis of Czech. In *27th International Conference on Text, Speech and Dialogue*, pages 279–290, Cham, Switzerland. Masaryk University, Springer.
- Milan Straka, Jana Straková, and Jan Hajič. 2019. [UDPipe at SIGMORPHON 2019: Contextualized Embeddings, Regularization with Morphological Categories, Corpora Merging](#). In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 95–103, Florence, Italy. Association for Computational Linguistics.
- Jana Straková, Eva Fučíková, Zdeňka Urešová, and Jan Hajič. 2026. Automatic Suggestions Help Extending Eventive Ontology: A Case Study on SynSemClass. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, Paris, France. European Language Resources Association.
- Jana Straková, Milan Straka, and Jan Hajič. 2014. [Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition](#). In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland. Association for Computational Linguistics.
- Zdeňka Urešová. 2012. Building the PDT-VALLEX valency lexicon. In *Proceedings of the 5th Corpus Linguistics Conference*, pages 1–18, Liverpool, UK. University of Liverpool.
- Zdeňka Urešová, Eva Fučíková, Cristina Fernández Alcaina, and Jan Hajič. 2025. Linking an Event-type Ontology to Morphosyntax of the Predicate-Argument Structure. *Dictionaries: Journal of the Dictionary Society of North America*, 46(1):207–227.
- Jens EL Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O’Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, Jan Hajič, James H. Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, Rosa Vallejos, and Nianwen Xue. 2021. [Designing a Uniform Meaning Representation for Natural Language Processing](#). *KI-Künstliche Intelligenz*, 35(3-4):343–360.
- Daniel Zeman and Jan Hajič. 2020. [FGD at MRP 2020: Prague Tectogrammatical Graphs](#). In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 33–39, Online. Association for Computational Linguistics.
- Šárka Zikánová, Jiří Mírovský, and Pavlína Synková. 2019. Explicit and Implicit Discourse Relations in the Prague Discourse Treebank. In *Proceedings of the 22nd International Conference on Text, Speech and Dialogue*, volume 11697 of *Lecture Notes in Computer Science*, pages 236–248, Cham / Heidelberg / New York / Dordrecht / London. University of West Bohemia, Springer International Publishing.
- Zdeněk Žabokrtský. 2011. Treex – an Open-source Framework for Natural Language Processing. In *Proceedings of the Information Technologies – Applications and Theory Conference*, pages 7–14.

### 13. Language Resource References

- Hajič, Jan and Bejček, Eduard and Bémová, Alevtina and Buráňová, Eva and Fučíková, Eva and Hajičová, Eva and Havelka, Jiří and Hlaváčová, Jaroslava and Homola, Petr and Ircing, Pavel and Kárník, Jiří and Kettnerová, Václava and Klyueva, Natalia and Kolářová, Veronika and Kučová, Lucie and Lopatková, Markéta and Mareček, David and Mikulová, Marie and Mírovský, Jiří and Nedoluzhko, Anna and Novák, Michal and Pajas, Petr and Panevová, Jarmila and Peterek, Nino and Poláková, Lucie and Popel, Martin and Popelka, Jan and Romportl, Jan and Rysová, Magdaléna and Semecký, Jiří and Sgall, Petr and Spoustová, Johanka and Straka, Milan and Straňák, Pavel and Synková, Pavlína and Ševčíková, Magda and Šindlerová, Jana and Štěpánek, Jan and Štěpánková, Barbora and Toman, Josef and Urešová, Zdeňka and Vidová Hladká, Barbora and Zeman, Daniel and Zikánová, Šárka and Žabokrtský, Zdeněk. 2024a. *Prague Dependency Treebank - Consolidated 2.0 (PDT-C 2.0)*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Charles University, Prague, Czech republic. PID <http://hdl.handle.net/11234/1-5813>.
- Hajič, Jan and Hlaváčová, Jaroslava and Mikulová, Marie and Straka, Milan and Štěpánková, Barbora. 2024b. *MorfFlex CZ 2.1 (2024-12-23)*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL). PID <http://hdl.handle.net/11234/1-5833>.
- Urešová, Zdeňka and Bémová, Alevtina and Fučíková, Eva and Hajič, Jan and Kolářová, Veronika and Mikulová, Marie and Pajas, Petr and Panevová, Jarmila and Štěpánek, Jan. 2024. *PDT-Vallex: Czech Valency lexicon linked to treebanks 4.5 (PDT-Vallex 4.5)*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL). PID <http://hdl.handle.net/11234/1-5814>.