

Meet UD_Czech-PDTC: A Large and Genre-Rich Treebank in Universal Dependencies

Marie Mikulová, Barbora Štěpánková, Daniel Zeman, Jan Štěpánek,
Milan Straka, Jan Hajič

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, Prague, Czechia
{mikulova,stepankova,zeman,stepanek,straka,hajic}@ufal.mff.cuni.cz

Abstract

Czech has been part of Universal Dependencies since its first release in 2015. It has also been one of the best represented languages, with the Prague Dependency Treebank being order of magnitude larger than most other UD treebanks. More recently, three other datasets from the Prague family were added and the annotations thoroughly revisited, forming the “Prague Dependency Treebank-Consolidated” (PDT-C). In comparison to the original PDT, PDT-C is more than twice as large, but it is also much more diverse in terms of genres and domains. In this paper, we describe the conversion of the new resource to Universal Dependencies. While the two annotation schemes are relatively similar at the first sight, there are numerous small differences in topology of the dependency structures and in granularity of the POS and relation type inventories. We demonstrate a selection of such differences on examples, discuss the diverging motivations, as well as ways to overcome the differences during conversion. We argue that while PDT is less “universal” and more tightly bound to one language, its multi-layer annotation is rich and provides all information needed for basic UD trees, and much more.

Keywords: universal dependencies, Czech, morphology, syntax, treebank

1. Introduction

An important application and promotion of an annotated corpus is its conversion into another formalism. The significance of such conversion has several aspects: (i) it can highlight similarities and differences in the frameworks, (ii) emphasize typological differences among languages in terms of their ability to incorporate or exclude specific linguistic phenomena, and (iii) integration into various concepts demonstrates the robustness and universality of the chosen format.

In our contribution, we aim to demonstrate the richness of linguistic annotation present in the Prague Dependency Treebank (Mikulová et al., 2026), linking morphology, syntax, and semantics. This is evidenced, among other things, by the fact that the treebank is used for conversions into various frameworks (e.g., Uniform Meaning Representation (Lopatková et al., 2024; Štěpánek et al., 2025), Minimal Recursion Semantics (Jakob et al., 2010), CorefUD (Nedoluzhko et al., 2022), Meaning Representation Parsing format used in CoNLL shared tasks (Oepen et al., 2019, 2020), the Penn Discourse Treebank format (Mírovský and Synková, 2026)). Here, we focus on its conversion to Universal Dependencies (de Marneffe et al., 2021) and elaborate the aspects (i)–(iii) listed above.

The **UD_Czech-PDTC** treebank (available first in UD version 2.16; Zeman et al., 2025) is a UD

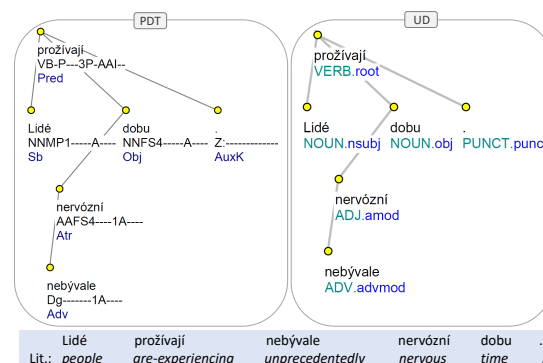


Figure 1: PDT sentence representation and its conversion to UD.

conversion of the Prague Dependency Treebank – Consolidated, released in 2024 (PDT-C 2.0; Hajič et al., 2024), and with 3440K words is now one of the largest treebanks in UD, containing texts of various genres. In addition to the pilot Prague Dependency Treebank of journalistic texts (enriched version from 2006; Hajič et al., 2006, converted to the first UD 1.0 version in 2015; Nivre et al., 2015), the consolidated PDT-C release includes three other PDT-corpora of spoken, translated, and user-generated content, all of which have undergone significant changes in surface syntax annotation, supplemented by complete manual annotation, using a uniform annotation scheme.

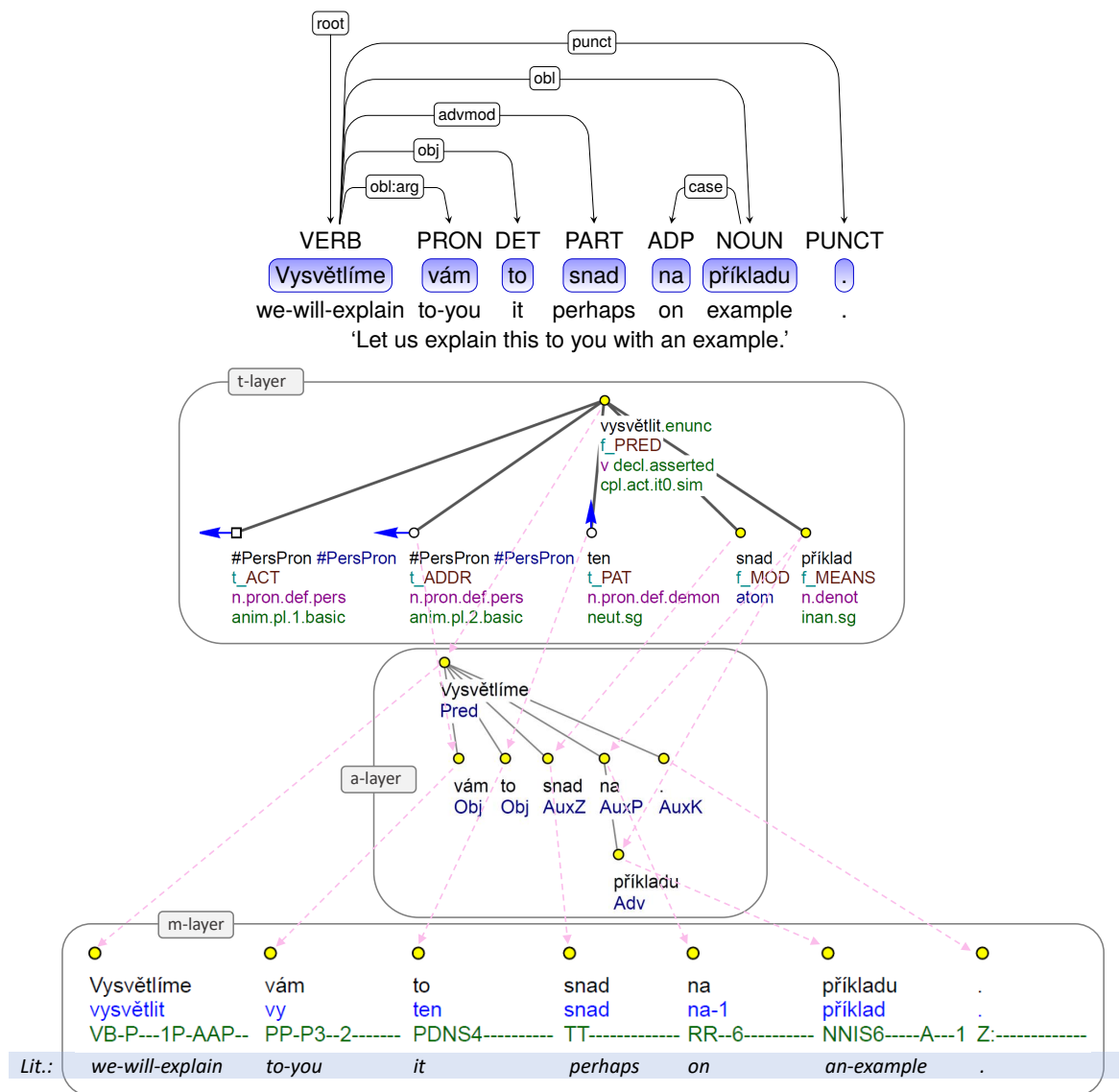


Figure 2: Representation of the Czech sentence (1) in UD (top) and PDT (bottom) annotation scheme.

In the contribution, we discuss the major differences between the two formalisms (theoretical foundations, approaches to syntax, morphology, and semantics). Using specific examples from various areas of language description (part-of-speech classification, coordination, ellipsis, semantic-pragmatic expressions), we discuss how these differences can influence data interpretation or linguistic research (the loss of specific phenomena, the adaptation of the format to new phenomena). Most importantly, we highlight that the morphosyntactic annotation in the PDT and UD frameworks is highly similar, as illustrated in Fig. 1.

The paper is organized as follows: the following subsections provide a brief description of both the frameworks (UD in 1.1, PDT in 1.2), the converted language (1.3), and history of PDT-to-UD conversions (1.4). The conversion is the focus of Sect. 2.

We discuss conversion of part-of-speech taxonomy (2.1), syntactic structure (2.2), and syntactic relations (2.3). Sect. 2.4 addresses the Enhanced UD. The conversion is summarized in Sect. 3. The Czech UDPipe model is presented in Sect. 4 and Sect. 5 concludes the paper.

1.1. Universal Dependencies

Universal dependencies (UD; de Marneffe et al., 2021) is a stunning project – framework for consistent morphosyntactic annotation across different languages. It is an open community effort, now with more than 700 contributors who created 319 treebanks in 179 languages. The general philosophy is to provide a universal inventory of categories and guidelines and to facilitate consistent annotation of similar constructions across languages, while allowing language-specific exten-

sions when necessary. This effort is a good basis for cross-linguistically consistent annotation of typologically diverse languages in a way that supports computational natural language understanding as well as broader linguistic studies. The detailed specification of UD annotation is available on the project website¹ and the overview can be also found in Nivre et al. (2020) or de Marneffe et al. (2021). Fig. 2 demonstrates the UD annotation scheme on the Czech example (1).

- (1) *Vysvětlíme vám to snad na příkladu.*
we-will-explain to-you it perhaps on an-example.
'Let us explain this to you with an example.'

The UD framework is based on a lexicalist view of syntax, which means that dependency relations hold between words, and that morphological features are encoded as properties of words with no attempt at segmenting words into morphemes. The **morphological specification** of a word in the UD scheme consists of three types of information: a lemma representing the base form of the word, a part-of-speech (POS) tag representing the grammatical category of the word, a set of features representing lexical and grammatical properties associated with the particular word form. For simplicity, only POS annotation is shown in Fig. 2.

Syntactic annotation consists of labeled dependency relations between words. The basic representation forms a tree rooted in one word, normally the main sentence predicate (the verb *vysvětlíme* 'we-will-explain' in Fig. 2), on which other words of the sentence (including punctuation) depend. The syntactic analysis gives priority to predicate-argument and modifier relations that hold directly between content words, as opposed to being mediated by function words, so all function words including copula are uniformly treated as dependents (cf. treatment of prepositional phrase *na příkladu* 'on example' in Fig. 2).

In addition to the basic representation, which is obligatory for all UD treebanks, it is possible to provide an **enhanced representation** (Nivre et al., 2020), which adds (and in a few cases changes) relations in order to give a more complete basis for semantic interpretation (details in Sect. 2.4).

1.2. Prague Dependency Treebank

The long-run project of Prague Dependency Treebank (PDT; Mikulová et al., 2026) is unique in its attempt to systematically cover and link different layers of language description including a rich semantic representation. Its hierarchical multi-layer architecture is based on the theory of Functional Generative Description (Sgall et al., 1986) and is

described in several detailed annotation manuals available from the project web site.²

The PDT annotation scheme is illustrated in Fig. 2 on the same Czech example (1) as we used for UD. In Fig. 2, each annotation layer of the PDT-system is displayed in a separate box. The links between the layers are indicated by the pink dashed arrows. The original raw text is stored at the lowest layer of the system (and it is not shown in Fig. 2). Above the raw text layer, there are three layers of annotations: morphological, surface syntactic, and deep syntactic.

Czech is a highly inflectional language (see Sect. 1.3). At the **morphological layer** (*m-layer* box in Fig. 2), each word form is described by a 15-character tag that specifies its lexical and grammatical properties. All tokens are also assigned a POS category within the tag (the first position).

Above the linearly structured m-layer, there are two syntactic layers, the **analytical** (*a-layer*) reflecting surface dependency structure and the **teletogrammatical** (*t-layer*) reflecting the deep syntactic structure. At the a-layer, which is closest to the UD representation, a syntactic structure is captured by a rooted directed tree structure with the specification of the head for each word and the assignment of a label describing dependency syntactic relation (such as subject, object, adverbial). As in UD, the root of the tree is typically the predicate of a sentence.

At the a-layer, a distinction is made between different groups of function words. The key division is between function words that operate within verbal complexes and encode morphosyntactic properties of verbs (i.e., auxiliaries), and those that are part of nominal groups (prepositions) or link clauses into a single unit (conjunctions). Auxiliaries are analyzed as dependents of the verb they "belong" to, whereas prepositions and conjunctions are treated as the heads of the nouns or clauses whose form they "govern"; cf. treatment of *na příkladu* 'on example' in Fig. 2.

The t-layer captures rich semantic annotations of a sentence: predicate-argument structure, semantic classification of adjuncts, semantic counterparts of morphological categories, topic-focus articulation, information structure, coreference, ellipsis, and discourse relations. The main difference between the two PDT syntactic layers is how words correspond to nodes: at the a-layer, every word of the raw text (including punctuation) is represented by a tree node and no additional nodes are allowed (this approach is the same as in UD), while the t-layer consists of nodes that only represent content words (with some exceptions such as coordination); function words such as prepositions, auxiliary verbs, etc. are not present. There is for exam-

¹<https://universaldependencies.org/>

²<https://ufal.mff.cuni.cz/pdt-c>

ple only one node for the prepositional phrase *na příkladu* ‘on example’ at t-layer annotation in Fig. 2. The contribution of function words to the meaning of the sentence is captured within the complex labels of the content word nodes – see many values attached to the nodes at the t-layer. At the t-layer, new nodes are also added for semantic units deleted on the surface; in Fig. 2 the restoration of a deletion is illustrated by the #PersPron (personal pronoun) node for the unexpressed subject (ACT) of the sentence predicate.

The latest version, **Prague Dependency Treebank – Consolidated 2.0** (Hajič et al., 2024)³ is a manually annotated, genre-diversified, consolidated release of the existing PDT-corpora of Czech data, uniformly annotated using the PDT scheme described above. PDT-C includes four datasets: Prague Dependency Treebank (written newspaper and journal texts); Czech part of Prague Czech-English Dependency Treebank (business news, translated from English), Prague Dependency Treebank of Spoken Czech (spoken data); PDT-Faust (user-generated texts). For the PDT-C 2.0 version, manual annotation at the a-layer is performed in those parts of the corpus that were previously annotated only by automatic tools. Unlike previous versions, PDT-C 2.0 provides fully manual annotation at all three annotation layers in all datasets, which ensures that its UD conversion has the quality required for official UD releases.

1.3. Czech

Czech is a West Slavic language spoken by ca. 10 million people, most of them living in Czechia. As a national language, it is also one of the official languages of the European Union. Czech is an inflectional language characterized by a complex system of nominal declension and verbal conjugation; its word order is very flexible, driven by pragmatic factors rather than syntax. It uses a Latin-based alphabet supplemented with diacritical marks for long vowels and palatalized consonants. In the field of NLP/CL, Czech cannot be considered a low-resource language, not least thanks to the Prague Dependency Treebank, which was in fact a pioneering effort and one of the earliest treebanks ever created. There is also a wealth of other corpora (esp. *Czech National Corpus*) and resources available, including morphological analyzers, valency lexicons, or the Czech *WordNet*.

1.4. History of PDT to UD Conversion

The PDT data has been part of UD since the very beginning. The first release of UD (v1.0; Nivre et al., 2015) in 2015 contained treebanks of 10

³<http://hdl.handle.net/11234/1-5813>

languages, including conversion of Prague Dependency Treebank 2.0 from 2006 (Hajič et al., 2006). In 2017, UD version 2.0 (Nivre and all, 2017) was published, which differs significantly from the previous releases. Changes in the guidelines **from v1 to v2** are summarized in Nivre et al. (2020), referring to release 2.5, covering 90 languages.

The consolidated release of **PDT-C 2.0**, containing 3440K words, appears in UD for the first time in version 2.16 (Zeman et al., 2025). Together with UD_German-HDT (Borges Völker et al., 2019), they are the largest treebanks in UD.

The first treebank featuring the **Enhanced UD** representation was the English Web Treebank (Schuster and Manning, 2016) in version 2.2 from 2018. PDT appeared in the enhanced UD format for the first time in version 2.6 (2020). The most recent conversion of PDT-C 2.0 also includes enhanced UD relations, which are based on the original manual annotation. **UD_Czech-PDTC** is one of the 19 UD treebanks that have all six types of enhancements defined in the guidelines.

2. PDT to UD Conversion

As shown in Fig. 1,⁴ the representation of a sentence in PDT and its counterpart in UD can be nearly identical, disregarding the different naming conventions for POS tags and syntactic relations. This similarity arises from the fact that the original UD proposal was heavily influenced by Indo-European languages and Czech belongs to this family. Both frameworks are based on dependency syntax, in contrast to, e.g., constituency-based syntax, which is traditional in American linguistics (Chomsky, 1957). They both treat the predicate as the head of the syntactic structure. In the basic surface syntactic representation (the a-layer in PDT and basic representation in UD), every word in the sentence, including punctuation, has a corresponding node in the tree, and no extra nodes are allowed. The core inventories of part-of-speech tags and syntactic relations also largely overlap (cf. POS annotation in Fig. 1: VERB in UD vs. V in PDT for verbs; NOUN vs. N for nouns, ADJ vs. A for adjectives, ADV vs. D for adverbs, PUNCT vs. Z for punctuation; and syntactic relations annotation: root in UD vs. Pred in PDT for the tree root; nsubj vs. Sb for subject; obj vs. Obj for object, advmod vs. Adv for adverbial).

⁴There are various approaches to visualizing dependency annotations. The standard visualization of PDT is provided by the TrEd tool (<https://ufal.mff.cuni.cz/tred/>). For simplicity, we display m-layer tags in a-layer tree. UD trees are typically shown as arcs over a linear sentence, which we mostly follow except for the side-by-side comparison of the two frameworks in Fig. 1.

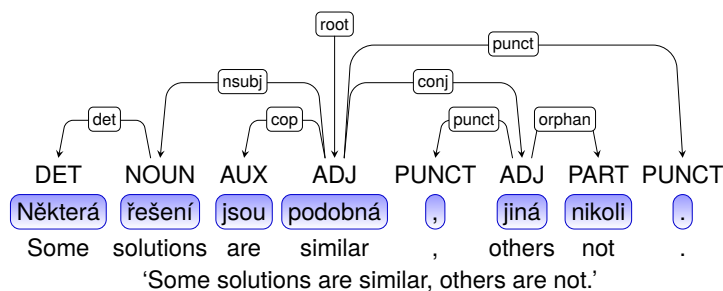


Figure 3: UD representation of sentence (2).

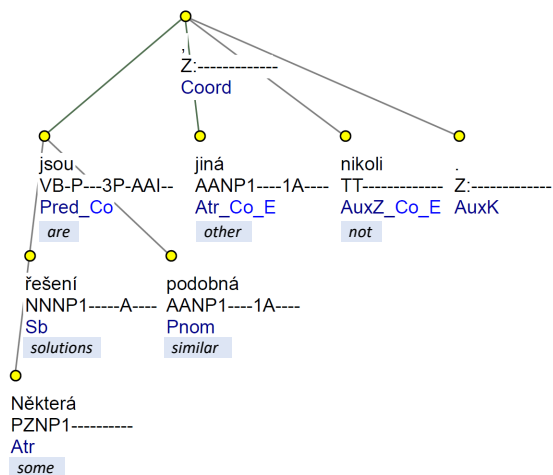


Figure 4: PDT representation of sentence (2).

The main difference is that PDT uses a multi-layer scheme separating linguistic information, while UD integrates all (mainly morphosyntactic) information into a single graph. In the following sections, we primarily compare the UD representation and PDT representation at the m-layer (the basis for POS and features conversion) and the a-layer (the basis for conversion of syntactic structure and relations).⁵ To illustrate the PDT-to-UD conversion process, we have selected several ‘clean’ cases where the conversion is clear and consistent, as well as several more complex instances where pragmatic decisions often need to be made that may result in either over- or under-interpretation of the original syntactic structure.

2.1. Part-of-Speech Categories

There are 17 universal POS categories in UD, including determiner (DET) and auxiliary (AUX), which are not part of the Czech/PDT POS taxon-

⁵Annotation at the m/a-layer in PDT may in some respects appear less fine-grained than in UD. However, there is also t-layer and differentiation of some distinctions introduced in UD is performed up to this layer.

omy.⁶ During the conversion, we adopted a strategy of tagging words that satisfy the UD definitions of determiners and auxiliaries, aiming for universality and facilitating cross-linguistic comparison.

2.1.1. Determiners

The notion of determiners (DET) is unknown in traditional Czech grammars (in Czech, there are no articles as in English). Words equivalent to English determiners are traditionally classified as pronouns, adjectives and/or numerals. Nevertheless, traditional pronouns can be divided to subclasses with distinct morphosyntactic behavior, namely those that resemble adjectives and those that do not. ‘Adjective-like pronouns’ can modify nominals and inflect for gender (and also number and case) to express agreement with the modified noun. Many of them can also stand alone as nominal heads. However, since UD v2 it is not necessary to distinguish them solely by context. We can pre-categorize words in the dictionary, and that is what we do in this case. If a ‘traditional pronoun’ can show gender agreement with a modified noun, we tag it DET regardless of whether it modifies or heads a nominal (e.g., the demonstrative *to* ‘the/it’ is tagged as a pronoun (P) in PDT, and it has always the DET category in UD – even if it functions as an object as in (1); cf. Fig. 2).

2.1.2. Auxiliaries

According to UD definition, an auxiliary (AUX) is a function word that accompanies the lexical verb and expresses grammatical distinctions not carried by the lexical verb, such as person, number, tense, mood, aspect, voice or evidentiality (so-called TAMVE markers). Since UD v2 the AUX category also includes the copula. In Czech, all these functions are fulfilled by the verb *být* ‘to be’.

In PDT, all occurrences of *být* ‘to be’ are uniformly tagged as verbs and the context-based dis-

⁶There is also PROP category used for proper names. In PDT, proper names are not addressed within the POS category but rather encoded as specific information in the lemma of the given word at the m-layer.

inction between auxiliary and full verb is left for the syntactic annotation. In the PDT to UD conversion, we took the side of pre-categorizing this verb as *AUX* in dictionary, on the grounds that only a small minority of cases could pass as full verbs by UD guidelines. Like in the PDT framework, the distinction is reflected in syntactic relations rather than the POS tag.

In **copular construction**, PDT treats the copula *být* ‘to be’ as the head due to the strong agreement principle in Czech; the nominal predicate is assigned the *P_{NOM}* relation. In UD, the relation is inverted and labeled *cop*. The different treatment of copula in PDT and UD is illustrated by the example *řešení jsou podobná* ‘solutions are similar’ (2) in Fig. 4 and 3, respectively.

- (2) *Některá řešení jsou podobná, jiná nikoli.*
 some solutions are similar, others not.
 ‘Some solutions are similar, others are not.’

2.2. Dependency Structure

In terms of capturing the dependency structure of a sentence, the main differences between the PDT and UD frameworks, aside from the treatment of function words,⁷ are coordination and ellipsis.

2.2.1. Coordination

Coordination is a complicated phenomenon in any dependency formalism (Popel et al., 2013). In PDT, a coordinating conjunction (cf. Fig. 6) or a punctuation mark (cf. Fig. 4) is made a “technical head” of the coordination and gets the label *Coord* as a marker of the coordination construction. The conjuncts are formally represented as its dependents and they each carry a flag (displayed as the *_Co* suffix in tree diagrams) to distinguish them from real dependents; cf. the PDT representation of (3) with two coordinate predicates in Fig. 6.

In UD, there is a different strategy to capture coordination. The first conjunct is treated as the parent (or “technical head”) of all following conjuncts via the *conj* relation. Coordinating conjunctions and punctuation delimiting the conjuncts are attached to their associated conjuncts using the *cc* and *punct* relations, respectively. See Fig. 5 depict the UD representations of sentence (3).

- (3) *Mají či nemají pravdu?*
 they-have or they-have-not truth?
 ‘Are they right or wrong?’

⁷The status of function words in various dependency frameworks is the subject of a special issue of *Linguistic Analysis* journal. The treatment of function words in the UD framework is described in de Marneffe et al. (2024) and in the PDT framework in Hajičová et al. (2024).

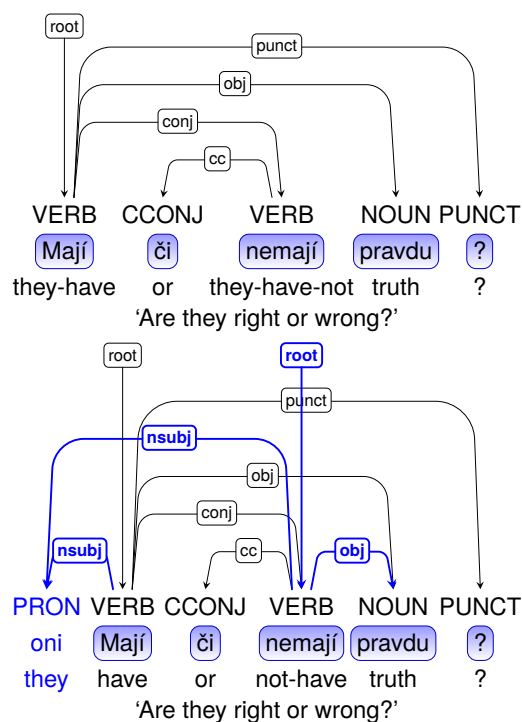


Figure 5: Basic (top) and enhanced (bottom) UD representation of sentence (3).

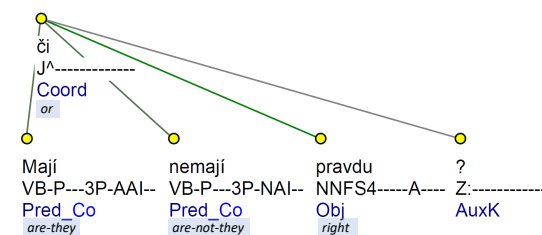


Figure 6: PDT representation of sentence (3).

Despite the different representation of coordination, the PDT to UD conversion is relatively straightforward. However, coordination may form very intricate structures when combined with ellipsis.

2.2.2. Ellipsis

Due to the adopted restrictions (esp. not allowing extra nodes to be added to the graph), the representation of ellipsis is problematic in both frameworks. When the governor of a dependent is omitted, it is impossible to construct a transparent dependency tree.

In PDT, an “orphan” dependent receives the label it would bear if the ellipsis were absent, but with an ellipsis flag (displayed as the *_E* suffix in diagrams), explicitly indicating that the node lacks its governor. The orphan is placed in the position that would have been occupied by the elided governor (recursively, if necessary, through further elided

governors higher up). See Fig. 4, there are two orphans in the second clause.⁸

In UD, the promoting strategy is applied: an orphan is “promoted” to the role of the governor and inherits the governor’s dependency label. This strategy is applied primarily in nominals. E.g., in (2), the noun *řešení* ‘solutions’ is elided in the nominal *jiná [řešení]* ‘other [solutions]’ in the second clause. The orphan *jiná* ‘other’ has been promoted to the role of the noun, i.e., to the role of the subject. However, in the second clause the predicate is also elided, thus the promoted subject *jiná* ‘other’ is consequently promoted to the head of the clause and, via a `conj` relation, it is conjoined with the predicate in the first clause. In such cases (when the elided element is a predicate and the promoted element is one of its arguments or adjuncts), the `orphan` relation is used to attach other orphans to the promoted head. This is the case of the negative particle *nikoli* ‘not’ in (2); cf. Fig. 3.

The PDT and UD frameworks adopt different approaches to representation of ellipsis. PDT provides more detailed information, in particular through the explicit marking of ellipsis and the preservation of the proper syntactic relation. This is a novelty in PDT-C 2.0 compared to earlier releases of the Prague treebanks, and it enables a more accurate conversion into UD.

2.3. Dependency Relations

The smallest overlap between the frameworks is in the inventory of syntactic relations. In PDT, syntactic relations are divided into surface syntactic (annotated at the a-layer, 23 types are distinguished) and deep syntactic ones (t-layer, over 50 types).

In UD, 37 dependency relations are defined. The relations are specified syntactically but, unlike in PDT, they are more related to dependent POS category. E.g., in UD, there are two types of relations for dependents that functionally correspond to adverbials: the `obl` relation is for nominal/prepositional phrases (such as *na příkladu* ‘on an-example’ in (1)) and `advmod` for adverbs (such as *názorně* ‘illustratively’ in *vysvětlit to názorně* ‘to-explain it illustratively’). In PDT, all adverbials are uniformly labeled as `Adv` relations (Fig. 2).

Some additional phenomena, especially deep syntactic ones, e.g., agent and patient in passives, are handled in UD by means of optional subtypes of the basic relations. The extent to which the conversion can model such phenomena is limited by the fact that it primarily relies on information from the a- and m-layers. E.g., passives are detected by the morphological annotation of passive partici-

⁸Note that ellipses are primarily resolved at the t-layer in PDT, allowing the node insertion (see more in Mikulová, 2014; Hajičová et al., 2015).

ples; if they have a dependent with `Case=Ins`, it is a possible candidate for the oblique agent, which is normally expressed as instrumental in Czech.

2.3.1. Pragmatic Elements

In both projects, some solutions have to be adopted for the words that are primarily characterised by pragmatic properties and are not typical dependents, i.e., words with dominant emphasizing, epistemic, assessment etc. function.

In the PDT framework, these words are POS-tagged as ‘particles’ following the Czech linguistic tradition.⁹ Although not understood as parts of surface syntax, they are placed at the a-layer with the label for pragmatic expression `AuxZ` (Mikulová et al., 2025).

In general UD guidelines, no special attention is paid to these expressions and they are mainly merged with adverbs (`ADV`). The type of dependency relation also varies – depending on the POS category of both the dependent and the parent, it typically involves the `advmod` dependency relation (indistinguishable from ordinary adverbials). Such mixing of adverbs and other expressions may be considered problematic (Cecchini, 2024).

In UD_Czech-PDTC, pragmatic expressions are labeled as follows: the category of POS is `PART`¹⁰ and for the syntactic relation `advmod:emph` is used (a specific variant of `advmod` signaling pragmatic item). See *snad* ‘perhaps’ in (1) and Fig. 2. This solution is consistent with the general principles of UD and also captures additional information from PDT.

2.4. Enhanced Universal Dependencies

UD also provides guidelines for so-called enhanced graphs. The main reason for adopting Enhanced UD is to make some of the implicit relations between words more explicit: especially in coordination or ellipsis constructions, the dependency path between two content words can be very long or not indicated at all. Therefore, the enhanced graph includes additional direct dependencies. Besides additional relations, enhanced graphs may also contain additional abstract nodes that represent elided material.

Cf. Fig. 5: in the basic tree, the dependency between the predicate *nemají* ‘they-have-not’ and the shared object *pravdu* ‘truth’ is not marked, whereas in the enhanced graph it is. There is also an abstract node for elided subject.

⁹This approach is widely used in Slavic and Central European linguistics, cf. Rozumko (2016).

¹⁰The UD guidelines define `PART` as a function word that does not fit the definitions of other parts of speech, e.g., the negation particle.

Test Data	Training Data		UPOS	XPOS	UFeats	AllTags	Lemmas	UAS	LAS
	PDT	PDT-C							
PDT-C: Overall	✓	✗	99.01	97.96	98.57	97.71	99.35	95.36	93.97
	✗	✓	99.27	98.45	98.92	98.26	99.57	96.20	95.06
PDT-C: Faust	✓	✗	98.91	96.14	96.51	95.76	98.12	86.26	82.74
	✗	✓	99.50	97.09	97.46	96.88	99.21	91.61	89.84
PDT-C: PCEDT	✓	✗	99.22	98.17	98.46	97.95	99.53	95.78	94.97
	✗	✓	99.68	98.89	99.00	98.76	99.78	97.07	96.40
PDT-C: PDT	✓	✗	99.04	98.19	98.81	97.96	99.42	96.18	94.89
	✗	✓	99.03	98.25	98.86	98.02	99.48	96.28	95.04
PDT-C: PDTSC	✓	✗	98.80	97.35	98.22	97.02	99.06	93.44	91.54
	✗	✓	99.50	98.66	99.09	98.53	99.63	95.55	94.33

Table 1: Comparison of UDPipe morphosyntactic performance in percents using either whole PDT-C or just PDT as training data, evaluated on whole PDT-C and also its four subsets.

In **UD_Czech-PDTC**, all 6 enhancement types defined by UD guidelines (Nivre et al., 2020) are present. When converting to UD, structural enhancements are mostly derivable from the annotation on the a-layer (e.g., the way coordination is represented allows explicit capturing of a dependent shared by the conjuncts: the shared dependent is captured in the PDT framework as a child of the head of the coordination structure. It does not carry the `_Co` flag, which signals that it is a shared dependent of the conjuncts, rather than another conjunct; cf. Fig. 6).

3. Conversion Highlights

A number of factors can be identified when transferring data from one framework to another. According to our analysis, the following three types seem to be the most crucial:

Framework. Both frameworks are based on the dependency concept, but their designs differ for certain phenomena. This may result from different conventions or from distinct focuses (the design of PDT is based on Czech, while UD aims for universality). This concerns the treatment of function words, coordination, and ellipsis (cf. Sect. 2.3). The differences here lie in the mode of representation rather than in the interpretation of the linguistic phenomena themselves, making conversion straightforward in such cases.

Language. Differences also arise from the language being converted—in our case, Czech. The language’s specific features must be mapped onto the universal UD categories in some way. E.g., Czech lacks articles, and more generally, determiners, but the `DET` category in UD is useful for distinguishing attributive pronouns (cf. Sect. 2.1). These differences are addressed individually during the conversion process, typically through explicit mapping rules or carefully curated rule lists.

Theory. Different linguistic traditions may influence the difference in frameworks. For instance, the distinction between adverbs and particles, common in the Czech linguistic tradition (and therefore also in PDT), is not universally accepted. However, UD enables such specifics to be captured, which can be beneficial for processing other languages or for language research in general.

4. UDPipe UD_Czech-PDTC Model

To quantify the effect of larger and more diverse morphosyntactic data, we train a UDPipe parser (Straka et al., 2019; Straka, 2018) using both the whole UD_Czech-PDTC 2.16 train data and also just their PDT subset. The evaluation of the two parsers on the whole test set and also its four subsets is presented in Table 1.

On the whole test set, using the complete training data reduces morphological tagging errors by 24%, lemmatization errors by 33%, and syntactic parsing errors by 18%. Given that performance on the PDT subset is improved only marginally, the error reduction in the other subsets (i.e., in other genres and domains) is therefore even higher.

We release the model trained on the UD_Czech-PDTC data under the CC BY-NC-SA license at <http://hdl.handle.net/11234/1-6117>.

5. Conclusion

We have presented the UD conversion of Prague Dependency Treebank – Consolidated 2.0, currently the largest manually annotated treebank of a Slavic language, and one of the largest UD treebanks in general. This is a significant milestone—until release 2.15, UD contained the pilot PDT (1.5M words), which has now been joined by three other treebanks from the Prague family (totaling

3.5M words). The expansion also leads to higher genre diversity (news, popular science, finance, spoken dialogues, user-generated content).

The annotation scheme of PDT is heavily influenced by Czech and the Czech linguistic school (e.g. the treatment of particles, pronouns vs. determiners, ellipsis). Unlike UD, PDT does not aspire to be “universal”, i.e., to suit languages typologically different from Czech without modification. Nevertheless, the two frameworks share many similarities, and the rich, multi-layer annotation of PDT provides enough information for most distinctions required in UD; therefore, the conversion can be performed in high quality. Optional features, relation subtypes and enhanced UD graphs are employed to preserve as much from the PDT annotation as possible.

UD_Czech-PDTC has led to more precise and robust parsing models. The UDPipe parser is currently used for annotating other resources, such as the parallel texts of InterCorp (Rosen, 2023), and Old Czech Texts (Zeman et al., 2023).

Limitations

While the core of PDT-C is richly annotated at all three layers (m-layer for morphology, a-layer for surface syntax and t-layer for deep syntax), about one fifth of the corpus lacks the t-layer. The conversion procedure thus cannot rely on t-layer being available (or, alternatively, we could convert only 80% of the data to UD). The t-layer could potentially help with certain phenomena, in particular with reconstruction of elided predicates in Enhanced UD, but this potential is not exploited at present.

Acknowledgments

The research reported here has been supported by the LINDAT/CLARIAH-CZ Research Infrastructure (<https://lindat.cz>), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

Bibliographical References

- Emanuel Borges Völker, Maximilian Wendt, Felix Hennig, and Arne Köhn. 2019. **HDT-UD: A very large Universal Dependencies treebank for German**. In *Proceedings of the Third Workshop on Universal Dependencies*, pages 46–57, Paris, France. Association for Computational Linguistics.
- Flavio Massimiliano Cecchini. 2024. **Let’s Do It Orderly: A Proposal for a Better Taxonomy of Adverbs in Universal Dependencies, and Beyond**. *The Prague Bulletin of Mathematical Linguistics*, 121:15–65.
- Noam Chomsky. 1957. *Syntactic Structures*. Mouton, The Hague.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. **Universal Dependencies**. *Computational Linguistics*, 47(2):255–308.
- Marie-Catherine de Marneffe, Joakim Nivre, and Daniel Zeman. 2024. **Function Words in Universal Dependencies**. *Linguistic Analysis*, 43(3-4):549–588.
- Eva Hajičová, Marie Mikulová, and Jarmila Panevová. 2015. **Reconstructions of Deletions in a Dependency-based Description of Czech: Selected Issues**. In *Proceedings of the Third International Conference on Dependency Linguistics*, pages 131–140. Uppsala University, Uppsala, Sweden.
- Eva Hajičová, Jarmila Panevová, Marie Mikulová, and Jan Hajič. 2024. **Function Words in Praguian Functional Generative Description**. *Linguistic Analysis*, 43(3-4):465–512.
- 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.
- Marie Mikulová. 2014. **Semantic Representation of Ellipsis in the Prague Dependency Treebanks**. In *Proceedings of the 26th Conference on Computational Linguistics and Speech Processing*, pages 125–138, Jhongli, Taiwan. Association for Computational Linguistics and Chinese Language Processing.
- Marie Mikulová, Jiří Mírovský, Milan Straka, Pavlína Synková, Barbora Štěpánková, Jan Štěpánek, and Jan Hajič. 2026. **Prague Dependency Treebank - Consolidated 2.0: Enriching a Complex Annotation Scheme**. In *Proceedings*

- of the Fifteenth Language Resources and Evaluation Conference, Palma de Mallorca, Spain. 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.
- 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, 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.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. [Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, 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.
- Martin Popel, David Mareček, Jan Štěpánek, Daniel Zeman, and Zdeněk Žabokrtský. 2013. [Coordination Structures in Dependency Treebanks](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 517–527, Sofia, Bulgaria. Association for Computational Linguistics.
- Alexandr Rosen. 2023. [The InterCorp parallel corpus with a uniform annotation for all languages](#). *Jazykovedný časopis / Journal of Linguistics*, 74(1):254–265.
- Agata Rozumko. 2016. Linguistic Concepts across Languages: The Category of Epistemic Adverbs in English and Polish. *Yearbook of the Poznan Linguistic Meeting*, 2(1):195–214.
- Sebastian Schuster and Christopher D. Manning. 2016. [Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 2371–2378, Portorož, Slovenia. European Language Resources Association.
- 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.
- Jan Štěpánek, Daniel Zeman, Markéta Lopatková, Federica Gamba, and Hana Hledíková. 2025. [Comparing Manual and Automatic UMRs for Czech and Latin](#). In *Proceedings of the Sixth International Workshop on Designing Meaning Representations*, pages 1–12, Prague, Czechia. Association for Computational Linguistics.
- 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, 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.
- Daniel Zeman, Pavel Kosek, Martin Březina, and Jiří Pergler. 2023. [Morphosyntactic annotation in Universal Dependencies for old Czech](#). *Jazykovedný časopis / Journal of Linguistics*, 74(1):214–222.

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. 2024. *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 Panevová, Jarmila and Hajičová, Eva and Sgall, Petr and Pajas, Petr and Štěpánek, Jan and Havelka, Jiří and Mikulová, Marie and Žabokrtský, Zdeněk and Ševčíková-Razímová, Magda and Urešová, Zdeňka. 2006. *Prague Dependency Treebank 2.0*. LDC2006T01, Linguistic Data Consortium, Philadelphia, PA, USA, ISLRN 942-053-729-014-3.
- Nivre, Joakim and all. 2017. *Universal Dependencies 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/11234/1-1983>.
- Nivre, Joakim and Bosco, Cristina and Choi, Jinho and de Marneffe, Marie-Catherine and Dozat, Timothy and Farkas, Richárd and Foster, Jennifer and Ginter, Filip and Goldberg, Yoav and Hajič, Jan and Kanerva, Jenna and Laippala, Veronika and Lenci, Alessandro and Lynn, Teresa and Manning, Christopher and McDonald, Ryan and Missilä, Anna and Montemagni, Simonetta and Petrov, Slav and Pyysalo, Sampo and Silveira, Natalia and Simi, Maria and Smith, Aaron and Tsarfaty, Reut and Vincze, Veronika and Zeman, Daniel. 2015. *Universal Dependencies 1.0*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Charles University, Prague, Czech republic. PID <http://hdl.handle.net/11234/1-1464>.
- Zeman, Daniel and Nivre, Joakim and Abrams, Mitchell and Ackermann, Elia and Adolphe, Jephthé and Aeppli, Noëmi and Aghaei, Hamid and Agić, Željko and Ahmadi, Amir and Ahrenberg, Lars and Ajede, Chika Kennedy and Akhundjanova, Arofat and Akkurt, Furkan and Aleksandravičiūtė, Gabrielė and Alfina, Ika and Algom, Avner and Alnajjar, Khalid and Alzetta, Chiara and Anastasopoulos, Antonios and Andersen, Erik and Andrews, Matthew and Antonsen, Lene and Aoyama, Tatsuya and Aplonova, Katya and Aquino, Angelina and Aragon, Carolina and Aranes, Glyd and Aranzabe, Maria Jesus and Arican, Bilge Nas and Arnardóttir, Pórunn and Arutie, Gashaw and Arwidarasti, Jessica Naraiswari and Asahara, Masayuki and Ásgeirsdóttir, Katla and Aslan, Deniz Baran and Asmazoğlu, Cengiz and Ateyah, Luma and Atmaca, Furkan and Attia, Mohammed and Atutxa, Aitziber and Augustinus, Liesbeth and Avelās, Mariana and Badmaeva, Elena and Bajorat, Jana and Balasubramani, Keerthana and Ballesteros, Miguel and Banerjee, Esha and Bank, Sebastian and Barbosa, Bryan Khelven da Silva and Barbu Mititelu, Verginica and Barkarson, Starkaður and Basile, Rodolfo and Basmov, Victoria and Batchelor, Colin and Bauer, John and Bedir, Seyyit Talha and Behzad, Shabnam and Belieni, Juan and Bémová, Alevtina and Bengoetxea, Kepa and Benli, İbrahim and Ben Moshe, Yifat and Benzerrak, Marie and Berg, Ansu and Berk, Gözde and Bhat, Riyaz Ahmad and Biagetti, Erica and Bick, Eckhard and Bielskienė, Agnė and Bilgin Taşdemir, Esma Fatima and Binici, Helin and Bjarnadóttir, Kristín and Blaschke, Verena and Blokland, Rogier and Böbel, Nina and Bobicev, Victoria and Boizou, Loïc and Bompolas, Stavros and Bonilla, Johnatan and Borges Völker, Emanuel and Börstell, Carl and Bosco, Cristina and Bouma, Gosse and Bowman, Sam and Boyd, Adriane and Braggaa, Anouck and Branco, António and Bras, Myriam and Brokaitė, Kristina and Bu, Lanni and Buráňová, Eva and Burchardt, Aljoscha and Cabeza, Carmen and Cáceres Arandia, Natalia and Campos, Marisa and Candito, Marie and Caron, Bernard and Caron, Gauthier and Carneiro, Catarina and Carvalho, Rita and Cassidy, Lauren and Castro, Maria Clara and Castro, Sérgio and Cavalcanti, Tatiana and Cebiroğlu Eryiğit, Gülşen and Cecchini, Flavio Massimiliano and Celano, Giuseppe G. A. and Çepani, Anila and Čěplö, Slavomír and Cesur, Neslihan and Cetin, Savas and Çetinoğlu, Özlem and

Chalub, Fabricio and Chamila, Liyanage and Chamoreau, Claudine and Chauhan, Shweta and Chen, Yifei and Chi, Ethan and Chika, Taishi and Cho, Yongseok and Choi, Jinho and Chontaeva, Bermet and Chun, Jayeol and Chung, Juyeon and Cignarella, Alessandra T. and Cinková, Silvie and Collomb, Aurélie and Çöltekin, Çağrı and Connor, Miriam and Corbetta, Claudia and Corbetta, Daniela and Costa, Francisco and Courtin, Marine and Crabbé, Benoît and Cristescu, Mihaela and Cvetkoski, Vladimir and Dahan, Netanel and Dale, Ingerid Løyning and Daniel, Philemon and Daoudi, Khensa and Dash, Bijayalaxmi and Dash, Satya Ranjan and Davidson, Elizabeth and de Alencar, Leonel Figueiredo and Dehouck, Mathieu and de Laurentiis, Martina and de Marnette, Marie-Catherine and Demir, Ahmet and de Paiva, Valeria and Derin, Mehmet Oguz and de Souza, Elvis and Diaz de Ilarraza, Arantza and Díaz Hernández, Roberto Antonio and Dickerson, Carly and Di Felippo, Ariani and Dinakaramani, Arawinda and Di Nuovo, Elisa and Dione, Bamba and Dirix, Peter and Do, Hoa and Dobrovoljc, Kaja and Döhmer, Caroline and Doyle, Adrian and Dozat, Timothy and Drojanova, Kira and Duran, Magali Sanches and Dwivedi, Puneet and Ebert, Christian and Eckhoff, Hanne and Eguchi, Masaki and Eiche, Sandra and Eiselen, Roald and Eli, Marhaba and Elkahky, Ali and Ephrem, Binyam and Erina, Olga and Erjavec, Tomaž and Esher, Louise and Es-lami, Soudabeh and Essaidi, Farah and Etienne, Aline and Evelyn, Wograine and Facundes, Sidney and Farkas, Richárd and Faryad, Ján and Favero, Federica and Ferdaousi, Jannatul and Fernanda, Marília and Fernandez Alcalde, Hector and Fethi, Amal and Foster, Jennifer and Francioni, Barbara and Fransen, Theodorus and Freitas, Cláudia and Fujita, Kazunori and Gajdošová, Katarína and Galbraith, Daniel and Galy, Edith and Gamba, Federica and Garcia, Marcos and García-Miguel, José María and Gärdenfors, Moa and Gaustad, Tanja and Genç, Efe Eren and Gerardi, Fabrício Ferraz and Gerdes, Kim and Gessler, Luke and Ginter, Filip and Godoy, Gustavo and Goenaga, Iakes and Gojenola, Koldo and Gökırmak, Memduh and Goldberg, Yoav and Goldin, Gili and Gómez Guinovart, Xavier and González Saavedra, Berta and Griciūtė, Bernadeta and Grioni, Matias and Grobol, Loïc and Grūzītis, Normunds and Guillaume, Bruno and Guiller, Kirian and Guillot-Barbance, Céline and Güngör, Tunga and Gurevich, Vladimir and Habash, Nizar and Hafsteinsson, Hinrik and Hahn, Michael and Hajič, Jan and Hajič jr., Jan and Hajičová, Eva and Hämäläinen, Mika and Hà Mỹ, Linh and Han,

Na-Rae and Hanifmuti, Muhammad Yudistira and Harada, Takahiro and Hardwick, Sam and Harris, Kim and Hassert, Naïma and Haug, Dag and Havelka, Jiří and Heinecke, Johannes and Hellwig, Oliver and Hennig, Felix and Hladká, Barbora and Hlaváčová, Jaroslava and Hociung, Florinel and Hoefels, Diana and Hohle, Petter and Howell, Nick and Huang, Yidi and Huerta Mendez, Marivel and Hwang, Jena and Ikeda, Takumi and Iliadou, Inessa and Ingason, Anton Karl and Ion, Radu and Irimia, Elena and Ishola, Ołájídé and Islamaj, Artan and Ito, Kaoru and Iurescia, Federica and Ivani, Jessica K. and Jagodzińska, Sandra and Jannat, Siratun and Jelínek, Tomáš and Jha, Apoorva and Jiang, Katharine and Job, Sylvanus and Jobanputra, Mayank and Johannsen, Anders and Jónsdóttir, Hildur and Jørgensen, Fredrik and Ju, Zhuoxuan and Juutinen, Markus and Kaşıkara, Hüner and Kabaeva, Nadezhda and Kahane, Sylvain and Kanayama, Hiroshi and Kanerva, Jenna and Kara, Neslihan and Karahóğa, Ritván and Kárník, Jiří and Kåsen, Andre and Kayadelen, Tolga and Kengatharaiyer, Sarveswaran and Kettnerová, Václava and Kharatyan, Lilit and Kirchner, Jesse and Klementieva, Elena and Klyachko, Elena and Kocharov, Petr and Köhn, Arne and Köksal, Abdullatif and Kolářová, Veronika and Kopacewicz, Kamil and Korkiakangas, Timo and Köse, Mehmet and Koshevoy, Alexey and Kote, Nelda and Kotsyba, Natalia and Kovačić, Barbara and Kovalevskaitė, Jolanta and Kowner, Emmanuelle and Krek, Simon and Krishnamurthy, Parameswari and Kübler, Sandra and Kučová, Lucie and Kuqi, Adrian and Kuyrukçu, Oğuzhan and Kuzgun, Aslı and Kwak, Sookyung and Kyle, Kris and Laan, Kābi and Laipala, Veronika and Lambertino, Lorenzo and Landau, Israel and Lando, Tatiana and Larasati, Septina Dian and Larrivée, Pierre and Lavrentiev, Alexei and Lee, John and Lê Hồng, Phương and Lenci, Alessandro and Lertpradit, Saran and Leung, Herman and Levina, Maria and Levine, Lauren and Li, Cheuk Ying and Li, Josie and Li, Keying and Li, Yixuan and Li, Yuan and Lim, KyungTae and Lima Padovani, Bruna and Lin, Yi-Ju Jessica and Lindén, Krister and Liu, Yang Janet and Liu, Zoey and Ljubešić, Nikola and Lobzhanidze, Irina and Loginova, Olga and Lopatková, Markéta and Lopes, Luce-lene and Luftiu, Edita and Lukashevskiy, Arsenii and Lusito, Stefano and Lutgen, Anne-Marie and Luthfi, Andry and Luukko, Mikko and Lyashevskaya, Olga and Lynn, Teresa and Macketanz, Vivien and Mahamdi, Menel and Maillard, Jean and Makarchuk, Ilya and Makazhanov, Aibek and Mambrini, Francesco

and Mandl, Michael and Manning, Christopher and Manurung, Ruli and Marşan, Būşra and Măranduc, Cătălina and Mareček, David and Marheinecke, Katrin and Markantonatou, Stella and Martínez Alonso, Héctor and Martín Rodríguez, Lorena and Martins, André and Martins, Cláudia and Mašek, Jan and Matsuda, Hiroshi and Matsumoto, Yuji and Mazzei, Alessandro and McDonald, Ryan and McGuinness, Sarah and Mehta, Maitrey and Ménard, Pierre André and Mendonça, Gustavo and Merhav, Hilla and Merzhevich, Tatiana and Meurer, Paul and Miekka, Niko and Mikulová, Marie and Milano, Emilia and Miletić, Aleksandra and Miller, Aaron and Min, Junghyun and Minerbi, Yael and Mírovský, Jiří and Mischenkova, Karina and Missilä, Anna and Mititelu, Cătălin and Mitrofan, Maria and Miyao, Yusuke and Mohapatra, Biswakalpita and Mojiri Foroushani, AmirHosseini and Molnár, Judit and Moloodi, Amirsaeid and Montemagni, Simonetta and More, Amir and Moreno Romero, Laura and Moretti, Giovanni and Mori, Shinsuke and Morioka, Tomohiko and Moro, Shigeki and Mortensen, Bjartur and Moskalevskiy, Bohdan and Muischnek, Kadri and Munro, Robert and Murawaki, Yugo and Mus, Nikolett and Müürisep, Kaili and Nainwani, Pinkey and Nakhlé, Mariam and Navarro Horñiacek, Juan Ignacio and Nedoluzhko, Anna and Neşpore-Bêrzkalne, Gunta and Nevaci, Manuela and Nguyễn Thị, Lương and Nguyễn Thị Minh, Huyền and Nikaido, Yoshihiro and Nikolaev, Vitaly and Nitisaraj, Rattima and Norrman, Victor and Nourian, Alireza and Novák, Michal and Nunes, Maria das Graças Volpe and Nurmi, Hanna and Ojala, Stina and Ojha, Atul Kr. and Óladóttir, Hulda and Olúòkun, Adédayò and Omura, Mai and Onwuegbuzia, Emeka and Ordan, Noam and Osenova, Petya and Östling, Robert and Ott, Annika and Øvrelid, Lilja and Oya, Masanori and Özateş, Şaziye Betül and Özçelik, Merve and Özgür, Arzucan and Öztürk Başaran, Balkız and Paccosi, Teresa and Pajas, Petr and Palmero Aprosio, Alessio and Panevová, Jarmila and Panova, Anastasia and Pardo, Thiago Alexandre Salgueiro and Parida, Shantipriya and Park, Hyunji Hayley and Partanen, Niko and Pascual, Elena and Passarotti, Marco and Patejuk, Agnieszka and Paulino-Passos, Guilherme and Pedonese, Giulia and Peeters, Oggi and Peljak-Łapińska, Angelika and Peng, Siyao and Peng, Siyao Logan and Pereira, Rita and Pereira, Sílvia and Perez, Cenel-Augusto and Perkova, Natalia and Perrier, Guy and Petrov, Slav and Petrova, Daria and Peverelli, Andrea and Phelan, Jason and Pierre-Louis, Claudel and Piitulainen, Jussi and Pinter, Yuval and Pinto, Clara and Pintucci, Ro-

drigo and Pirinen, Tommi A and Pitler, Emily and Plamada, Magdalena and Plank, Barbara and Plum, Alistair and Poibeau, Thierry and Ponomareva, Larisa and Popel, Martin and Poudjade, Clamença and Pretkalniņa, Lauma and Pretorius, Rigardt and Prévost, Sophie and Prokopidis, Prokopis and Przepiórkowski, Adam and Pugh, Robert and Puolakainen, Tiina and Purschke, Christoph and Pyysalo, Sampo and Qi, Peng and Querido, Andreia and Rääbis, Andriela and Rabinovich, Ella and Rademaker, Alexandre and Rahman, Mutee-u and Rahoman, Mizanur and Rama, Taraka and Ramasamy, Loganathan and Ramisch, Carlos and Ramos, Joana and Rashel, Fam and Rasooli, Mohammad Sadegh and Ravishankar, Vinit and Real, Livy and Rebeja, Petru and Reddy, Siva and Regnault, Mathilde and Rehm, Georg and Riabi, Arij and Riabov, Ivan and Rießler, Michael and Rimkutė, Erika and Rinaldi, Larissa and Rituma, Laura and Rizqiyah, Putri and Rocha, Luisa and Rögnvaldsson, Eiríkur and Roksandic, Ivan and Roman, Norton Trevisan and Romanenko, Mykhailo and Romanova, Natalia and Rosa, Rudolf and Roşca, Valentin and Roulon, Paulette and Rovati, Davide and Rozonoyer, Ben and Rudina, Olga and Rueter, Jack and Ruffolo, Paolo and Rúnarsson, Kristján and Rushiti, Rozana and Sadde, Shoal and Safari, Pegah and Sahala, Aleksí and Sahoo, Kalyanamalini and Sahoo, Saraswati and Saleh, Shadi and Salomoni, Alessio and Samardžić, Tanja and Sampanis, Konstantinos and Samson, Stephanie and Sánchez-Rodríguez, Xulia and Sanguinetti, Manuela and Saniyar, Ezgi and Särg, Dage and Sartor, Marta and Sarymsakova, Albina and Sasaki, Mitsuya and Saulíte, Baiba and Savary, Agata and Sawanakunanon, Yanin and Saxena, Shefali and Scannell, Kevin and Scarlata, Salvatore and Schang, Emmanuel and Schneider, Nathan and Schuster, Sebastian and Schwartz, Lane and Seddah, Djamé and Seeker, Wolfgang and Sellmer, Sven and Seraji, Mojgan and Ševčíková, Magda and Sgall, Petr and Shahzadi, Syeda and Shen, Mo and Shimada, Atsuko and Shin, Gyu-Ho and Shirasu, Hiroyuki and Shishkina, Yana and Shohibussirri, Muh and Shvedova, Maria and Sibille, Jean and Siewert, Janine and Sigurðsson, Einar Freyr and Silva, João and Silveira, Aline and Silveira, Natalia and Silveira, Sara and Simi, Maria and Simionescu, Radu and Simkó, Katalin and Šimková, Mária and Símonarson, Haukur Barri and Simov, Kiril and Sitchinava, Dmitri and Sither, Ted and Smith, Aaron and Soares-Bastos, Isabela and Solberg, Per Erik and Sollberger, Dolores and Sonnenhauser, Barbara and Sourov, Shafi and Speransky, Nina and

Sprugnoli, Rachele and Stamou, Vivian and Steingrímsson, Steinþór and Stella, Antonio and Štěpánek, Jan and Štěpánková, Barbora and Stephen, Abishek and Straka, Milan and Strass, Omer and Strickland, Emmett and Strnadová, Jana and Suhr, Alane and Sulestio, Yogi Lesmana and Sulubacak, Umut and Sung, Hakyung and Suzuki, Shingo and Swanson, Daniel and Szántó, Zsolt and Taguchi, Chihiro and Taji, Dima and Talamo, Luigi and Tamburini, Fabio and Tan, Mary Ann C. and Tanaka, Takaaki and Tanaya, Dipta and Tavoni, Mirko and Teker, Nursena and Tella, Samson and Tellier, Isabelle and Testori, Marinella and Thomas, Guillaume and Tiraş, Tarık Emre and Tollersrud, Thea and Tonelli, Sara and Torga, Liisi and Toribio, Lucas and Toska, Marsida and Trosterud, Trond and Trukhina, Anna and Tsarfaty, Reut and Tulchynska, Kira and Türk, Utku and Tyers, Francis and Þórðarson, Sveinbjörn and Þorsteinsson, Vilhjálmur and Uematsu, Sumire and Untilov, Roman and Urešová, Zdeňka and Uria, Larraitz and Uszkoreit, Hans and Utka, Andrius and Vagnoni, Elena and Vajjala, Sowmya and Vak, Socrates and Vakirtzian, Socrates and van der Goot, Rob and Vanhove, Martine and van Niek-erk, Daniel and van Noord, Gertjan and Varga, Viktor and Vedenina, Uliana and Venturi, Giulia and Vergez-Couret, Marianne and Vidová Hladká, Barbora and Villemonte de la Clergerie, Eric and Vincze, Veronika and Vissamsetty, Anishka and Vlasova, Natalia and Vligouridou, Eleni and Wakasa, Aya and Wallenberg, Joel C. and Wallin, Lars and Walsh, Abigail and Wang, John and Washington, Jonathan North and Weissweiler, Leonie and Wendt, Maximilian and Widmer, Paul and Wigderson, Shira and Wijono, Sri Hartati and Wille, Vanessa Berwanger and Williams, Seyi and Winkler, Miriam and Wintner, Shuly and Wirén, Mats and Wittern, Christian and Witzlack-Makarevich, Alena and Woldemariam, Tsegay and Wong, Tak-sum and Wróblewska, Alina and Wu, Qishen and Yako, Mary and Yamashita, Kayo and Yamazaki, Naoki and Yan, Chunxiao and Yang, Xiulin and Yasuoka, Koichi and Yavrumyan, Marat M. and Yenice, Arife Betül and Yilandiloğlu, Enes and Yıldız, Olcay Taner and Yu, Zhuoran and Yuliawati, Arlisa and Žabokrtský, Zdeněk and Zahra, Shorouq and Zeldes, Amir and Zhou, He and Zhu, Hanzhi and Zhu, Yilun and Zhuravleva, Anna and Ziane, Rayan and Znotiņš, Artūrs. 2025. *Universal Dependencies 2.16*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Charles University, Prague, Czech republic. PID <http://hdl.handle.net/11234/1-5901>.

6. Appendix

The PDT-to-UD conversion software is open-source, available at <https://github.com/ufal/hamledt> and <https://github.com/ufal/treeex>. Its central part is the conversion of morphological tags and dependency labels. The former is documented at <https://universaldependencies.org/tagset-conversion/cs-pdtc-uposf.html>. Table 2 gives an idea about the dependency correspondences, although it must be noted that individual label mappings often depend on additional conditions and may be accompanied by transformations of the tree structure.

PDT	UD
Adv	advmod, obl, advcl
Apos	appos
Attr	amod, nummod, det, nmod, acl, flat
Atv	advcl:pred
AuxC	mark, fixed
AuxG	punct
AuxK	punct
AuxO	discourse
AuxP	case, fixed
AuxR	expl:pass
AuxT	expl:pv
AuxV	aux, aux:pass
AuxX	punct
AuxY	mark, cc, case, advmod, compound, fixed
AuxZ	cc, advmod:emph
Coord	conj
Denom	root, parataxis
Obj	obj, iobj, obl:arg, ccomp, xcomp
Partl	discourse
Pnom	cop
Pred	root, parataxis
Sb	nsubj, nsubj:pass, csubj, csubj:pass
Vocat	vocative

Table 2: Approximate correspondence of dependency relations.