***D. Zeman***

**Enhanced Universal Dependencies:
The current state and outlook[[1]](#footnote-1)**

**Abstract.** Universal Dependencies (UD) is a multilingual collection of corpora featuring morphological and syntactic annotation in a unified style. We discuss an optional layer of deep-syntactic annotation in UD, called Enhanced Universal Dependencies. We survey the existing enhanced repre­sen­tation as of release 2.8 and consider two possible future expansions: semi-automatic addition of existing enhancement types to new languages, and addition of new enhancement types.

**Keywords.** Dependency syntax, deep syntax, multilingual corpora, gapping, coordination, coreference.

1. Introduction

Universal Dependencies[[2]](#footnote-2) (UD) [Nivre et al. 2020; de Marneffe et al. 2021] is an international community project that strives to define a unified morpho-syntactic annotation scheme applicable to all natural languages, and to collect corpora (treebanks) annotated following that scheme. It started with the first version of annotation guidelines in 2014, and with the first release of 10 treebanks in January 2015; after six years, release 2.8 of UD boasts about 202 treebanks for 114 languages from 24 different families. Some treebanks are just tiny samples of less than thousand tokens while others contain over a million tokens; the total size of the collection amounts to 27 million.

UD has become an indispensable resource for research on multi-lingual natural language processing, especially morphological tagging and syntactic parsing. It has been also used in many linguistic studies, in particular in linguistic typology. Two large CoNLL shared tasks were organized in 2017 and 2018 to evaluate parsing systems on UD data [Zeman et al. 2018].

The morphological annotation in UD includes the lemma, universal part-of-speech tag (UPOS), morphological feature-value pairs, and possibly another tag from a treebank-specific tag set (XPOS). The UPOS tag must be picked from a fixed set of 17 categories; any finer distinctions, if desirable, are encoded in the morphological features (in addition to universally defined features, treebanks may also use language-specific values and features). All UD treebanks must have at least the UPOS tags manually checked. Lemmas, features and XPOS are optional and in a few UD corpora they have been assigned automatically.

The *basic* syntactic representation is a rooted dependency tree where every word/node (except the root) has one parent node. Each relation in the tree is labeled with its type; while the main types come from a fixed set of 37 universal relations, it is possible to define language-specific subtypes. Like UPOS tags, the parent nodes and the relation types are manually checked in all UD treebanks.

Examples of UD basic trees with UPOS tags are given in Figures 1 (English) and 2 (Russian), respectively. The sentences are parallel and so is their syntactic annotation: Relations between content words are identical in both structures, although English also has a number of additional relations between content and function words, which are not present in Russian. Such parallelism is possible thanks to the fact that content words are attached higher in UD, and function words (articles, prepositions, auxiliaries etc.) are normally attached as leaf nodes — an approach that is relatively uncommon in dependency frameworks outside UD.

**

*Fig. 1.* Basic UD tree of the English sentence
*The house has been sold to a foreigner.*

**

*Fig. 2.* Basic UD tree of the Russian sentence *Дом продан иностранцу.* “The house has been sold to a foreigner.”

1. Enhanced Representation

Besides the basic syntactic representation, UD may optionally contain an *enhanced* dependency structure, which is still a directed and rooted graph, but not necessarily a tree. The enhanced UD layer was proposed by [Schuster and Manning 2016] but its first official specification appeared in the version 2 of the UD guidelines [Nivre et al. 2020: §3.4]. The purpose of the enhanced representation is to facilitate downstream language understanding tasks by making certain relations explicitly annotated. In most cases, the enhanced graph is only a moderate modification of the basic tree. There are six types of enhancements defined in the guidelines (a UD treebank may annotate only some types and ignore the others):

1. gapping (empty nodes for elided predicates);
2. parent of coordination (propagated relations to non-first conjuncts);
3. shared dependent of coordination (propagated relations from non-first conjuncts);
4. external subject of a controlled or raised verb;
5. relative clause (modified noun attached instead of the relative pronoun, thus forming a directed cycle);
6. case information in the relation label.

**

*Fig. 3.* Enhanced UD graph of the Ukrainian sentence *Чоловіки повинні носити брюки, а жінки – панчохи.* “Men should wear pants and women should wear stockings.”

Figure 3 illustrates three enhancement types. There are two ‘empty’ nodes that represent elided predicates in a gapping construction, *повинні* “should” and *носити* “wear”. Propagation of coordination parent results in the second root relation, pointing to the empty node representing the second instance of *повинні*. And finally, both *чоловіки* “men” and *жінки* “women” have two incoming relations each, making them subject not only of *повинні* but also of *носити*. Figure 4 illustrates enhanced representation of relative clauses; note the directed cycle between *szamponu* “shampoo” and *myje* “washes”. For more details on the six enhancement types, see [Droganova and Zeman 2019; Nivre et al. 2020].

**

*Fig. 4.* Enhanced UD graph of the Polish sentence *A gdzie szukać szamponu, który myje?* “And where to look for shampoo that works?”

Some of the enhancements can be computed deterministically from the basic tree, as they really just foreground information that is already there. This is the case with coordination parents and case-enhanced relation labels. Other types will benefit from manual disambiguation but they still can be approximated algorithmically using relatively simple heuristics. At the other end of the scale, distinguishing between shared dependents of coordination and private dependents of the first conjunct clearly requires extra knowledge and may be sometimes hard even for a human annotator. Several tools have been proposed that use heuristics to compute some of the enhancements from the basic tree; see [Nivre et al. 2018] for a comparison. Among the UD treebanks that currently have some enhanced dependency structures, little annotation (if any at all)[[3]](#footnote-3) has been added manually. In some cases, the treebanks were converted from non-UD annotation schemes where the extra information was available. In others, heuristic enhancers were employed. Table 1 shows statistics of enhancement types in UD 2.8, which is the most recent release at the time of writing. Overall, 30 treebanks of 18 languages have at least one enhancement type, and 15 treebanks of 8 languages have all six types.

*Table 1.* Overview of enhancements in UD 2.8. Number of nodes includes empty nodes. Enhanced relations per 1,000 nodes: **G** = gapping relations (i.e., to or from empty nodes); **P** = relations propagated from parents of coordination; **S** = relations propagated to shared dependents of coordination; **X** = relations to controlled external subjects; **R** = relations added in relative clauses (usually 2 for each clause); **C** = relation labels enhanced with case information.

| **Treebank** | **Nodes** | **G** | **P** | **S** | **X** | **R** | **C** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Arabic PADT** | 282,460 | 2 | 48 | 10 |  | 16 | 286 |
| **Belarusian H.** | 305,406 | 4 | 44 | 0 | 6 | 16 | 148 |
| **Bulgarian BT.** | 156,149 |  | 7 | 4 | 3 | 16 | 118 |
| **Chukchi HSE** | 6,207 | 15 |  |  |  |  | 1 |
| **Czech CAC** | 495,497 | 15 | 65 | 22 | 5 | 24 | 184 |
| **Czech FicTree** | 167,371 | 11 | 55 | 19 | 5 | 21 | 127 |
| **Czech PDT** | 1,509,052 | 11 | 45 | 11 | 7 | 23 | 173 |
| **Czech PUD** | 18,623 | 4 | 39 | 1 | 7 | 26 | 172 |
| **Dutch Alpino** | 208,747 | 2 | 14 | 5 | 6 | 18 | 130 |
| **Dutch LassyS.** | 98,242 | 5 | 25 | 8 | 2 | 12 | 138 |
| **English EWT** | 254,857 | 1 | 21 | 9 | 14 | 11 | 128 |
| **English GUM** | 134,553 | 2 | 26 | 8 | 13 | 11 | 143 |
| **English GU.R.** | 16,286 |  | 19 | 11 | 18 | 14 | 123 |
| **English PUD** | 21,183 | 1 | 21 | 7 | 12 | 16 | 150 |
| **Estonian EDT** | 438,175 | 4 | 0 | 0 |  | 15 | 42 |
| **Estonian EWT** | 68,968 | 6 |  |  |  | 14 | 53 |
| **Finnish PUD** | 15,817 | 1 |  |  |  |  |  |
| **Finnish TDT** | 202,453 | 6 | 33 | 32 | 5 | 0 |  |
| **Italian ISDT** | 298,380 | 1 | 24 | 5 | 5 | 20 | 169 |
| **Latvian LVTB** | 252,961 | 13 | 36 | 32 | 9 |  | 158 |
| **Lithuanian A.** | 70,051 | 0 | 77 | 27 | 6 | 11 | 233 |
| **Polish LFG** | 130,967 |  | 12 | 8 | 7 | 0 | 85 |
| **Polish PDB** | 350,036 |  | 41 | 20 |  |  |  |
| **Polish PUD** | 18,389 |  | 37 | 20 |  |  |  |
| **Russian Syn.** | 1,107,741 | 4 |  |  |  |  | 3 |
| **Slovak SNK** | 106,184 | 5 | 39 | 7 | 4 | 15 | 151 |
| **Swedish PUD** | 19,085 | 2 | 22 | 8 | 11 | 28 | 158 |
| **Swedish Tal.** | 96,859 | 2 | 34 | 9 | 11 | 22 | 149 |
| **Tamil TTB** | 9,581 |  | 25 |  |  | 7 | 271 |
| **Ukrainian IU** | 122,324 | 10 | 51 | 8 | 9 | 17 |  |

Similarly to the basic dependencies, there are parsing models that can generate the enhanced graphs for previously unseen text. Some successful parsers take advantage of the fact that many enhancements can be guessed based on the basic tree and combine a tree parsing model with enhancing heuristics. Enhanced UD parsers have been evaluated in two shared tasks run in connection with the IWPT 2020 and 2021 conference [Bouma et al. 2020].

1. What Is Next?

As Table 1 clearly shows, the enhanced representation, being optional, is only available for a fraction of the UD treebanks, and it does not grow as quickly as the data with the basic representation. [Droganova and Zeman 2018] note that this is unlikely to change, as more complex annotation requires more annotation effort, and it is thus difficult to get sufficient manpower to annotate data in a new language. They propose to at least apply a heuristic enhancer to all UD treebanks after each release and make this data available to the users. In addition, they propose heuristics to normalize syntactic alternations such as passive vs. active diathesis; they call the resulting data Deep UD, to distinguish it from the Enhanced UD defined in the official UD guidelines. Various other “enhanced-plus” variants have been proposed by [Schuster and Manning 2016] and others, but they are not (yet?) approved as a part of UD. Even the existing guidelines sometimes leave room for multiple interpretations, leading to enhancement ‘sub-types’ that only appear in some treebanks.

The treatment of relative clauses can be extended to attributively used participles, as in the French example in Figure 5, where *fusée* “rocket” is modified by the participle *pouvant* “able”, and at the same time it is also annotated as the external subject of the participle (as well as its complement infinitive *menacer* “threaten”).

**

*Fig. 5.* Enhanced UD graph of the French sentence *L’Irak ne dispose d’aucune fusée pouvant menacer ses voisins.* “Iraq has no rockets that could threaten its neighbors.”

In pro-drop languages, empty nodes could help restore the co­reference of controlled subjects if the main subject is missing, as the empty node *(on)* is used in Figure 6.

**

*Fig. 6.* Enhanced UD graph of the Czech sentence *Příště se chystá prošetřit hospodaření fondu kultury.* “Next time they are going to investigate the management of the culture fund.”

Finally, empty nodes in enhanced UD have been used to show the attachment of constituents that are incorporated in the verb and lack a node in the basic representation, like the Chukchi adverb *пытӄы* “again” in Figure 7.

**

*Fig. 7.* Enhanced UD graph of the Chukchi sentence *Гым ытԓьата гэнапытӄынԓыватԓен.* “My mother was breast-feeding me again.”

1. Conclusion

We have presented the current state of the *enhan­ced repre­sen­tation* in Universal Dependencies. Being an optional and more com­plex annotation layer, it is only available for a fraction of the UD treebanks. Fortunately, a significant part of it can be computed or estimated with simple heuristics from the basic representation.

The guidelines for enhanced graphs are not considered as frozen as the basic guidelines in the UD community, and some details are still being elaborated as more languages are added. There is room for future additions of new variants of existing enhancement types or even completely new types with the same general motivation: to make otherwise implicit syntactico-semantic relations explicit and thus more easily accessible for language understanding applications.

**References**

1. *Bouma G., Seddah D., Zeman D.* (2020). Overview of the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies. In: Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies, pp. 151–161, ACL, Stroudsburg, PA, USA, <https://www.aclweb.org/anthology/2020.iwpt-1.16/>.

2. *de Marneffe M.-C., Manning C., Nivre J., Zeman D.* (2021). Universal Dependencies. In: Computational Linguistics, online first access at <https://doi.org/10.1162/coli_a_00402>.

3. *Droganova K., Zeman D.* (2019). Towards Deep Universal Dependencies. In: Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019), pp. 144–152, Paris, France, <https://www.aclweb.org/anthology/W19-7717/>.

4. *Nivre J., de Marneffe M.-C., Ginter F., Hajič J., Manning C., Pyysalo S., Schuster S., Tyers F., Zeman D.* (2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In: Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020), pp. 4034–4043, ELRA, Marseille, France, <https://www.aclweb.org/anthology/2020.lrec-1.497/>.

5. *Nivre J., Marongiu P., Ginter F., Kanerva J., Montemagni S., Schuster S., Simi M.* (2018). Enhancing Universal Dependency Treebanks: A Case Study. In: Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), pp. 102–107, Bruxelles, Belgium, <https://www.aclweb.org/anthology/W18-6012/>.

6. *Schuster S., Manning C.* (2016). Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), pp. 2371–2378, ELRA, Portorož, Slovenia, [https://www.aclweb.org/‌anthology/L16-1376/](https://www.aclweb.org/anthology/L16-1376/).

7. *Zeman D., Hajič J., Popel M., Potthast M., Straka M., Ginter F., Nivre J., Petrov S.* (2018). CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pp. 1–21, Bruxelles, Belgium, <https://www.aclweb.org/anthology/K18-2001/>.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Daniel Zeman**

Charles University, Faculty of Mathematics and Physics, ÚFAL (Czechia)

***E-mail: zeman@ufal.mff.cuni.cz***

1. This work was supported by the Grant No. GX20-16819X (LUSyD) of the Czech Science Foundation (GAČR). [↑](#footnote-ref-1)
2. <https://universaldependencies.org/> [↑](#footnote-ref-2)
3. Exact information about the origin of enhanced annotation in individual treebanks is not available. [↑](#footnote-ref-3)