From Raw Text to Enhanced Universal Dependencies: the Parsing Shared Task at IWPT 2021

Gosse Bouma* **Djamé Seddah†** **Daniel Zeman°**

*University of Groningen, Centre for Language and Cognition
†INRIA Paris
°Charles University in Prague, Faculty of Mathematics and Physics, ÚFAL
g.bouma@rug.nl, djame.seddah@inria.fr
zeman@ufal.mff.cuni.cz

Abstract

We describe the second IWPT task on end-to-end parsing from raw text to Enhanced Universal Dependencies. We provide details about the evaluation metrics and the datasets used for training and evaluation. We compare the approaches taken by participating teams and discuss the results of the shared task, also in comparison with the first edition of this task.

1 Introduction

Universal Dependencies (UD) (Nivre et al., 2020) is a framework for cross-linguistically consistent treebank annotation that has so far been applied to 114 languages. UD defines two levels of annotation, the basic trees and the enhanced graphs (EUD) (Schuster and Manning, 2016).

There are several good parsers that can predict the basic trees (including tokenization and morphology) for previously unseen text (Straka et al., 2016; Qi et al., 2020). Two large shared tasks on basic UD parsing were organized at CoNLL (Zeman et al., 2017, 2018). Enhanced UD parsing attracted comparatively less attention until the shared task organized at IWPT 2020 (Bouma et al., 2020). The present paper describes a second instance of that task, organized as a part of the 17th International Conference on Parsing Technologies[1] (IWPT), collocated with ACL-IJCNLP 2021. Like in the previous year, the evaluation was done on datasets covering 17 languages from four language familiies.

This paper is a follow-up of the overview paper of the previous instance of the shared task (Bouma et al., 2020). To make the paper self-contained, we include updated versions of some sections of that paper, in particular describing the enhanced annotation format, the task, and the evaluation metric. The data section now documents the modifications we made to the data from UD release 2.7.

[1] https://iwpt21.sigparse.org

2 Motivation

The basic dependency annotation in the Universal Dependencies format introduces labeled edges between nodes that represent tokens in the input string, where each node is a dependent of exactly one other node, with the exception of the node token. While this tree structure supports many downstream tasks, there are also phenomena that are hard to capture using single-parent edges only. The enhanced dependency layer therefore supports richer annotation where nodes may have more than one parent, and where additional 'empty' nodes represent elided material that is not overtly expressed in the input string. The enhanced level can be used to account for a range of linguistic phenomena (see Section 3) and to support downstream applications that rely on the semantic interpretation of the input.

There are now a number of treebanks that include enhanced dependency annotation. Furthermore, the recent shared tasks on dependency parsing and subsequent work have shown that considerable progress has been made in multilingual dependency parsing. For enhanced dependency parsing, there are additional challenges. The enhanced representation is a connected directed graph, possibly containing cycles, while the bulk of dependency parsing work still focuses on rooted trees. The set of labels to be predicted is also much larger, as some enhanced dependency labels incorporate the lemma of certain dependents.

On the other hand, it has been shown that much of the enhanced annotation can be predicted on the basis of the basic UD annotation (Nyblom et al., 2013; Schuster et al., 2017; Nivre et al., 2018). Moreover, most state-of-the-art work in

dependency parsing uses a graph-based approach, where the assumption that the output must form a tree is only used in the final step from predicted links to final output. And finally, work on deep-syntax and semantic parsing has shown that accurate mapping of strings into rich graph representations is possible (Oepen et al., 2014, 2015, 2019, 2020) and could even lead to state-of-the-art performance for downstream applications as shown by the results of the Extrinsic Parsing Evaluation shared task (Oepen et al., 2017).

The previous IWPT shared task (Bouma et al., 2020) reflected this development quite well: some submissions took the way of direct text-to-graph mapping, some of them predicted a rooted tree and then employed heuristics to enhance it; and one submission encoded graphs as trees, then used a tree parser to predict them. Since it was the first task of its kind on large scale multilingual Enhanced Dependencies parsing and some teams may not have been able to successfully implement all their ideas in time (or new ideas may have occurred after seeing what other teams had done), a second round of the task is a natural next step to see whether we can do even better.

3 Enhanced Universal Dependencies

UD version 2[2] states that apart from the morphological and basic dependency annotation layers, strings may be annotated with an additional, enhanced, dependency layer, where the following phenomena can be captured:

- Gapping. To support a linguistically more satisfying treatment of ellipsis, empty nodes can be introduced to represent missing predicates in gapping constructions.
- Parent of coordination. Incoming relations are propagated from the parent of the coordination structure to each conjunct.
- Shared dependent of coordination. Outgoing relations are propagated from each conjunct to a shared dependent, e.g., a shared subject or object of coordinate verbs.
- Control and raising constructions. The external subject of `xcomp` dependents, if present, can be explicitly marked.
- Relative clauses. The antecedent noun of a relative clause is annotated as a dependent of a node within the relative clause (thus introducing a cycle) and the relative pronoun is annotated as a `ref` dependent of the antecedent noun.
- Case information. Selected dependents (in particular `obl` and `nmod`), if they are marked by morphological case and/or by an adpositional case dependent, can now be labeled as `obl:marker` or `nmod:marker` where `marker` is the lemma of the case dependent and/or the value of the morphological feature `Case`.

All enhancements are optional, so a UD treebank may contain enhanced graphs with one type of enhancement and still lack the other types.

4 Data

The evaluation was done on 17 languages from 4 language families: Arabic, Bulgarian, Czech, Dutch, English, Estonian, Finnish, French, Italian, Latvian, Lithuanian, Polish, Russian, Slovak, Swedish, Tamil, Ukrainian. The language selection is driven simply by the fact that at least partial enhanced representation is available for the given language.

Training and development data were based on the UD release 2.7 (Zeman et al., 2020) but for several treebanks the enhanced annotation is richer than in UD 2.7. Besides improvements in the officially released versions of the individual treebanks, a few other things have changed in comparison to the IWPT 2020 task. The English data now includes the GUM treebank (its enhanced annotation was not present in UD 2.7 but it was being prepared for UD 2.8 and it was ready in time for the shared task). As in 2020, we include two French treebanks whose enhanced annotation is still not included in the official UD releases, but the annotation is more conservative this year, omitting the extra labels for diathesis neutralization (Candito et al., 2017) and surface vs deep syntax markers. Still, some enhancements in French go slightly beyond the official UD guidelines (see below for details). In Polish, we now harmonize the relation subtypes in the three treebanks so that merging them into one dataset is no longer an issue. Finally, we omit the Chukchi treebank, which is new in UD 2.7 and has enhanced graphs, but the graphs

[2] https://universaldependencies.org/u/overview/enhanced-syntax.html

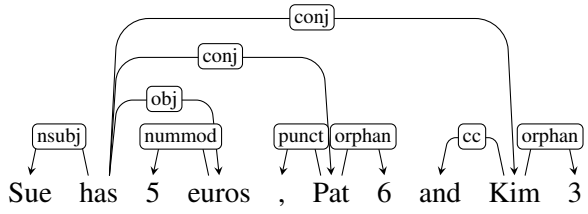


Figure 1: A basic tree of a gapping structure.

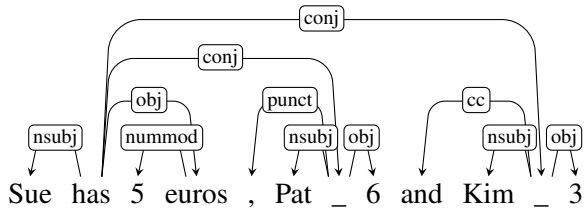


Figure 2: The correct enhanced graph of the gapping structure from Figure 1. "_" are empty nodes.

are there only to provide empty nodes to capture incorporated modifiers (rather than gapping); furthermore, the treebank is too small and has no training data.

There are 13 treebanks of 7 languages in UD 2.7 that contain all types of enhancements: Czech (CAC, FicTree, PDT, and PUD), Dutch (Alpino and LassySmall), English (EWT and PUD), Italian (ISDT), Lithuanian (ALKSNIS), Slovak (SNK), and Swedish (Talbanken and PUD). For the remaining languages, we applied simple heuristics and added at least some enhancements for the purpose of the shared task, but these annotations are not yet part of the regular UD releases. We only applied our heuristics to the missing enhancement types; we did not attempt to modify the enhancements provided by the data providers. Table 1 gives an overview of enhancements in individual treebanks.

The enhancements differ in how easily and accurately they can be inferred from the basic UD annotation:

- Enhancing relation labels with case information is deterministic. We apply it to the relations `obl`, `nmod`, `advcl` and `acl`. If they have a `case` or `mark` dependent, we add its lowercased lemma (for fixed multiword expressions or for multiple `case`/`mark` dependents we glue the lemmas with the "_" character). For `obl` and `nmod` we further examine the `Case` feature and add its lowercased value, if present.

Treebank	UD 2.7	Task
Arabic PADT	GPS RC	GPS RC
Bulgarian BTB	PSXRC	PSXRC
Czech CAC	GPSXRC	GPSXRC
Czech FicTree	GPSXRC	GPSXRC
Czech PDT	GPSXRC	GPSXRC
Czech PUD	GPSXRC	GP XRC
Dutch Alpino	GPSXRC	GPSXRC
Dutch LassySmall	GPSXRC	GPSXRC
English EWT	GPSXRC	GPSXRC
English GUM		GPSXRC
English PUD	GPSXRC	GPSXRC
Estonian EDT	GPS R	GPS RC
Estonian EWT	G	GP RC
Finnish PUD	GP	GP RC
Finnish TDT	GPSX	GPSXRC
French FQB		PSXR
French Sequoia		PSXR
Italian ISDT	GPSXRC	GPSXRC
Latvian LVTB	GPSX C	GPSXRC
Lithuanian ALKS.	GPSXRC	GPSXRC
Polish LFG	PSX C	PSXRC
Polish PDB	PS	GPSXRC
Polish PUD	PS	GPSXRC
Russian SynTagRus	G	GP XRC
Slovak SNK	GPSXRC	GPSXRC
Swedish PUD	GPSXRC	GPSXRC
Swedish Talbanken	GPSXRC	GPSXRC
Tamil TTB	PS	PS RC
Ukrainian IU	GPSXR	GPSXRC

Table 1: New annotation for the shared task. Abbreviations: `G` = gapping; `P` = parent of coordination; `S` = shared dependent of coordination; `X` = external subject of controlled verb; `R` = relative clause; `C` = case-enhanced relation label.

- Linking the parent of coordination to all conjuncts is deterministic.

- Recognizing and transforming relative clauses is easy if relative pronouns can be recognized. This can be tricky in languages where the same pronouns can be used relatively (Figure 3) and interrogatively (Figure 4). We cannot recognize all instances of the latter case reliably; fortunately they do not seem to be too frequent.

- External subjects of `xcomp` clauses are subjects, objects or oblique dependents of the matrix clause. To find them, we need to know

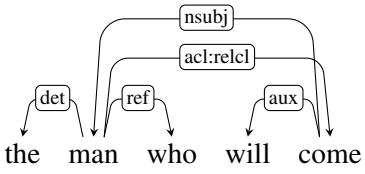


Figure 3: Enhanced graph of a relative clause.

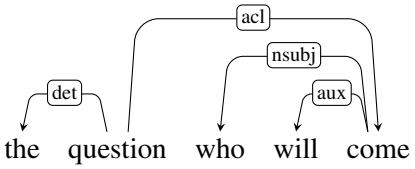


Figure 4: Enhanced graph of an interrogative clause.

whether the governing verb has subject or object control. We use language-specific verb lists, which can resolve many cases, but not all. If a verb is not on any list, we skip it.

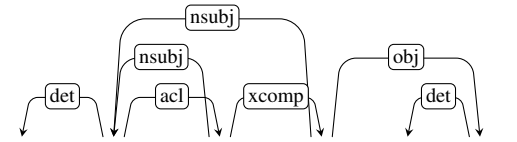
- Gapping can be easily identified by the presence of the `orphan` relation in the basic tree, insertion of empty nodes is thus trivial. However, we do not know the type of the relation between the empty node and the orphaned dependents. Figure 2 shows a graph where each empty node has one `nsubj` and one `obj` dependent. We cannot infer these labels from the basic tree (Figure 1), so we use `dep` instead.

- Linking conjuncts to shared dependents cannot be done reliably because we cannot know whether a dependent should be shared (this may be sometimes difficult even for a human annotator!) Therefore we do not attempt to add this enhancement to the datasets that do not have it.

Although the UD releases distinguish several different treebanks for some languages, for the purpose of the shared task evaluation we merged all test sets of each language. We wanted to promote robust parsers that are not tightly tied to one particular dataset. Merging treebanks of one language was possible because for almost all languages it holds that treebanks participating in the present task are maintained by the same team, hence no significant treebank-specific annotation decisions are expected. The exceptions are English and Polish but there should not be any significant divergence in these languages either. In English, the GUM corpus is maintained by other people than EWT and PUD; nevertheless, the corpora use the same set of relations, and there are ongoing efforts to harmonize the way the relations are used. In Polish, the LFG treebank uses a different set of relation subtypes than PDB and PUD; however, this year we removed the subtypes that are not used in all three treebanks, so it should be possible to train a parser on one treebank and successfully apply it

Treebank	**basic**	**lab**	**add**	**rem**
Arabic PADT	301399	27	7	1
Bulgarian BTB	156151	12	4	1
Czech CAC	494383	18	13	2
Czech FicTree	167056	13	11	2
Czech PDT	1506484	17	10	2
Czech PUD	18610	17	8	2
Dutch Alpino	208540	13	5	1
Dutch LassySmall	98044	14	5	1
English EWT	254829	13	6	1
English GUM	134476	14	6	1
English PUD	21176	15	6	1
Estonian EDT	437769	22	2	1
Estonian EWT	56399	18	8	1
Finnish PUD	15813	19	3	1
Finnish TDT	202291	18	10	1
French FQB	24135	0	2	0
French Sequoia	70567	0	5	0
Italian ISDT	298344	17	6	1
Latvian LVTB	219955	16	11	2
Lithuanian ALKSNIS	70047	23	12	1
Polish LFG	130968	9	3	0
Polish PDB	350036	16	9	1
Polish PUD	18389	18	9	1
Russian SynTagRus	1106296	17	7	1
Slovak SNK	106097	15	7	1
Swedish PUD	19076	16	7	1
Swedish Talbanken	96819	15	8	1
Tamil TTB	9581	27	3	0
Ukrainian IU	122094	16	10	1
total	6696809	17	8	1

Table 2: Comparing the impact of enhancements in the shared task treebanks where 'basic' is the number of basic dependencies (i.e., the number of words in the treebank) and the rest is given as a percentage of 'basic': 'lab' are enhanced dependencies that differ from a basic dependency only in label; 'add' are new enhanced dependencies (not only label but also the parent node differs from basic); 'rem' are basic dependencies that were removed from the enhanced graph.



"anglers who came to clean the banks"

Figure 5: Participial adnominal clauses in French are treated similarly to relative clauses: The modified noun is attached as a subject of the participle (and here also of the `xcomp` infinitive controlled by the participle).

to another.

Table 2 shows that the effect of enhancements differs quite a bit between the various languages. For instance, the percentage of basic dependencies that have a different label in the enhanced graph (mostly because of adding the case information to `obl` and other relations), ranges from 0 to 27%. Enhanced dependencies that introduce truly novel edges are rarer. In the table they are again expressed relatively to the number of basic dependencies, and the figure varies between 2 and 13%. Up to 2% basic edges are omitted in the enhanced graph.

There are slight differences in how individual languages implement particular enhancement types. Some languages follow earlier proposals for enhanced relation subtypes that are not supported by the current UD guidelines, e.g., external subjects are labeled `nsubj:xsubj`, antecedents of relative clauses are `nsubj:relsubj` or `obj:relobj`, the "case" information is extended to showing conjunction lemma with conjuncts (`conj:and`, `conj:or` etc.) Empty nodes are occasionally used for other ellipsis types than gapping or stripping. The adding of relations from relative clauses to modified nouns is further extended in French to infinitival and participial adnominal clauses, as in Figure 5.[3]

Upon completion of the shared task, the data has been made publicly available at the permanent address `http://hdl.handle.net/11234/1-3728`.

5 Task

As in the previous dependency parsing shared tasks, participants were expected to go from raw, untokenized strings to full dependency annotation. The evaluation focused on the enhanced annotation layer, but the participants were encouraged to predict all annotation layers, and the evaluation of the other layers is available on the shared task website.[4] The task was open, in the sense that participants were allowed to use any additional resources they deemed fit (with the exception of UD 2.7 test data) as long as this was announced in advance and the additional resource was freely available to everybody.

The submitted system outputs had to be valid CoNLL-U files; if a file was invalid, its score would be zero.[5] The official UD validation script[6] was used to check validity, although only at 'level 2', which means that only basic file format was checked and not the annotation guidelines (e.g., an unknown relation label would not render the file invalid). Constraints that have to be met at this level are that there must be at least one root node and every node must be reachable via a directed path from at least one root node (*rootedness* and *connectedness*), that the enhanced graph can contain cycles, but not self-loops (a node depending on itself), and that dependency labels can only contain characters from a limited set.

In addition to CoNLL-U validity, we also required that systems do not alter any non-whitespace characters when processing the input. This is a pre-requisite for the evaluation, where system-predicted tokens must be aligned with gold-standard tokens; files with modified word forms would be rejected.

6 Evaluation Metrics

The main evaluation metric is ELAS (*labeled attachment score on enhanced dependencies*), where ELAS is defined as F_1-score over the set of enhanced dependencies in the system output and the gold standard. Complete edge labels are taken into account, i.e. `obl:on` differs from `obl`. A second metric is EULAS, which differs from ELAS in that only the universal part of the dependency relation label is taken into account. Relation subtypes are ignored, i.e., `obl:on`, `obl:auf`, and `obl` are treated as identical.

Another issue we address is the evaluation of

[3] See (Candito et al., 2017) for details of the other enhancements they added (controlled-adjectives, causative constructions, etc.)

[4] https://universaldependencies.org/iwpt21/

[5] https://universaldependencies.org/format.html

[6] https://universaldependencies.org/release_checklist.html#validation

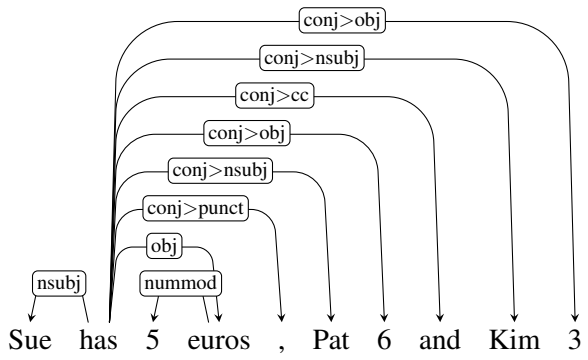


Figure 6: The enhanced graph from Figure 2 after collapsing empty nodes and reflecting the paths in dependency labels.

empty nodes. A consequence of the treatment of gapping and ellipsis is that some sentences contain additional nodes (numbered 1.1 etc.). It is not guaranteed that gold and system agree on the position in the string where these should appear, but the information encoded by these additional nodes might nevertheless be identical. Thus, such empty nodes should be considered equal even if their string index differs. To ensure that this is the case, we have opted for a solution that basically compiles the information expressed by empty nodes into the dependency label of its dependents. I.e. if a dependent with dependency label `L2` has an empty node `i2.1` as parent which itself is an `L1` dependent of `i1`, its dependency label will be expanded into a path `i1:L1>L2`. This preserves the information that the dependent was an `L2` dependent of something that was itself an `L1` dependent of `i1`, while at the same time removing the potentially conflicting `i2.1` (Figure 6).[7]

Finally, to analyze results, we computed ELAS scores per phenomenon. This should be seen as a diagnostic only, and is intended to gain further insights into the capability of various systems to deal with challenging phenomena, such as the proper analysis of phenomena occurring in the context of coordination and ellipsis.

7 Approaches

The predominant approach to obtaining the enhanced dependency graph is to use a biaffine function, i.e., predicting for each pair of nodes how likely it is that they are in a parent-child relation. There is wide variety in the way the final annotation graph is obtained, and ensuring that the result is valid (i.e. connected). GREW (Guillaume and Perrier, 2021) uses manually constructed rewrite rules to map basic UD into EUD, while FASTPARSE (Anderson and Gómez-Rodríguez, 2021) reformulate the task as a sequence-labeling task.

For the initial stages of the analysis (sentence splitting, tokenization, lemmatization, POS-tagging) most teams use Stanza (Qi et al., 2020) or Trankit (Van Nguyen et al., 2021) or similar methods. In a post-evaluation experiment, the DCU-EPFL team (Barry et al., 2021) obtained improved scores using Trankit instead of Stanza, while the TGIF team (Shi and Lee, 2021) uses a variation of the Trankit and Stanza systems to obtain the best pre-processing results, especially for sentence-splitting.

A wide variety of monolingual and multilingual pre-trained language models is used, with XML-R (Conneau et al., 2020) being the most popular. The ShanghaiTech system (Wang et al., 2021) learns an input representation from a combination of pre-trained language models where the various representations are concatenated into a single vector and masking is used to learn a weighting for various components of the combined vector. Both COMBO (Klimaszewski and Wróblewska, 2021) and UNIPI (Attardi et al., 2021) use a method that learns weights for the scores obtained from various layers of the BERT model to be used as input for the biaffine parser.

Most teams reduce the number of edge labels during training by de-lexicalizing edge labels. Dependency paths involving an empty node are usually also replaced by concatenating the path labels into a single path, as is also done in the evaluation script, thereby removing the need to predict empty nodes.

8 Results

Table 3 gives scores for LAS, EULAS, and ELAS macro-averaged over languages.[8] The 'baseline' is simply copying the UD annotation to EUD, but note that this is a strong baseline as it assumes perfect UD input, something that clearly is not the case for automated systems. Nevertheless,

[7] If there are multiple empty nodes in the sentence, we lose the information which orphans were siblings and which were not. On the other hand, multiple empty nodes in one sentence are extremely rare.

[8] More detailed results (per language and treebank, unofficial results) are available on the website of the shared task, https://universaldependencies.org/iwpt21/Results.html

most systems perform well above the baseline for ELAS. The NUIG submission was incomplete, in that the results for some languages were missing.[9] The submissions of TGIF and ShanghaiTech contain dummy annotations for all annotation layers except EUD, so no LAS is provided.

LAS and ELAS correlate strongly, with ELAS generally being 3-4% lower than LAS, except for DCU-EPFL, whose ELAS beats LAS. The best system in the first edition of this shared task (Bouma et al., 2020) obtained a ELAS of 84.50, while the current highest scoring system obtains an ELAS of 89.24. The average of ELAS of the top-5 was 78.75 for the first edition, while the current top-5 has an average of 86.14. The higher scores are most likely both due to more uniform annotations across treebanks as described in section 4 and improvements in approaches.

Team	LAS	EULAS	ELAS
baseline	100.00	96.28	79.87
TGIF	n/a	90.16	89.24
ShanghaiTech	n/a	88.49	87.07
RobertNLP	89.18	88.00	86.97
Combo	87.84	85.20	83.79
Unipi	87.25	85.24	83.64
DCU-EPFL	82.65	84.47	83.57
Grew	85.77	84.07	81.58
Fastparse	71.72	68.78	65.81
Nuig	39.78	31.63	30.03

Table 3: Evaluation results on the test data, macro-averaged over languages. LAS is the evaluation of the basic dependency annotation, while EULAS and ELAS evaluate the enhanced graph.

Table 4 gives the highest ELAS per language. Again, we see considerable improvements for all languages compared to the best ELAS for that language in the first edition of the shared task. The only exception is English, but it should be noted that for English the GUM treebank was added to this years data, so that results are not really comparable.

For the first edition of this task (Bouma et al., 2020) we provided a qualitative evaluation, where scores were computed per treebank, while taking into account that some treebanks do not include all enhancements stated in the guidelines in their enhanced layer. This year, as the annotation is considerable more uniform across treebanks, we decided to concentrate on performance per enhancement type. We used a script that labeled each edge in the enhanced annotation as belonging to one of the phenomena or enhancement types listed in Table 5. ELAS per phenomenon are given in Table 6. Note that the classification script assumes that basic UD annotation is also provided. For systems that only provide dummy labels and relations in their basic annotation (TGIF and ShanghaiTech), scores for some of the phenomena can therefore not be computed in a meaningful way and we replaced the score with 'n/a'. Table 6 illustrates that some systems do not take gapping (G) and treatment of orphans (O) into account. Also, scores for coordination (P and S), controlled subjects (X) and relatives (R) differ quite a bit among systems. While some of the phenomena are relatively rare in the data, it seems that to do well on the task, a system needs to perform reasonably well on all the phenomena listed here.

[9]No system description paper was submitted for NUIG.

Language	2020	2021
Arabic	77.82	82.26
Bulgarian	90.73	93.63
Czech	87.51	92.24
Dutch	85.14	91.78
English[1]	88.94	88.19
Estonian	84.54	88.38
Finnish	89.49	91.75
French[2]	86.23	91.73
Italian	91.54	93.31
Latvian	84.94	90.23
Lithuanian	77.64	86.06
Polish	84.64	91.46
Russian	90.69	94.01
Slovak	88.56	94.96
Swedish	85.64	89.90
Tamil	64.23	65.58
Ukrainian	87.22	92.78

Table 4: Best ELAS per language for 2020 and 2021. All best scores for 2021 were obtained by TGIF except for Arabic (ShanghaiTech). [1]*: English compares the score for the EWT and PUD treebanks (2020) with EWT+PUD+GUM (2021).* [2]*: French compares the scores between the 2021 more simple annotation scheme and the 2020 more complex original proposal.*

B	basic	this enhanced edge is identical to an edge in the basic tree (including the label)
C	cased	case-enhanced relation (the relation with the shorter label may or may not exist in the basic tree)
L	relabeled	the same two nodes are also connected in the basic tree but the label is different and the difference does not look like a case enhancement
G	gapping	the parent or the child is an empty node; the edge was added because of gapping
O	orphan	basic relation missing from enhanced graph because it was replaced by a relation to/from an empty node (the basic edge is not necessarily labeled `orphan`)
P	coparent	shared parent of coordination, relation propagated to a non-first conjunct
S	codepend	shared dependent of coordination, relation propagated from a non-first conjunct
X	xsubj	relation between a controlled predicate and its external subject
R	relcl	relation between a node in a relative clause and the modified nominal; also the `ref` relation between the modified nominal and the coreferential relative pronoun
W	relpron	basic relation incoming to a relative pronoun is missing from enhanced graph because it was replaced by the `ref` relation
M	missing	basic relation is missing from the enhanced graph but none of the above reasons applies
E	enhanced	this enhanced edge does not exist in the basic tree and none of the above reasons applies

Table 5: Classification of enhanced dependencies according to phenomenon and enhancement type.

Phenom'n	Combo	DCU_EPFL	Fastparse	Grew	RobertNLP	ShanghaiTech	TGIF	Unipi
B	90.86	89.13	78.32	88.00	91.56	n/a	n/a	90.19
C	83.28	80.17	61.03	76.79	83.10	n/a	n/a	82.30
L	0.00	0.02	0.00	0.00	0.03	0.01	0.00	0.05
G	21.81	0.00	0.00	12.57	0.00	56.55	58.39	0.00
O	29.84	0.00	0.00	15.81	0.00	n/a	n/a	0.00
P	60.63	73.48	26.39	62.09	64.78	75.91	79.61	61.24
S	38.02	59.07	0.71	40.92	64.19	65.40	69.22	57.64
X	64.29	84.41	3.37	71.00	86.82	85.96	88.09	84.75
R	64.73	84.42	1.53	65.21	85.38	85.67	85.08	82.42
W	88.17	87.06	0.00	81.50	90.63	n/a	n/a	90.76
M	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
E	0.51	0.51	0.02	0.00	0.00	0.09	0.09	1.77

Table 6: ELAS per phenomenon. Scores are micro-averaged, i.e. computed for the concatenation of all treebanks. Note that for systems that only provide dummy annotations for basic UD, some of the scores cannot be computed in a meaningful way. The NUIG system was not included as it lacked results for some languages.

9 Conclusions

The second edition of the shared task for parsing into enhanced universal dependencies shows improvements at various levels. First of all, the same set of languages was included as for the first edition, but now we were using treebanks of UD release 2.7 (Zeman et al., 2020). This EUD annotation of this release is more consistent and according to guidelines than the data of release 2.5, but we still had to harmonize some of the annotations so that differences in annotation would not have a negative effect on system performance.

Second, the requirement that submitted annotations should be minimally valid according to the guidelines, was now more easily met by all participating teams. Teams ensured that graphs would be connected, for instance, by applying several heuristics that introduce the minimal amount of additional edges to meet connectedness.

Third, while the best performing system in the first shared task used a method that pre-compiled the enhanced annotation graph into a tree, compatible with basic UD, and used a standard dependency parsing algorithm for learning to produce such annotations, almost all systems in this years shared task went for a graph-based approach. There still is quite a bit of variation in the way the graph is constructed though, with some systems first producing a tree, and then adding ad-

ditional edges, where others try to produce the graph directly. At the same time, most systems do apply some form of pre-compilation to make the data more suitable for learning. In particular, case-enhanced dependency labels are replaced by de-lexicalized labels that can be easily reconstructed in postprocessing. Similarly, most teams adopt a method that removes 'empty' nodes and instead expresses the information in incoming and outgoing edges from these nodes in the form of complex dependency labels (as is done in the evaluation script as well).

Finally, a very positive outcome of this evaluation is that scores have increased considerably, not only for the top performing system, but also for the top-5 systems. In particular, lower performance now seems to be restricted to languages for which very limited amounts of data is available, and, as Table 4 shows, the best system obtains an ELAS of over 90% for 11 of the 17 languages included in the evaluation.

Acknowledgments

We heartily thank everyone involved in the development of the Enhanced UD treebanks and who made this shared task possible.

This work has been partially supported by the LUSyD project, grant 20-16819X of the Czech Science Foundation (GAČR). The second author was partly funded by two French National Research Agency projects, PARSITI (ANR-16-CE33-0021) and SoSweet (ANR-15-CE38-0011).

References

Mark Anderson and Carlos Gómez-Rodríguez. 2021. Splitting EUD graphs into trees: A quick and clatty approach. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies*.

Giuseppe Attardi, Daniele Sartiano, and Maria Simi. 2021. Biaffine dependency and semantic graph parsing for enhanced universal dependencies. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies*.

James Barry, Alireza Mohammadshahi, Joachim Wagner, Jennifer Foster, and James Henderson. 2021. The dcu-epfl enhanced dependency parser at the iwpt 2021 shared task. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies*.

Gosse Bouma, Djamé Seddah, and Daniel Zeman. 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*, pages 151–161, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marie Candito, Bruno Guillaume, Guy Perrier, and Djamé Seddah. 2017. Enhanced UD dependencies with neutralized diathesis alternation. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 42–53, Pisa, Italy.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 84408451.

Bruno Guillaume and Guy Perrier. 2021. Graph Rewriting for Enhanced Universal Dependencies. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies*.

Mateusz Klimaszewski and Alina Wróblewska. 2021. Combo: a new module for EUD parsing. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies*.

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 International Conference on Language Resources and Evaluation (LREC 2020)*, pages 4027–4036, Paris, France. European Language Resources Association.

Joakim Nivre, Paola Marongiu, Filip Ginter, Jenna Kanerva, Simonetta Montemagni, Sebastian Schuster, and Maria Simi. 2018. Enhancing universal dependency treebanks: A case study. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 102–107.

Jenna Nyblom, Samuel Kohonen, Katri Haverinen, Tapio Salakoski, and Filip Ginter. 2013. Predicting conjunct propagation and other extended Stanford Dependencies. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pages 252–261, Praha, Czechia. Matfyzpress.

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.

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.

Stephan Oepen, Jari Björne, Richard Johansson, Emanuele Lapponi, Filip Ginter, Erik Velldal, and Lilja Øvrelid. 2017. The 2017 Shared Task on Extrinsic Parser Evaluation (EPE 2017).

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinková, Dan Flickinger, Jan Hajič, and Zdeňka Urešová. 2015. SemEval 2015 task 18: Broad-coverage semantic dependency parsing. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Dan Flickinger, Jan Hajič, Angelina Ivanova, and Yi Zhang. 2014. Semeval 2014 task 8: Broad-coverage semantic dependency parsing. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 63–72.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, page 101108, Seattle, WA, USA.

Sebastian Schuster, Eric De La Clergerie, Marie Candito, Benoît Sagot, Christopher D. Manning, and Djamé Seddah. 2017. Paris and Stanford at EPE 2017: Downstream evaluation of graph-based dependency representations.

Sebastian Schuster and Christopher D. Manning. 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)*, Paris, France. European Language Resources Association.

Tianze Shi and Lillian Lee. 2021. TGIF: Tree-Graph Integrated-Format Parser for Enhanced UD with Two-Stage Generic- to Individual-Language Finetuning. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies*.

Milan Straka, Jan Hajič, and Jana Straková. 2016. Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Minh Van Nguyen, Viet Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, page 8090.

Xinyu Wang, Zixia Jia, Yong Jiang, and Kewei Tu. 2021. Enhanced universal dependency parsing with automated concatenation of embeddings. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies*.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 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*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Ika Alfina, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Maria Jesus Aranzabe, Hórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Kepa Bengoetxea, Gözde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Ethan Chi, Yongseok Cho, Jinho Choi, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin,

Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Marília Fernanda, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Fabrício Ferraz Gerardi, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Eva Huber, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Sarveswaran K, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sookyoung Kwak, Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yuan Li, KyungTae Lim, Krister Lindén, Nikola Ljubešić, Olga Loginova, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lng Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Arzucan Özgür, Balkız Öztürk Başaran, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Eiríkur Rögnvaldsson, Mykhailo Romanenko, Rudolf Rosa, Valentin Roca, Davide Rovati, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Benoît Sagot, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Einar Freyr Sigurðsson, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Steinþór Steingrímsson, Antonio Stella, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Samson Tella, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, Hanzhi Zhu, and Anna Zhuravleva. 2020. Universal dependencies 2.7. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Daniel Zeman, Martin Popel, Milan Straka, Jan

Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gökırmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.