Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1114–1124
Marseille, 20-25 June 2022

Out-of-Domain Evaluation of Finnish Dependency Parsing

Jenna Kanerva and Filip Ginter
TurkuNLP, Department of Computing
University of Turku, Finland
{jmnybl, figint}@utu.fi

Abstract

The prevailing practice in the academia is to evaluate the model performance on in-domain evaluation data typically set aside from the training corpus. However, in many real world applications the data on which the model is applied may very substantially differ from the characteristics of the training data. In this paper, we focus on Finnish out-of-domain parsing by introducing a novel UD Finnish-OOD out-of-domain treebank including five very distinct data sources (web documents, clinical, online discussions, tweets, and poetry), and a total of 19,382 syntactic words in 2,122 sentences released under the Universal Dependencies framework. Together with the new treebank, we present extensive out-of-domain parsing evaluation utilizing the available section-level information from three different Finnish UD treebanks (TDT, PUD, OOD). Compared to the previously existing treebanks, the new Finnish-OOD is shown include sections more challenging for the general parser, creating an interesting evaluation setting and yielding valuable information for those applying the parser outside of its training domain.

Keywords: Finnish, Universal Dependencies, Treebank, Parsing, Out-of-domain

1. Introduction

During software development and model performance evaluation, the prevailing practice in the academia is to evaluate the model performance on an in-domain dataset. This typically means that the model is evaluated on a test section set aside from the training corpus, therefore the test dataset sharing the same properties as the data used to train the model. However, in many real world applications this may not be the case. The data on which the model is applied in its actual use in downstream applications may in practice very substantially differ from the characteristics of the training data.

In this paper, we focus on Finnish dependency parsing in the Universal Dependencies (UD) scheme. When both trained and tested on the UD dataset, the state of the art is approaching human performance (Virtanen et al., 2019). Consequently, the Finnish parser is in active use in the academia as well as in the commercial industry, and applied in numerous downstream tasks and domains as a text normalization and pre-processing component. In some cases, however, these application domains substantially differ in their characteristics from the training corpus and there is no hard evidence as to the effect on the parser performance.

To address this question, we selected and annotated a new treebank meant solely for out-of-domain evaluation of the models trained on the UD Finnish-TDT dataset. This new UD Finnish-OOD treebank allows us to quantify the parsing performance in various downstream applications, and to better understand the limits of generalization exhibited by the most recent dependency parsing methodology. In addition to introducing the new dataset, we carry out several Finnish out-of-domain parsing experiments, where in addition to the presented Finnish-OOD treebank, we use the existing section-level metadata in order to carry out section level performance evaluation. We show the new Finnish-OOD treebank being more challenging compared to the different sections existing in the current treebanks available for Finnish in the UD collection.

2. Data Sources

The three UD treebanks presently available for Finnish, namely Finnish-TDT (Haverinen et al., 2014; Pyysalo et al., 2015), Finnish-PUD (Zeman et al., 2017), and Finnish-FTB (Voutilainen et al., 2010), represent 6 different text genres based on the UD genre classification: blogs, fiction, grammar examples, legal, news and Wikipedia.

The UD Finnish-TDT is a general domain treebank containing 202,453 syntactic words (15,136 sentences) from 10 different text sources: Wikipedia articles, online fiction, JRC-Acquis legislation, popular online blogs, EuroParl speeches, grammar examples, Wikinews, university news, economy news, and student magazine articles. The treebank is divided into training, development and test set, the training set of the TDT treebank being the primary training data used throughout all experiments reported in this study. The UD Finnish-PUD is the Finnish part of the parallel UD treebank collection annotating the same underlying text translated for multiple languages. It includes 15,317 words (1,000 sentences) from two text sources: Wikipedia and news. The Finnish-PUD is used as external test set in this study. The UD Finnish-FTB is a treebank of grammar book examples annotated as a separate effort, independently of the other two Finnish UD treebanks and its annotation is not compatible in many important details with the abovementioned treebanks. Since these incompatibilites would mask any interesting differences, we do not use the FTB treebank

in this study. Nevertheless, the *Grammar examples* section of the TDT treebank is in fact sampled from the FTB text material, and therefore the FTB treebank text domain is represented in our experiments.

In order to build the new out-of-domain treebank for Finnish UD parsing we consider four text genres defined in UD v2.8 genre classification (see e.g (Zeman et al., 2021), (Müller-Eberstein et al., 2021)) but absent from the previously available Finnish treebanks: medical, poetry, social and web. Under these four genres, we include documents from five different text sources: (1) clinical nursing narratives of hospital patients for the medical domain, (2) web documents manually identified to contain poems or song lyrics for poetry, (3) discussion forum messages and (4) tweets for the social domain, and (5) randomly sampled documents from a general internet crawl for web. Of these, especially the nursing narratives, poetry, and tweets differ very substantially from any text in the training data. Yet, in particular the clinical domain and tweets represent a typical application domain for the Finnish parser.

In the following, we describe the data collection and cleaning procedure for each data source separately.

2.1. Medical – Clinical Nursing Narratives

In clinical nursing narratives the patient's visit in the hospital is recorded in a free text narrative, where the document is amended by nurses throughout their shifts to describe the patient's condition, treatments and status development during the stay in the hospital. These nursing records are meant for medical professionals to help in clinical decision making, and due to the fact that the records are targeted to professionals, the text is heavy on special terminology. Additionally, the nature of nursing narratives substantially differ from general language use, nursing narratives oftentimes including only critical facts expressed in a sentence without a main verb rather than carefully edited sentences. Laippala et al. (2014) described the key characteristics in nursing narratives including frequent misspellings, abbreviations, domain terminology, telegraphic writing style and non-standard syntactic structures.

Suominen et al. (2009) collected a corpus of nursing narratives of Finnish intensive care unit patients in the Turku University Hospital during years 2005-2006. In this work the nursing narratives of selected patients (those whose stay in the intensive care unit was at least 5 days) where extracted from the hospital's electronic patient records. Due to the obvious considerations on personal health data, this corpus is not available online. However, later on, a small section of this corpus consisting of 8 full narratives was manually anonymized and made openly available with annotation in the Stanford Dependencies (SD) scheme (Haverinen et al., 2009; Haverinen et al., 2010). While the dependency relations were manually annotated in this clinical Finnish treebank, the segmentation, morphology and lemma annotation layers were only automatically predicted. In this work, we sample two full narratives (939 sentences) from the original clinical treebank, and manually re-annotate them into UD scheme, this time including manual annotation of all layers, i.e. morphological tags, lemmas, and syntax.

The original clinical treebank data is available only with automatic segmentation, and since we do not have access to the original corpus used to obtain the anonymized records, the information on original text is lost. In order to include at least a minimal support towards testing segmentation models on the medical data, we apply manual detokenization, where e.g. punctuation markers are reconnected with the previous tokens when applicable, and thus the text is "corrected" to reflect orthographic standards in the original corpus. Misspellings and other segmentation issues that we could assume not to be introduced by the original tokenizer were left as-is. By doing this, we recognize the issue of the detokenized data not fully reflecting the possible variation of misspellings in cases where it was unclear whether the error was introduced by the original writer or the automatic tokenizer.

2.2. Poetry – Poems and Song Lyrics

In our poetry subsection, we rely on web documents manually identified to include poems or song lyrics. These documents are drawn from the FinCORE corpus (Laippala et al., 2019), where a random sample of Finnish web crawled documents are manually labeled for their text register, using 8 top-level labels (narrative, opinion, informational description, interactive discussion, how-to/instructional, informational persuasion, lyrical, and spoken) and several subcategories. We extract all documents manually labeled as lyrical in the FinCORE corpus, denoting a top-level category including both poems and song lyrics. At the time of data collection, we were able to identify 6 documents (144 sentences) with the lyrical label.

2.3. Social — Tweets

The Finnish tweets were downloaded between years 2016-2018 using the Twitter streaming API [1]. During downloading, we keep only tweets labeled as Finnish by the Twitter's language recognition (available in the tweet json). However, we observed the downloaded dataset to include a large number of tweets incorrectly labeled as Finnish, and therefore we manually identified all Finnish tweets from a small sample of 1250 tweets randomly selected among all downloaded tweets. This manual curation step discarded over 50% of all sampled tweets, indicating the language identification labels not being accurate enough for selecting Finnish tweets.[2] Finally, 130 randomly sampled

[1] We used the Tweepy Library `https://github.com/tweepy/tweepy`.

[2] All sampled tweets with their manually annotated labels are available at `https://github.com/TurkuNLP/finnish-tweets-lang-identification` for any

tweets from the curated dataset proceeded into the manual morphosyntactic annotation.

Likely due to change in Twitter character limits in 2017, the main text field in the downloaded tweet json sometimes contains a truncated version of the tweet. Similarly for retweets, the main text field includes a retweet marker (`RT @USERNAME:`), and the actual tweet can become truncated. In both cases, we always extract the full tweet text rather than the truncated one. This also has the property of not including the retweet marker as part of the extracted text, but retaining the information in the corpus metadata. This strives to mimic the textual content of a tweet as the user would see it through the online interface.

2.4. Social – Discussion Forum Messages

The second subset of social network data is gathered from the Suomi24 corpus[3], containing all messages posted in the Finnish Suomi24 online discussion forum between years 2001 and 2017. Historically, it has been one of the largest social network forums in Finland and covers a broad range of discussion topics including language with wide range of different writing styles and formality. From this dataset, we randomly sample 51 different messages for manual annotation, where messages may be anything between quick reactions to previous messages to longer posts on any number of different topics.

2.5. Web – Random Sample of the Internet Crawl

For the web domain, we take a random sample of 30 documents from the Finnish Internet Parsebank (Luotolahti et al., 2015). Five documents manually determined to be machine translated, thus, including many incomprehensible sentences, were replaced with new documents during sampling. Due to many web documents being quite long, each document was truncated after 25 sentences in order to avoid overly long documents biasing the web data towards particular topics. Furthermore, unnatural repetition appearing in some documents was removed (e.g. repeating quotations blocks) to avoid artificially skewing the evaluation statistics, and in these cases, more sentences were taken from the same document until the 25 sentence limit was reached.

3. Treebank Annotation

The data was annotated by a single annotator with a long-term experience in Finnish UD treebanking and the sole maintainer of the UD Finnish-TDT corpus. In the annotation, the Universal Dependencies guidelines were used as adapted in the Finnish-TDT corpus, thus making the corpus suitable for out-of-domain experiments especially for models trained with the Finnish-TDT treebank and making the new corpus fully compatible with UD Finnish-TDT in the numerous analysis choices and guideline interpretations. The dataset is natively annotated into the UD scheme, including fully manual analysis of all relevant layers (segmentation, morphology, lemmas and dependency syntax).

While some of the new text sources can be quite straightforwardly annotated using the general guidelines, some of the domains need domain-specific choices, as there are no established prior guidelines for some of the constructions. By far the most difficult domains in this work were poetry due to its specialities in sentence segmentation, and tweets due to including tokens limited to social media texts (e.g. hashtags and mentions), not appearing in the Finnish-TDT treebank. In addition to these, the medical domain posed interesting challenges in its specific medical terminology, while discussion forum messages and web documents did not substantially differentiate from the general domain texts in terms of annotation, and therefore, did not require adaptations to the general guidelines.

Next, we will discuss the annotation process separately for the poetry, tweets and clinical nursing narratives, as well as discuss the most relevant related work supporting the annotation decisions made during the annotation.

3.1. Clinical Nursing Narratives

While some of the medical terms used in the clinical nursing narratives are easily understandable to readers without professional medical knowledge, some terms require domain-specific understanding in order to correctly determine their meaning. Clearly, the annotation of morphological features and dependency relations for such terms is difficult for a person working outside the domain. While many of such terms are available in different medical dictionaries, especially highly abbreviated versions of medical terms are oftentimes difficult to find. In order to support the corpus annotation, we start the annotation process by translating all domain-specific terms into a general language with the help of a trained nurse. These translations are included as additional annotation in the MISC field of the CoNLL-U file, where the translations could be provided on word-to-word basis (`Gen=Translation`). An informative description of a concept is included instead in the MISC field in cases where a word-to-word translation is not feasible (`Gen_desc=Description`).

In general, the medical domain is quite rare in UD treebanks, in addition to ours, only 6 UD treebanks are reported as including medical texts. In fact, in all of these 6 treebanks (Czech-CAC (Raab, 2008), French-Sequoia (Candito and Seddah, 2012), Kiche-IU (Tyers and Henderson, 2021), Persian-Seraji (Seraji et al., 2016), Romanian-RRT (Mititelu, 2018), and Romanian-SiMoNERo (Mititelu and Mitrofan, 2020)),

later experiments on language identification of Finnish tweets.

[3] `http://urn.fi/urn:nbn:fi:lb-2020021802`

the medical texts are reported to be based on scientific or technical writings from the field of medicine, thus being carefully edited, official publications. In contrast, the clinical nursing narratives used in our corpus are quickly drafted notes written to other professionals and not meant to be publicly shared, making the nature of our medical texts very different from other UD treebanks. However, after dealing with the terminology, the rest of the annotation work was quite straightforward.

3.2. Poetry

While annotating texts from the poetry genre, one feature clearly standing out was the usage of capitalization and line breaks to articulate the layout of the text (indicating rhythm) rather than following standard structure of dividing text into paragraphs and sentences. In some documents, this resulted in having long text passages without punctuation characters indicating the standard sentence or clause structures.

Similar to medical, poetry is also among one of the rarest genres in UD. In addition to our treebank only 6 datasets are reported as including it: Belarusian-HSE [4], Breton-KEB (Tyers and Henderson, 2021), Latin-UDante (Cecchini et al., 2020), Old French-SRCMF (Stein and Prévost, 2013), Romanian-Nonstandard (Mărănduc and Bobicev, 2017), and Russian-Taiga (Shavrina and Shapovalova, 2017). While the segmentation of poetry texts is not explicitly mentioned in the studies or UD specifications, we follow a similar principle that seems to be the consensus in other UD treebanks, based on our understanding of the examples available in the papers, as well as inspecting annotated sentences in the released datasets.

We segment the texts into sentences following the existence of the sentence-final punctuation rather than capitalization or single line breaks, as oftentimes the text after a single newline was evidently a continuum of the previous sentence (fitting e.g. dependency relations `obl`, `advcl`, `acl`, or `conj`). In such cases where the sentence continuation was semantically ambiguous (full stop could have been easily used to break the sentence), these segments are connected with the parataxis relation marking for side-by-side clauses without coordination, subordination or argument relation. An exception, where we follow line breaks rather than punctuation, is made with double newlines (indicating paragraph or stanza boundary), where the sentence boundary is annotated even without an explicit sentence-final punctuation.

3.3. Tweets

Tweets include several characteristics rather unique to limited social media channels, the most common being mentions (`@username`) and hashtags (`#hashtag`), while also URLs and emoticons are substantially more frequent in tweets than in many other genres in the Finnish treebanks. Also, due to the character limits in social media platforms, tweets are rather short documents typically including only one or two short sentences. Likely due to this reason, in many treebanks including Twitter data, tweets are considered to be single sentence units and further sentence splitting is not applied (see e.g. the data releases of Italian-PoSTWITA (Sanguinetti et al., 2017), Italian-TWITTIRO (Cignarella et al., 2019), or Tweebank by Liu et al. (2018)). However, based on manual annotation 35% of the tweets in our sample include more than one sentence, 72% of sentences in these multi-sentence tweets containing a predicate in the main clause, thus indicating the individual sentences more often being real sentence-like units rather than short noun phrases. As the CoNLL-U format supports indicating document structure as metadata, we do not want to artificially analyse tweets as single sentences, when similar text passages in any other genre would be segmented into multiple sentences. Therefore, we consider a tweet as a small document which is further segmented into sentences as necessary. However, special tokens (mentions and hashtags) as well as plain interjections (e.g. *Wonderful!*) in the beginning or end of the tweet are kept together with the corresponding sentence. Regarding interjections, a similar approach is applied also in the Finnish-TDT treebank, thus not making deviation to the original annotation scheme. Regarding token segmentation, we treat mentions and hashtags as single tokens, where the special characters (`@` or `#`) are simply part of the main token. Otherwise standard tokenization guidelines are applied.

While there are several studies involving UD annotation on tweets (see e.g. Sanguinetti et al. (2017), Liu et al. (2018), Bhat et al. (2018), and Blodgett et al. (2018)), there does not seem to be a clear consensus regarding the annotation of tokens specific to Twitter or other social media platforms. Mentions are usernames appearing typically at the beginning of the sentence to mark dialogue participant in addressed speech, or occasionally replacing a normal content-bearing word in the sentence, usually when referring to an entity which would otherwise be a proper name (e.g. person or company name). Hashtags have a similar distinction, where most of the hashtags are used as a list of topical keywords appearing in the beginning or at the end of the sentence, however, some can be used as normal content-bearing words to replace any normal token in the sentence. In Figure 1 we illustrate a typical tweet taken from the corpus.

While Sanguinetti et al. (2017) and Bhat et al. (2018) annotated mentions with `SYM` part-of-speech tag, Liu et al. (2018) used `PROPN`, however, all agreeing of using `vocative` dependency relation for those mentions appearing in the beginning of the tweet to address the dialogue participant. As mentions are references to Twitter usernames and thus can be considered as proper

[4] `https://github.com/UniversalDependencies/UD_Belarusian-HSE`

Peliriippuvuudesta tuli viimein #oikea #sairaus – WHO lisäsi virallisiin tautiluokituksiin #peliaddiktio #peliriippuvuus #WHO #tautiluokitus https://t.co/P8wSQZzW45

Gaming disorder finally became a #real #disease – WHO added (it) to the official classification of diseases #gamingdisorder #gamingaddiction #WHO #classificationofdiseases https://t.co/P8wSQZzW45

Figure 1: An example of a typical tweet including hashtags both as replacing normal, content-bearing words (#oikea/#real and #sairaus/#disease) as well as listing topical keywords at the end.

names, we opted for labeling all mentions with `PROPN` on the part-of-speech level, while the syntactic relation depends on how the token is used. For mentions addressing the dialogue participant we follow the other treebanks by annotating them with the `vocative` dependency relation, while those used as content-bearing words are annotated with their corresponding function in the sentence (e.g. subject or object).

Hashtags are annotated in various ways in the released treebanks. Sanguinetti et al. (2017) and Bhat et al. (2018) analyse all hashtags as symbols (`SYM`), whereas Liu et al. (2018) uses `X` for topical hashtags, while annotating content-bearing hashtags as any normal tokens. In terms of relations, both Sanguinetti et al. (2017) and Liu et al. (2018) use the corresponding relations in the sentence for content-bearing hashtags, while Bhat et al. (2018) and Blodgett et al. (2018) do not distinguish content-bearing hashtags from the topical ones. The topical hashtags are annotated as `parataxis` (Sanguinetti et al., 2017; Blodgett et al., 2018), or `discourse` (Liu et al., 2018; Bhat et al., 2018). We opted for analysing hashtags with their corresponding part-of-speech tags when the token is an actual Finnish word (i.e. `#beautiful` would be an adjective and `#forest` a noun). However, in some cases giving a real part-of-speech analyse for a hashtag is not meaningful, this would be the case for example with foreign words or tokens artificially joining several words together (e.g. `#thisisbeautiful`). For these, the `X` part-of-speech tag is used in the same manner as would be done with similar regular tokens as well. In the relation annotation, we annotate topical hashtags with the `discourse` relation, while content-bearing hashtags receive annotation regarding its real syntactic function in the sentence.

Due to the choices done during the text preprocessing, retweet markers often appearing in Twitter corpora (such as `RT` in the beginning of a tweet), do not appear in our data. Regarding URLs and emoticons quite frequently occurring in the corpus, we follow the general Finnish-TDT annotation standards, where both are annotated as symbols (`SYM`) in the part-of-speech level. While in Finnish-TDT emoticons are always attached with the `discourse` relation to the sentence root, the relation and attachment of URLs depend on the sentence context. However, most of the URLs appearing in tweets are sentence-final referential items, which do not hold any content-bearing function, we use the same `discourse` relation for such URLs as well.

4. Treebank Statistics

The statistics of the Finnish-OOD corpus are summarized in Table 1, where the section-specific document, sentence and syntactic word counts are plotted together with the two other corpora annotated using the same guidelines and used later in the parsing experiments, Finnish-TDT and Finnish-PUD. The total size of the Finnish-OOD corpus is 19,382 syntactic words (2,122 sentences), where *syntactic word* is the basic element of syntactic annotation in Universal Dependencies. The different subsections vary in size between 2,005 (poetry) and 6,906 words (web documents). The whole corpus is released as test data only.

For comparison, among the 217 test sets in the present Universal Dependencies release 2.9, the average length is 17,946 words and median length is 11,385 words. This makes the Finnish-OOD with its 19,382 words an average UD test set in length, in fact considerably above the median length, ranking 53th out of 217. In terms of full UD treebanks (not only their test sets), Finnish-OOD still contains more total words than 81 of the 217 UD treebanks.

5. Out-of-domain Parsing

In this section we report on dependency parsing experiments, where the parser trained on the Finnish-TDT treebank is tested both on its in-domain data (Finnish-TDT) and out-of-domain data using the newly introduced Finnish-OOD and the existing Finnish-PUD datasets. First, we measure off-the-shelf parsing performance on these datasets in order to report baseline performance directly comparable to other studies, and later perform several detailed section-wise analyses. Additionally, since the Finnish-TDT treebank preserves metadata about the original text sources, we also carry out "leave section out" experiments across the 10 sections of the Finnish-TDT treebank, obtaining further out-of-domain parsing experiments. These allow us to gauge the benefit of the new dataset compared to what was available previously.

The parsing experiments are carried out using the Turku Neural Parser Pipeline (Kanerva et al., 2018), which is a full parsing pipeline with parsing accuracy at the level of present state-of-the-art for UD Finnish parsing. Updated from its original release, the current pipeline consist of a segmentation module based on the UDPipe implementation (Straka and Straková, 2017), custom part-of-speech and morphological feature tagger including separate POS and feature classification

	Train			Dev			Test		
Section	**Doc.**	**Sent.**	**Words**	**Doc.**	**Sent.**	**Words**	**Doc.**	**Sent.**	**Words**
Finnish-TDT									
Wikipedia	160	1,799	25,109	20	200	2,890	20	270	3,936
Fiction	51	2,202	26,342	7	221	2,785	7	316	3,732
Legal	23	914	19,130	3	85	1,938	3	142	2,892
Blogs	61	1,356	16,773	8	259	3,348	8	166	2,219
EuroParl	64	872	16,298	8	94	1,674	8	116	1,986
Grammar examples	—	1,601	13,608	—	200	1,623	—	201	1,771
Wikinews	80	921	11,953	10	92	1,086	10	107	1,256
University news	40	765	10,644	5	86	1,342	5	91	1,243
Economy news	40	854	10,499	5	63	821	5	85	1,136
Student magazines	19	933	12,668	2	64	823	2	61	928
Total	—	**12,217**	**163,024**	—	**1,364**	**18,330**	—	**1,555**	**21,099**
Finnish-PUD									
Wikipedia							251	625	9,901
News							146	375	5,916
Total							**397**	**1,000**	**15,817**
Finnish-OOD									
Web documents							30	584	6,906
Clinical							2	939	5,330
Online discussions							51	263	3,071
Tweets							130	192	2,070
Poetry							6	144	2,005
Total							**218**	**2,122**	**19,382**

Table 1: Section-specific statistics for Finnish TDT, PUD and OOD treebanks in terms of document, sentence and token counts. Sections in each treebank are sorted in descending order based on the test set token count.

layers on top of shared pre-trained embeddings, graph-based bi-affine parser of (Dozat et al., 2017) based on its implementation in Diaparser[5], and a sequence-to-sequence lemmatizer by Kanerva et al. (2020). Out of these four components, the tagger and parser utilize the pre-trained FinBERT language model by (Virtanen et al., 2019), while the segmenter and lemmatizer modules do not rely on any pre-training.

In Table 2 we report the baseline experiments, where the parser trained on the full TDT corpus training set is evaluated on its own test set (TDT) as well as the two external test sets (PUD and OOD). When applying the model to the PUD dataset, the parsing performance does not decrease, the LAS performance actually being +1pp higher compared to the original TDT test set. Similar observations are reported in multiple other studies as well (see e.g. Zeman et al. (2017)), suggesting the PUD test set being easier compared to the TDT test set. Additionally, one must take into account the fact that although we treat PUD as external, separately constructed treebank, the sections included in PUD have a major domain overlap between those in TDT (namely Wikipedia for PUD Wikipedia and Wikinews, economy news and university news for PUD news). Therefore, the PUD dataset cannot be considered as out-of-domain data for TDT trained models (and was, in fact, never meant to be an out-of-domain test set in the first place). On the contrary to PUD, the parsing performance drastically decreases on the newly introduced Finnish-OOD dataset, LAS decreasing over 13pp from 91.00 to 77.50.

Next, we set out to study this further by breaking down the data section-by-section in each of the three treebanks, and carrying out the "leave section out" experiments also across the 10 different sections of the Finnish-TDT treebank. In these "leave section out" experiments, the trained model has never seen data from the particular section during model training, thus demonstrating the out-of-domain parsing performance on the TDT treebank also. In respect of the two PUD sections (Wikipedia and news), we report numbers when leaving all corresponding TDT sections out during training, while in OOD the model is trained on full TDT data as there is no domain overlap between the treebank sections.

The section-wise results are shown in Table 3. In terms of the OOD sections (web documents, clinical, online discussions, tweets, and poetry), quite unsurprisingly the two best performing out-of-domain sections are web documents and online discussions in terms of parsing accuracy (LAS metric), those sections not very substantially differing from the general data in terms of data annotation, and thus likely closest to the genres seen during the model training as well. In addition to the treebank data, the pre-trained FinBERT model used as starting point in parser fine-tuning, was trained on a

[5] https://github.com/Unipisa/diaparser

Treebank	Tokens	Sent.	Words	UPOS	UFeats	Lemmas	UAS	LAS
TDT	99.6	87.2	99.6	97.9	96.7	95.8	93.0	91.0
PUD	99.6	91.3	99.6	98.0	97.1	95.3	94.0	92.1
OOD	97.6	65.5	97.5	92.5	91.9	91.1	81.6	77.5

Table 2: Baseline parsing experiments for the parser trained on the TDT data, and tested on its own test set (TDT) as well as two external test sets (PUD, OOD).

Source	Domain	Tokens	Sentences	Words	UPOS	UFeats	Lemmas	UAS	LAS
TDT	University news	99.9	81.3	99.9	98.5	98.1	94.1	95.6	93.8
TDT	Student magazines	100.0	100.0	100.0	98.1	96.9	96.3	95.8	93.2
PUD	News	99.8	94.1	99.8	98.1	96.4	95.8	94.0	92.0
TDT	EuroParl	99.9	94.9	99.9	98.5	98.0	98.3	93.7	91.9
PUD	Wikipedia	99.5	87.9	99.5	97.6	96.8	94.9	93.2	91.2
TDT	Economy news	99.9	75.8	99.9	98.1	97.8	97.3	91.7	89.8
TDT	Wikipedia	99.1	89.2	99.1	96.8	96.2	93.4	91.8	89.5
TDT	Wikinews	99.6	81.0	99.6	98.6	96.3	92.6	91.5	89.5
TDT	Blogs	98.9	83.0	98.9	96.8	94.6	94.5	91.5	89.4
TDT	Fiction	99.7	93.2	99.6	96.5	94.3	93.3	90.8	88.4
OOD	Web documents	99.3	80.3	99.3	96.3	95.2	94.3	89.1	86.4
TDT	Legal	99.0	45.2	99.0	97.2	95.8	95.5	88.1	85.9
TDT	Grammar examples	99.8	71.4	99.8	96.2	93.9	94.8	88.5	85.7
OOD	Online discussions	98.1	86.2	98.1	94.0	93.8	93.0	87.9	83.9
OOD	Poetry	99.6	55.5	99.6	95.2	94.7	94.6	80.3	76.1
OOD	Tweets	92.2	57.8	92.1	83.5	82.6	81.4	73.9	69.5
OOD	Clinical	96.4	53.1	96.4	89.2	89.2	88.5	72.0	66.1
AVERAGE		98.9	78.2	98.8	95.8	94.7	93.7	88.8	86.0

Table 3: Out-of-domain parsing performance of a model trained on UD Finnish-TDT and tested on all available test sets. All tests are out-of-domain, i.e. if the test set originates from the TDT treebank, the relevant section is removed from the training data. Similarly, for the PUD test set sections, the corresponding domain was removed from the training data. The results are sorted by LAS and color-coded by difference from the average.

large collection of web and discussion forum data. During pre-training, the FinBERT language model used 3.3B tokens of Finnish including discussion forum data (52%), web crawl (33%), and news (15%). Therefore, although these two genres are out-of-domain in terms of parser training, the parser was exposed to these genres through language model pre-training.

On the other end of the scale in terms of LAS is the clinical domain text, nearly 26pp below the in-domain performance, an accuracy level which is likely too low for practical applications. The parsing accuracy on tweets is about 20pp below the in-domain performance, also a very substantial drop. Other measures, such as the accuracy of POS and morphological tagging, and lemmatization, on the other hand, do not exhibit nearly as substantial drop as the syntactic tree accuracy. Especially lemmatization, which is an important step in search and indexing type of applications, sees a comparatively moderate absolute drop in performance across the various OOD subdomains.

When comparing the different sections from all three treebanks, it's clear that the sections selected for the Finnish-OOD are in general more difficult that the ones present in TDT and PUD even when evaluated in "leave section out" manner. With the exception of OOD web documents having higher LAS than TDT legal and grammar examples, the OOD sections locate to the bottom of the table when sorted in terms of LAS in descending order.

Finally, in Table 4 we compare the in-domain and out-of-domain parsing performance across all sections in the Finnish-TDT corpus by reporting the evaluation performance for both in-domain model where the corresponding section is present in the training data, as well as out-of-domain model, where the section is removed from the training data. In this way we are able to estimate the pure out-of-domain parsing effect, removing the effect of some domains being naturally more difficult to parse than others. While many of the sections do not express substantial differences between the in-domain and out-of-domain performance, especially the legal domain significantly suffers in the out-of-domain setting. Interestingly, when comparing the in-domain performance between different sections, the legal domain receives the second best LAS performance, suggesting the section not being particularly difficult in general but likely the legal text significantly standing out from the other data sources included in the corpus.

6. Conclusion

In this work, we introduced a dedicated out-of-domain, manually annotated test set for UD Finnish parser evaluation including data from five distinct text sources previously absent from the UD Finnish treebanks. The

	F1	Tokens	Sentences	Words	UPOS	UFeats	Lemmas	UAS	LAS
Student magazines	**Δ**	**0.0**	**0.0**	**0.0**	**-0.6**	**-0.2**	**-0.1**	**0.8**	**0.5**
	OOD	100.0	100.0	100.0	98.1	96.9	96.3	95.8	93.2
	ID	100.0	100.0	100.0	98.7	97.1	96.4	95.0	92.7
Blogs	**Δ**	**-0.8**	**-1.2**	**-0.7**	**-1.1**	**-1.4**	**-1.8**	**0.0**	**0.1**
	OOD	98.9	83.0	98.9	96.8	94.6	94.5	91.5	89.4
	ID	99.7	84.2	99.7	97.9	96.1	96.4	91.5	89.2
University news	**Δ**	**-0.1**	**-6.3**	**-0.1**	**0.0**	**0.0**	**-2.5**	**0.1**	**0.1**
	OOD	99.9	81.3	99.9	98.5	98.1	94.1	95.6	93.8
	ID	100.0	87.6	100.0	98.5	98.2	96.5	95.5	93.7
EuroParl	**Δ**	**-0.1**	**0.0**	**-0.1**	**-0.1**	**0.0**	**0.1**	**0.1**	**-0.6**
	OOD	99.9	94.9	99.9	98.5	98.0	98.3	93.7	91.9
	ID	99.9	94.9	99.9	98.6	98.0	98.2	93.6	92.5
Fiction	**Δ**	**0.1**	**2.3**	**0.1**	**-0.8**	**-1.2**	**-1.9**	**-0.8**	**-1.0**
	OOD	99.7	93.2	99.6	96.5	94.3	93.3	90.8	88.4
	ID	99.6	90.9	99.5	97.3	95.5	95.2	91.6	89.4
Economy news	**Δ**	**0.0**	**-0.6**	**0.0**	**-0.2**	**1.0**	**-0.4**	**-0.7**	**-1.2**
	OOD	99.9	75.8	99.9	98.1	97.8	97.3	91.7	89.8
	ID	99.9	76.4	99.9	98.3	96.8	97.8	92.4	91.0
Wikipedia	**Δ**	**-0.1**	**-5.2**	**-0.1**	**-0.4**	**-0.5**	**0.2**	**-1.1**	**-1.2**
	OOD	99.1	89.2	99.1	96.8	96.2	93.4	91.8	89.5
	ID	99.2	94.4	99.2	97.2	96.7	93.2	92.9	90.7
Wikinews	**Δ**	**0.4**	**-2.9**	**0.4**	**0.5**	**-0.4**	**-0.6**	**-1.0**	**-1.2**
	OOD	99.6	81.0	99.6	98.6	96.3	92.6	91.5	89.5
	ID	99.2	83.9	99.2	98.1	96.7	93.2	92.5	90.7
Grammar examples	**Δ**	**-0.3**	**-11.2**	**-0.3**	**-0.8**	**-1.4**	**-1.6**	**-3.3**	**-3.5**
	OOD	99.8	71.4	99.8	96.2	93.9	94.8	88.5	85.7
	ID	100.0	82.5	100.0	97.0	95.3	96.4	91.9	89.2
Legal	**Δ**	**-0.7**	**-27.4**	**-0.7**	**-1.9**	**-2.5**	**-1.8**	**-6.8**	**-7.5**
	OOD	99.0	45.2	99.0	97.2	95.8	95.5	88.1	85.9
	ID	99.7	72.6	99.7	99.1	98.3	97.3	94.9	93.4

Table 4: Parsing performance on the various sections of the UD Finnish TDT treebank. OOD refers to an out-of-domain run, where the section is removed from the training data, while ID refers to an in-domain run, where the section is present in the training data. Their difference Δ then directly shows the absolute loss in parsing performance on each section, as if it were an out of domain section. Since in-domain and out-of-domain numbers can be compared, the fact that some sections have a higher overall parsing performance than others does not affect the results. The sections are sorted by ΔLAS.

selection mirrors practical use cases seen for Finnish dependency parsing in the academia as well as in the industry. In terms of its size, this test set is comparable to other test sets in UD, with its 19,382 syntactic words being considerably above the median UD test set size. Our parsing experiments on this dataset demonstrate that, indeed, syntactic parsing performance can substantially degrade on several domains and the OOD test set now allows us to quantify the effect. On the other hand, we were also able to establish that the effect is at its strongest specifically when measuring the accuracy of the syntactic tree (LAS metric) and is notably less pronounced on the tagging and lemmatization tasks, which have a number of applications in their own right.

Together, these parsing experiments and the Finnish-OOD test set comprise the broadest evaluation of a Finnish state of the art syntactic parser across numerous domains, giving valuable knowledge for all applying the parser outside its training domain in various real-life applications. The new Finnish-OOD treebank is available through the official data releases of the Universal Dependencies framework.

7. Acknowledgements

We would like to thank Akseli Leino, Hans Moen and Henry Suhonen for their help with handling the medical terminology. Computational resources were provided by CSC – the Finnish IT Center for Science, and the research was supported by the Academy of Finland.

8. Bibliographical References

Bhat, I. A., Bhat, R. A., Shrivastava, M., and Sharma, D. M. (2018). Universal dependency parsing for Hindi-English code-switching. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL) 2018*.

Blodgett, S. L., Wei, J., and O'Connor, B. (2018). Twitter universal dependency parsing for African-American and mainstream American English. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425.

Candito, M. and Seddah, D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *TALN 2012 - 19e conférence sur le Traitement Automatique des Langues Naturelles*, Grenoble, France, June.

Cecchini, F. M., Sprugnoli, R., Moretti, G., and Passarotti, M. (2020). Udante: First steps towards the

Universal Dependencies treebank of Dante's Latin works. In *CLiC-it*.

Cignarella, A. T., Bosco, C., and Rosso, P. (2019). Presenting TWITTIRO-UD: An Italian Twitter treebank in Universal Dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 190–197.

Dozat, T., Qi, P., and Manning, C. D. (2017). Stanford's graph-based neural dependency parser at the CoNLL 2017 shared task. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30.

Haverinen, K., Ginter, F., Laippala, V., and Salakoski, T. (2009). Parsing clinical Finnish: Experiments with rule-based and statistical dependency parsers. In Kristiina Jokinen et al., editors, *Proceedings of NODALIDA'09, Odense, Denmark*, pages 65–72.

Haverinen, K., Ginter, F., Viljanen, T., Laippala, V., and Salakoski, T. (2010). Dependency-based PropBanking of clinical Finnish. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 137–141, Uppsala, Sweden, July. Association for Computational Linguistics.

Haverinen, K., Nyblom, J., Viljanen, T., Laippala, V., Kohonen, S., Missilä, A., Ojala, S., Salakoski, T., and Ginter, F. (2014). Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, 48:493–531. Open access.

Kanerva, J., Ginter, F., Miekka, N., Leino, A., and Salakoski, T. (2018). Turku neural parser pipeline: An end-to-end system for the CoNLL 2018 shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics.

Kanerva, J., Ginter, F., and Salakoski, T. (2020). Universal Lemmatizer: A sequence-to-sequence model for lemmatizing Universal Dependencies treebanks. *Natural Language Engineering*, pages 1–30.

Laippala, V., Viljanen, T., Airola, A., Kanerva, J., Salanterä, S., Salakoski, T., and Ginter, F. (2014). Statistical parsing of varieties of clinical Finnish. *Artificial Intelligence in Medicine*, 61(3):131–136. Text Mining and Information Analysis of Health Documents.

Laippala, V., Kyllönen, R., Egbert, J., Biber, D., and Pyysalo, S. (2019). Toward multilingual identification of online registers. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 292–297, Turku, Finland, September–October. Linköping University Electronic Press.

Liu, Y., Zhu, Y., Che, W., Qin, B., Schneider, N., and Smith, N. A. (2018). Parsing tweets into Universal Dependencies. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL) 2018*.

Luotolahti, J., Kanerva, J., Laippala, V., Pyysalo, S., and Ginter, F. (2015). Towards universal web parsebanks. In *Proceedings of the International Conference on Dependency Linguistics (Depling'15)*, pages 211–220. Uppsala University.

Mititelu, V. B. and Mitrofan, M. (2020). The Romanian medical treebank – SiMoNERo. In *The 15th Edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing*.

Mititelu, V. B. (2018). Modern syntactic analysis of Romanian. *Clasic şi modern în cercetarea filologică românească actuală*.

Mărănduc, C. and Bobicev, V. (2017). Non standard treebank Romania – Republic of Moldova in the Universal Dependencies. In *Proceedings of the Conference on Mathematical Foundations of Informatics (MFOI'2017)*, pages 111–116.

Müller-Eberstein, M., van der Goot, R., and Plank, B. (2021). How universal is genre in Universal Dependencies? In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT 2021) at SyntaxFest*.

Pyysalo, S., Kanerva, J., Missilä, A., Laippala, V., and Ginter, F. (2015). Universal Dependencies for Finnish. In *Proceedings of NoDaLiDa 2015*, pages 163–172. NEALT.

Raab, J. (2008). The Czech academic corpus 2.0 guide. *The Prague Bulletin of Mathematical Linguistics*, 89(2008):41–96.

Sanguinetti, M., Bosco, C., Mazzei, A., Lavelli, A., and Tamburini, F. (2017). Annotating Italian social media texts in Universal Dependencies. In *Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 229–239. Linköping University Electronic Press.

Seraji, M., Ginter, F., and Nivre, J. (2016). Universal dependencies for Persian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2361–2365.

Shavrina, T. and Shapovalova, O. (2017). To the methodology of corpus construction for machine learning: Taiga syntax tree corpus and parser. In *Proceedings of the International Conference CORPORA-2017*.

Stein, A. and Prévost, S. (2013). Syntactic annotation of medieval texts. *New methods in historical corpora*, 3:275.

Straka, M. and Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Suominen, H., Lundgrén-Laine, H., Salanterä, S., Karsten, H., and Salakoski, T. (2009). Information flow in intensive care narratives. In *2009*

IEEE International Conference on Bioinformatics and Biomedicine Workshop, pages 325–340.

Tyers, F. M. and Henderson, R. (2021). A corpus of K'iche' annotated for morphosyntactic structure. In *Proceedings of the First Workshop on NLP for Indigenous Languages of the Americas (Americas-NLP)*.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., and Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076*.

Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinková, S., Hajič jr., J., Hlaváčová, J., Kettnerová, V., Urešová, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C. D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., de Paiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, Ç., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H. F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonça, G., Lando, T., Nitisaroj, R., and Li, J. (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*, Vancouver, Canada, August. Association for Computational Linguistics.

9. Language Resource References

Voutilainen, A., Purtonen, T., Leisko-Järvinen, S., Kumlander, M., Linden, K., Nissinen, M., and Hardwick, S. (2010). Finntreebank 1.

Zeman, D., Nivre, J., Abrams, M., Ackermann, E., Aepli, N., Aghaei, H., Agić, Ž., Ahmadi, A., Ahrenberg, L., Ajede, C. K., Aleksandravičiūtė, G., Alfina, I., Antonsen, L., Aplonova, K., Aquino, A., Aragon, C., Aranzabe, M. J., Arıcan, B. N., Arnardóttir, Ĥ., Arutie, G., Arwidarasti, J. N., Asahara, M., Aslan, D. B., Ateyah, L., Atmaca, F., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Balasubramani, K., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Barkarson, S., Basmov, V., Batchelor, C., Bauer, J., Bedir, S. T., Bengoetxea, K., Berk, G., Berzak, Y., Bhat, I. A., Bhat, R. A., Biagetti, E., Bick, E., Bielinskienė, A., Bjarnadóttir, K., Blokland, R., Bobicev, V., Boizou, L., Borges Völker, E., Börstell, C., Bosco, C., Bouma, G., Bowman, S., Boyd, A., Braggaar, A., Brokaitė, K., Burchardt, A., Candito, M., Caron, B., Caron, G., Cassidy, L., Cavalcanti, T., Cebiroğlu Eryiğit, G., Cecchini, F. M., Celano, G. G. A., Čéplö, S., Cesur, N., Cetin, S., Çetinoğlu, Ö., Chalub, F., Chauhan, S., Chi, E., Chika, T., Cho, Y., Choi, J., Chun, J., Cignarella, A. T., Cinková, S., Collomb, A., Çöltekin, Ç., Connor, M., Courtin, M., Cristescu, M., Daniel, P., Davidson, E., de Marneffe, M.-C., de Paiva, V., Derin, M. O., de Souza, E., Diaz de Ilarraza, A., Dickerson, C., Dinakaramani, A., Di Nuovo, E., Dione, B., Dirix, P., Dobrovoljc, K., Dozat, T., Droganova, K., Dwivedi, P., Eckhoff, H., Eiche, S., Eli, M., Elkahky, A., Ephrem, B., Erina, O., Erjavec, T., Etienne, A., Evelyn, W., Facundes, S., Farkas, R., Fernanda, M., Fernandez Alcalde, H., Foster, J., Freitas, C., Fujita, K., Gajdošová, K., Galbraith, D., Garcia, M., Gärdenfors, M., Garza, S., Gerardi, F. F., Gerdes, K., Ginter, F., Godoy, G., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., González Saavedra, B., Griciūtė, B., Grioni, M., Grobol, L., Grūzītis, N., Guillaume, B., Guillot-Barbance, C., Güngör, T., Habash, N., Hafsteinsson, H., Hajič, J., Hajič jr., J., Hämäläinen, M., Hà Mỹ, L., Han, N.-R., Hanifmuti, M. Y., Hardwick, S., Harris, K., Haug, D., Heinecke, J., Hellwig, O., Hennig, F., Hladká, B., Hlaváčová, J., Hociung, F., Hohle, P., Huber, E., Hwang, J., Ikeda, T., Ingason, A. K., Ion, R., Irimia, E., Ishola, Ọ., Ito, K., Jelínek, T., Jha, A., Johannsen, A., Jónsdóttir, H., Jørgensen, F., Juutinen, M., K, S., Kaşıkara, H., Kaasen, A., Kabaeva, N., Kahane, S., Kanayama, H., Kanerva, J., Kara, N., Katz, B., Kayadelen, T., Kenney, J., Kettnerová, V., Kirchner, J., Klementieva, E., Köhn, A., Köksal, A., Kopacewicz, K., Korkiakangas, T., Kotsyba, N., Kovalevskaitė, J., Krek, S., Krishnamurthy, P., Kuyrukçu, O., Kuzgun, A., Kwak, S., Laippala, V., Lam, L., Lambertino, L., Lando, T., Larasati, S. D., Lavrentiev, A., Lee, J., Lê H`ông, P., Lenci, A., Lertpradit, S., Leung, H., Levina, M., Li, C. Y., Li, J., Li, K., Li, Y., Lim, K., Lima Padovani, B., Lindén, K., Ljubešić, N., Loginova, O., Luthfi, A., Luukko, M., Lyashevskaya, O., Lynn, T., Macketanz, V., Makazhanov, A., Mandl, M., Manning, C., Manurung, R., Marşan, B., Mărănduc, C., Mareček, D., Marheinecke, K., Martínez Alonso, H., Martins, A., Mašek, J., Matsuda, H., Matsumoto, Y., Mazzei, A., McDonald, R., McGuinness, S., Mendonça, G., Miekka, N., Mischenkova, K., Misirpashayeva, M., Missilä, A., Mititelu, C., Mitrofan, M., Miyao, Y., Mojiri Foroushani, A., Molnár, J., Moloodi, A., Montemagni, S., More, A., Moreno Romero, L., Moretti, G., Mori, K. S., Mori, S., Morioka, T., Moro, S., Mortensen, B., Moskalevskyi, B., Muischnek, K., Munro, R., Murawaki, Y., Müürisep, K., Nainwani, P., Nakhlé, M., Navarro Horñiacek, J. I., Nedoluzhko, A., Nešpore-Bērzkalne, G., Nevaci, M., Nguy˜ên Thị, L., Nguy˜ên Thị Minh, H., Nikaido, Y., Nikolaev, V., Nitisaroj, R., Nourian, A., Nurmi, H., Ojala, S., Ojha, A. K., Olúòkun, A., Omura, M., Onwuegbuzia, E., Osenova, P., Östling, R., Øvrelid, L., Özateş, Ş. B., Özçelik, M., Özgür, A., Öztürk Başaran, B., Park, H. H., Par-

tanen, N., Pascual, E., Passarotti, M., Patejuk, A., Paulino-Passos, G., Peljak-Łapińska, A., Peng, S., Perez, C.-A., Perkova, N., Perrier, G., Petrov, S., Petrova, D., Phelan, J., Piitulainen, J., Pirinen, T. A., Pitler, E., Plank, B., Poibeau, T., Ponomareva, L., Popel, M., Pretkalniņa, L., Prévost, S., Prokopidis, P., Przepiórkowski, A., Puolakainen, T., Pyysalo, S., Qi, P., Rääbis, A., Rademaker, A., Rama, T., Ramasamy, L., Ramisch, C., Rashel, F., Rasooli, M. S., Ravishankar, V., Real, L., Rebeja, P., Reddy, S., Rehm, G., Riabov, I., Rießler, M., Rimkutė, E., Rinaldi, L., Rituma, L., Rocha, L., Rögnvaldsson, E., Romanenko, M., Rosa, R., Roșca, V., Rovati, D., Rudina, O., Rueter, J., Rúnarsson, K., Sadde, S., Safari, P., Sagot, B., Sahala, A., Saleh, S., Salomoni, A., Samardžić, T., Samson, S., Sanguinetti, M., Sanıyar, E., Särg, D., Saulīte, B., Sawanakunanon, Y., Saxena, S., Scannell, K., Scarlata, S., Schneider, N., Schuster, S., Schwartz, L., Seddah, D., Seeker, W., Seraji, M., Shen, M., Shimada, A., Shirasu, H., Shishkina, Y., Shohibussirri, M., Sichinava, D., Siewert, J., Sigursson, E. F., Silveira, A., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Skachedubova, M., Smith, A., Soares-Bastos, I., Spadine, C., Sprugnoli, R., Steingrímsson, S., Stella, A., Straka, M., Strickland, E., Strnadová, J., Suhr, A., Sulestio, Y. L., Sulubacak, U., Suzuki, S., Szántó, Z., Taji, D., Takahashi, Y., Tamburini, F., Tan, M. A. C., Tanaka, T., Tella, S., Tellier, I., Testori, M., Thomas, G., Torga, L., Toska, M., Trosterud, T., Trukhina, A., Tsarfaty, R., Türk, U., Tyers, F., Uematsu, S., Untilov, R., Urešová, Z., Uria, L., Uszkoreit, H., Utka, A., Vajjala, S., van der Goot, R., Vanhove, M., van Niekerk, D., van Noord, G., Varga, V., Villemonte de la Clergerie, E., Vincze, V., Vlasova, N., Wakasa, A., Wallenberg, J. C., Wallin, L., Walsh, A., Wang, J. X., Washington, J. N., Wendt, M., Widmer, P., Williams, S., Wirén, M., Wittern, C., Woldemariam, T., Wong, T.-s., Wróblewska, A., Yako, M., Yamashita, K., Yamazaki, N., Yan, C., Yasuoka, K., Yavrumyan, M. M., Yenice, A. B., Yıldız, O. T., Yu, Z., Žabokrtský, Z., Zahra, S., Zeldes, A., Zhu, H., Zhuravleva, A., and Ziane, R. (2021). Universal Dependencies 2.8.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.