Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation, pages 33–38
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Universal Dependency Treebanks for Low-Resource Indian Languages: The Case of Bhojpuri

Atul Kr. Ojha, Daniel Zeman
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
shashwatup9k@gmail.com, zeman@ufal.mff.cuni.cz

Abstract

This paper presents the first dependency treebank for Bhojpuri, a resource-poor language that belongs to the Indo-Aryan language family. The objective behind the Bhojpuri Treebank (BHTB) project is to create a substantial, syntactically annotated treebank which not only acts as a valuable resource in building language technological tools, also helps in cross-lingual learning and typological research. Currently, the treebank consists of 4,881 annotated tokens in accordance with the annotation scheme of Universal Dependencies (UD). A Bhojpuri tagger and parser were created using machine learning approach. The accuracy of the model is 57.49% UAS, 45.50% LAS, 79.69% UPOS accuracy and 77.64% XPOS accuracy. The paper describes the details of the project including a discussion on linguistic analysis and annotation process of the Bhojpuri UD treebank.

Keywords: Bhojpuri, Indian languages, parser, low-resource languages, Universal Dependencies, UDPipe

1. Introduction

Bhojpuri is an Indian language that belongs to the Indo-Aryan language group. It is spoken in the western part of Bihar, north-western part of Jharkhand, and the Purvanchal region of Uttar Pradesh. The number of speakers according to the present Census of India[1] is considerably large at 50,579,447. It should be noted that Bhojpuri is spoken not just in India but also in other countries such as Nepal, Trinidad, Mauritius, Guyana, Suriname, and Fiji (Verma, 2003; Ojha, 2019). Since Bhojpuri was considered a dialect of Hindi for a long time, it did not attract much attention from linguists and hence remains among the many lesser known and less-resourced languages of India.

With the rise in language technology for Indian languages, significant developments have been achieved in major Indian languages but contributions towards research in the lesser-known/low-resourced languages remain minimal. Most parsers and treebanks have been developed for the scheduled (official) languages; the non-scheduled and lesser known languages still have a long way to go. In its endevour to fill this gap, the present paper discusses the creation and development of Bhojpuri Universal Dependency (UD) treebank and parser. UD has been acknowledged as an emerging framework for cross-linguistically consistent grammatical annotation. There is an open community with over 300 contributors, who have produced 157 treebanks in 90 languages to date (as per the latest release of UD–v2.5[2] (Zeman et al., 2019). The primary aim of this project is to facilitate multilingual parser development. The system will also take into account cross-lingual learning and perform parsing research from the perspective of language typology. The syntactic part of the annotation scheme can be seen as an evolution of (universal) Stanford dependencies (De Marneffe et al., 2006; De Marneffe and Manning, 2008; De Marneffe et al., 2014), while the lexical and morphological part builds on the Google universal part-of-speech tags (Petrov et al., 2012), and the Interset Interlingua (Zeman, 2008) for morpho-syntactic tagsets.

Section 2. discusses the language resources that have been created so far for the Bhojpuri language. While Bhojpuri has some considerable efforts in progress, it has no dependency treebank and parser. Section 3. discusses methodology to develop the Bhojpuri treebank. Section 4. presents a linguistic study and annotation of the Bhojpuri Dependency treebank while Sections 5. and 6. discuss the development of the Bhojpuri parser and evaluate its results. The final section ends with concluding remarks and future work.

2. Literature Review

In 2013, a consortium was formed under the leadership of IIIT Hyderabad to start a project sponsored by TDIL, Government of India, and called The Development of Dependency Treebank for Indian Languages.[3] The fundamental objective of this project was to resurrect annotation work towards monolingual and parallel treebanks for languages such as Hindi, Marathi, Bengali, Kannada, and Malayalam. To accomplish this treebank model, the Pāṇinian Kāraka Dependency annotation scheme was followed (Bharati et al., 2006). The annotation scheme was previously also utilized to annotate data in Telugu, Urdu, and Kashmiri (Begum et al., 2008; Husain et al., 2010; Bhat, 2017).

Within the Universal Dependencies framework, as of UD release 2.5, treebanks and parsers are available for Sanskrit, Hindi, Urdu, Marathi, Tamil, and Telugu (Zeman et al., 2019; Ravishankar, 2017; Straka and Straková, 2019).

[1] http://www.censusindia.gov.in/2011Census/C-16_25062018_NEW.pdf

[2] https://universaldependencies.org/

[3] http://meity.gov.in/content/language-computing-group-vi

NLP research in Bhojpuri has led to the development of a statistical POS tagger (Ojha et al., 2015; Singh and Jha, 2015), a machine-readable dictionary (Ojha, 2016), a language identification tool (Kumar et al., 2018), a Sanskrit-Bhojpuri machine translation system (Sinha and Jha, 2018), and more recently an English-Bhojpuri machine translation system (Ojha, 2019). Nevertheless, there is no prior work on Bhojpuri treebanking and parsing.

3. Data and Methodology

The data for the treebank has been extracted from the Bhojpuri Language Technological Resources (BHLTR) project[4] (Ojha, 2019). The data has been selected from the news and non-fiction domains. For this project, we use 5000 sentences (105,174 tokens) which were previously annotated with part-of-speech tags now used for language-specific tags representation in the XPOS. Out of 5000 sentences and 105,174 tokens, 254 sentences and 4881 tokens have been manually annotated at present, and released in UD 2.5.

The Bhojpuri Treebank (BHTB)[5] follows the annotation guidelines of Universal Dependencies for part-of-speech categories, morphological features, and dependency relations. Since Bhojpuri is closely related to Hindi and there is already a Hindi treebank in UD, we followed the Hindi tagset wherever possible. Besides, the universal part-of-speech tagset (UPOS), UD also permits a secondary tagset (XPOS), which is language-specific and typically follows an established pre-UD practice. We use the Bureau of Indian Standards (BIS) POS tagset[6] here. This is a generic tagset for annotating Indian languages. The XPOS tags were already present in our input data and we obtained the UPOS tags through automatic conversion from XPOS. In addition to UPOS, we use 16 lexical and inflectional features defined in UD. The details of used morphological features, UPOS tags and UD relations and their statistics are demonstrated in Tables 1, 2 and 3.

4. Linguistic Analysis of Bhojpuri Dependency Treebank

As discussed earlier, we followed UD v2 guidelines to annotate BHTB. We mention below some Bhojpuri constructions and their linguistics analysis under UD.

- **Nominal Predicate with Copula**
  Figure 1 is an example of a nominal predicate with a copula. In the example, the copula हऽ *(ha)* is preceded by the nominal predicate देश *(deśa)*. In accordance with UD, the nominal देश is the root of the sentence and the verbal copula हऽ is attached to it via the relation `cop`, as shown in the figure.

[4] `https://github.com/shashwatup9k/bho-resources`

[5] https://github.com/UniversalDependencies/UD_Bhojpuri-BHTB

[6] `http://tdil-dc.in/tdildcMain/articles/134692Draft%20POS%20Tag%20standard.pdf`

Morph. Features	Description	Count
AdpType	Adposition type	726
Aspect	Aspect	242
Case	Case	3007
Echo	Echo word or a reduplicative	9
Foreign	Foreign word	5
Gender	Gender	2916
Mood	Mood	37
Number	Number	3144
NumType	Numeral type	84
Person	Person	2485
Polite	Politeness	103
Poss	Possessive	1
PronType	Pronominal type	163
VerbForm	Form of verb or deverbative	293
Voice	Voice	231

Table 1: Statistics of morphological features used in the BHTB

UPOS Tags	Description	Count
ADJ	Adjective	183
ADP	Adposition	720
ADV	Adverb	18
AUX	Auxiliary	256
CCONJ	Coordinating conjunction	112
DET	Determiner	256
INTJ	Interjection	4
NOUN	Noun	1361
NUM	Numeral	110
PART	Particle	135
PRON	Pronoun	230
PROPN	Proper noun	352
PUNCT	Punctuation	504
SCONJ	Subordinating conjunction	86
VERB	Verb	553
X	Other	1

Table 2: Statistics of UPOS tags used in the BHTB

- **Verbal Predicates**
  We discuss three types of verbal predicates in this section: a simple verb construction, a conjunct verb construction and a compound verb construction. A simple verb or a verb with auxiliary is called simple verb construction. An example is given in Figure 2 where a verb गइल *(gaila)* is combined with an auxiliary रहन *(rahana)*. In UD, the verb is tagged as the `root` of the sentence and the auxiliary is related to it via the relation `aux`. In Figure 3, the sentence has both a conjunct verb and a compound verb. The conjunct verb is formed by combining a noun नजर *(najara)* with the compound verb चलि गइल *(cali gaila)*. Moreover, the compound verb is formed by combining

UD Relations	Description	Count
acl	Clausal modifier of noun	82
advcl	Adverbial clausal modifier	57
advmod	Adverbial modifier	11
amod	Adjectival modifier of noun	160
aux	Auxiliary verb	224
case	Case marker	661
cc	Coordinating conjunction	15
ccomp	Clausal complement	46
clf	Classifier	3
compound	Compound	1191
conj	Non-first conjunct	96
cop	Copula	2
csubj	Clausal subject	9
dep	Unspecified dependency	11
det	Determiner	118
discourse	Discourse element	7
fixed	Non-first word of fixed expression	9
flat	Non-first word of flat structure	1
goeswith	Non-first part of broken word	1
iobj	Indirect object	18
list	List item	10
mark	Subordinating marker	89
nmod	Nominal modifier of noun	678
nsubj	Nominal subject	192
nummod	Numeric modifier	41
obj	Direct object	93
obl	Oblique nominal	245
punct	Punctuation	504
root	Root	254
xcomp	Open clausal complement	55

Table 3: UD relations used in BHTB. Out of 37 relations defined in the UD guidelines, we currently use 30

the main verb चलि *(cali)* with the light verb गइल *(gaila)*. In UD, the main verb चलि is marked as `root` and the noun नजर is related to it via the compound relation. The light verb गइल, on the other hand, is related to the main verb via the relation `aux`.

- **Coordination**
  Figure 4 illustrates an example of a coordinate construction. The conjuncts आईं *(āīṁ)* and सपरिवार आईं *(saparivāra āīṁ)* are conjoined through the conjunction आ *(ā)*. In UD, the first conjunct serves as the technical head and the second conjunct is attached to it via the relation `conj`. The coordinating conjunction आ is related to the following conjunct via the relation `cc`.

- **Types of Clauses**
  In UD, subordinate clauses are sub-categorised in five types: clausal subject (`csubj`), clausal complement (`ccomp`), open clausal complement (`xcomp`), adverbial clausal modifier (`advcl`), clausal modifier of noun (`acl`). We found all five types of clauses in Bhojpuri. We demonstrate here an example of a clausal modifier of noun. In Figure 5, the clause जवना के सभे आनन्द लिहल *(javanā ke sabhe ānanda lihala)* is a modifier of the noun गवनई *(gavanaī)* which is shown by the tag `acl`. Interestingly, the clausal modifier is displaced from its base position and right adjoined to the verb. This is a common property of many Indo-Aryan languages.

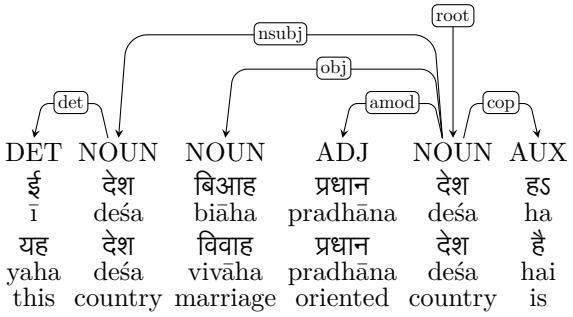

Figure 1: A parallel copular sentence in Bhojpuri (lines 2–3) and Hindi (lines 4–5).

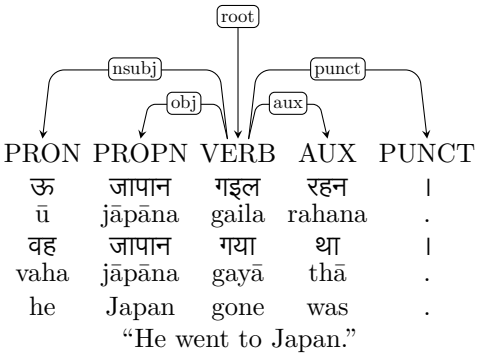

Figure 2: A parallel sentence with simple verb construction in Bhojpuri (lines 2–3) and Hindi (lines 4–5)

5. Development of a Bhojpuri Parser

Initially, we conducted an experiment where the parser was trained on the Hindi UD treebank (HDTB) and applied to Bhojpuri. While the parsing quality suffers from the differences between the two languages, this is partially counterbalanced by the fact that the Hindi treebank is large: there are 16,647 training sentences (351,704 tokens). This experiment was evaluated on the first 50 manually annotated sentences (650 tokens) in Bhojpuri. We found that the Hindi parser gives only 56.77% UAS, 45.61% LAS, and 52.35% UPOS tagging accuracy, respectively (as shown in Table 4). Along with this, the tokenization accuracy of the Hindi model is only 89.15%.

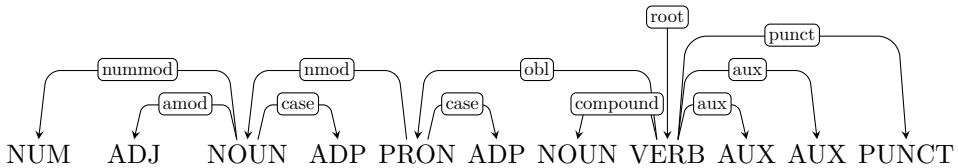

Figure 3: A parallel sentence with compound verb construction in Bhojpuri (lines 2–3) and Hindi (lines 4–5)

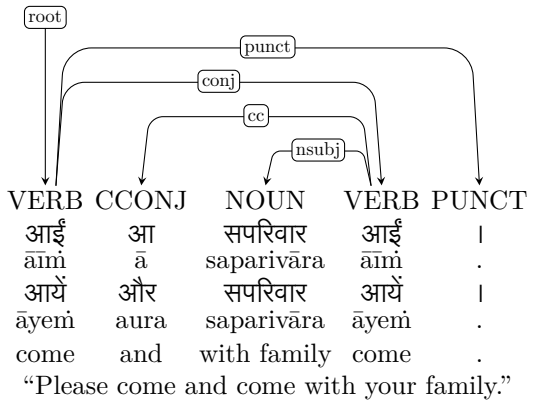

Figure 4: A parallel sentence with coordination in Bhojpuri (lines 2–3) and Hindi (lines 4–5)

Since accuracy was very low, to improve it, we conducted three experiments solely based on Bhojpuri data to build a Bhojpuri parser. In all the three experiments, we used the UDPipe open source tool (Straka and Straková, 2017). In all the cases, we used cross-validation 90:10 average. We used UDPipe's default epoch size and learning rate, while the other hyperparameters were randomized.

However, the three experiments differ in the data size: first experiment was conducted on 1000 tokens, the second experiment was conducted on 1500 tokens, and the third experiment was conducted on 4880 tokens. The results are discussed below in the evaluation section.

Tokenization F_1	**UPOS**	**UAS**	**LAS**
89.15%	52.35%	56.77%	45.61%

Table 4: Accuracy of a UDPipe model trained on the Hindi UD treebank (HDTB) and applied to the first 50 Bhojpuri sentences.

6. Evaluation

The results of the three experiments with the Bhojpuri parser are shown in Table 5 and in Figure 6. In terms of labeled attachment score, the results show that the third experiment, which was trained on 4880 tokens, produced slightly better result in comparison to the previous two experiments. The third experiment also performed better in comparison to the result of the Hindi parser that was used to tag the Bhojpuri data. This result is interesting. Even though Hindi is a closely related language and its parser is trained on a large amount of data (351,704 tokens), the Hindi parser couldn't perform any better on the Bhojpuri data. This shows that the same parser should not be used on two different languages even when they are closely related. We need to develop a separate and robust parser.

	XPOS	**UPOS**	**UAS**	**LAS**
Experiment 1	66.67%	69.86%	39.73%	31.96%
Experiment 2	66.95%	60.17%	45.76%	35.59%
Experiment 3	77.64%	79.69%	57.49%	45.50%

Table 5: UDPipe accuracy of the conducted experiments

7. Conclusion and Future Work

This paper reports the development of the very first dependency treebank for Bhojpuri language using the annotation scheme of Universal Dependency (UD). The primary aim behind undertaking this project is to facilitate dependency treebank for Bhojpuri which is one of the low-resourced Indian languages. Currently, the Bhojpuri treebank consists of 4,881 tokens. This paper discussed the annotation guidelines used, the annotation process, and statistics of the used tags/UD relations. It also presented the linguistic analysis of the Bhojpuri treebank using examples from the language. Additionally, this paper presented a Bhojpuri parser which has been trained on UDPipe tool. The accuracy of the developed model is 57.49% UAS, 45.50% LAS, 79.69% UPOS and 77.64% XPOS.

In the near future, we plan to extend BHTB up to 5,000 sentences and develop a parallel Bhojpuri-Hindi treebank. Along with this we will improve and develop a robust Bhojpuri parser using a neural model.

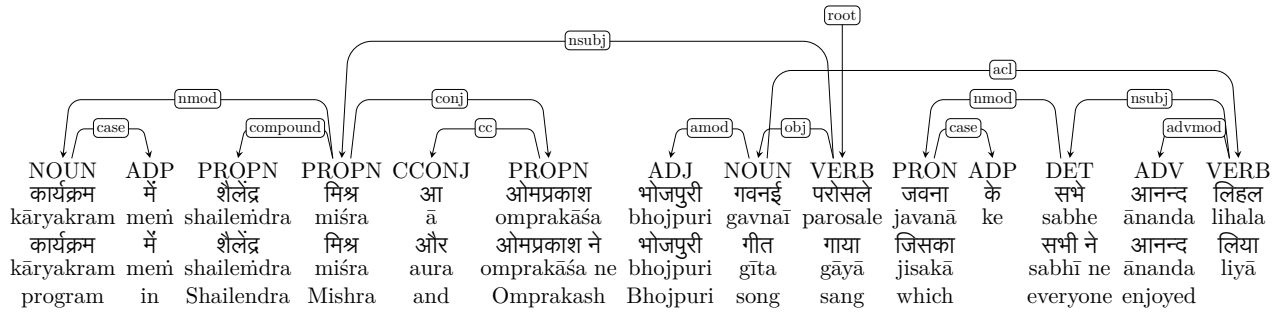

Figure 5: A parallel sentence with clauses in Bhojpuri (lines 2–3) and Hindi (lines 4–5)

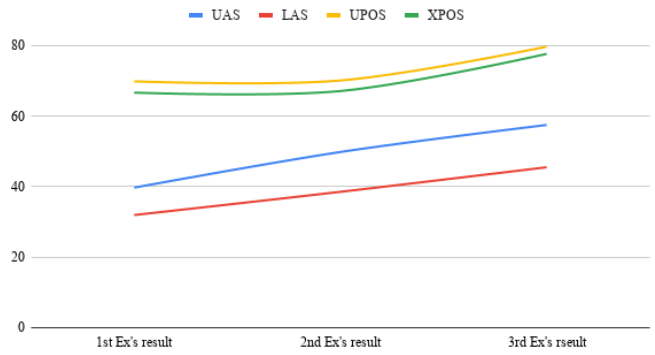

Figure 6: Learning curve of the Bhojpuri models.

8. Acknowledgements

This work has been supported by LINDAT/CLARIAH-CZ and Khresmoi, the grants no. LM2018101 and 7E11042 of the Ministry of Education, Youth and Sports of the Czech Republic, and FP7-ICT-2010-6-257528 of the European Union.

9. References

Begum, R., Husain, S., Dhwaj, A., Sharma, D. M., Bai, L., and Sangal, R. (2008). Dependency annotation scheme for Indian languages. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Bharati, A., Chaitanya, V., and Sangal, R. (2006). *Natural Language Processing: A Paninian Perspective*. Prentice-Hall of India, New Delhi, India.

Bhat, R. A. (2017). *Exploiting linguistic knowledge to address representation and sparsity issues in dependency parsing of Indian languages*. Ph.D. thesis, PhD thesis, International Institute of Information Technology, India.

De Marneffe, M.-C. and Manning, C. D. (2008). The Stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pages 1–8. Association for Computational Linguistics.

De Marneffe, M.-C., MacCartney, B., Manning, C. D., et al. (2006). Generating typed dependency parses from phrase structure parses. In *LREC*, volume 6, pages 449–454.

De Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal Stanford dependencies: A cross-linguistic typology. In *LREC*, volume 14, pages 4585–4592.

Husain, S., Mannem, P., Ambati, B. R., and Gadde, P. (2010). The ICON-2010 tools contest on Indian language dependency parsing. *Proceedings of ICON-2010 Tools Contest on Indian Language Dependency Parsing, ICON*, 10:1–8.

Kumar, R., Lahiri, B., Alok, D., Ojha, A. K., Jain, M., Basit, A., and Dawar, Y. (2018). Automatic identification of closely-related Indian languages: Resources and experiments. In Girish Nath Jha, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Ojha, A. K., Behera, P., Singh, S., and Jha, G. N. (2015). Training & evaluation of POS taggers in Indo-Aryan languages: a case of Hindi, Odia and Bhojpuri. In *the proceedings of 7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 524–529.

Ojha, A. K. (2016). Developing a machine readable multilingual dictionary for Bhojpuri-Hindi-English.

Ojha, A. K. (2019). *English-Bhojpuri SMT System: Insights from the Karaka Model*. Ph.D. thesis, Ph D thesis, Jawaharlal Nehru University, New Delhi, India.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Ravishankar, V. (2017). A Universal Dependencies treebank for Marathi. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 190–200.

Singh, S. and Jha, G. N. (2015). Statistical tagger for Bhojpuri (employing support vector machine). In *2015 International Conference on Ad-*

vances in Computing, Communications and Informatics (ICACCI), pages 1524–1529. IEEE.

Sinha, S. and Jha, G. N. (2018). Issues in conversational Sanskrit to Bhojpuri MT. In Girish Nath Jha, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

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.

Straka, M. and Straková, J. (2019). Universal dependencies 2.4 models for UDPipe (2019-05-31). LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Verma, M. K. (2003). Bhojpuri. *The Indo-Aryan Languages*, pages 515–537.

Zeman, D., Nivre, J., Abrams, M., Aepli, N., Agić, Ž., Ahrenberg, L., Aleksandravičiūtė, G., Antonsen, L., Aplonova, K., Aranzabe, M. J., Arutie, G., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Basmov, V., Batchelor, C., Bauer, J., Bellato, S., Bengoetxea, K., Berzak, Y., Bhat, I. A., Bhat, R. A., Biagetti, E., Bick, E., Bielinskienė, A., Blokland, R., Bobicev, V., Boizou, L., Borges Völker, E., Börstell, C., Bosco, C., Bouma, G., Bowman, S., Boyd, A., Brokaitė, K., Burchardt, A., Candito, M., Caron, B., Caron, G., Cavalcanti, T., Cebiroğlu Eryiğit, G., Cecchini, F. M., Celano, G. G. A., Čéplö, S., Cetin, S., Chalub, F., Choi, J., Cho, Y., Chun, J., Cignarella, A. T., Cinková, S., Collomb, A., Çöltekin, Ç., Connor, M., Courtin, M., Davidson, E., de Marneffe, M.-C., de Paiva, V., de Souza, E., Diaz de Ilarraza, A., Dickerson, C., Dione, B., Dirix, P., Dobrovoljc, K., Dozat, T., Droganova, K., Dwivedi, P., Eckhoff, H., Eli, M., Elkahky, A., Ephrem, B., Erina, O., Erjavec, T., Etienne, A., Evelyn, W., Farkas, R., Fernandez Alcalde, H., Foster, J., Freitas, C., Fujita, K., Gajdošová, K., Galbraith, D., Garcia, M., Gärdenfors, M., Garza, S., Gerdes, K., Ginter, F., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., González Saavedra, B., Griciūtė, B., Grioni, M., Grūzītis, N., Guillaume, B., Guillot-Barbance, C., Habash, N., Hajič, J., Hajič jr., J., Hämäläinen, M., Hà Mỹ, L., Han, N.-R., Harris, K., Haug, D., Heinecke, J., Hennig, F., Hladká, B., Hlaváčová, J., Hociung, F., Hohle, P., Hwang, J., Ikeda, T., Ion, R., Irimia, E., Ishola, Ọ., Jelínek, T., Johannsen, A., Jørgensen, F., Juutinen, M., Kaşıkara, H., Kaasen, A., Kabaeva, N., Kahane, S., Kanayama, H., Kanerva, J., Katz, B., Kayadelen, T., Kenney, J., Kettnerová, V., Kirchner, J., Klementieva, E., Köhn, A., Kopacewicz, K., Kotsyba, N., Kovalevskaitė, J., Krek, S., Kwak, S., Laippala, V., Lambertino, L., Lam, L., Lando, T., Larasati, S. D., Lavrentiev, A., Lee, J., Lê Hồng, P., Lenci, A., Lertpradit, S., Leung, H., Li, C. Y., Li, J., Li, K., Lim, K., Liovina, M., Li, Y., Ljubešić, N., Loginova, O., Lyashevskaya, O., Lynn, T., Macketanz, V., Makazhanov, A., Mandl, M., Manning, C., Manurung, R., Mărănduc, C., Mareček, D., Marheinecke, K., Martínez Alonso, H., Martins, A., Mašek, J., Matsumoto, Y., McDonald, R., McGuinness, S., Mendonça, G., Miekka, N., Misirpashayeva, M., Missilä, A., Mititelu, C., Mitrofan, M., Miyao, Y., Montemagni, S., More, A., Moreno Romero, L., Mori, K. S., Morioka, T., Mori, S., Moro, S., Mortensen, B., Moskalevskyi, B., Muischnek, K., Munro, R., Murawaki, Y., Müürisep, K., Nainwani, P., Navarro Horñiacek, J. I., Nedoluzhko, A., Nešpore-Bērzkalne, G., Nguyễn Thị, L., Nguyễn Thị Minh, H., Nikaido, Y., Nikolaev, V., Nitisaroj, R., Nurmi, H., Ojala, S., Ojha, A. K., Olúòkun, A., Omura, M., Osenova, P., Östling, R., Øvrelid, L., Partanen, N., Pascual, E., Passarotti, M., Patejuk, A., Paulino-Passos, G., Peljak-Łapińska, A., Peng, S., Perez, C.-A., Perrier, G., Petrova, D., Petrov, S., 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., Ramasamy, L., Rama, T., Ramisch, C., Ravishankar, V., Real, L., Reddy, S., Rehm, G., Riabov, I., Rießler, M., Rimkutė, E., Rinaldi, L., Rituma, L., Rocha, L., Romanenko, M., Rosa, R., Rovati, D., Roșca, V., Rudina, O., Rueter, J., Sadde, S., Sagot, B., Saleh, S., Salomoni, A., Samardžić, T., Samson, S., Sanguinetti, M., Särg, D., Saulīte, B., Sawanakunanon, Y., Schneider, N., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shen, M., Shimada, A., Shirasu, H., Shohibussirri, M., Sichinava, D., Silveira, A., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Smith, A., Soares-Bastos, I., Spadine, C., Stella, A., Straka, M., Strnadová, J., Suhr, A., Sulubacak, U., Suzuki, S., Szántó, Z., Taji, D., Takahashi, Y., Tamburini, F., Tanaka, T., Tellier, I., Thomas, G., Torga, L., Trosterud, T., Trukhina, A., Tsarfaty, R., Tyers, F., Uematsu, S., Urešová, Z., Uria, L., Uszkoreit, H., Utka, A., Vajjala, S., van Niekerk, D., van Noord, G., Varga, V., Villemonte de la Clergerie, E., Vincze, V., Wallin, L., Walsh, A., Wang, J. X., Washington, J. N., Wendt, M., Williams, S., Wirén, M., Wittern, C., Woldemariam, T., Wong, T.-s., Wróblewska, A., Yako, M., Yamazaki, N., Yan, C., Yasuoka, K., Yavrumyan, M. M., Yu, Z., Žabokrtský, Z., Zeldes, A., Zhang, M., and Zhu, H. (2019). Universal dependencies 2.5. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Zeman, D. (2008). Reusable tagset conversion using tagset drivers. In *LREC*, volume 2008, pages 28–30.