Proceedings of the WILDRE-6 Workshop @LREC2020, pages 84–89
Marseille, 20 June 2022

Universal Dependency Treebank for Odia Language

Shantipriya Parida[1], Kalyanamalini Sahoo[2], Atul Kr. Ojha[3], Saraswati Sahoo[4], Satya Ranjan Dash[5] and Bijayalaxmi Dash[6]
[1]Silo AI, Helsinki, Finland
[2]University of Lille, France
[3]Insight Centre for Data Analytics, DSI, NUI, Galway, Ireland
[4]Institute of Mathematics and Applications, India
[5]KIIT University, Bhubaneswar, India
[6]Ravenshaw University, Cuttack, India
shantipriya.parida@silo.ai, kalyanamalini.shabadi@univ-lille.fr, atulkumar.ojha@insight-centre.org,
sahoosaraswati455@gmail.com, sdashfca@kiit.ac.in, rudrabijayalaxmi@gmail.com

Abstract

This paper presents the first publicly available treebank of Odia, a morphologically rich low resource Indian language. The treebank contains approx. 1082 tokens (100 sentences) in Odia selected from "Samantar", the largest available parallel corpora collection for Indic languages. All the selected sentences are manually annotated following the "Universal Dependency (UD)" guidelines. The morphological analysis of the Odia treebank was performed using machine learning techniques. The Odia annotated treebank will enrich the Odia language resource and will help in building language technology tools for cross-lingual learning and typological research. We also build a preliminary Odia parser using a machine learning approach. The accuracy of the parser is 86.6% Tokenization, 64.1% UPOS, 63.78% XPOS, 42.04% UAS and 21.34% LAS. Finally, the paper briefly discusses the linguistic analysis of the Odia UD treebank.

Keywords: Universal Dependency, Odia UD Treebank, UPOS tags

1. Introduction

Odia (earlier known as Oriya) is an Indian language belonging to the Indo-Aryan branch of the Indo-European language family. It is the predominant language of the Indian state of Odisha. Odia is written in Odia script, which is a Brahmic script. There are 37 million Odia speakers in India.[1] Odia is one of the many official languages of India and is designated as a Classical language.

Odia is an agglutinative language (Sahoo, 2001), and hence, a morphologically rich language. Odia's verb morphology is rich with a three-tier tense system, person, number, and honorific markers. The prototypical word order is subject-object-verb (SOV) (Parida et al., 2020a; Parida et al., 2020b). Odia nominal morphology differentiates between plural and singular numbers; case marking on nouns; first, second, and third-person pronouns. But it does not have grammatical gender marking, which reduces the complexities of learning the language. Odia language allows Noun-verb, Adjective-verb, and Verb-verb compounding but does not allow elision. It has 28 consonants, 6 vowels, 9 diphthongs, and 4 semivowel phonemes. Most vowels can be short or long, and care must be taken to remember that the length of the vowel changes the word meaning completely. Odia's vocabulary is influenced by Sanskrit and also a little influence from Arabic, Persian, and Austronesian languages as the Kalinga empire (Odisha's ancient name) was connected to different other kingdoms.[2] Odia language lacks online content and resources for natural language processing (NLP) research.

Unlike Treebanks of widely accepted languages such as English, Mandarin, Hindi, and Spanish for Natural Language Processing applications, applications based on low resource language like Odia is stagnated due to low resources. This paper is one step toward providing resources for such a low resource language. To start with we have worked on making a treebank in the Odia language. This project will surely help the Odia community and NLP researchers in providing resources for NLP applications.

2. Odia Language Grammar

Odia is an SOV language. Usually, a simple sentence begins with a subject and ends with a finite verb. The major word classes found in Odia are nouns, pronouns, verbs, adjectives, and postpositions. Certain minor categories like classifiers, complementizers, and conjunctions are also found. The objects occur between the subject and the verb, the Indirect Object precedes the Direct Object. The modifier precedes the item it modifies: the adjective precedes the substantive it qualifies, and the adverb precedes the verb. Although scrambling is allowed, usually, the word-order sticks to the V-final constructions except for poetic inversion (Sahoo, 2001).

Declension Odia has two numbers: singular and plural; and three persons: 1st person, 2nd person, and 3rd person. The subject NP agrees with the verb in person, number, and honorific. Honorificity goes along with person and number and it is marked in various word classes like nouns, pronouns, verbs, and, interestingly enough, with some of the post-positions that function as genitive, locative, and ablative markers. Generally, the person-number suffixes also go together.

There are eight cases in Odia: nominative, accusative, in-

[1] https://censusindia.gov.in/2011Census/Language_MTs.html

[2] https://www.nriol.com/indian-languages/oriya-page.asp

strumental, dative, ablative, genitive, locative, and vocative. Except for the nominative case, all the other cases are marked morphologically.
Phonologically, there is no distinction in the form of a word in masculine, feminine, or neuter gender in Odia. E.g. *pua* 'son' (masc), *jhia* 'daughter' (fem), *phaLa* 'fruit' (neuter). But there are certain cases, where one finds such differences between the masculine and the feminine form of the words phonologically. E.g. *chhaatra* 'male student', *chhaatri* 'female student'.

Pronouns Odia pronouns are shown in Figure 1.

(Sahoo, 2001)

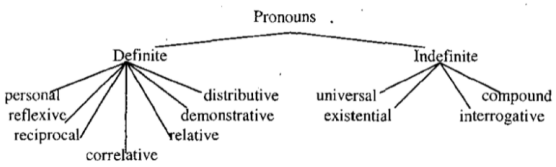

Figure 1: Odia Pronouns

- Personal: *mun* 'I', *tu* 'you', *tume* 'you', *aapaNa* 'you', *se* 'she' / 'he'
- Reflexive: *se nije* 'he himself / she herself'
- Reciprocal: *paraspara* 'each other'
- Correlative: *jie* 'who (ever)' —*se* 'he' / 'she'
- Relative: *je* 'whoever', *jaahaaku* 'whomsoever'
- Demonstrative: *eha / ehi* 'this', *eguDika / eguDaaka* 'these', sehi 'that' and seguDika / seguDaaka 'those'
- Distributive: *pratyeka* 'each' / 'every'
- Universal: *samaste* 'all'
- Existential: *jaNe* 'one person', *goTe* 'a'/ 'one'
- Interrogative: *kie* 'who', *kaahaaku* 'whom',
- Compound: *kehi jaNe / kie jaNe* 'somebody'

Case morphemes Odia case morphemes are shown in Table 1.

Case	Singular	Plural/[+Hon] sg
Nominative (NOM)	-	-e
Accusative (ACC)	*ku*	*nku, maananku*
Instrumental (INST)	*re, dwaaraa, dei*	*re, dwaaraa, dei*
Dative (DAT)	*ku*	*nku, maananku*
Ablative (ABL)	*ru, Thaaru*	*MaanankaThaaru*
Genitive (GEN)	*ra*	*nkara, maanankara*
Locative (LOC)	*re, Thaare*	*MaanankaThaare*
Vocative (VOC)	*he, bho*	

Table 1: The Case morphemes in Odia

Postpositional words The following postpositional words are used to express different case relations.

- *aagare* 'before'
- *pare* 'after'
- *kari* 'by'
- *nimitte* 'for
- *parjyante* 'up to'
- *paain* 'for'
- *prati* 'to', 'against'
- *baahaara* 'out', 'outside'
- *byatita* 'without'
- *binaa* 'without'
- *boli* 'because of', 'literally speaking', e.g. *goli boli goTe pilaa thilaa* 'there was a child called Goli'
- *bhitare* 'in', 'inside'
- *majhire* 'inside', 'in the midst of'
- *laagi* 'for' e.g. *raatidina laagi* 'for day and night'
- *sahite* 'with'

Conjunctions Conjunction markers include *o / eban / aau* 'and', *kimbaa / abaa/athabaa* 'or', *madhya* 'also', *tathaapi* 'still', *kintu* 'but' etc.

Classifiers A classifier is a noun-related element but has no independent nominal reading. Having insufficient referential or predicative content, it is not fully lexical. *-Taa* 'one$_{[+def]}$', *Topaa* 'drop', *muThaa* 'fist', *gochhaa* 'bundle', *jaNa* 'one$_{[+Human]}$', *paTa* 'slice', *asaraa* 'shower', *menchaa*, etc. are usually are identified as classifiers.

Complex verbs Complex verb constructions like the combination of a verbal with a nominal (N-V sequences), and the combination of a verbal with a verbal (V-v sequences) are found in Odia.

Serial verbs Odia is a verb serializing language. A series of verbs along with their complements and adjuncts (if any) can occur in a single clause having a common subject. Very often, the series of verbs have a common object too.

Verbal Nouns Many verbal nouns are found in Odia, such as *chaasa* 'ploughing', *chaaDa* 'release', *maajaNaa* 'bath', *rahaNi* 'stay' *bikaa* 'selling', *baahuDaa* 'return'. Some verbal nouns have been borrowed from Sanskrit, e.g. *anubhaba* 'feeling', *bidroha* 'revolution', *prabesha* 'entrance', *sthiti* 'existence', etc. which are used along with a light verb in Odia.

Copular sentences Copular constructions are usually sentences with a subject and a predicate. The predicate may be either a noun (nominal predicate) or an adjective (adjectival predicate).

Adverbs Like English, Odia also has Time, Place, and Manner adverbials.

Finite Verbal Forms Agreement features contribute to the finiteness of a verbal form in Odia. All the finite verbal forms have an agreement in concord with the subject NP. The agreement features of the verbal form are marked for the person, number, and honorific of the subject NP.

The Infinitive In Odia, the infinitival form is realized by the verbal ending – *ibaaku* 'to do'.

The Conditional affix *-ile* (or *–le*) The morpheme *-ile* (or *–le*) functions as a conditional marker. It is suffixed to the bare verbal root. It is nonfinite as it does not carry any Agr feature and thus can co-occur in a verbal form irrespective of person, number, or gender of the subject.

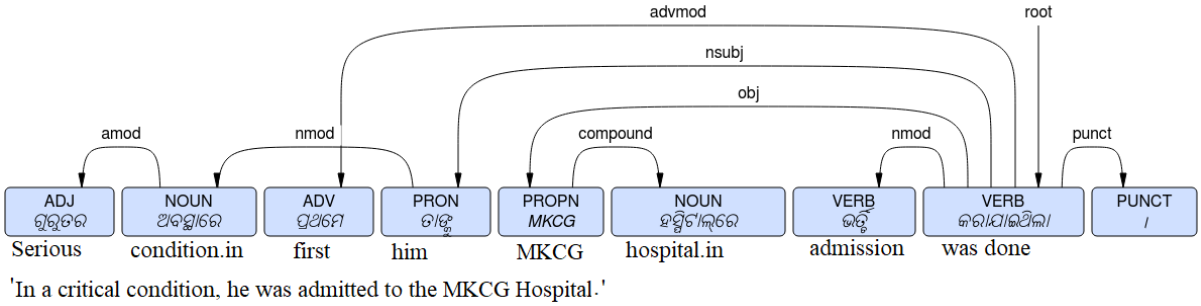

Figure 2: Passivization with Adverbial modifier construction in UD Odia

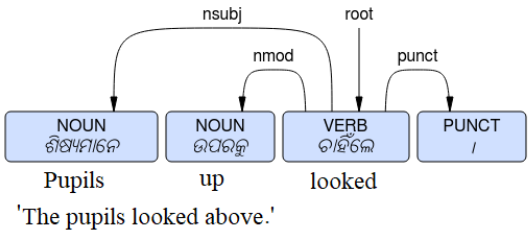

Figure 3: Finite intransitive verb construction in UD Odia

3. Related Work

Under the leadership of IIIT-Hyderabad, a consortium was formed in 2013 to start a project sponsored by TDIL (Government of India), called Development of Dependency Treebank for Indian Languages.[3] This project aimed to restore annotation work in monolingual treebanks for various languages such as Hindi, Marathi, Bengali, Kannada, and Malayalam. To achieve this model, the Pāṇinian Kāraka Dependency scheme was followed (Begum et al., 2008; Husain et al., 2010; Bhat, 2017; Ojha and Zeman, 2020). The same annotation scheme was used to annotate data in Telugu, Urdu, and Kashmiri.

NLP research of Odia has led to development of a statistical POS tagger (Ojha et al., 2015), neural network based POS tagger (Das and Patnaik, 2014), POS tagging using Support Vector Machine (SVM) (Das et al., 2015), a shallow parsing tool [4], and English-Odia machine translation system (Parida et al., 2020a).

Within the Universal Dependencies framework, as of UD release 2.8, treebanks and parsers are available for Bhojpuri, Hindi, Marathi, Sanskrit, Tamil, Telugu and Urdu (Zeman and et al., 2021). Nevertheless, there is no prior work on Odia dependency treebanking and parser.

4. Data and Methodology

To collect Odia text, we used *Samanantar*, the largest parallel corpora collection for 11 Indian languages (Ramesh et al., 2021). The parallel corpora collection includes English-Odia parallel text that covers many domains. We selected the Odia sentences of word length between 5 to 15 words per sentence. For annotation, all selected sentences are converted into CoNLL-U format consisting of 10 fields (Buchholz and Marsi, 2006). The fields are "ID", "Word", "Lemma", "UPOS", "XPOS", "FEATS", "HEAD", "DEPREL", "DEPS", and "MISC". The "UPOS" tags are based on the universal POS tags [5] following the UD guidelines, version 2. For "XPOS", we annotated according to Bureau of Indian Standards (BIS) Part of Speech (POS) tagset [6] guideline released by the department of information technology ministry of communications & information technology, the government of India. The guideline includes a POS tagset for the Odia language. The dependency relations were marked on Universal dependency tags which is an updated version of Stanford Dependencies (de Marneffe et al., 2014). Out 17 UPOS tags, we use 15 UPOS tag in this dataset, while out of 37 dependency tags, we use only 24 tags (see the Table 2 & 3). The annotation task was performed by 6 native Odia speakers including 2 linguists.

UPOS Tags	UPOS description	Statistics
NOUN	Noun	570
VERB	Verb	234
PUNCT	Punctuation	192
PROPN	Proper noun	170
ADJ	Adjective	102
ADP	Adposition	82
DET	Determiner	75
PRON	Pronoun	55
CCONJ	Coordinating conjunction	48
ADV	Adverb	45
NUM	Numeral	26
PART	Particle	23
AUX	Auxiliary	13
SCONJ	Subordinating conjunction	7
SYM	Symbol	1

Table 2: Statistics of used UPOS Tags in the Odia treebank

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

[4] http://calts.uohyd.ac.in/calts/sptil-parser.html

[5] https://universaldependencies.org/u/pos/

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

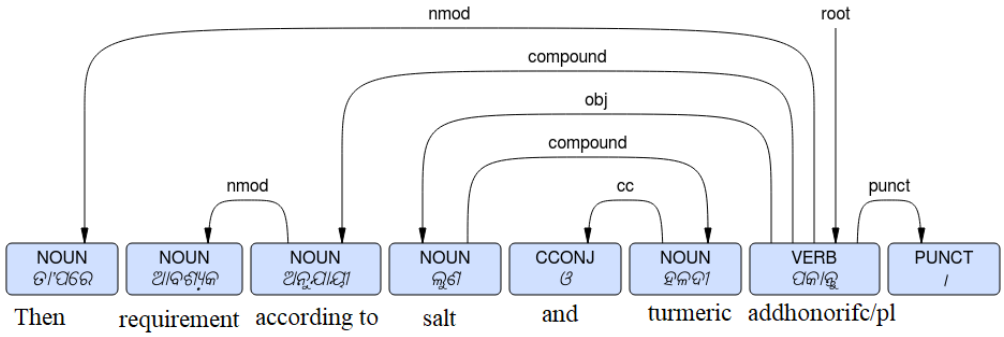

Figure 4: Finite verb in imperative sentence in UD Odia

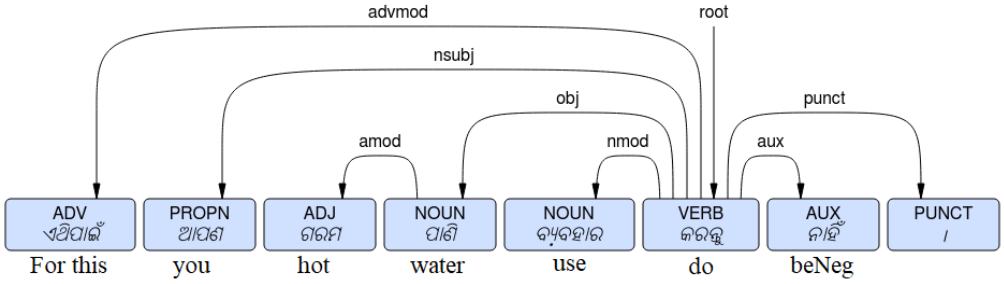

Figure 5: Main verb construction in UD Odia

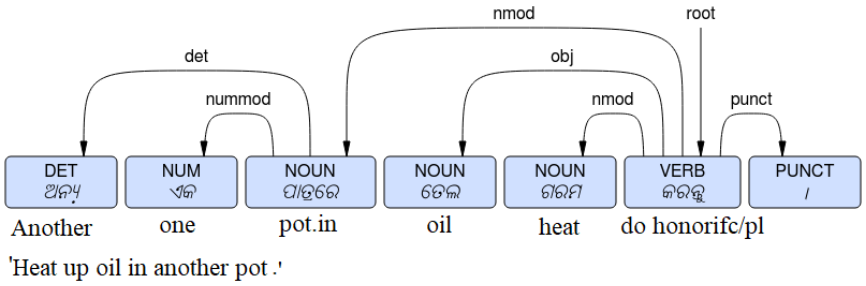

Figure 6: Finite verb with Noun modifier construction in UD Odia

5. Experiment and Results

As mentioned earlier, the Odia treebank was manually annotated using the UD annotation framework. In this, we have built Odia parser on 2026 tokens using the UDPipe open-source tool (Straka and Straková, 2017). We conducted our experiment in two parts. The first experiment was conducted on 50 sentences, while the second experiment was conducted on the rest of the dataset. We used a cross-validation 90:10 average for the data splitting where the batch size, learning rate, and dropout were 50, 0.005, and 0.10, respectively; while the other hyperparameters were randomized. The results are demonstrated in Table 4: Due to the small size of the data, the parser's accuracy is very low except on Tokenization.

6. Linguistic Analysis

We are providing few sample tree constructions along with their linguistics analysis in Figures 2 to 6

In Figure 2, *'karaajaaithilaa'* is a finite verb. So, it forms the root. The adjective *'guruttara'* modifies the noun *'abasthaare'*. The main verb has *'taanku'* as the external argument (the subject) and *'MKCG hospitalre'* as internal argument (the object) of it. It has the adverbial modifier *'prathame'*.

In Figure 3, the finite intransitive verb *'chaahinle'* 'wanted' is the root of the sentence. It takes *'shishyamaane'* 'pupils'

UD Relations	Description	Statistics
advmod	Adverbial modifier	67
advmod	Locative adverbial modifier	4
amod	Adjectival modifier of noun	109
aux	Auxiliary verb	9
case	Case marker	1
cc	Coordinating conjunction	1
ccomp	Clausal complement	2
compound	Compound	85
conj	Non-first conjunct	2
cop	Copula	1
det	Determiner	72
fixed	Non-first word of fixed expression	63
flat	non-first word of flat structure	51
goeswith	Non-first part of broken word	6
iobj	Indirect object	72
mark	Subordinating marker	52
nmod	Nominal modifier of noun	287
nsubj	Nominal subject	136
nummod	Numeric modifier	33
obj	Direct object	122
obl	Oblique nominal	1
punct	Punctuation	192
root	Root	174
xcomp	Open clausal complement	1

Table 3: UD relations used in Odia trebank

Tokenization	UPOS	XPOS	UAS	LAS
81.82%	48.25%	45.0%	36.62%	16.91%
86.6%	64.1%	63.78%	42.04%	21.34%

Table 4: Results of Odia Parser

as the subject argument. Being intransitive, it does not take any object or internal argument.

In Figure 4, *'pakaantu'* 'put' is the finite verb, which forms the root. Being an imperative sentence, the subject noun is not realized, and *'luNa o haLadi'* 'salt and turmeric' functions as the object of the sentence. *'taa pare'* 'after that' functions as the adverbial modifier. *'aabashyaka anujaayi'* 'as per the requirement' functions as a nominal modifier for *'luNa o haLadi'* 'salt and turmeric'.

In Figure 5, the main verb *'karantu'* 'do Pl/Honorific ' takes *'aapaNa'* 'you' as the subject noun and *'paaNi'* 'water' as the object. It is a negative sentence, and the negative auxiliary *'naahin'* 'be-Neg' occurs at the end of the sentence. The object *'paaNi'* 'water' is modified by the adjective *'garama'* 'hot'. *'ethipaain'* 'because of this' functions as the adverbial modifier for the sentence.

In Figure 6, *'karantu'* 'do Pl/Honorific ' which is a finite verb, forms the root of the sentence. the det *'anya'* 'another' and the numeral modifier *'eka''*one' modify the locative modifier *'paatrare'* 'in a pot'. The nsubject *'aapaNa'* 'you Pl/Honorific ' is not realized in the sentence. The main verb *'karantu'* 'do Pl/Honorific ' takes *'tela'* 'oil' as the object and *'paatrare'* 'in a pot' as the locative modifier.

7. Conclusion and Future Work

We presented the first UD Odia treebank aimed for linguistic research and applications in NLP, primarily for POS tagging, parser, semantic analyzer, and machine translation. Also, we built a preliminary Odia parser using the UDPipe tool. The accuracy of the Odia parser is 86.6% Tokenization, 64.1% UPOS, 63.78% XPOS, 42.04% UAS and 21.34% LAS.

Future research direction includes: *i)* enrich the Odia treebank with more annotated data for training, development, and validation, *ii)* including lemma for the Odia tokens, *iii)* perform detail morphological analysis, and *iv)* experiment with neural network based models for performance evaluation.

8. Acknowledgements

Atul Kr. Ojha would like to acknowledge the EU's Horizon 2020 Research and Innovation programme through the ELEXIS project under grant agreement No. 731015.

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*.

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.

Buchholz, S. and Marsi, E. (2006). Conll-x shared task on multilingual dependency parsing. In *Proceedings of the tenth conference on computational natural language learning (CoNLL-X)*, pages 149–164.

Das, B. R. and Patnaik, S. (2014). A novel approach for odia part of speech tagging using artificial neural network. In *Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2013*, pages 147–154. Springer.

Das, B. R., Sahoo, S., Panda, C. S., and Patnaik, S. (2015). Part of speech tagging in odia using support vector machine. *Procedia Computer Science*, 48:507–512.

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 *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

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.

Ojha, A. K. and Zeman, D. (2020). Universal Dependency Treebanks for Low-Resource Indian Languages: The Case of Bhojpuri. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 33–38.

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.

Parida, S., Bojar, O., and Dash, S. R. (2020a). Odiencorp: Odia–english and odia-only corpus for machine translation. In *Smart Intelligent Computing and Applications*, pages 495–504. Springer.

Parida, S., Dash, S. R., Bojar, O., Motlicek, P., Pattnaik, P., and Mallick, D. K. (2020b). OdiEnCorp 2.0: Odia-English parallel corpus for machine translation. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 14–19, Marseille, France, May. European Language Resources Association (ELRA).

Ramesh, G., Doddapaneni, S., Bheemaraj, A., Jobanputra, M., AK, R., Sharma, A., Sahoo, S., Diddee, H., Kakwani, D., Kumar, N., et al. (2021). Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *arXiv preprint arXiv:2104.05596*.

Sahoo, K. (2001). *Oriya verb morphology and complex verb constructions*. NTNU Trondheim.

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.

Zeman, D. and et al. (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.