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

Description

Resource Building
Syntactically annotated treebank
UD Framework
4881 annotated tokens
ML-based Tagger and Parser
Data Source: BLTR
Domain: news and non-fiction
5000 sentences (105,174 tokens)
254 sentences(4881 tokens) manually annotated
XPOS and UPOS tags
Support: Hindi Treebank | BIS Tagset

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

Statistics of morphological features

Bhojpuri

Indo_Aryan Language
Bihar | Jharkhand | Uttar Pradesh
Nepal | trinindad | Mauritius | Guyana | Suriname | Fiji
Speakers: 50,579,447
Resource Poor Language for ML

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

Statistics of UPOS tags

Accuracy

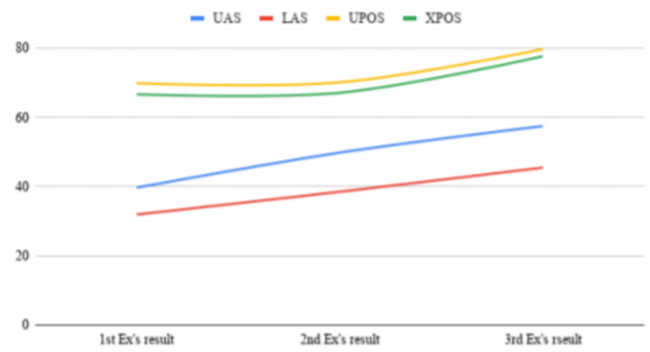
57.49% UAS | 45.50% LAS
79.69% UPOS | 77.64% XPOS

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

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

	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%

UDPipe accuracy of the conducted experiments



Learning curve of the Bhojpuri models

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

UD relations. Out of 37 we use 30

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.