Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5705–5711
Marseille, 20-25 June 2022

Universal Dependencies for Punjabi

Aryaman Arora
Georgetown University
Washington, D.C., USA
aa2190@georgetown.edu

Abstract

We introduce the first Universal Dependencies treebank for Punjabi (written in the Gurmukhi script) and discuss corpus design and linguistic phenomena encountered in annotation. The treebank covers a variety of genres and has been annotated for POS tags, dependency relations, and graph-based Enhanced Dependencies. We aim to expand the diversity of coverage of Indo-Aryan languages in UD.

Keywords: Punjabi, Universal Dependencies, syntax, POS tagging, treebanking

1. Introduction

Universal Dependencies (UD) is a community project that maintains a standard scheme for the annotation of grammar (including part-of-speech tags, syntactic relations, and morphological features) in a cross-lingually consistent manner, and releases treebanks for many languages annotated with the UD schema (Nivre et al., 2016, 2020). UD is currently on its version 2.9 release, comprised of 217 treebanks covering 122 languages.

However, UD coverage is biased towards high-resource languages in NLP, especially in South Asia where only Hindi and Urdu have treebanks with greater than 100k tokens. South Asian languages with tens of millions of speakers (e.g. Punjabi, Kannada, Gujarati, Pashto) often do not have treebanks yet. UD treebanked data is useful for upstream tasks in NLP, including semantic parsing (Reddy et al., 2017) and natural language understanding (Schuster and Manning, 2016), and syntactic corpora are useful for linguists studying language typology and change (Levshina, 2019).

In this paper, we present a new UD treebank for Punjabi, a language of South Asia with over 100 million native speakers but otherwise underesearched in computational linguistics and NLP. Punjabi has official status at the regional level in both India and Pakistan. We describe the composition of the corpus, thorny linguistic issues that are relevant for UD annotation of other South Asian languages, and future plans for Punjabi NLP. We especially explore the annotation of complex syntactic phenomena which are relevant to many other languages.

2. Related treebanks

The Indo-Aryan (IA) languages are unevenly represented in UD, with large automatically-converted (and thus of imperfect quality) treebanks for the two major IA languages Hindi (Tandon et al., 2016) and Urdu (Ehsan and Butt, 2020), and a smattering of smaller manual treebanks for lesser-studied languages such as Marathi (Ravishankar, 2017), Bhojpuri (Ojha and Zeman, 2020), Sanskrit (Hellwig et al., 2020; Dwivedi and Zeman, 2017), Ashokan Prakrit (Farris and Arora, 2021), and code-mixed Hindi–English (Bhat et al., 2018). Besides UD, there is a substantial body of work on syntactic annotation using Paninian karaka formalisms for many South Asian languages (Bhatt et al., 2009; Bhat and Sharma, 2012; Nallani et al., 2020).

Genre	Doc.	Sent.	Tok.
`misc`	—	71	1,664
`news`	4	89	1,614
`editorial`	2	80	1,570
`blog`	1	33	806
`nonfiction`	1	26	566
Total	7	299	6,220

Table 1: Data in the Punjabi UD corpus by genre. Columns are ‘documents’, ‘sentences’, and ‘tokens’.

3. Corpus

There is a dearth of broad-coverage reference corpora for South Asian languages. There are plenty of multilingual corpora that include Punjabi, e.g. IndicCorp (Kakwani et al., 2020), EMILLE (McEnery et al., 2000; Baker et al., 2002), and PMIndia (Haddow and Kirefu, 2020), but these are heavily biased towards news and other formal-register texts and, in the case of IndicCorp, scraped without regard to document structure.

Noting that other South Asian treebanking projects, such as the Hindi Dependency Treebank (Bhatt et al., 2009) are also biased towards news, we collected texts manually from diverse sources in an effort to more broadly represent modern usage of Punjabi. The variety represented is standard Majhi Punjabi in the Gurmukhi script.

Table 1 shows the breakdown of the corpus into various genres. The `misc` portion included randomly sampled sentences from IndicCorp and the FLORES-101 low-resource machine-translation dataset (Goyal et al., 2021). These were selected as a pilot when making basic annotation decisions.

Later, the corpus will also incorporate non-fiction, poetry/song, and social media texts. Eventually, a Shahmukhi-script corpus for Punjabi (as used in Pakistan) will also be needed to begin studying dialectal variation.

POS	Count	%
NOUN	1362	21.9%
ADP	1064	17.1%
VERB	713	11.5%
PROPN	554	8.9%
PUNCT	522	8.4%
AUX	441	7.1%
ADJ	422	6.8%
PRON	332	5.3%
DET	187	3.0%
CCONJ	162	2.6%
ADV	145	2.3%
NUM	116	1.9%
PART	96	1.5%
SCONJ	93	1.5%
X	4	0.1%
INTJ	4	0.1%
SYM	3	0.0%

Table 2: Distribution of POS tags in the treebank.

3.1. Annotation process

So far, the treebank has been annotated by a single heritage speaker of Punjabi, the first author of this paper, over the month of December 2021. The tool UD Annotatrix (Tyers et al., 2017) was used to manually tokenise, quickly connect and label dependency relations, and validate output CONLLU formatting. Each document is named by its genre, source, and a unique one-word identifier, e.g. `news_bbc_rajnikanth_{n}` is the *n*-th sentence in a news article from the BBC about South Indian actor Rajnikanth's entry into politics.

We relied on reference dictionaries (RCPLT, 2021; Singh, 1895) and grammars (Bhatia, 1993; Gill and Gleason, 2013) to design the annotation guidelines, and also referred to other treebanks (particularly for Hindi). The Universal Dependencies community also helped discuss some linguistic issues in annotation.

4. Part-of-speech tags

UD for Punjabi uses the entire Universal Part-of-Speech (POS) tagset. There do exist Punjabi-specific POS tagsets and taggers (Gill et al., 2009; Sharma and Lehal, 2011), but since we wanted to make initial progress on dependency annotation, we began with using the UD tagset directly. The distribution of POS tags in the annotated corpus is listed in table 2.

Demonstratives. Punjabi has demonstrative pronouns that can stand alone: ਇਹ *é* 'this' and ਉਹ *ó* 'that'. These, along with some other pronouns, can also behave as determiner modifiers in a noun phrase. We elected to tag these as PRON if standing as an independent NP or DET if modifying a noun.

Infinitives. Infinitives in Punjabi, like English gerunds, are morphologically verbs but can syntactically behave as nouns. Given the need for morphological features to be indicated, and precedent in other Indo-Aryan treebanks, we label them as VERBs. This can lead to some messy situations, such as when an infinitive behaving as a noun takes case-marking. We POS-tag the case marker as ADP but use the deprel `mark` in such cases.

5. Syntactic relations

We discuss some of the linguistically interesting issues in the annotation of syntactic relations. Our main guidance was from precedent in the other Indo-Aryan UD treebanks as well as the Punjabi grammar of Bhatia (1993). Table 3 shows the distribution of deprel tags in the treebank.

Deprel	#	%
case	928	14.9%
punct	512	8.2%
obl	451	7.3%
nsubj	410	6.6%
nmod	379	6.1%
aux	310	5.0%
root	299	4.8%
amod	270	4.3%
obj	250	4.0%
conj	247	4.0%
det	184	3.0%
mark	177	2.8%
cc	174	2.8%
compound:lvc	165	2.7%
flat	150	2.4%
advcl	136	2.2%
compound	131	2.1%
advmod	126	2.0%
compound:svc	97	1.6%
discourse	96	1.5%
cop	96	1.5%
nummod	91	1.5%
acl	78	1.3%
xcomp	63	1.0%
fixed	60	1.0%
appos	58	0.9%
ccomp	57	0.9%
parataxis	49	0.8%
acl:relcl	48	0.8%
aux:pass	40	0.6%
nsubj:pass	31	0.5%
compound:redup	27	0.4%
iobj	15	0.2%
obl:agent	7	0.1%
orphan	3	0.0%
reparandum	3	0.0%
csubj	1	0.0%
dislocated	1	0.0%

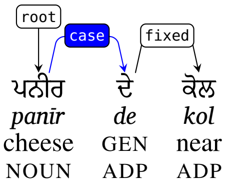
Table 3: Distribution of dependency relations in the treebank.

5.1. Case

Punjabi has case markers which are written joined on pronouns, and in that case often morphologically opaque, e.g. the independent genitive case marker is ਦਾ *dā* but the genitive 1SG pronoun is ਮੇਰਾ *merā*. The Marathi UD corpus, facing a similar phenomenon, elected to treat these as multi-word tokens with the case separated into a separate word. For Punjabi, we decided not to split these morphologically opaque pronouns, since it is trivial to automate such a split, and it has not been studied whether splitting these case clitics is useful for upstream tasks like syntactic parsing.

Postpositions in Punjabi are often multiword, formed with the genitive (or sometimes the ablative) case attached. We decided to mark the first token of a multi-word postposition with `case`, with subsequent tokens attached to that one with `fixed`; this is shown in (1).

(1) *'near the cheese'*

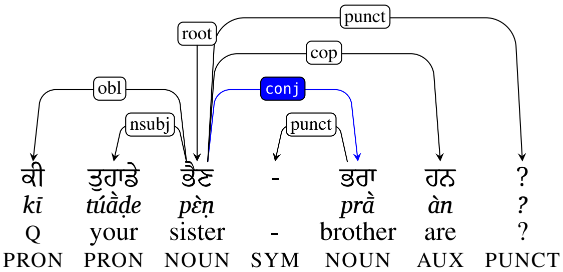


The Hindi Dependency Treebank (HDTB) and Urdu Dependency Treebank (UDTB) treat multiword postpositions as two separate `case` dependents to the noun, while Hindi Parallel Universal Dependencies (PUD) agrees with our annotation. Our analysis makes it clear that the expression is a single unit and makes it easier to query case marker counts, but having two separate case dependents may be more linguistically sensible at the cost of losing ordering information.

5.2. Compounds

Punjabi has a very productive system of compound-creation. True compounds, where one noun (generally the preceding one) is dependent to another, take the deprel `compound`. Another class of compounds, called *dvandva* in traditional grammatical descriptions, is headless. We treat them as coordination without an explicit conjunction, with the first as the head of a `conj` relation.

(2) *Do you have any siblings?*

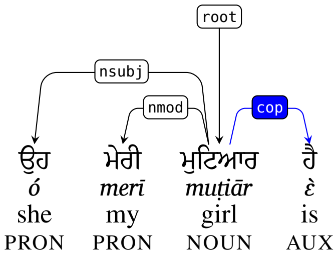


Finally, reduplicative compounds (e.g. ਵਾਰ-ਵਾਰ *vār-vār* 'again and again') use the subtype `compound:redup` with the first element as the head.

5.3. Copulae

The only copula in Punjabi is ਹੋਣਾ *hoṇā* 'to be', syntactically annotated as in (3). While other verbs, such as ਰਹਿਣਾ *rèṇā* 'to remain, continue to be' and ਬਣਨਾ *baṇnā* 'to become', do take predicative complements in a syntactically similar manner, for those we used the deprel `xcomp` instead to conform with other treebanks.

(3) *'She is my girl.'*

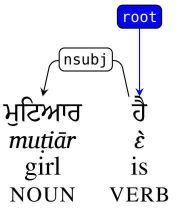


However, the copula does not necessarily have to take two arguments. When it takes only one argument, it serves an existential function; unlike in English *there is*, Punjabi does not use a dummy pronoun. The UD guidelines have been subject to much debate on how to handle the various functions of copulae across languages, which have been delineated along these categories:

- Equation (aka identification): "she is my mother"
- Attribution: "she is nice"
- Location: "she is in the bathroom"
- Possession: "the book is hers"
- Benefaction: "the book is for her"
- Existence: "there is food (in the kitchen)"

Ideally, if the same construction is used for all 6 of these categories, then UD should annotate them all the same way. But in e.g. Czech, which has a situation parallel to Punjabi, the existential copula is treated as a normal verb with an `nsubj` argument. To conform with UD standards we did this in Punjabi as well, but the lack of uniformity across categories, especially when using the same construction, is problematic. The current annotation is shown in (4).

(4) *'There is a girl.'*



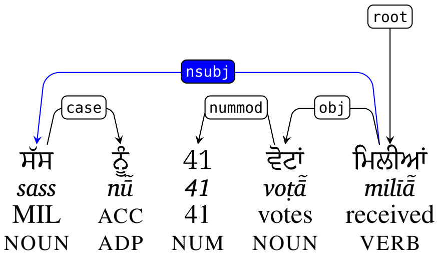
5.4. Core verbal arguments

The relations `nsubj`, `obj`, and `iobj` indicate core arguments to verbs. We annotate `iobj` only when the direct object is also present, per UD guidelines.

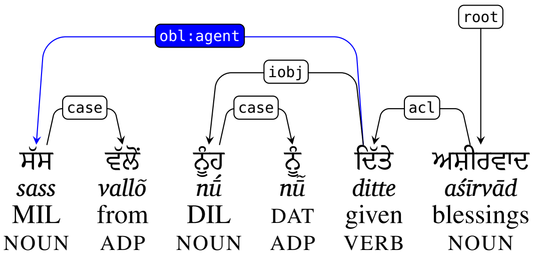
Non-nominative subjects. It has been well-established in the linguistic literature that South Asian languages can have subjects that are not in ergative/nominative case (Bhaskararao and Subbarao, 2004). For example, Punjabi has many verbs with experiencer semantics that take a subject in the dative case. However, other IA treebanks do not consistently label non-nominative subjects with the relation `nsubj`, even though such subjects pass cross-lingual criteria for subjecthood.

In this treebank, we used `nsubj` for dative and locative subjects as in (5), and `obl:agent` for semantic subjects demoted to oblique position through passivisation, which take the case markers ਕੋਲੋਂ *kolõ*, ਵੱਲੋਂ *vallõ*, etc. as in (6).

(5) *'The mother-in-law got 41 votes.'*



(6) *'The blessings given by the mother-in-law to the daughter-in-law.'*

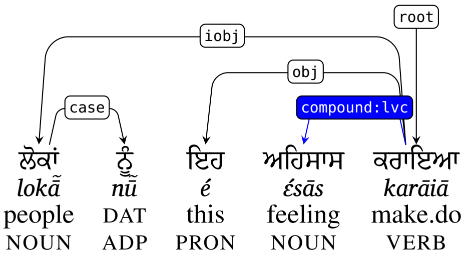

root
obl:agent
iobj
case
case
acl
ਸੱਸ ਵੱਲੋਂ ਨੂੰਹ ਨੂੰ ਦਿੱਤੇ ਅਸ਼ੀਰਵਾਦ
sass vallõ nū́ nū̃ ditte aśīrvād
MIL from DIL DAT given blessings
NOUN ADP NOUN ADP VERB NOUN

5.5. Complex predicates

Noun/adjective–verb concatenation. A thorny construction for South Asian syntacticians is the noun–verb and adjective–verb concatenation (also called 'conjunct verb', 'noun/adjective–verb complex', etc.). These are nominals that are verbalised with the concatenation of semantically vacuous light verb, e.g. ਕਰ 'to do'. 28% of the verbs in this Punjabi treebank are involved in this construction.[1]

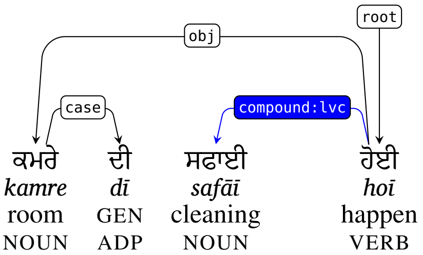
Although some formalisms have tried a noun-headed analysis of such constructions, e.g. TAG (Vaidya et al., 2014), we follow UD precedent in Hindi–Urdu and Chinese (Poiret et al., 2021) and adopt a verb-headed analysis as in (7).

(7) *'People were made to realise this.'*



We also treat genitive objects as verbal arguments rather than dependents of the nominal, shown in (8). This is supported by work on Hindi (Mohanan, 1994), but unfortunately differences in Punjabi are unstudied.

(8) *'The room was cleaned.'*

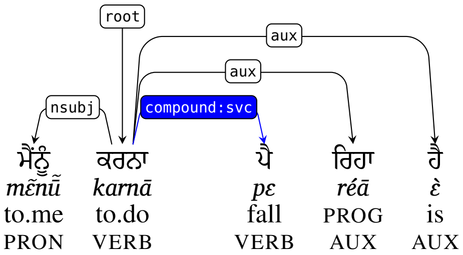


Also note the new subtyped relation `compound:lvc`, which has been used in other typologically-similar languages that have such light verb constructions.

An open question is to what extent the nominal component of this construction can take dependents (as the verb–object construction in Chinese can). Traditional syntactic research on IA languages is divided (Mohanan, 1994; Bhatia, 1993), so a Punjabi treebank will be useful here.

Verb–verb concatenation. Punjabi also has a considerable inventory of aspectual light verbs that modify main verbs, with two verbs describing a single event, a phenomenon called verb–verb concatenation (also 'compound verb', 'serial verb construction', 'verb–verb complex', etc.), shown in (9).

(9) *'I'm being forced to do it.'*



For these we use `compound:svc` rather than the `aux` relation favoured by the Hindi and Urdu treebanks. These are morphologically verbs with complete paradigms, which makes them unsuitable to be treated as auxiliaries. Following UD aims, the main content verb, i.e. the first one, is the head.

5.6. Relative clauses

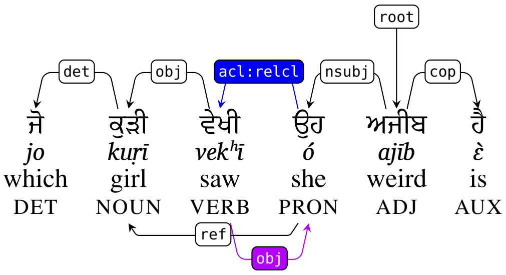
Punjabi, like other Indo-Aryan languages, has relative clause movement resulting in complex non-projective trees as shown in figure 1. Relative clauses can move to clause-final position or pre-clausally in a correlative construction (Bhatt, 2003).

We also use Enhanced Dependencies in these constructions to connect the head of the relative clause to its referent relative pronoun, using the deprel `ref`. Note that a pronominal head may refer to a determiner relative inside the relative clause (see the second figure in figure 1); we make that the child of the `ref` edge but it is actually the entire NP that is being referred to. Finally, we add an edge from the relative clause to its head indicating the relation that the head would have head in the non-relativised equivalent (for example, a relativised subject gets the edge `nsubj`).

Enhanced Dependencies have not been used in treebanks for related languages before, so we attempted to replicate the English guidelines.

Headless relative clauses. The head of a relative clause is not obligatory, in which case we label the clause-internal relative pronoun as the head and have an Enhanced edge with its intended syntactic relation in the clause. However, when the relative pronoun behaves as a determiner this approach cannot produce a well-formed tree. (10) shows our analyses for two contrasting trees, the first with a pronoun governing the relative clause, and the second without a head.

(10) *'The girl who I saw is weird.'*



[1] There are 561 verbs in the treebank, of which 74 are non-main verbs dependent in verb-verb concatenations. Thus, out of 487 main verbs, 135 have `compound:lvc` dependents.

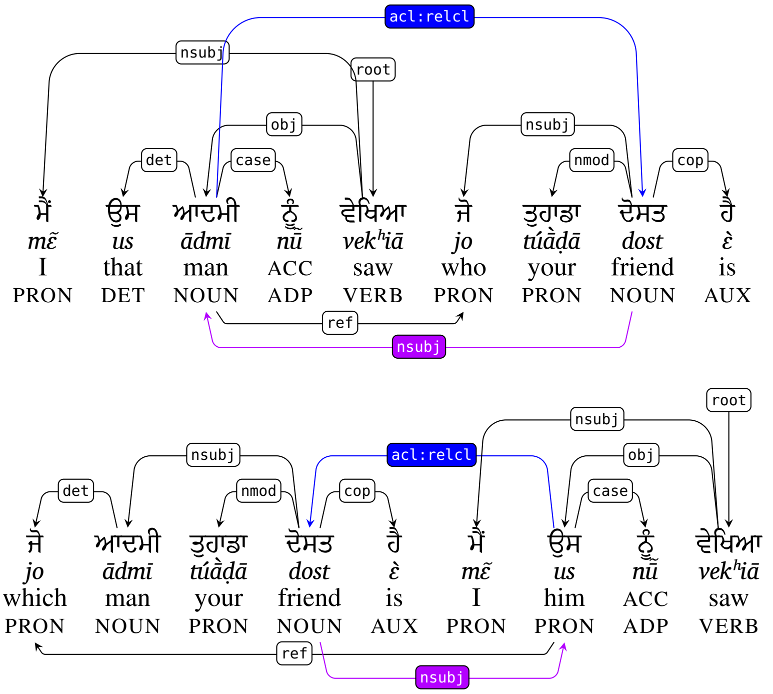
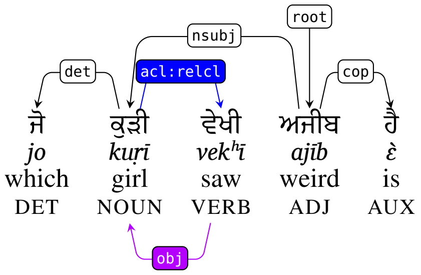


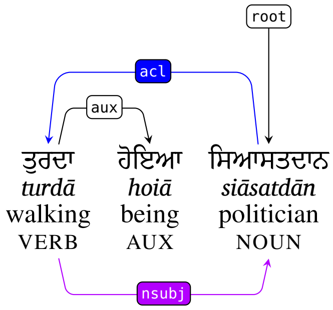
Figure 1: *'I saw the man who is your friend.'* with both clause-final and pre-clausal relative clause movement.



In the somewhat formal texts that comprise our treebank, no headless relative clauses were encountered. However, informal texts (e.g. spontaneous speech) that we annotate in the future will probably feature this construction.

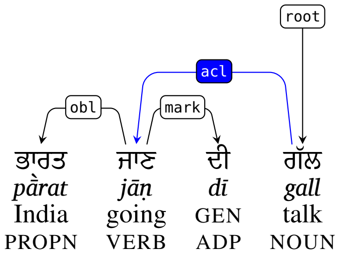
Non-finite relative clauses. Punjabi also has non-finite relative clauses that obligatorily precede their head, shown in (11). These also do not have a relative pronoun internally, so we use the bare acl relation (as in the Hindi and Urdu treebanks).

(11) *'a walking politician'*



Nouns with infinitival verbal dependents also make use of the bare acl relation, as in (12). In this case, the head noun is not an argument to the dependent clause, so there is no edge from the verb to the noun in the Enhanced layer.

(12) *'talk of going to India'*



6. Conclusion

We introduced a new Universal Dependencies-annotated treebank for Punjabi in the Gurmukhi script. We discussed some linguistic issues in annotation, with respect to both POS tagging and a wide variety of syntactic relations. Future work will be adding morphological feature annotations, conducting an interannotator study by double-annotating some portions of the corpus to increase quality, and training a parser for Punjabi. In the long-term, much work remains to be done to increase the coverage of Indo-Aryan languages in UD (e.g. Punjabi in the Shahmukhi script, Saraiki, Hindko)—we hope that this work is a step towards that.

Acknowledgements

Thanks to Amir Zeldes, whose Computational Corpus Linguistics class at Georgetown led to the creation of this Punjabi UD treebank, and to Nathan Schneider and the three anonymous reviewers for their helpful feedback.

7. Bibliographical References

References

Baker, Paul, Hardie, Andrew, McEnery, Tony, Cunningham, Hamish, and Gaizauskas, Rob (2002). EMILLE, a 67-million word corpus of indic languages: Data collection, mark-up and harmonisation. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. European Language Resources Association (ELRA), Las Palmas, Canary Islands - Spain.

Bhaskararao, P. and Subbarao, K. V., editors (2004). *Nonnominative Subjects*, volume 1. John Benjamins Publishing Company, Netherlands.

Bhat, Irshad, Bhat, Riyaz A., Shrivastava, Manish, and Sharma, Dipti (2018). Universal Dependency parsing for Hindi-English code-switching. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 987–998. Association for Computational Linguistics, New Orleans, Louisiana. doi:10.18653/v1/N18-1090.

Bhat, Riyaz Ahmad and Sharma, Dipti Misra (2012). Dependency treebank of Urdu and its evaluation. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 157–165. Association for Computational Linguistics, Jeju, Republic of Korea.

Bhatia, Tej K. (1993). *Punjabi: A cognitive-descriptive grammar*. Routledge, London and New York.

Bhatt, Rajesh (2003). Relativization strategies in Indo-Aryan.

Bhatt, Rajesh, Narasimhan, Bhuvana, Palmer, Martha, Rambow, Owen, Sharma, Dipti, and Xia, Fei (2009). A multi-representational and multi-layered treebank for Hindi/Urdu. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 186–189. Association for Computational Linguistics, Suntec, Singapore.

Dwivedi, Puneet and Zeman, Daniel (2017). Universal Dependencies for Sanskrit: A pilot study. Preprint.

Ehsan, Toqeer and Butt, Miriam (2020). Dependency parsing for Urdu: Resources, conversions and learning. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5202–5207. European Language Resources Association, Marseille, France.

Farris, Adam and Arora, Aryaman (2021). For the purpose of curry: A UD treebank for Ashokan Prakrit. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 49–61. Association for Computational Linguistics, Sofia, Bulgaria.

Gill, Harjeet Singh and Gleason, Henry A. (2013). *A Reference Grammar of Punjabi*. Punjabi University, Patiala. Revised edition by Mukhtiar Singh Gill.

Gill, Mandeep Singh, Lehal, Gurpreet Singh, and Joshi, Shiv Sharma (2009). Part-of-speech tagging for grammar checking of Punjabi. *Linguistics Journal*, 4(1).

Goyal, Naman, Gao, Cynthia, Chaudhary, Vishrav, Chen, Peng-Jen, Wenzek, Guillaume, Ju, Da, Krishnan, Sanjana, Ranzato, Marc'Aurelio, Guzman, Francisco, and Fan, Angela (2021). The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation.

Haddow, Barry and Kirefu, Faheem (2020). PMIndia – a collection of parallel corpora of languages of India.

Hellwig, Oliver, Scarlata, Salvatore, Ackermann, Elia, and Widmer, Paul (2020). The treebank of vedic Sanskrit. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5137–5146. European Language Resources Association, Marseille, France.

Kakwani, Divyanshu, Kunchukuttan, Anoop, Golla, Satish, N.C., Gokul, Bhattacharyya, Avik, Khapra, Mitesh M., and Kumar, Pratyush (2020). IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961. Association for Computational Linguistics, Online. doi:10.18653/v1/2020.findings-emnlp.445.

Levshina, Natalia (2019). Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, 23(3):533–572. doi:doi:10.1515/lingty-2019-0025.

McEnery, Anthony, Baker, Paul, Gaizauskas, Rob, and Cunningham, Hamish (2000). EMILLE: building a corpus of South Asian languages. In *Proceedings of the International Conference on Machine Translation and Multilingual Applications in the new Millennium: MT 2000*. University of Exeter, UK.

Mohanan, Tara (1994). *Argument structure in Hindi*. Center for the Study of Language (CSLI).

Nallani, Sneha, Shrivastava, Manish, and Sharma, Dipti (2020). A fully expanded dependency treebank for Telugu. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 39–44. European Language Resources Association (ELRA), Marseille, France.

Nivre, Joakim, de Marneffe, Marie-Catherine, Ginter, Filip, Goldberg, Yoav, Hajič, Jan, Manning, Christopher D., McDonald, Ryan, Petrov, Slav, Pyysalo, Sampo, Silveira, Natalia, Tsarfaty, Reut, and Zeman, Daniel (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666. European Language Resources Association (ELRA), Portorož, Slovenia.

Nivre, Joakim, de Marneffe, Marie-Catherine, Ginter, Filip, Hajič, Jan, Manning, Christopher D., Pyysalo, Sampo, Schuster, Sebastian, Tyers, Francis, and Zeman, Daniel (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043. European Language Resources Association, Marseille, France.

Ojha, Atul Kr. and Zeman, Daniel (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. European Language Resources Association (ELRA), Marseille, France.

Poiret, Rafaël, Wong, Tak-Sum, Lee, John, Gerdes, Kim, and Leung, Herman (2021). Universal Dependencies for Mandarin Chinese. *Language Resources and Evaluation*, pages 1–38.

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

RCPLT (2021). *Punjabi–English Dictionary*. Punjabi University, Patiala.

Reddy, Siva, Täckström, Oscar, Petrov, Slav, Steedman, Mark, and Lapata, Mirella (2017). Universal semantic parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 89–101. Association for Computational Linguistics, Copenhagen, Denmark. doi:10.18653/v1/D17-1009.

Schuster, Sebastian and Manning, Christopher D. (2016). Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2371–2378. European Language Resources Association (ELRA), Portorož, Slovenia.

Sharma, Sanjeev Kumar and Lehal, Gurpreet Singh (2011). Using Hidden Markov Model to improve the accuracy of Punjabi POS tagger. In *2011 IEEE International Conference on Computer Science and Automation Engineering*, volume 2, pages 697–701. doi:10.1109/CSAE.2011.5952600.

Singh, Maya (1895). *The Panjabi Dictionary*. Munshi Gulab Singh & Sons, Lahore.

Tandon, Juhi, Chaudhry, Himani, Bhat, Riyaz Ahmad, and Sharma, Dipti (2016). Conversion from paninian karakas to Universal Dependencies for Hindi dependency treebank. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 141–150. Association for Computational Linguistics, Berlin, Germany. doi:10.18653/v1/W16-1716.

Tyers, Francis M., Sheyanova, Mariya, and Washington, Jonathan North (2017). UD annotatrix: An annotation tool for Universal Dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 10–17. Prague, Czech Republic.

Vaidya, Ashwini, Rambow, Owen, and Palmer, Martha (2014). Light verb constructions with 'do' and 'be' in Hindi: A TAG analysis. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 127–136. Association for Computational Linguistics and Dublin City University, Dublin, Ireland. doi:10.3115/v1/W14-5816.