Proceedings of the WILDRE-6 Workshop @LREC2020, pages 60–67
Marseille, 20 June 2022

Bengali and Magahi PUD Treebank and Parser

Pritha Majumdar[1], Deepak Alok[1], Akanksha Bansal[1],
Atul Kr. Ojha[2,1], John P. McCrae[2]
[1]Panlingua Language Processing LLP, India, [2]DSI, National University of Ireland Galway, Ireland
panlingua@outlook.com, {atulkumar.ojha, john.mccrae}@insight-centre.org

Abstract

This paper presents the development of the Parallel Universal Dependency (PUD) Treebank for two Indo-Aryan languages: Bengali and Magahi. A treebank of 1,000 sentences has been created using a parallel corpus of English and the UD framework. A preliminary set of sentences was annotated manually - 600 for Bengali and 200 for Magahi. The rest of the sentences were built using the Bengali and Magahi parser. The sentences have been translated and annotated manually by the authors, some of whom are also native speakers of the languages. The objective behind this work is to build a syntactically-annotated linguistic repository for the aforementioned languages, that can prove to be a useful resource for building further NLP tools. Additionally, Bengali and Magahi parsers were also created which is built on machine learning approach. The accuracy of the Bengali parser is 78.13% in the case of UPOS; 76.99% in the case of XPOS, 56.12% in the case of UAS; and 47.19% in the case of LAS. The accuracy of Magahi parser is 71.53% in the case of UPOS; 66.44% in the case of XPOS, 58.05% in the case of UAS; and 33.07% in the case of LAS. This paper also includes an illustration of the annotation schema followed, the findings of the Parallel Universal Dependency (PUD) treebank, and it's resulting linguistic analysis.

Keywords: Indian languages, Bengali, Magahi, Parallel Universal Dependency Treebank, parser

1. Introduction

Sentence parsing is one of the trickiest, yet essential components in the field of Natural Language Processing (NLP). Parsing not only enables better understanding of a sentence structure, but is also useful in the development of various NLP applications like machine translation, and information retrieval. In this paper, we aim to discuss Parallel Universal Dependency (PUD) treebank and parser for two Indian languages, namely, Bengali and Magahi. Bengali, also referred as Bangla, is mostly spoken in the Indian regions of West Bengal, Assam, and Tripura and is the mother tongue of about 97.2 million speakers as per the 2011 Census Report of India.[1] It is one of the 22 scheduled Indian languages and the national language of Bangladesh. Magahi is an Eastern Indo-Aryan language spoken mostly in the Indian states of Bihar and certain areas of Jharkhand. Linguistically, both the languages belong to the Indo-Aryan language family, follow a Subject-Object-Verb (SOV) construction and are head-final languages with a relatively free word order. Like several other Indian languages, they also follow the post-position trait. They are also nominative-accusative languages that allow pro-drop of all arguments, contain complex verb constructions, rich classifiers, differential object marking and has no assigned gender. The verbs show only person agreement with the subject and no agreement with number and gender. However, a distinctive feature of Bengali is that the copula or verb linking the subject and predicate is often found missing in this language.

The objective of UD is to automatically analyze dependency structure of sentences, create multilingual parsers for cross lingual learning, and conduct parsing research for typologically diverse languages under a common framework.[2] The annotation scheme is based on Stanford dependencies (De Marneffe and Manning, 2008), Google universal part of speech, (Petrov et al., 2011) and the Interset Interlingua for morphosyntactic tag sets (Zeman, 2008). Currently, the UD project contains more than 217 treebanks for 122 languages belonging to 24 language families.[3]

It would however be ignorant to state that Indian languages have not progressed at all in the field of NLP (See section-2). The extent and the contribution of their work will be discussed in the later sections of the paper. Nevertheless, discussion on development of PUD treebank and a parser for the above stated languages with English as the source language is what is aimed in this paper.

Section (2), provides an overview of the linguistic resources that have been created for Indian languages. Section (3), discusses the experiment and the data size used to build the parser for the respective languages. Section (4), demonstrates the

[1] https://censusindia.gov.in/2011Census/Language-2011/Statement-1.pdf

[2] https://universaldependencies.org/introduction.html

[3] https://universaldependencies.org/

development of the parser and its outcome. Section (5), illustrates the cross-lingual study of the UD treebank built in this project. The paper concludes with closing remarks and plans for future work.

2. Related Work

In 1991, an initiative was taken up by Technology Development for Indian languages (TDIL), to develop tools for POS tagging, frequency count, spell checkers, and morphological processing for all Indian national (official/scheduled) languages. Thus, a corpus of 3 million words was created for every Indian national language by the end of 1994, including Bengali (Dash, 2004). As per (Kumar et al., 2018), a corpus of 0.17 million sentences is available in Magahi. In 2013, Indian Institute of Information Technology, Hyderabad (IIIT-H) developed monolingual and parallel Pāṇinian Kāraka Dependency (PD) treebanks for Indian languages Hindi, Bengali, Marathi, Kannada and Malayalam.[4] The same was then utilized to annotate Telugu, Urdu, Kashmiri (Bhat, 2017) and a dependency parser for Hindi, Telugu, Bengali, Urdu and Kashmiri was created. Presently, IIIT-H[5] is developing PD treebanks for Indian languages, Bengali, Kannada, Hindi and Malayalam.

As of UD release 2.9, 217 treebanks including the Indian languages, Bhojpuri, Marathi, Hindi, Sanskrit, Tamil, Telugu and Urdu are available(Zeman, 2021).[6] In addition, Magahi UD and Bengali PUD treebanks have been recently reported by (Raj et al., 2021) and (Majumdar, 2021) at *Workshop on Parsing and its Applications for Indian Languages* (in progress) and *Widening NLP workshop 2021* respectively. However, the Bengali UD treebank that was reported at WiNLP has been modified to improve the quality of translation and annotation for the requirements of this paper. Nonetheless, there has been no prior work/resources in Magahi PUD Treebank.

3. Data & Methodology

In order to build the Bengali and Magahi PUD, the English sentences have been taken as the source text from the English PUD (Zeman et al., 2017) which was further translated into the respective target languages, preparing it for the corresponding treebank annotation. In this study both manual and automatic annotation schemas were followed. The sentence alignment is 1-1 but occasionally a sentence-level segment actually consists of two real sentences. The data has been collected from the news domain and Wikipedia. The corpus was then annotated for parts of speech, which was further divided into universal parts of speech (UPOS) and language specific parts of speech (XPOS), and dependency relations. The Bureau of Indian Standards (BIS) tagset[7] has been used for language specific POS tags (Choudhary and Jha, 2014; Ojha and Zeman, 2020). Out of the 37 universal dependency relations, 29 deprel have been used in Bengali and 31 in Magahi (The statistics are given in Table 2). All the 17 UPOS have been used for both Bengali and Magahi (The statistics are given in Table 1). The number of tokens reported for the Bengali and Magahi sentences are 13,110 and 7,575 respectively. Eventually, a total of 50 Bengali sentences were made available for validation to three inter-annotators - native speakers of the language. A kappa score of 0.942613, per dependency, was thus derived. However, we could not do inter-annotators agreement for Magahi.

UPOS Tags	Description	Bengali Statistics	Magahi Statistics
NOUN	Noun	3815	1775
VERB	Verb	1590	780
PUNCT	Punctuation	1720	755
PROPN	Proper noun	805	430
ADJ	Adjective	1275	495
ADP	Adposition	815	1275
DET	Determiner	615	305
PRON	Pronoun	764	214
CCONJ	Coordinating conjunction	385	155
ADV	Adverb	440	150
NUM	Numeral	226	231
PART	Particle	130	110
AUX	Auxiliary	904	783
SCONJ	Subordinating conjunction	270	150
SYM	Symbol	55	33

Table 1: Statistics of used UPOS Tags in the Bengali and Magahi PUD treebank

4. Development of Bengali & Magahi Parser

As mentioned earlier, the Bengali and Magahi treebank was manually annotated using the UD annotation framework. Both, Bengali and Magahi parsers were built on 600 and 200 sentences. The experiment was conducted in two steps.

- **Bengali:** Experiment-1 was run on 200 sentences while Experiment-2 was conducted on

[4] https://www.meity.gov.in/content/language-computing-group-vi

[5] https://kcis.iiit.ac.in/LT/

[6] http://hdl.handle.net/11234/1-4611

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

UD Relations	Description	Bengali Statistics	Magahi Statistics
advmod	Adverbial modifier	530	195
amod	Adjectival modifier of noun	908	324
aux	Auxiliary verb	555	385
case	Case marker	780	1215
cc	Coordinating conjunction	234	102
ccomp	Clausal complement	156	92
compound	Compound	1210	792
conj	Non-first conjunct	276	108
cop	Copula	12	51
det	Determiner	585	175
fixed	Non-first word of fixed expression	180	115
flat	non-first word of flat structure	120	102
goeswith	Non-first part of broken word	-	1
iobj	Indirect object	42	12
mark	Subordinating marker	286	165
nmod	Nominal modifier of noun	995	429
nsubj	Nominal subject	1193	22
nummod	Numeric modifier	186	193
obj	Direct object	179	57
obl	Oblique nominal	860	556
punct	Punctuation	1710	808
root	Root	1000	1000
xcomp	Open clausal complement	230	88

Table 2: Statistics of used UD relations in Bengali and Magahi PUD trebank

600 sentences with the aid of UDPipe open source tool (Straka and Straková, 2017). Cross-validation with an average of 90:10 was used for data splitting where the batch size, learning rate, dropout and embedding size were 50, 0.005, 0.10, 200 respectively, while the other hyper-parameters were randomized for each experiment.

- **Magahi:** Magahi's Experiment-1 was run on 200 sentences using UDPipe similar to Bengali. The data spliting and training features were also the same. Experiment-2 was built on multilingual multi-task model with the aid of UDify open source tool (Kondratyuk and Straka, 2019). We used the same Magahi sentences. In this experiment, the training configurations and features were default including data splitting, batch size, epoch, learning rate, and multilingual BERT layer.

The results are demonstrated in Table 3:

Language	Experiment Details	UPOS	XPOS	UAS	LAS
Bengali	Experiment 1	51.25	62.04	30.23	35.13
Bengali	Experiment 2	78.13	76.09	56.12	47.19
Magahi	Experiment 1	68.08	69.18	34.0	41.74
Magahi	Experiment 2	71.53	66.44	58.05	33.07

Table 3: Results (%) of the Bengali and Magahi Parser

5. Cross-lingual Analysis of Bengali & Magahi PUD

In this section, an extensive linguistic illustration of the language pairs following the annotation schema of the UD v2 guidelines is discussed.

5.1. Nominals

The nominals are divided into three categories in UD - the core arguments, the non-core arguments, and the nominal dependents. These include nsubj (nominal subject), obj (object), iobj (indirect object) under core arguments. The non-core dependents include obl (oblique), vocative, expel (expletive), and dislocated. Lastly, the nominal dependents include nmod (nominal modifier), appos (appositional modifier), and nummod (numeric modifier). In core arguments, nsubj and obj are the most frequently used relations followed by the non-core argument obl. With respect to the nominal dependents, nmod with its subtype nmod:poss is the most frequent followed by the dependency relation nummod.

Figure 1 and Figure 2, showcase the presence of the nominal relations nsubj, obj, iobj, obl. The verb জানিয়েছেন '*janiyechen*' is the root of the sentence. The noun প্রত্যক্ষদর্শী '*protokkhodorshi*' is the nsubj, and the noun পুলিশ '*police*' is the iobj, both of which are dependent on the root. In the lower clause, the noun এপ্রিল *'april'* acts as the obl and the noun ব্যক্তি '*byekti*' acts as the obj, both dependent on the verb করেছিল '*korechilo*', which is further dependent on the root via ccomp relation. In the same way, in Magahi, the noun गबाह ‘*gabaah*’ ‘witness’ is the nsubj of the root ‘told’, and the noun पुलिस ‘*pulis*’ ‘police’ is the iobj. On the other hand, in the lower clause पीड़ित आदमी अप्रील में संदिग्ध आदमी पर हमला कैलकई हल ‘*piiRit aadamii april meN sandigdh aadamii par hamalaa kailkai hal*’ ‘the victim had attacked the suspect in April’, which is dependent on the root ‘told’ via ccomp relation, the noun अप्रील ‘*april*’ ‘April’, has obl relation with the lower verb हमला कैलकई ‘*hamalaa kailkai*’ ‘attack’. There are also nsubj and obj in the lower clause, the nominal पीड़ित आदमी ‘*piiRit aadamii*’ ‘victim’ and संदिग्ध ‘*sandigdh*’ ‘suspect’ respectively.

5.2. Clauses

In UD, clauses are categorized into five- csubj (clausal subject), ccomp (clausal complement), xcomp (open clausal complement), advcl (adverbial clause modifier), and acl (adnominal clause). The four relations ccomp, xcomp, advcl, and acl are frequently found in both Bengali and Magahi, leaving the csubj. We illustrate xcomp and advcl relations here (see figure 1 and figure 2 where ccomp relation is mentioned.)

The following illustrations, Figure 3, and Figure 4 showcase an example of a clausal construction in Bengali and Magahi respectively. In both the languages, the verb kill মেরে *‘mere* in Bengali and मारे ‘*maare*’ in Magahi, which are dependent on the verb root, carry the xcomp relation. The verb চেষ্টা করার *‘chesta korar*, ‘try do-PRESENT-CONTINUOUS in Bengali and परयास करे *paraaas kare*, ‘try do’ in Magahi, have advcl relation with the verb kill.

5.3. Predicates

In this section, we will discuss two types of predicate constructions - simple verb construction and compound verb construction. The compound verb construction can further be subdivided into serial verbs and light verb constructions since Indian languages are rich in compound formation. A simple verb construction contains only the verb, which in UD terms is often referred to as the root, and sometimes combines the verb with an auxiliary. A light verb compound construction is formed by combining the main verb (taken as the root) and a corresponding noun/adjective which is dependent on the root. A serial verb compound formation is the amalgamation of the main verb and a corresponding serial verb, which is again dependent on the main verb.

Figure 5 and Figure 6 showcase an example of a simple verb construction in Bengali and Magahi respectively, wherein the verb মনে ‘*mone*’ acts as the root and is combined with the auxiliary হয়ে ‘*hoye*’ in Bengali and होबऽ ‘*hobe*’ acts as the root and is combined with the auxiliary हे ‘*he*’ in Magahi.

The Figure 1 and Figure 3 illustrated, also showcase a light verb compound construction in Bengali, wherein the noun অভিযোগে *‘obhijoge’* ‘complaints’ is dependent on the verb করা ‘*kora*’ ‘to-do’ and carries the compound:lvc relation in Figure 3. The noun আক্রমণ *akromon* ‘attack’ is dependent on the verb করেছিল ‘*korechilo* ‘did’ and carries the compound:lvc relation in Figure 1. In the same way, the Figure 2 and Figure 4 illustrated, showcase a light verb compound construction in Magahi, wherein the noun आरोप *‘aaropa’* ‘complaints’ is dependent on the verb लगाबल ‘*lagaabala*’ ‘place’ and carries the compound:lvc relation in Figure 4 and the noun हमला *hamalaa* ‘attack’ is dependent on the verb कैलकई ‘*kailkai* ‘did’ and carries the compound:lvc relation in Figure 2. Compound verb formation is a very common construction found in both Bengali and Magahi.[8]

There is also the presence of another type of predicate construction in UD wherein the adjective or noun acts as the root. In Magahi, it is seen that the noun/adjective is combined with the copula, wherein the corresponding noun/adjective acts as the root of the sentence. However, since Bengali lacks copular construction, the noun/adjective itself acts as the root and the other relations are further dependent on it. Figure 7 showcases such a construction in Bengali, wherein the adjective গুমোট *‘gumot*’ ‘stuffy’ acts as the root of the sentence, and Figure-8 illustrates it in Magahi, wherein the adjective उबाऊ ‘*ubaauu* ‘stuffy’ is a root of the sentence.

5.4. Coordination

The figures, 9 and 10, showcase examples of coordination constructions. In UD, a conjunct (conj) is a relation between two elements which are connected by a coordinating conjunction (cc). The first conjunct serves as the head and the second conjunct is related to the first through the cc.

The example, Figure 9, showcases a coordination relation in Bengali, wherein the noun স্পনসরশিপ ‘*sponsorship*’ acts as the first conjunct and the noun বিজ্ঞাপন ‘*biggapon*’ acts as the second conjunct and are joined with the coordinating conjunction এবং ‘*ebong*’. Similarly, Figure 10 showcases a coordination relation in Magahi. The noun परयोजन ‘*pariyojanaa*’ ‘sponsorship’ is the first conjunct and the noun बिज्ञापन ‘*bigyaapana*’ ‘advertising’ is the second conjunct, which are conjoined by a coordinating conjunction आउ ‘*aau*’ ‘and’.

[8] There could be a different view on which element acts as a root in such a construction (e.g., a noun/adjective is a root and the verb depends on it). We have assumed that the verb is a root and a noun/adjective depends on it. Arguing here in favor of our view will take us in a different direction. Also, there is a space constraint.

Conclusion and Future work

This paper presented an attempt in developing a PUD treebank and a parser for the Indian languages, Bengali and Magahi. Currently, the treebanks consist of 1, 000 sentences. The annotation schema, tags used, and linguistic analysis have also been discussed in the sections above. The built Bengali and Magahi PUD treebank will be publicly released in the UD repository.[9],[10]

In the near future, we plan to encode the morphological information in the same PUD treebank for better usage of the built resource. Additionally, a plan to develop a robust parser on 1,000 parallel annotated sentences using zero-shot and to build an enhanced quality of Machine Translation models and NLP tools will also be aimed. Finally, an attempt will be made in increasing the number of sentences to a minimum of 100 for inter-annotator agreement in both the languages to achieve a better understanding of the quality of annotation.

Acknowledgements

Pritha Majumdar, Deepak Alok and Akanksha Bansal would like to acknowledge Panlingua Language Processing LLP for supporting the creation of these treebanks both academically & financially. Atul Kr. Ojha and John P. McCrae would like to acknowledge the EUs Horizon 2020 Research & Innovation programme through the ELEXIS project under grant agreement No. 731015.

6. Bibliographical References

Bhat, R. A. (2017). Exploiting linguistic knowledge to address representation and sparsity issues in dependency parsing of indian languages.

Choudhary, N. and Jha, G. N. (2014). Creating multilingual parallel corpora in indian languages. In Zygmunt Vetulani et al., editors, *Human Language Technology Challenges for Computer Science and Linguistics*, pages 527–537, Cham. Springer International Publishing.

Dash, N. S. (2004). Language corpora: present indian need. In *Proceedings of the SCALLA 2004 Working Conference*, pages 5–7. Citeseer.

De Marneffe, M.-C. and Manning, C. D. (2008). Stanford typed dependencies manual. Technical report, Technical report, Stanford University.

Kondratyuk, D. and Straka, M. (2019). 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China, November. Association for Computational Linguistics.

Kumar, R., Lahiri, B., Alok, D., Ojha, A. K., Jain, M., Basit, A., and Dawer, Y. (2018). Automatic identification of closely-related indian languages: Resources and experiments.

Majumdar, P. (2021). Bengali parallel universal dependency treebank.

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, Marseille, France, May. European Language Resources Association (ELRA).

Petrov, S., Das, D., and McDonald, R. (2011). A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Raj, M., Ratan, S., Ritesh, K., Alok, D., and Ojha, A. K. (2021). Developing universal dependencies treebanks for magahi and braj.

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., Popel, M., Straka, M., Hajic, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinkova, S., Hajic jr., J., Hlavacova, J., Kettnerová, V., Uresova, 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., dePaiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, c., 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., Mendonca, 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*, pages 1–19, Vancouver, Canada, August. Association for Computational Linguistics.

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

Zeman, Daniel; et al., . (2021). Universal dependencies 2.9. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[9] `https://github.com/UniversalDependencies/UD_Bengali-PUD`

[10] `https://github.com/UniversalDependencies/UD_Magahi-PUD`

7. Appendix

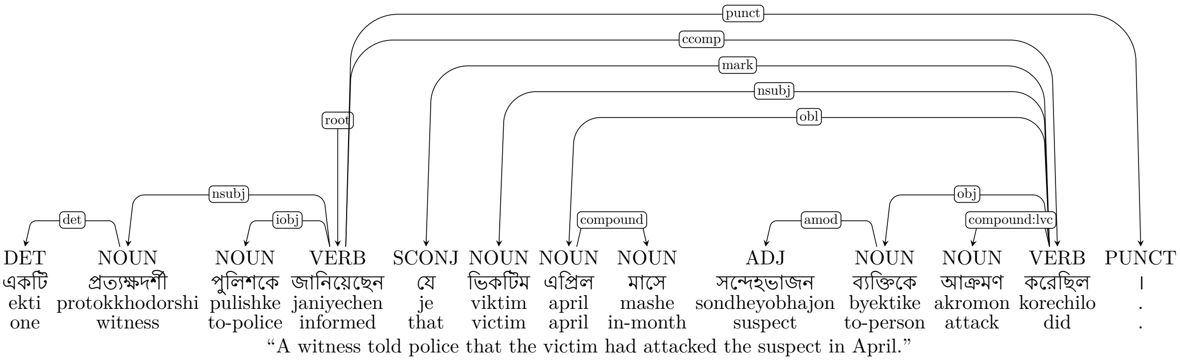
punct
ccomp
mark
nsubj
obl
root
nsubj
obj
det
iobj
compound
amod
compound:lvc
DET NOUN NOUN VERB SCONJ NOUN NOUN NOUN ADJ NOUN NOUN VERB PUNCT
একটি প্রত্যক্ষদর্শী পুলিশকে জানিয়েছেন যে ভিকটিম এপ্রিল মাসে সন্দেহভাজন ব্যক্তিকে আক্রমণ করেছিল ।
ekti protokkhodorshi pulishke janiyechen je viktim april mashe sondheyobhajon byektike akromon korechilo .
one witness to-police informed that victim april in-month suspect to-person attack did .
“A witness told police that the victim had attacked the suspect in April.”

Figure 1: A parallel Bengali construction illustrating the nominal relations

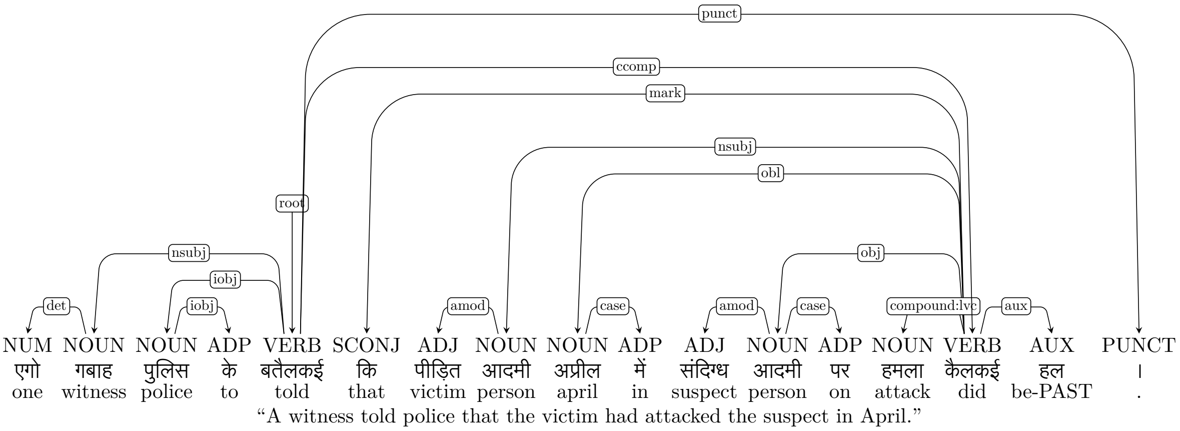
punct
ccomp
mark
nsubj
obl
root
nsubj
obj
iobj
det
iobj
amod
case
amod
case
compound:lvc
aux
NUM NOUN NOUN ADP VERB SCONJ ADJ NOUN NOUN ADP ADJ NOUN ADP NOUN VERB AUX PUNCT
एगो गबाह पुलिस के बतैलकई कि पीड़ित आदमी अप्रील में संदिग्ध आदमी पर हमला कैलकई हल ।
one witness police to told that victim person april in suspect person on attack did be-PAST .
“A witness told police that the victim had attacked the suspect in April.”

Figure 2: A parallel Magahi construction illustrating the nominal relations

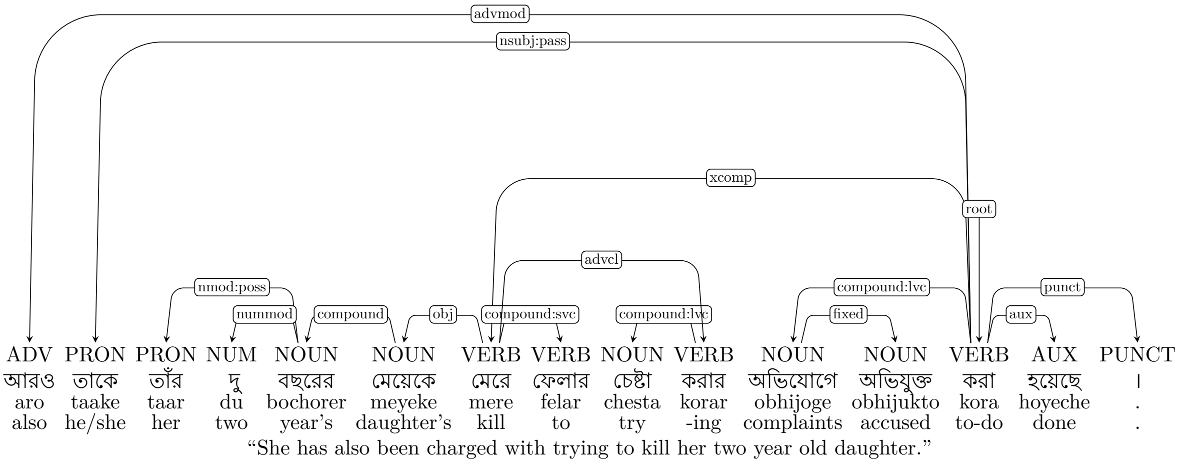
advmod
nsubj:pass
xcomp
root
advcl
nmod:poss
compound:lvc
punct
nummod
compound
obj
compound:svc
compound:lvc
fixed
aux
ADV PRON PRON NUM NOUN NOUN VERB VERB NOUN VERB NOUN NOUN VERB AUX PUNCT
আরও তাকে তাঁর দু বছরের মেয়েকে মেরে ফেলার চেষ্টা করার অভিযোগে অভিযুক্ত করা হয়েছে ।
aro taake taar du bochorer meyeke mere felar chesta korar obhijoge obhijukto kora hoyeche .
also he/she her two year’s daughter’s kill to try -ing complaints accused to-do done .
“She has also been charged with trying to kill her two year old daughter.”

Figure 3: A parallel Bengali construction illustrating the clausal relation

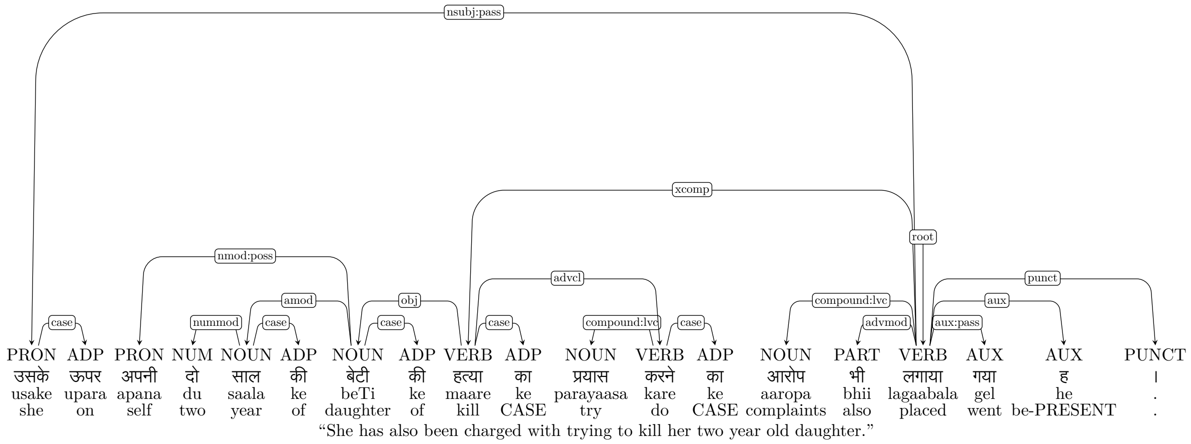
nsubj:pass
xcomp
root
nmod:poss
advcl
amod
obj
compound:lvc
punct
case
nummod
case
case
case
compound:lvc
case
advmod
aux:pass
aux
PRON ADP PRON NUM NOUN ADP NOUN ADP VERB ADP NOUN VERB ADP NOUN PART VERB AUX AUX PUNCT
उसके ऊपर अपनी दो साल की बेटी की हत्या का प्रयास करने का आरोप भी लगाया गया ह ।
usake upara apana du saala ke beTi ke maare ke parayaasa kare ke aaropa bhii lagaabala gel he .
she on self two year of daughter of kill CASE try do CASE complaints also placed went be-PRESENT .
"She has also been charged with trying to kill her two year old daughter."

Figure 4: A parallel Magahi construction illustrating the clausal relation

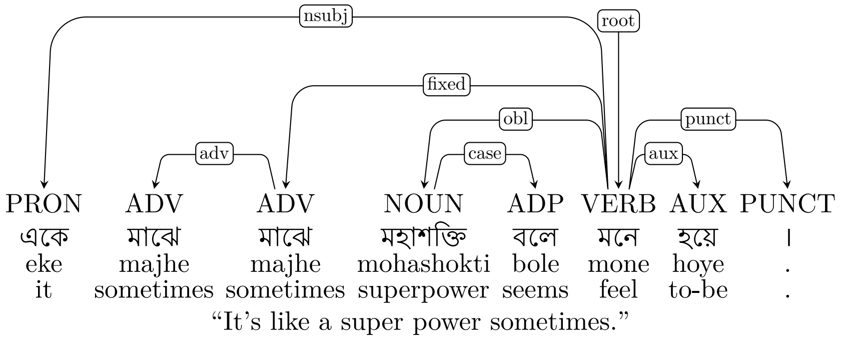
nsubj
root
fixed
obl
punct
adv
case
aux
PRON ADV ADV NOUN ADP VERB AUX PUNCT
একে মাঝে মাঝে মহাশক্তি বলে মনে হয়ে ।
eke majhe majhe mohashokti bole mone hoye .
it sometimes sometimes superpower seems feel to-be .
"It's like a super power sometimes."

Figure 5: A parallel Bengali construction illustrating the simple verb construction.

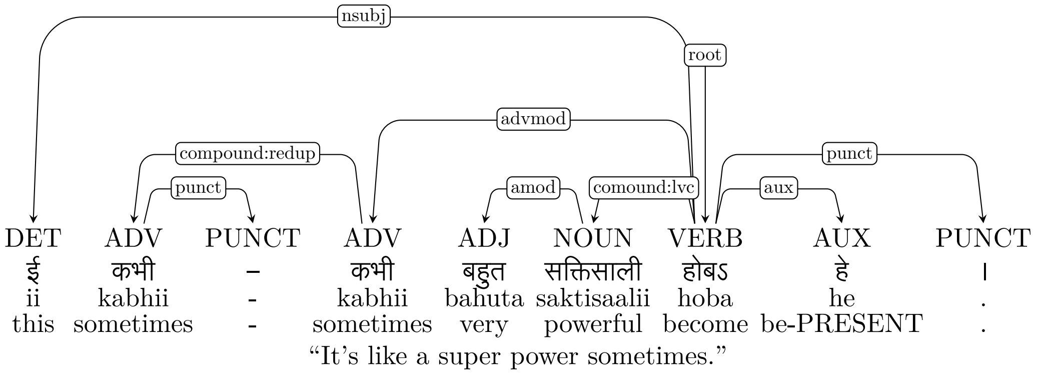
nsubj
root
advmod
compound:redup
punct
punct
amod
comound:lvc
aux
DET ADV PUNCT ADV ADJ NOUN VERB AUX PUNCT
ई कभी – कभी बहुत सक्तिसाली होबऽ हे ।
ii kabhii - kabhii bahuta saktisaalii hoba he .
this sometimes - sometimes very powerful become be-PRESENT .
"It's like a super power sometimes."

Figure 6: A parallel Magahi construction illustrating a simple verb construction

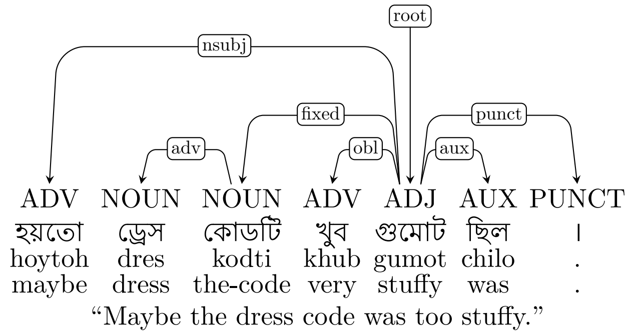
root
nsubj
fixed
punct
adv
obl
aux
ADV NOUN NOUN ADV ADJ AUX PUNCT
হয়তো ড্রেস কোডটি খুব গুমোট ছিল ।
hoytoh dres kodti khub gumot chilo .
maybe dress the-code very stuffy was .
“Maybe the dress code was too stuffy.”

Figure 7: A parallel Bengali construction illustrating the non-verbal predicate construction.

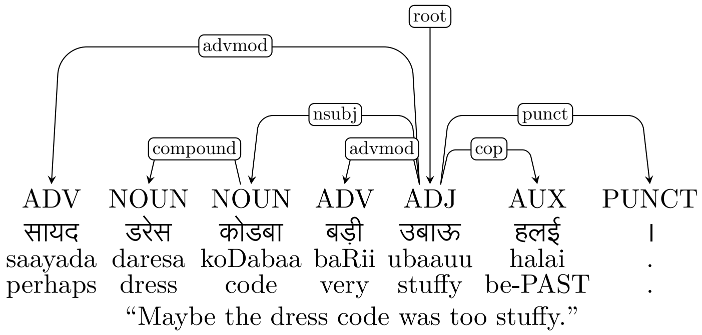
root
advmod
nsubj
punct
compound
advmod
cop
ADV NOUN NOUN ADV ADJ AUX PUNCT
सायद डरेस कोडबा बड़ी उबाऊ हलई ।
saayada daresa koDabaa baRii ubaauu halai .
perhaps dress code very stuffy be-PAST .
“Maybe the dress code was too stuffy.”

Figure 8: A parallel Magahi construction illustrating a non-verbal predicate

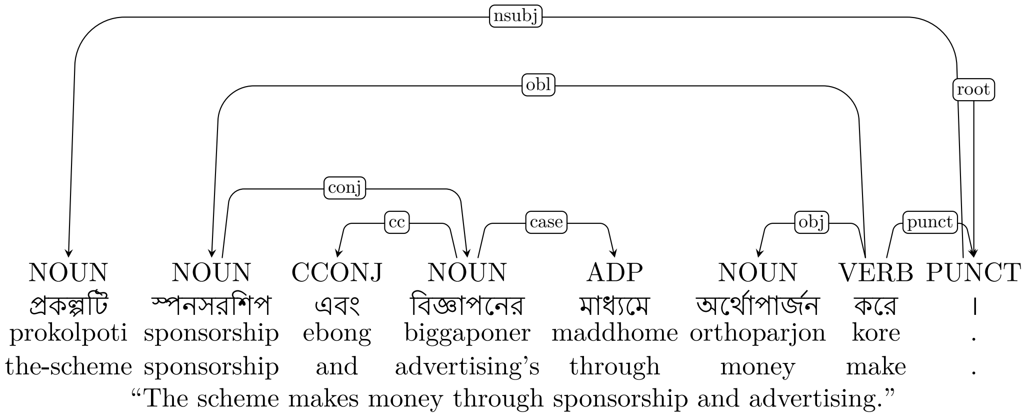
nsubj
obl
root
conj
cc
case
obj
punct
NOUN NOUN CCONJ NOUN ADP NOUN VERB PUNCT
প্রকল্পটি স্পনসরশিপ এবং বিজ্ঞাপনের মাধ্যমে অর্থোপার্জন করে ।
prokolpoti sponsorship ebong biggaponer maddhome orthoparjon kore .
the-scheme sponsorship and advertising’s through money make .
“The scheme makes money through sponsorship and advertising.”

Figure 9: A parallel Bengali construction illustrating the nominal coordinating relation.

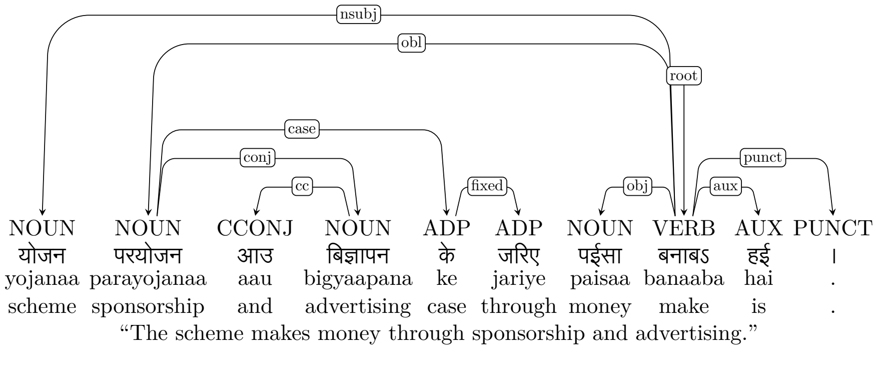
nsubj
obl
root
case
punct
conj
cc
fixed
obj
aux
NOUN NOUN CCONJ NOUN ADP ADP NOUN VERB AUX PUNCT
योजन परयोजन आउ बिज्ञापन के जरिए पईसा बनाबऽ हई ।
yojanaa parayojanaa aau bigyaapana ke jariye paisaa banaaba hai .
scheme sponsorship and advertising case through money make is .
“The scheme makes money through sponsorship and advertising.”

Figure 10: A parallel Magahi construction illustrating the nominal coordinating relation.