PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 120 APRIL 2023 31–46

Universal Dependencies for Malayalam

Abishek Stephen, Daniel Zeman

ÚFAL, Faculty of Mathematics and Physics, Charles University

Abstract

Treebanks can play a crucial role in developing natural language processing systems and to have a gold-standard treebank data it becomes necessary to adopt a uniform framework for the annotations. Universal Dependencies (UD) aims to develop cross-linguistically consistent annotations for the world's languages. The current paper presents the essential pivots of a UD-based syntactically annotated treebank for Malayalam. Sentences extracted from the IndicCorp corpus were manually annotated for morphological features and dependency relations. Language-specific properties are discussed which shed light on many of the grammatical areas in the Dravidian language syntax which needs to be examined in depth. This paper also discusses some pertaining issues in UD taking into consideration the Dravidian languages and provides insights for further improvements in the existing treebanks.

1. Introduction

A treebank is a collection of syntactically and (or) semantically annotated language data. Most treebanks are developed using a combination of manual and automated processes (Kakkonen, 2006). Many treebanks based on dependency grammar (Jiang and Liu, 2015; de Marneffe and Nivre, 2019) have been developed recently. The Prague Dependency Treebank (Hajič et al., 2020) of Czech is one of the largest dependency treebanks. But with the growing need for multilingual language systems and better cross-linguistic evaluations, a uniform framework is needed. Universal Dependencies (UD) (de Marneffe et al., 2021) is a framework for consistent annotation of natural language data (parts of speech, morphological features, and syntactic dependencies) across different human languages. UD is an open community effort with

 Corresponding author: stephen@ufal.mff.cuni.cz
Cite as: Abishek Stephen, Daniel Zeman. Universal Dependencies for Malayalam. The Prague Bulletin of Mathematical Linguistics No. 120, 2023, pp. 31–46. doi: 10.14712/00326585.026.

over 500 contributors who have produced 243 treebanks in 138 languages so far.[1] Currently, UD has treebanks for 10 Indian languages among which there are 3 Dravidian languages including Malayalam, the others being Telugu and Tamil. The Malayalam treebank[2] is a step forward to do better comparative evaluation of syntactic properties of the Dravidian languages and also with other unrelated languages. The following sections of the paper describe how the treebank was developed elaborating on the challenges and the resorts taken.

2. Malayalam

Malayalam is a Dravidian language spoken primarily in the south-Indian state of Kerala. Malayalam is an agglutinating language like the other Dravidian languages. There are 35 million native-speakers of Malayalam in India. Malayalam has borrowed freely from other languages, especially from Sanskrit. That accounts for lemmas of many content words resembling those in Sanskrit. The canonical word order in Malayalam is SOV. Unlike Tamil or Telugu, Malayalam lacks verbal agreement, i.e., the verb does not encode the person, number and gender of the subject (nor those of object or any other argument). We have a three-way distinction of tense in Malayalam, i.e., present, past and future. Additionally, Malayalam has perfective and imperfective aspects along with a number of different moods. Nominalized verb forms are very frequent and so are cleft constructions. Core arguments are marked by the morphological cases nominative (subject) and accusative (object). Core arguments are bare noun phrases without adpositions. Subjects are suppressed when verbs are passivized.

3. Data

The first 20 annotated sentences are Malayalam equivalents of the examples from the Cairo CICLing Corpus.[3] With a preference for texts from different genres in order to get hold of different and unique syntactic constructions, the rest is taken from the Malayalam part of IndicCorp (Kakwani et al., 2020). IndicCorp is a freely available corpus for Indian languages, developed by scraping of web sources comprising of news articles, magazines and books. The corpus contains a single large text file with automatic sentence segmentation, one sentence per line. The publicly released version is randomly shuffled and untokenized.[4] The size of the Malayalam part of IndicCorp exceeds 50 million sentences. The Malayalam treebank currently contains 218 sentences / 2403 words, to be released in UD 2.12.

[1] `https://universaldependencies.org/`

[2] Currently just a small sample of Malayalam grammatical examples.

[3] `https://github.com/UniversalDependencies/cairo`

[4] Available at `https://ai4bharat.iitm.ac.in/corpora`.

4. Methodology

We process the sentences in small batches. After the initial batches, annotation guidelines specific to Malayalam are refined depending on phenomena encountered. After each batch we also retrain a model for tagging and parsing and use it to pre-annotate the next batch, which is then manually corrected in two steps: First, the annotator verifies the annotation of every word including its attachment in the dependency tree, and modifies the annotation where needed. Second, automatic tools are employed to identify errors and inconsistencies, which are then manually corrected. We do not have at our disposal multiple Malayalam-speaking annotators who could annotate the same span and then compare the results. Script-based quality checking should at least partially compensate for this shortcoming.

Manual annotation (including corrections of tokenization and occasionally sentence segmentation) is done in the CoNLL-U Editor (Heinecke, 2019).

4.1. Preprocessing

Unicode NFC normalization is applied to all input sentences. For example, some texts represent the long *ō* (`MALAYALAM VOWEL SIGN OO`, `U+D4B`) as the sequence of *ē* (`MALAYALAM VOWEL SIGN EE`, `U+D47`) and *ā* (`MALAYALAM VOWEL SIGN AA`, `U+D3E`); both representations result in the same glyph. The normalization makes sure to convert them to `U+D4B`, which is the canonical representation. In addition to NFC, we also normalize a few sequences that are used as an alternative representation of so-called chillu letters. These letters are specific syllable-closing variants of certain consonants and they do not have analogy in other Indian scripts. The alternative encoding uses a standard consonant followed by viram (`U+D4D`) and `ZERO WIDTH JOINER` (`U+200D`); we convert any such sequence to the Unicode point dedicated to the resulting chillu consonant.

Furthermore, we generate sentence-level English translation with the help of Google Translate and we use a script[5] to add Latin transliteration of whole sentences as well as of individual word forms. This step is repeated after annotation to also provide transliteration of lemmas.

4.2. Tokenization

In Malayalam, words are delimited by whitespace characters or punctuation. Multiword tokens are relatively common in Malayalam. In the following situations, we understand orthographic tokens as corresponding to multiple syntactic words and split them:

- The copula ആക്കുക (*āk*) 'to be' is written as a suffix of the nominal or adjectival predicate. However, sometimes it is suffixed to another word in the clause, in-

[5] `https://github.com/dan-zeman/translit/blob/main/conllu_translit.pl`

dicating that it is a clitic rather than a derivational morpheme that would derive a verb from a noun/adjective.
- The quotative particle or the complementizer എന്ന് (*enn*) 'that' usually occurs as a suffix of the verb or the copula. Given that we split the copula as a syntactic word, we split the complementizer as well. (Also, it increases parallelism with languages where complementizers are independent words, and avoids having to define a language-specific feature for verb with complementizer.)
- The coordinating clitics -ഉം (*-uṁ*) 'and' and -ഓ (*-ō*) 'or' are written together with conjuncts but analyzed as separate syntactic words.
- In orthography sometimes the object and the verb of a sentence occur as a multiword token. For example, in the sentence പെൺകുട്ടി തന്റെ സുഹൃത്തിന് കത്തെഴുതി (*peṇkuṭṭi tanṟe suhrttin katteḻuti*) 'The girl wrote a letter to her friend', കത്ത് (*katt*) 'letter' and എഴുതി (*eḻuti*) 'wrote' occur as a multiword token and are split.

4.3. Annotation

The selected sentences from the IndicCorp were added to the CoNLL-U Editor. The editor commands were thereby used to carry out the annotations. Splitting of tokens and/or paragraphs[6] were done in the editor itself.

4.4. Validation and Feature Checking

The official UD validation script[7] verifies the CoNLL-U file format as well as data conformity with the general UD annotation guidelines. It can also check permitted feature-value combinations for individual part-of-speech categories in the given language, dependency relation subtypes, and lemmas of auxiliary verbs. We have provided Malayalam-specific definitions for these tests.

While the validator can exclude certain universally defined feature values from Malayalam data, and it can allow feature values separately for individual POS categories, we want to specify more detailed rules that go beyond this. For example, the UD validator knows that `Gender` is relevant for pronouns in Malayalam, but we want to make sure that it occurs only with third-person personal pronouns. The UD validator checks that `Tense` does not occur with anything but verbs (and auxiliaries), but we want to be more specific, allow it for indicative forms and disallow it for imperative and necessitative forms. Moreover, we want to increase consistency by requiring that all verbs in indicative have a non-empty value of `Tense`. Tests of this sort are implemented in the Udapi-Python tool[8] (Popel et al., 2017) in the processing block

[6]While normally a line in the corpus corresponds to one sentence, some lines were sequences of multiple sentences.

[7]`https://universaldependencies.org/release_checklist.html#validation`

[8]`http://udapi.github.io/`

`ud.ml.MarkFeatsBugs`. In the future we envisage similar language-specific tests also for the dependency relations.

4.5. UDPipe

Manual annotation is a laborious task, especially if all morphological features have to be filled out for every word in a morphologically rich language. We thus use UDPipe 1.2[9] (Straka and Straková, 2017), a trainable tool that can tokenize text, tag it and parse it in the UD style. Obviously, the output of UDPipe is not perfect, so we must invest significant manual effort anyway, but at least part of the annotation can be guessed correctly by UDPipe's model.

After annotating the first 30 sentences (which was done without the help of UDPipe), we used these sentences as training data and trained a simple model (with the default configuration). This model was then used to parse 100 sentences from IndicCorp. As expected, the accuracy was quite bad, but at least the tool could guess the approximate word segmentation and prepare the data in the CoNLL-U format. In the next round we carefully polished annotation of the new sentences until it passed the UD validation and all additional consistency tests defined by us. A new UDPipe model, trained on 130 hand-annotated sentences, was significantly better and could predict some annotations correctly. We will repeat this process with new batches of manually verified data and we expect the model to gradually improve and make fewer errors.

5. Part-of-Speech Tagging

The current version of the treebank contains 16 part-of-speech tags including `SYM` and `X` (see POS frequencies in Table 1); the only category missing from the current data is interjections. For the POS tagging the morphological cues were predominantly used. But in some cases the syntactic context was considered to capture the word category in a better way. For instance, the quotative particle എന്ന് (*enn*) 'that' is tagged `PART` where it is used as a 'quotative marker' and `SCONJ` where it is used as complementizer.

AUX: The copula verbs ആക് (*āk*) 'be' and ഉണ്ട് (*uṇṭ*) 'be' are tagged `AUX`. Additionally, the modal auxiliary verbs കഴിയുക (*kaḻiyuka*) 'can, be able to' and വേണം (*vēṇaṁ*) 'want' are also tagged `AUX`.

CCONJ: The particle -ഉം (*-uṁ*) 'and' that serves as a conjoining element for nouns and verbs is tagged `CCONJ` along with പക്ഷേ (*pakṣē*) 'but' and the particle -ഓ (*-ō*) 'or'. In Malayalam, the third person plural pronouns അവർ (*ivar*) 'they' and അവ (*iva*) 'these' can act as a conjunction if realized as എന്നിവർ (*ennivar*) and എന്നിവ (*enniva*). These forms are also tagged `CCONJ`.

[9] `https://ufal.mff.cuni.cz/udpipe/1`

POS	count	POS	count	POS	count	POS	count
ADJ	230	CCONJ	93	PART	58	SCONJ	25
ADP	38	DET	39	PRON	84	SYM	1
ADV	99	NOUN	720	PROPN	260	VERB	282
AUX	113	NUM	42	PUNCT	317	X	2

Table 1. Frequencies of POS tags.

SCONJ: In Malayalam, the reported speech is marked with a quotative particle (Asher and Kumari, 1997). The quotative particle എന്ന് (*enn*) when used as the complementizer is tagged `SCONJ`. Malayalam has only one sentence-final complementizer.

PART: The particle -ഉം (*-uṁ*) when used as an emphasizing element (rather than conjunction) is tagged `PART`. The quotative particle എന്ന് (*enn*) and its variant എന്ന (*enna*)[10] used in adnominal clauses are also tagged `PART`.

6. Morphological Features

The inherent gender of nouns[11] determines which personal pronoun can refer to the noun, and it is sometimes reflected as agreement on adjectives. It is not reflected on verbs (unlike in related Tamil). We do not annotate the gender of nouns in data but we do so for third-person pronouns with one of three values: `Masc`, `Fem` or `Neut`. Like `Gender`, `Animacy` is also an inherent feature of nominal words (`NOUN`, `PROPN`, and `PRON`). It has two values: `Anim` and `Inan`. Animacy is grammatically relevant because inanimate nouns may occur without accusative marking -എ (*-e*) when used as direct objects (cf. examples (1a) and (1b) below). Animates include nouns denoting persons and in some cases animals, or trees. Animacy aligns with gender only partially. Masculine and feminine third person pronouns refer to persons and are perceived as animate. Neuter pronouns can be animate if referring to animals or plants, and inanimate otherwise. For inanimates, the accusative form is equal to the nominative അത് (*at*) 'it', while for animates it uses a separate form അതിനെ (*atine*) 'it'. We annotate the animacy of third person neuter pronouns but we omit the feature for other personal pronouns.

In example (1) we can see how the accusative case assignment based on animacy of the objects plays a vital role in disambiguating the subject and the object. The example

[10]The quotative particle is realized as the relative particle എന്ന (*enna*) in relative clauses. It is referred to as 'relative particle' in Asher and Kumari (1997).

[11]There is a tendency that masculine nouns end in -അൻ (*-an*) and feminine nouns in -ഇ (*-i*). For example, male thief is കള്ളൻ (*kaḷḷan*) and female thief is കള്ളി (*kaḷḷi*). However this type of classification cannot be generalized (Asher and Kumari, 1997).

(1d) shows that if the object and subject both are animate, then the object needs to be marked accusative, otherwise it will not be possible to distinguish the subject and object in the sentence (because both SOV and OSV word orders are possible).

(1) a. ñān oru vaṇṭi vāṅṅi
I.NOM one car buy.PAST
'I bought a car'
b. ñān oru vaṇṭi(y)-e vāṅṅi
I.NOM one car-ACC buy.PAST
'I bought a car'
c. ñān avan-e viliccu
I.NOM he-ACC call.PAST
'I called him'
d. *ñān avan viliccu
I.NOM he call.PAST
'I called he'

Case has 13 possible values: `Nom`, `Acc`, `Gen`, `Dat`, `Ins`, `Loc`, `Abl`, `All`, `Cmp`, `Com`, `Ben`, `Cau`, `Voc`. Malayalam is an agglutinative language and many spatiotemporal and/or case-like morphemes are analyzed as postpositions. The `Case` feature occurs with the nominal words, i.e., `NOUN`, `PROPN`, `PRON`, `NUM` and also with nominalized verb forms. Nominalized verb forms are frequently used where the verbs take the nominalizing suffix -ത് (*-t*) (Asher and Kumari, 1997). These verb forms are marked as `VerbForm=Vnoun` and are morphologically marked for case, tense and polarity. In cleft constructions, they occur along with the copula ആക് (*āk*), which is postposed to the focused element. In example (2) we can see how nominalization works in Malayalam.

(2) a. avan at śariyāyi paṟaññu
he.NOM that correctly say.PAST
'He said it correctly'
b. avan paṟaññat śariy-āṇ
he.NOM say.PAST.NML correct-be.PRES
'What he said was correct.'
c. avan paṟaññat-āṇ śari
he.NOM say.PAST.NML-be.PRES correct
'*What he said* was correct.'

Example (2a) is a simple declarative clause with a finite verb. (2b) shows the nominalized construction and (2c) is a cleft construction.

7. Dependency Relations

The main dividing lines in the taxonomy of dependency relations in UD are between the core arguments of clausal predicates, non-core dependents of clausal predicates, and dependents of nominals.[12]

7.1. Core and Non-Core Dependents

According to the UD taxonomy, core arguments are subjects and objects. But this limits the treatment only to those constituents that are morphologically marked with the nominative and accusative case.[13] The non-core dependents or the oblique dependents are those arguments with coding strategies not used by the core arguments (Zeman, 2017). In world's languages, certain predicates would take dependents occupying the subject and object positions and not marked as nominative and accusative respectively. For example, in the Czech sentence *Martin hýbá nábytkem* 'Martin moves the furniture', the noun *nábytek* 'furniture' takes the instrumental case, although the verb *hýbat* 'to move' selects it as an argument. On being passivized the object remains in the instrumental case (Zeman, 2017). Similar examples from other languages show that what is traditionally regarded as 'objects' or 'subjects' in these languages may be coded with cases similar to the oblique dependents.

7.1.1. Non-Nominative Subjects

The constituent ordering in morphologically rich languages can be different from the *typical* ordering of nominative constituents preceding the non-nominative constituents and it is largely semantically predictable (Bayer, 2004). For example in German we do find instances where a dative argument occurs with certain predicates which may or may not have any nominative arguments.

(3) a. Mir ist kalt
me.DAT is cold
'I am cold'

b. Mir war schlecht
me.DAT was bad
'I was sick'

Data from Sigurðsson (2004) for Icelandic also shows similar constructions. In Icelandic the non-nominative subjects (NNS) are referred to as *quirky* subjects (Sigurðsson, 1992) as they pass the tests for subjecthood.

[12] https://universaldependencies.org/u/dep/

[13] This follows from the normal treatment of A and P arguments in primary transitive clauses (Andrews, 2007) in Malayalam. The nominative and accusative cases identify nominal arguments. For open and closed clausal dependents the core vs. non-core distinction is trickier (Przepiórkowski and Patejuk, 2018).

(4) a. Þeim er kalt
them.DAT is cold
'They are freezing'

b. Henni fór fram
her.DAT went forth
'She got better'

This type of pre-verbal dative arguments in German and Icelandic look similar, nevertheless they are syntactically different from each other (Fischer, 2004): In German they are just oblique dependents, while in Icelandic there is evidence that they behave like subjects, that is, core arguments.

Similar dative experiencer *subjects* in Kannada (Amritavalli, 2004) and Hindi (Mahajan, 2004) originate in unaccusative contexts, i.e., the nature of the predicates decides the origin of these non-nominative subjects. In Malayalam, the dative experiencer constructions occur with predicates that express possession and mental or physical experience.

(5) a. avalkk oru vīṭ uṇṭ
her.DAT one house is
'She has a house'

b. enikk viśakkunnu
me.DAT hunger.PRES
'I am hungry'

The dative case of the dative NPs in Malayalam is an inherent or a semantic case (Jayaseelan, 2004a) and there can be more than one case relation for an argument as in (6):

(6) a. enikk kaḻiy-illa, ninn-e nokk-ān
me.DAT be.able-NEG you-ACC look.after-INF
'I cannot look after you' (Jayaseelan, 2004a)

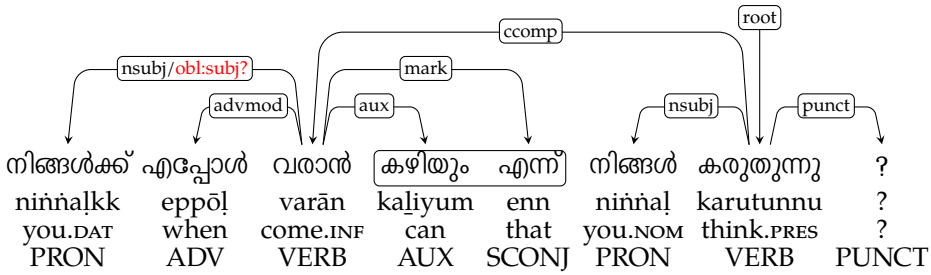
b. enn-ekkoṇṭu kaḻiy-illa, ninn-e nokk-ān
me.INSTR be.able-NEG you-ACC look.after-INF
'I cannot look after you' (Jayaseelan, 2004a)

With the verb നോക്കുക (*nōkkuka*) 'look after', we can have the dative and instrumental alternation on the *subject* argument. Both the sentences in (6) have the same semantic reading. With a different verb having different semantics this is not possible:

(7) a. enikk ninne iṣṭam alla
me.DAT you-ACC liking NEG
'I don't like you'
(Jayaseelan, 2004a)

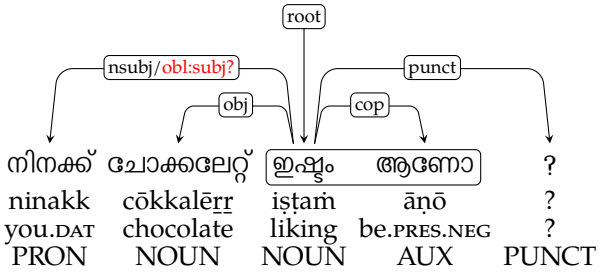
b. *enn-ekkoṇṭu ninne iṣṭam alla
me.INSTR you-ACC liking NEG
'I don't like you'
(Jayaseelan, 2004a)

Hence, Jayaseelan (2004a) concludes that dative NP is an oblique argument, not a subject as the case-marking of the verb's oblique arguments are semantically determined. Zeman (2017) has shown that the non-nominative arguments in Russian and Czech do not behave like *typical* subjects and should be treated as oblique arguments



'When do you think you can come?'

Figure 1. An example of a non-nominative subject.

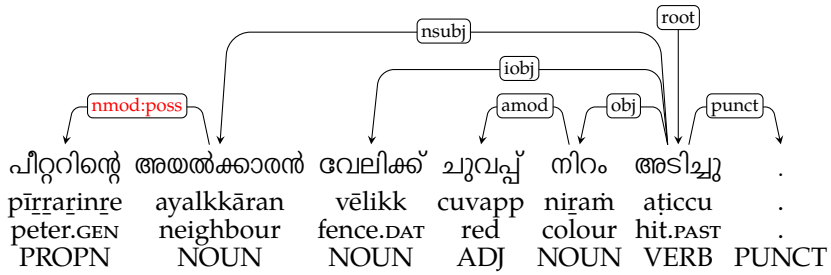


'Do you like chocolate?'

Figure 2. An example of a non-nominative subject.

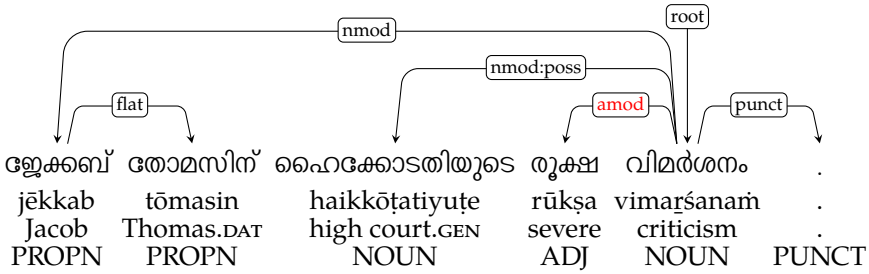
marked with the dependency relation `obl:arg`. However, the existing UD treebanks for the Dravidian languages, i.e., Tamil MWTT (Krishnamurthy and Sarveswaran, 2021) and Telugu MTG (Rama and Vajjala, 2018) treat the non-nominative arguments as core dependents marking them as `nsubj:nc`. We have tentatively also used the `nsubj` relation for non-nominative subjects in the current version of the Malayalam UD treebank, hence all Dravidian UD treebanks are compatible. However, we regard the question as open and do not exclude the possibility of re-analyzing them as oblique dependents in the future—preferably throughout the Dravidian family.

Example annotation of NNS in Malayalam is shown in Figures 1 and 2. Since the UD taxonomy of core vs. non-core dependents is an ongoing discussion in the UD community we may revert the NNS to oblique dependents and label them with a new subtype `obl:subj`. The goal here is to achieve a consistent explanation of the NNS constructions across the Dravidian languages.



'Peter's neighbor painted the fence red.'

Figure 3. An example of genitive modification.



'High Court criticizes Jacob Thomas.'

Figure 4. An example of an adjective modifying a nominal.

7.2. Nominal Dependents

nmod: We have used the `nmod` relation to mark the attributes of nouns or noun phrases. The label `nmod:poss` is used for the genitive complements (Figure 3).

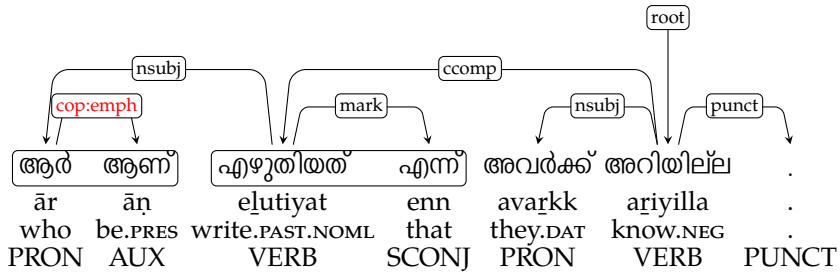
amod: This relation is used for all the non-clausal adjectival attributes of nouns or pronouns (Figure 4).

7.3. Other Dependency Relations

Here we discuss the other relations, mainly the subtypes[14] of various dependency relations that are used for Malayalam.

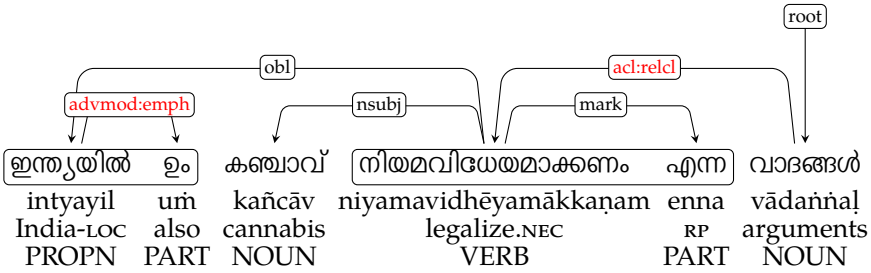
cop:emph is a special relation capturing the *focus* in a phrase. In cleft constructions, the verb is nominalized and the copula is postposed to the focused element (Figure 5).

[14]Subtypes are language-specific and optional.



'They do not know who wrote it.'

Figure 5. An example of a copula used to emphasize the focused constituent in cleft constructions. More literally, the sentence says 'Who is it (whose) writing (it was), that they know-not.'



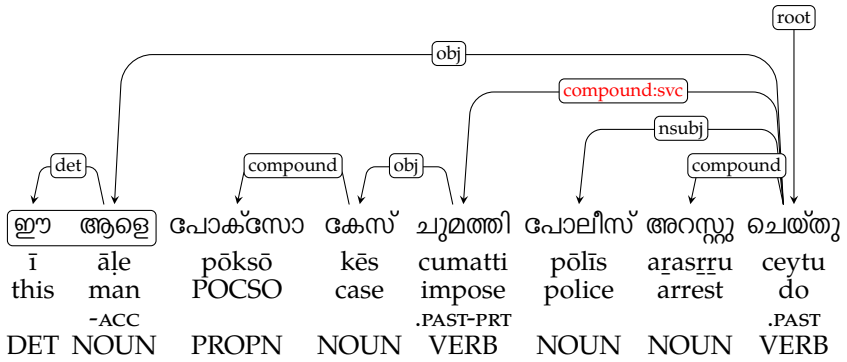
'Arguments to legalize cannabis in India'

Figure 6. An example of the -ഉം (uṁ) particle emphasizing a nominal. This sentence also serves as an example of a relative clause.

`advmod:emph`: The particle -ഉം (*uṁ*) (which is also the coordinating clitic) is used as an emphasizing element and to differentiate it from the `cc` dependency relation, `advmod:emph` is used (Figure 6).

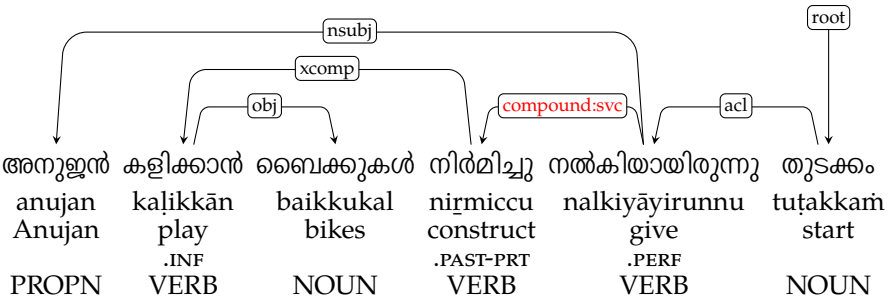
`compound:svc`: Serial verb constructions are the sequence of verbs and their (shared) complements.[15] The verbs in these constructions are not separated by any overt marker of coordination or subordination. In most of the cases, the verbs are lexicalized and cannot be separated by any intervening material. The final verb is usually finite and the preceding verbs are non-finite and resemble the past participle forms (Jayaseelan, 2004b) (Figures 7 and 8).

[15] `https://universaldependencies.org/u/dep/compound-svc.html`



'He was arrested by the police on a POCSO case.'

Figure 7. An example of a serial verb construction.

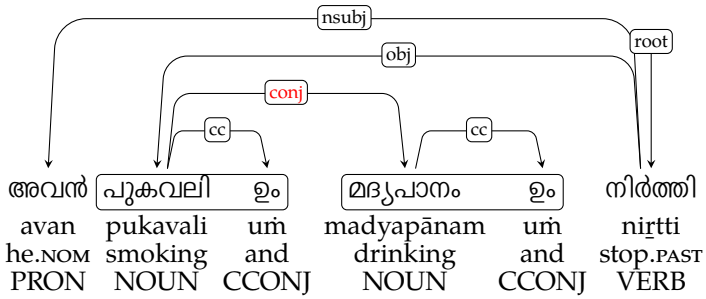


'It started with making bikes for Anujan to play with'

Figure 8. An example of a serial verb construction.

`acl:relcl`: The relative clause formation requires the relative particle എന്ന (*enna*). This subtype can be also applied to the sentential relative clauses but there are no such examples in the treebank yet. The participial relative clauses are treated as dependents of nominals and are labelled with the dependency relation `acl:relcl`; an example can be seen in Figure 6.

`conj`: According to UD guidelines, coordination receives asymmetric treatment, i.e., the leftmost conjunct is the technical head and all other conjuncts 'depend' on it. For head-final languages it may cause a problem as discussed for Japanese and Korean in Kanayama et al. (2018). This depends on how the case marking happens in these languages. If both the conjuncts are case-marked then the left-headed `conj`



'He stopped smoking and drinking'

Figure 9. An example of coordination.

relation works fine but if the mechanism of case assignment to the conjuncts happens in some other way that might disrupt the phrasal units, then the existing principle of left-headed `conj` may pose some challenges. In Malayalam, we see that the left-headedness does not cause any problems. Malayalam uses multiple `cc` relations in a coordination unit because the coordinating clitics -ഉം (*-uṁ*) 'and' and -ഓ (*-ō*) 'or' are appended to each of the conjuncts (Figure 9).

8. Conclusion

This paper presents the properties of a new UD-based treebank for Malayalam. We have discussed the annotation process along with elaborating on the various choices of the dependency relations. The UD treebanks of the Dravidian languages need to adopt a consistent annotation for syntactically similar constructions. We have illustrated various ways in which many syntactic phenomena in Malayalam have been tackled based on the existing UD guidelines. In the subsequent releases of the treebank, the annotations may undergo subtle improvements.

Acknowledgements

This work was supported by the grants 20-16819X (LUSyD) of the Czech Science Foundation; and LM2023062 (LINDAT/CLARIAH-CZ) of the Ministry of Education, Youth, and Sports of the Czech Republic. In addition, the first author was supported by the Scholarship of the Ministry of Education, Youth and Sports in Support of Foreign Nationals' Study at Public Institutions of Higher Education in the Czech Republic (promulgated on 28 January 2014 under ref. No. MSMT-44726/2013).

Bibliography

Amritavalli, R. Experiencer datives in Kannada. In *Non-nominative Subjects: Volume 1*, pages 1–24. 2004. doi: 10.1075/tsl.60.03amr.

Andrews, Avery D. The Major Functions of the Noun Phrase. In Shopen, Timothy, editor, *Language Typology and Syntactic Description. Volume I: Clause Structure*, pages 132–223. Cambridge University Press, 2007. doi: 10.1017/CBO9780511619427.003.

Asher, R. E. and T. C. Kumari. *Malayalam*. Routledge Descriptive Grammars. Routledge, London, 1997. doi: 10.4324/9781315002217.

Bayer, Josef. Non-nominative subjects in comparison. In *Non-nominative Subjects: Volume 1*, page 49–76. 2004. doi: 10.1075/tsl.60.05bay.

de Marneffe, Marie-Catherine and Joakim Nivre. Dependency Grammar. *Annual Review of Linguistics*, 5(1):197–218, 2019. doi: 10.1146/annurev-linguistics-011718-011842. URL https://doi.org/10.1146/annurev-linguistics-011718-011842.

de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, June 2021. doi: 10.1162/coli_a_00402. URL https://aclanthology.org/2021.cl-2.11.

Fischer, Susann. The diachronic relationship between quirky subjects and stylistic fronting. In *Non-nominative Subjects: Volume 1*, page 193–212. 2004. doi: 10.1075/tsl.60.11fis.

Hajič, Jan, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Fučíková, and et al. Prague Dependency Treebank – Consolidated 1.0 (PDT-C 1.0), 2020. URL http://hdl.handle.net/11234/1-3185. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Heinecke, Johannes. ConlluEditor: a fully graphical editor for Universal Dependencies treebank files. In *Universal Dependencies Workshop 2019*, Paris, 2019. doi: 10.18653/v1/W19-8010. URL https://github.com/Orange-OpenSource/conllueditor/.

Jayaseelan, K. The possessor — experiencer dative in Malayalam. In *Non-nominative Subjects: Volume 1*, page 227–244. 2004a. doi: 10.1017/CBO9781139003575.006.

Jayaseelan, K. The Serial Verb Construction in Malayalam. In *Clause Structure in South Asian Languages*, pages 67–91. Springer Netherlands, 01 2004b. ISBN 978-1-4020-2719-2. doi: 10.1007/978-1-4020-2719-2_3.

Jiang, Jingyang and Haitao Liu. Review of Lucien Tesnière, Elements of structural syntax, translated by Timothy Osborne and Sylvain Kahane, Amsterdam & Philadelphia, PA: John Benjamins, 2015. *Journal of Linguistics*, 51(3):705–709, 2015. ISSN 0022-2267. URL https://www.jstor.org/stable/26570750.

Kakkonen, Tuomo. Dependency treebanks: methods, annotation schemes and tools. In *Proceedings of the 15th Nordic Conference of Computational Linguistics (NODALIDA 2005)*, pages 94–104, Joensuu, Finland, May 2006. University of Joensuu, Finland. URL https://aclanthology.org/W05-1714.

Kakwani, Divyanshu, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 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, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.445. URL https://aclanthology.org/2020.findings-emnlp.445.

Kanayama, Hiroshi, Na-Rae Han, Masayuki Asahara, Jena Hwang, Yusuke Miyao, Jinho Choi, and Yuji Matsumoto. Coordinate Structures in Universal Dependencies for Head-final Languages. pages 75–84, 01 2018. doi: 10.18653/v1/W18-6009.

Krishnamurthy, Parameswari and Kengatharaiyer Sarveswaran. Towards Building a Modern Written Tamil Treebank. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 61–68, Sofia, Bulgaria, December 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.tlt-1.6.

Mahajan, Anoop K. On the origin of non-nominative subjects. In *Non-nominative Subjects: Volume 1*, page 283–299. 2004. doi: 10.1075/tsl.60.16mah.

Popel, Martin, Zdeněk Žabokrtský, and Martin Vojtek. Udapi: Universal API for Universal Dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 96–101, Gothenburg, Sweden, 2017.

Przepiórkowski, Adam and Agnieszka Patejuk. Arguments and Adjuncts in Universal Dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3837–3852, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1324.

Rama, Taraka and Sowmya Vajjala. A Dependency Treebank for Telugu. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pages 119–128, Prague, Czechia, 2018. URL https://aclanthology.org/W17-7616.pdf.

Sigurðsson, Halldor Armann. The case of quirky subjects. Workingpaper, Department of Scandinavian Languages, Lund University, 1992.

Sigurðsson, Halldor Armann. Icelandic non-nominative subjects. In *Non-nominative Subjects: Volume 2*, page 137–159. 2004. doi: 10.1075/tsl.61.09sig.

Straka, Milan and Jana Straková. 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, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/K17-3009. URL http://www.aclweb.org/anthology/K/K17/K17-3009.pdf.

Zeman, Daniel. Core Arguments in Universal Dependencies. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 287–296, Pisa, Italy, September 2017. Linköping University Electronic Press. URL https://aclanthology.org/W17-6532.

Address for correspondence:
Abishek Stephen
stephen@ufal.mff.cuni.cz
ÚFAL MFF UK
Malostranské náměstí 25, Praha, CZ-11800, Czechia