Sinhala Dependency Treebank (STB)

Chamila Liyanage[*], **Kengatharaiyer Sarveswaran**[+], **Thilini Nadungodage**[*] and **Randil Pushpananda**[*]

[*]University of Colombo School of Computing, Sri Lanka
[*]{cml,hnd,rpn}@ucsc.cmb.ac.lk
[+]Department of Computer Science, University of Jaffna, Sri Lanka
[+]sarves@univ.jfn.ac.lk

Abstract

This paper reports the development of the first dependency treebank for the Sinhala language (STB). Sinhala, which is morphologically rich, is a low-resource language with few linguistic and computational resources available publicly. This treebank consists of 100 sentences taken from a large contemporary written text corpus. These sentences were annotated manually according to the Universal Dependencies framework. In this paper, apart from elaborating on the approach that has been followed to create the treebank, we have also discussed some interesting syntactic constructions found in the corpus and how we have handled them using the current Universal Dependencies specification.

1 Introduction

Integrating linguistic information, specifically syntactic information, into language processing tools and applications improves accuracy. This has been proven for applications such as machine translators (Habash, 2007; Li et al., 2017) and natural language understanding (McCord et al., 2012; Ohta et al., 2006). It is also shown that explicitly integrating syntactic and semantic information for training pre-trained models such as Bidirectional Encoder Representations from Transformer (BERT) improves the model's performance (Zhou et al., 2020), even though some of the linguistic information will automatically be learned during the model training. This constitutes evidence that data annotated with syntactic information are essential for the development of NLP applications. In addition, linguists also use linguistically annotated data and computational tools to do linguistic analysis. Therefore, they also require linguistic resources.

Like other Indic languages (Bhattacharyya et al., 2019), Sinhala is also a low-resource language with a few publicly available resources. de Silva (2019) has surveyed available tools and resources in the Sinhala language and reported that no parsers or syntactically annotated treebanks are available for Sinhala. However, some Parts of Speech (POS) and Name Entity Recognition (NER) data are available. In addition, other resources like parallel corpora (Guzmán et al., 2019; Fernando et al., 2022) are also available.

This paper reports the development of the first-ever treebank with syntactic annotations for the Sinhala language. These annotations are added according to the Universal Dependencies framework.

2 The Sinhala Language

The Sinhala language is an Indo-Aryan language spoken by about 20 million people worldwide. It is one of the two official languages in Sri Lanka, spoken by 75% of its population. Tamil, Sanskrit and Pali have influenced the Sinhala language. Although Tamil is from a different language family called Dravidian, Sinhala has been in contact with it for a long time. The Portuguese, the Dutch, and the English colonized and stayed in Sri Lanka for centuries. Therefore, the influence of the languages spoken by them can be seen in Sinhala; several daily words have been borrowed from Portuguese and Dutch. Further, Sinhala has linguistic similarities with languages like Hindi, Bengali, Panjabi, and Marathi *etc.* spoken in India and Divehi, which is primarily spoken in the Maldives.

Sinhala is a diglossic language which appears in two distinct varieties: Spoken Sinhala and Written Sinhala, also known as Colloquial Sinhala and Literary Sinhala, respectively. Significant differences in these two styles are marked in all levels of the language, including lexical and syntactic levels (Gair, 1968). Sinhala is a relatively free word order language, though its unmarked word order is SOV. Different word orders are also possible with discourse–pragmatic effects (Liyanage et al., 2012). As with most Indo-Aryan languages,

Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023), pages 17 - 26

Sinhala is also an agglutinative language in which a single nominal element can be inflected for several forms to indicate the grammatical features of the case, number, gender, definiteness and animacy, and a verbal element can be conjugated for that of tense, number, gender, person, and volition (Karunatillake, 2009).

Although no work is reportedly done on developing a treebank for Sinhala, Liyanage and Wijeratne (2017) have discussed a dependency-based annotation schema for the Sinhala language, which has not proceeded to develop a treebank. Further, Prasanna (2021) has also analyzed the dependency relations of the Sinhala language from a theoretical perspective.

3 Treebank Development

In this section, we have outlined the steps we followed to create the Sinhala treebank.

3.1 Our approach

In accordance with the Universal Dependencies (UD), the treebank annotation includes lemma, POS, morphological features, and dependency relations. The sentence annotation is performed manually, with the authors serving as the primary annotators. The process of creating the annotated treebank involved the following steps.

1. Data for the annotation was selected from a Sinhala text corpus.
2. Selected data were preprocessed and tokenized.
3. An annotation guideline was developed by considering the peculiarities of Sinhala.
4. POS, Morphology, and Dependency annotations were done manually.
5. Identified issues in the annotation were reanalyzed and fixed.
6. A conversion tool specifically developed for this work was used to provide Latin transliteration for all sentences.

When designing the annotation guideline, we referred to the dependency-based annotation schema developed for the Sinhala language (Liyanage and Wijeratne, 2017) and Indian languages (Begum et al., 2008). Further, we referred to a couple of treebanks, including Hindi Treebank (HDTB) (Tandon et al., 2016), Modern Written Tamil Treebank (MWTT) (Krishnamurthy and Sarveswaran, 2021), and Marathi Treebank (UFAL) (Ravishankar, 2017).

3.2 Data Selection

The sentences for the development of the treebank were selected from the 10 million words contemporary text corpus of UCSC. This corpus contains literary or written Sinhala texts, including novels and short stories by renowned Sinhala writers. Further, it includes Sinhala translations, critiques, and texts from mainstream Sinhala newspapers such as Silumina, Dinamina, Lankadeepa, and Lakbima. Therefore, this corpus can be considered a collection of contemporary written Sinhala and thus selected as the primary source to extract and select a set of sentences.

In the sentence selection process, the first step was to categorize all the sentences in the corpus based on the number of words in each sentence. Concise entries of one to five-word entries in the corpus are mostly the newspaper headings and topics of the writings, which cannot be considered complete sentences. Further, based on a corpus study on the UCSC's 10M word corpus, Prasanna (2021) reports that the average sentence length of Sinhala sentences is 8 to 10 words, and thus in this work, we only considered the sentences with 6 to 10 words. As a first step, we selected 500 such sentences, then eliminated colloquial and erroneous sentences to filter 100 sentences to be annotated with the UD annotations.

3.3 Word Segmentation and Lemmatization

Word segmentation is a challenge in the Sinhala writing system. This has been discussed among Sinhala linguists for decades and reported in several reforms from 1959 to 2015. The issue is still not fully resolved, and writers use varying styles in their writing. For instance, according to the word segmentation reform by the Educational Publications Department of Sri Lanka (EPD, 2014), the particle ය (ya) occurs in the finite verbs should be written without any spaces. Contrarily, it should be written separately as per the reform by the National Institute of Education (NIE, 2015). Thus, the lexical entry ගිහින්ය *giyēya* is correct in accordance with the reform by EPD (2014); in contrast, it is incorrect, and ගිහින් ය *giyē ya*, the form segmented is correct according to the reform by NIE (2015). However, in accordance with the statis-

tics of the UCSC's 10 million words Sinhala text corpus, ගියේය *giyēya* shows 2,341 occurrences, whereas ගියේ ය *giyē ya* occurs for 2,666 times. Therefore, both lexical entries should be preserved and represented. Further, data for annotation were extracted from a text corpus, and it is worth keeping the original text as it occurs in the corpus. Accordingly, we did not follow any reforms and kept the sentences without tokenization.

Lemmatization in Sinhala is also challenging as the language is rich in morphology. When morpho-phonemic changes happen in words, it is tough to identify the lemma of a particular word. For instance, the Sinhala verb root කර *kara* 'do' becomes කරයි *karaji* do.non-past.3sg and කරති *karati* do.non-past.3pl, where markers suffixed to the lemma. However, when the verb becomes past the respective forms, become කළේය *kaḷēya* and කළෝය *kaḷōya* where the verb root has become කළ *kaḷa*. Therefore, the regular suffix stripping will not always work for Sinhala like in other morphologically rich Indic languages.

3.4 Sinhala Script and Transliteration

Sinhala script is an abugida or alphasyllabary script in which consonant-vowel sequences are written as single units, and the script is written from left to right. The script consists of 20 vowels and 40 consonants. Although the old Sinhala writing system uses some complex character combinations, in this research, we use only the character combinations used in the contemporary Sinhala writing system. Further, in the annotation, we followed the ISO 15919 standard to do the transliteration of text. In order to do this, we created a script[1].

3.5 Part-of-Speech Tagging

Although there are 17 tags in the Universal Parts-of-Speech (POS) tagset, we have used 13 POS tags in this treebank. There were no occurrences of INTJ (interjection), SCONJ (subordinating conjunction), SYM (symbol), and X (other) found in our data. The distribution of the POS tags in the treebank is given in Table 1.

[1]The tool is available at the https://subasa.lk/ website and can be accessed through the following URL - https://subasa.lk/services/si_en_transliteration/Real_Time_Transliteration.html

POS Label	Count	%
ADJ	50	5.7
ADP	24	2.7
ADV	36	4.1
AUX	47	5.3
CCONJ	6	0.7
DET	23	2.6
NOUN	308	35.0
NUM	4	0.5
PART	93	10.6
PRON	44	5.0
PROPN	38	4.3
PUNCT	100	11.4
VERB	107	12.2

Table 1: Distribution of POS tags in the treebank.

3.6 Morphological Features

As a morphologically rich agglutinative language, significant linguistic information are stacked in the morphology of a word in Sinhala. We have done this annotation manually in the treebank. Morphological verb features include mood, tense, aspect, voice, evident, polarity, person, and verb form. We include the morphological features of gender, number, case, definiteness, and degree for nouns. Although animacy is not a common grammatical feature in Sinhala, it can change the morphological suffix used to mark the definiteness. Therefore, we have incorporated animacy as a feature for nouns.

For adjectives, we use degree, verbForm and tense as features. Since Sinhala is a head-final language, no relative clauses occur in the language. Instead, participial forms occur in clausal modifiers, and the head of such constructions, which we treat as adjectives, were adopted features of verbform and tense. Further, the features of number, case, gender, and person were adopted for PronType.

The current version of the treebank consists of 54 unique morphological feature pairs, and the feature-value pairs that have more than 50 occurrences are tabulated in Table 2.

3.7 Syntactic Annotation

Syntactic annotations also were done manually based on the annotation guideline and the previous work. However, we faced some challenges when identifying dependency relations, which are elaborated on in the following sections. As shown in

Feature	Value	Count	%
Number	Sing	229	12.4
Gender	Neut	210	11.4
Case	Nom	175	9.5
Definite	Def	140	7.6
Case	Acc	99	5.4
AdpType	Post	94	5.1
VerbForm	Fin	68	3.7
Number	Plur	65	3.5
Mood	Ind	62	3.4
VerbForm	Part	55	3.0
Gender	Masc	54	2.9
Definite	Ind	51	2.8

Table 2: List of top morphological feature-value pairs that have more than 50 occurrences in the treebank.

Table 3, the treebank consists of 24 syntactic relations out of 37 relations that are documented in the Universal Dependencies specification. Apart from these 24 primary relations, ten sub-relations have also been identified in the data. It is interesting to note that there are more nominal subjects than the given sentences. Also, a significant number of *compound:lvc* relations are also found in the treebank. This may be due to the fact that a significant number of verbs are formed from nouns by adding a verbaliser. However, this requires more linguistic analysis. Further, there are also a significant number of *nmod* found as Sinhala. Annotation of extended dependency features will be done in the future.

3.8 Head Initial vs Head Final

Sinhala is considered a head-final language, which means that the head of a phrase or sentence appears last. However, in flat multi-word expressions, the semantic head appears first in Sinhala, whereas it comes last in English. For example, in the Sinhala phrase සුමිත් මහතා *sumit mahatā*, සුමිත් *sumit* is the semantic head and appears first, while මහතා *mahatā* appears last. In contrast, in the English equivalent "Mr. Sumith" the semantic head "Sumith" appears last, while the honorific noun "Mr." appears first. In the context of this work, the head-final approach is used for some constructions, while the head-first approach is applied specifically to flat names and complex predicates.

DEPREL Label	Count	%
nsubj	109	12.4
punct	100	11.4
root	100	11.4
dep	69	7.8
case	53	6.0
nmod	53	6.0
advmod	43	4.9
obj	42	4.8
aux	38	4.3
amod	36	4.1
compound	29	3.3
det	24	2.7
obl	24	2.7
flat	19	2.2
csubj	17	1.9
acl	16	1.8
cc	6	0.7
conj	3	0.3
cop	2	0.2
mark	2	0.2
xcomp	2	0.2
advcl	1	0.1
ccomp	1	0.1
nummod	1	0.1
compound:lvc	39	4.4
compound:svc	14	1.6
nmod:poss	11	1.3
obl:lmod	10	1.1
obl:tmod	9	1.0
compound:prt	3	0.3
advmod:emph	1	0.1
aux:pass	1	0.1
det:poss	1	0.1
nmod:tmod	1	0.1

Table 3: Distribution of dependency relations in the treebank.

Occurrence type	LVC	SVC	Com
CP Finite verbs	28	09	02
CP Gerunds	09	00	00
CP Participles	06	00	00
CP With No WS	08	02	00

Table 4: CP occurrences in the treebank.

Sentence type	Count
S with non-complex predicates	26
S with complex predicates	41
S with non-verbal predicates	33

Table 5: Types of predication in the treebank

4 Discussions

This section outlines some of the interesting syntactic constructions found in the treebank. Some of these may not be common in other languages.

4.1 Predicates in Sinhala

Many of the sentences in this treebank are with a verbal predicate. As mentioned in the distribution of sentences in Table 5, 67 sentences are with verbal predicates. However, only 26 of these are with simple verbs, whereas the rest of the 41 sentences consist of complex predicates.

4.1.1 Complex Predicates

Light verb constructions are common in Sinhala; specifically, they can be found in noun-verb, adjective–verb and particle-verb constructions. There are two verbs that function as light verbs in Sinhala: කර *kara*, the volitive indicator and වෙ *ve*, the involitive indicator. Further, similar to most South Asian Languages, Sinhala also has verb-verb compounds, which involve collocations of two verbs (Slade and Aronoff, 2020). The other type of complex predicate in Sinhala is the phrasal verb, which is formed with nouns accompanied by verbs, except for the two light verbs mentioned above. For instance, පාඩම් කරයි *pāḍam karayi* study.non-past.3sg in Figure 6 is a CP in Sinhala with a light verb construction which has developed to a complex construction පාඩම් කර ගනියි *pāḍam kara ganiyi* get-studied.non-past.3sg in Figure 7.

Some UD treebanks such as Hindi (Tandon et al., 2016) and Punjabi (Arora, 2022) use the carrier of grammatical functions, which is the second token of the compound as the head of the complex predicates. However, we treated the first token or the semantic head of the complex predicate as the head of the relation, as used by Krishnamurthy and Sarveswaran (2021), since the second token only carries the grammatical functions.

Sinhala complex predicate constructions can be divided into three categories: i) `Head` + `LVC`[2], ii) `Head` + `SVC`[3], and iii) `Head` + `Com`[4]. To differentiate from the other two, the second element of category 3 was annotated as a compound. Table 4 lists the occurrences of all three constructions found in the treebank.

Aux	Function	Example
tibe	`Aspct-perf`	*dalvā tibe* have lit
æta	`Aspct-perf`	*dalvā æta* have lit
	`Aspct-prosp`	*dalvanu æta* will be lit
næta	`Aspct-perf-neg`	*dalvā næta* have not lit
siṭi	`Aspct-prog`	*dalvamin siṭi* {be}lighting
pavati	`Aspct-prog`	*dalvamin pavati* {be}lighting
yutu	`Modal-nec`	*dælviya yutu* should be lit
hæki	`Modal-pot`	*dælviya hæki* can be lit
laba	`Pasv-NonPast`	*dalvanu laba* light
lada	`Pasv-Past`	*dalvana lada* lit

Table 6: Auxiliaries in the Sinhala language.

4.1.2 Auxiliary Verbs

The auxiliaries in Sinhala can be treated for several functions. They include aspectual (`Aspect`), modal and passive (`Pass`) auxiliaries. Further, the roles of the aspectual auxiliaries can be perfect (`perf`), progressive (`prog`) or prospective (`prosp`), and that of modal auxiliaries be either necessitative (`nec`) or potential (`pot`). Moreover, two passive auxiliaries occur for `past` and `non-past` in Sinhala. Except ලද *lada*, the passive-past auxiliary, all the other auxiliaries occur in the treebank. Auxiliaries in the Sinhala language are exemplified in Table 6 using the verb stem දැල්ව- *dalva* (light-up).

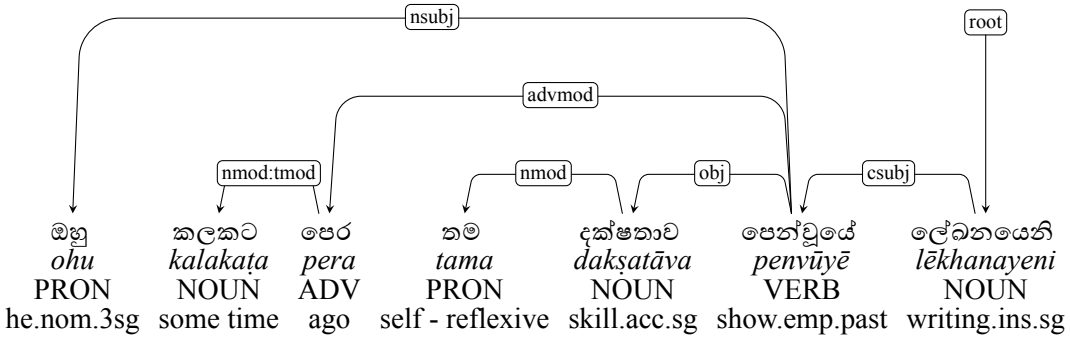
4.1.3 Non-verbal Predicates

According to Gair and Paolillo (1988), a wide range of sentences in Sinhala lacks overt verbal predication. As given in Table 5, the treebank consists of 33 sentences with non-verbal predicates.

[2]Light Verb Construction

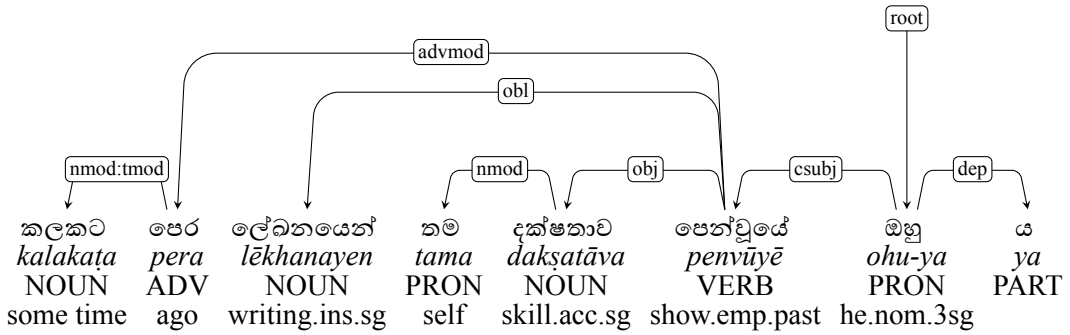
[3]Serial Verb Construction

[4]Compound



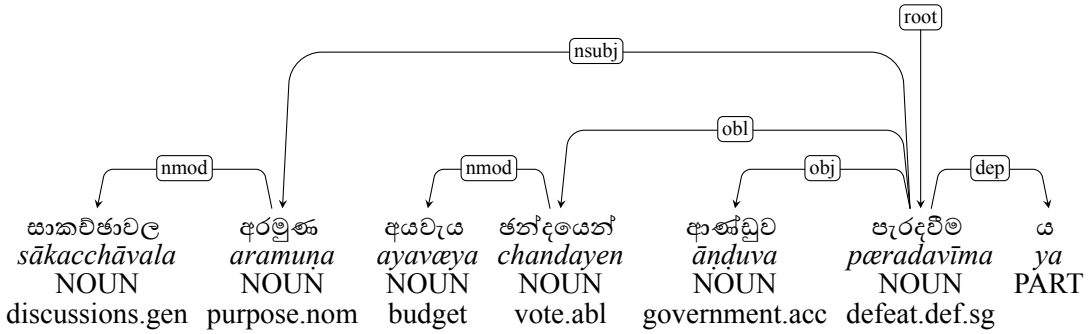
'It was through writing that he demonstrated his talent some time ago.'

Figure 1: Dependency relations in a focus construction



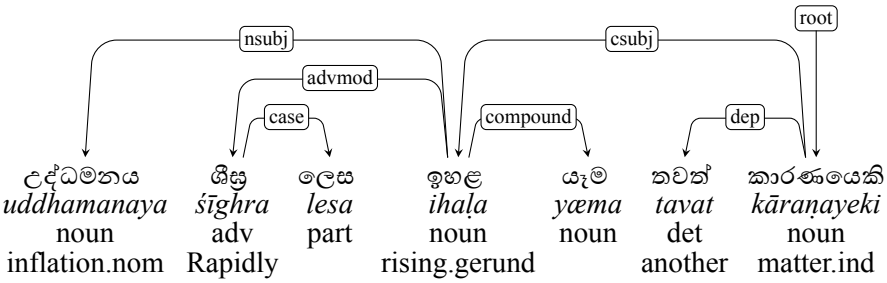
'It was he who demonstrated his talent in writing some time ago.'

Figure 2: Dependency relations with shifted emphasis for the sentence in Figure 1



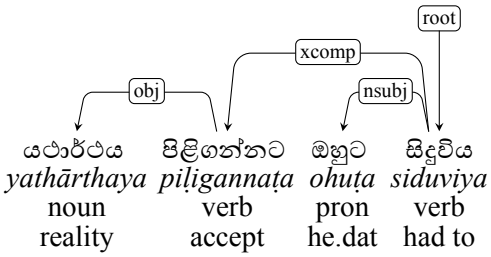
'The purpose of the discussions is to defeat the government in the budget vote.'

Figure 3: Dependency relations in a topic-comment construction



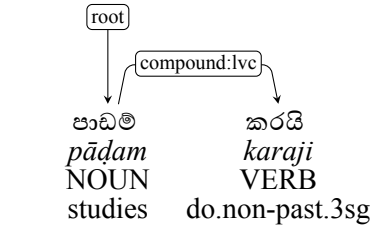
'Rapidly rising inflation is another matter.'

Figure 4: csubj in a topic-comment construction



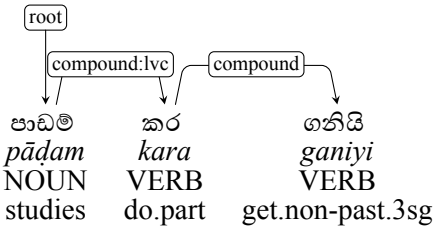
'He had to accept the reality'

Figure 5: A sentence with a clausal complement



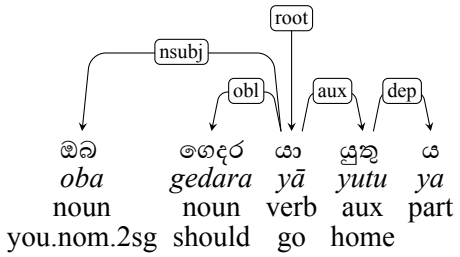
'study'

Figure 6: A Noun+LVC Construction



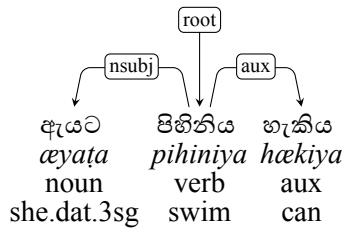
'get studied'

Figure 7: A Noun+LVC(compound) Construction



'You should go home'

Figure 8: A sentence with a modal auxiliary



'She can swim'

Figure 9: A sentence with a dative subject

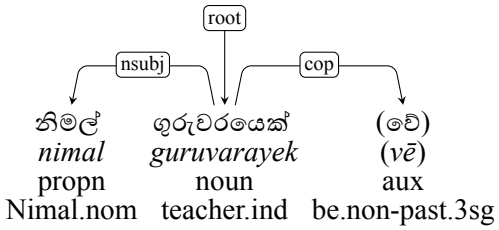
Sentences with non-verbal predicates can further be classified into the following three types based on their syntactic structure.

i. Focus Constructions: Gair and Sumangala (1991) and Slade (2011) state that there are several methods for creating focus constructions in Sinhala, one of which is the use of an emphatic form. The treebank contains numerous sentences employing this technique. Figure 1 displays a sentence where the focus is placed on a noun, which serves as the root of the sentence. The verb acts as the clausal subject and is the direct dependent of the root, with all other elements dependent on it. When a sentence is transformed into a focus construction, the main verb adopts an emphatic form (Gair and Paolillo, 1988). In the sentence of Figure 1 පෙන්ව- *penva*, the verb root has changed into the emphatic form පෙන්වූයේ *penvūyē* and has become the head of the clausal subject. The lexical item that is being focused on, which serves as the root of the sentence, is often followed by the emphatic form. Since Sinhala word order is relatively free, there are occasions where the emphasized lexical item may appear first. However, the emphatic form always depends on the emphasized lexical item. For instance, if the focus is placed on the lexical item ඔහු *ohu*, which serves as the nominal subject of the sentence in Figure 1, the sentence will transform into the sentence depicted in Figure 2, where ඔහු *ohu* is followed by an emphatic form.

ii. Copula Constructions: Sinhala is a language with zero copula; the only be verb වේ *ve:* or වෙයි *veji*, which have the same lexical root, comes in the copula position in literary Sinhala. Unlike in English, copula in Sinhala can be elided, which will not affect the syntactic structure. For instance, Figure 10 is a copula construction in Sinhala. The copula can be replaced with the sentence ending particle ය *ya* as an indication of the sentence ending. Further, Figure 11 shows a sentence with a null copula, but still, the sentence is a complete one. This particular construction is also in literary form. Interestingly, although there are no copula, a suffix 'i'[5] is used to mark the predication.

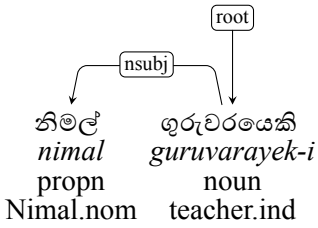
iii. Topic-Comment Constructions: In topic-comment constructions, the nominal subject depends on the nominal predicate, which is exemplified in Figure 3.

[5] 'i' marker is not discussed in the Sinhala literature. However, based on the analysis of several constructions, we concluded that 'i' marks the predication in this particular case. However, this requires more linguistic exploration.



'Nimal is a teacher.'

Figure 10: Copula construction



'Nimal is a teacher.'

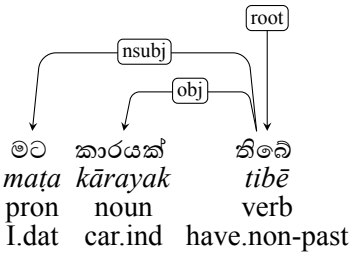
Figure 11: A zero copula construction

4.2 Core Arguments in Sinhala

As discussed in 4.1.3, nonverbal predicates are common in Sinhala; therefore, relatively more clausal subjects can be seen in the data. These clausal subjects predominantly occur in focus constructions compared to topic-comment constructions. For instance, Figure 1 is a focus construction and occurs `csubj`. However, both Figure 3 and Figure 4 are topic-comment constructions where Figure 4 consists of a `csubj` but not in Figure 3. Further, Figure 5 depicts a construction with `xcomp` along with a `nsubj`.

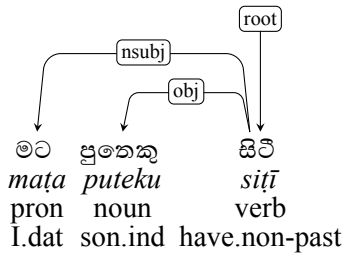
Sinhala also has non-canonical subjects with dative case marking which are referred to as dative subjects (Chandralal, 2010). According to Prasanna (2021) dative subjects can be found in a variety of sentence constructions, including involitive doers, possessive subjects, Abilitative Subjects, etc. Figure 9 illustrates the occurrence of dative subjects along with potential[6] modal verbs. In addition, dative subjects can occur in sentences with possessive verbs. Sinhala has two such possessive verbs: සිටි *siti* — with animate objects and තිබෙ *tibe* — with inanimate objects. The respective constructions are shown in Figure 12 and 13. Apart from functioning as possessive verbs, these two can also function as aspectual auxiliaries as given in Table 6.

[6]The term `potential` is borrowed from the Universal Dependencies annotation documentation - https://universaldependencies.org/u/feat/all.html#Pot



'I have a car.'

Figure 12: Dative subject with an inanimate object



'I have a son'

Figure 13: Dative subject with an animate object

5 Issues and Challenges

This section outlines some of the challenges we encountered during the linguistic analysis and annotation.

5.1 Lack of morphological feature labels

In Sinhala, ගුරුවරයෙකි *guruvarayek-i* and ගුරුවරයෙක් *guruvarayek* (teacher.Ind) refer to the same lexical element and can function as nonverbal predicates. The suffix '-i' that we have identified as the predicate marker cannot be marked with the existing features set available in the Universal Dependencies or UniMorph. Therefore, we introduced a new feature called `predicate` with the value 'yes'[7] to mark whether a word is a predicate or not.

5.2 Challenges with Dependency Annotation

The particle '*-ya*' in Literary Sinhala has been described as a predicative marker by Gair and Karuṇātilaka (1974); however, it can more accurately be identified as a sentence-ending marker. It is semantically empty but marks the end of the sentence, as shown in Figure 2 and Figure 3. When a predicate is accompanied by an auxiliary, the particle '*-ya*' can be written either together with the `AUX` or as a separate token following the `AUX`. As shown

[7]Here we followed the UD specification to define the feature `predicate` and the value 'yes'

in Figure 8 '*-ya*' that appears after the AUX must be marked as a dependent of the AUX. However, the Universal Dependencies (UD) schema does not allow auxiliaries to have children, so dependents of AUX are not permitted in the current UD specification.

6 Conclusion

We have reported the development of the first treebank for the Sinhala language, which is annotated using the Universal Dependencies framework. As a first attempt, we have annotated 100 sentences taken from a contemporary Sinhala text corpus. Apart from the data selection and the annotation process, we have also given analyses for the interesting constructions found in the data and explained how we had captured them using the current Universal Dependencies specification.

Acknowledgements

This research was made possible with the support of the UCSC research fund. The authors express their gratitude to Dr. Ruvan Weerasinghe, a senior lecturer at UCSC, for his support and encouragement in making this research successful. Authors also extend their appreciation to Prof. W.M. Wijeratne and Ms. Lakshika Madushani from the Department of Linguistics at the University of Kelaniya for their support. Lastly, the authors thank Mr. Vincent Halahakone for helping with the language corrections.

References

Aryaman Arora. 2022. Universal Dependencies for Punjabi. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5705–5711.

Rafiya Begum, Samar Husain, Arun Dhwaj, Dipti Misra Sharma, Lakshmi Bai, and Rajeev Sangal. 2008. Dependency annotation scheme for Indian languages. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Pushpak Bhattacharyya, Hema Murthy, Surangika Ranathunga, and Ranjiva Munasinghe. 2019. Indic language computing. *Communications of the ACM*, 62(11):70–75.

Dileep Chandralal. 2010. Sinhala. *Sinhala*, pages 1–312.

Nisansa de Silva. 2019. Survey on Publicly Available Sinhala Natural Language Processing Tools and Research. *arXiv preprint arXiv:1906.02358*.

EPD. 2014. *Sinhala lēkhana vyavahāraya - upadēśa samgrahaya*. Educational Publications Department, Sri Lanka.

Aloka Fernando, Surangika Ranathunga, Dilan Sachintha, Lakmali Piyarathna, and Charith Rajitha. 2022. Exploiting bilingual lexicons to improve multilingual embedding-based document and sentence alignment for low-resource languages. *Knowledge and Information Systems*.

James Gair and Lelwala Sumangala. 1991. What to focus in sinhala. In *The proceedings of the Eastern States Conference on Linguistics (ESCOL)*, volume 91, pages 93–108.

James W Gair. 1968. Sinhalese diglossia. *Anthropological Linguistics*, pages 1–15.

James W Gair and Ḍabliv Es Karuṇātilaka. 1974. *Literary Sinhala*. South Asia Program and Department of Modern Languages and Linguistics.

James W Gair and John C Paolillo. 1988. Sinhala nonverbal sentences and argument structure. *Cornell working papers in linguistics*, 8:39–77.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES Evaluation Datasets for Low-Resource Machine Translation: Nepali–English and Sinhala–English. 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 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Nizar Habash. 2007. Syntactic Preprocessing for Statistical Machine Translation. In *Proceedings of Machine Translation Summit XI: Papers*, pages 215–222, Copenhagen, Denmark. Association for Computational Linguistics.

WS Karunatillake. 2009. *Sinhala bhasha vyakaranaya*. M. D. Gunasena Co. Ltd, Sri Lanka.

Parameswari Krishnamurthy and Kengatharaiyer Sarveswaran. 2021. 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.

Junhui Li, Deyi Xiong, Zhaopeng Tu, Muhua Zhu, Min Zhang, and Guodong Zhou. 2017. Modeling Source Syntax for Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–697, Vancouver, Canada. Association for Computational Linguistics.

Chamila Liyanage, Randil Pushpananda, Dulip Lakmal Herath, and Ruvan Weerasinghe. 2012. A computational grammar of Sinhala. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 188–200. Springer.

Chamila Liyanage and WM Wijeratne. 2017. Developing a Dependency Tag Set for Sinhala: Procedure and Issues. In *Proceedings of Third International Conference on Linguistics in Sri Lanka (ICLSL 2017)*. Unviersity of Kelaniya, Sri Lanka.

Michael C McCord, J William Murdock, and Branimir K Boguraev. 2012. Deep Parsing in Watson. *IBM Journal of Research and Development*, 56(3.4):3–1.

NIE. 2015. *Sinhala lēkhana rītiya - New Edition*. National Institute of Education, Sri Lanka.

Tomoko Ohta, Yusuke Miyao, Takashi Ninomiya, Yoshimasa Tsuruoka, Akane Yakushiji, Katsuya Masuda, Jumpei Takeuchi, Kazuhiro Yoshida, Tadayoshi Hara, Jin-Dong Kim, et al. 2006. An intelligent search engine and GUI-based efficient MEDLINE search tool based on deep syntactic parsing. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 17–20.

Warahena Liyanage Chamila Prasanna. 2021. *An Exploration of Dependency Grammar for Sinhala Language*. Unpublished MPhil thesis, University of Kelaniya.

Vinit Ravishankar. 2017. A Universal Dependencies treebank for Marathi. In *Proceedings of the 16th international workshop on treebanks and linguistic theories*, pages 190–200.

Benjamin Slade and Mark Aronoff. 2020. Verb Concatenation in Asian Linguistics. In *Oxford Research Encyclopedia of Linguistics*.

Benjamin Martin Slade. 2011. *Formal and philological inquiries into the nature of interrogatives, indefinites, disjunction, and focus in Sinhala and other languages*. University of Illinois at Urbana-Champaign.

Juhi Tandon, Himani Chaudhry, Riyaz Ahmad Bhat, and Dipti Misra Sharma. 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.

Junru Zhou, Zhuosheng Zhang, Hai Zhao, and Shuailiang Zhang. 2020. LIMIT-BERT : Linguistics Informed Multi-Task BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4450–4461, Online. Association for Computational Linguistics.