Aalamaram: A Large-Scale Linguistically Annotated Treebank for the Tamil Language

A M Abirami[1*], Wei Qi Leong[2,3*], Hamsawardhini Rengarajan[2,3*], D Anitha[1], R Suganya[4], Himanshu Singh[5], Kengatharaiyer Sarveswaran[6,7], William Chandra Tjhi[2,3], Rajiv Ratn Shah[5]

[1]Thiagarajar College of Engineering, Madurai, India
{abiramiam,anithad}@tce.edu

[2]AI Singapore, Singapore
[3]National University of Singapore, Singapore
{weiqi,hamsa,wtjhi}@aisingapore.org

[4]Vellore Institute of Technology, Chennai, India
suganya.ramamoorthy@vit.ac.in

[5]Indraprastha Institute of Information Technology, Delhi, India
{himanshu17291,rajivratn}@iiitd.ac.in

[6]University of Jaffna, Sri Lanka
[7]University of Konstanz, Germany
sarves@univ.jfn.ac.lk

Abstract

Tamil is a relatively low-resource language in the field of Natural Language Processing (NLP). Recent years have seen a growth in Tamil NLP datasets in Natural Language Understanding (NLU) or Natural Language Generation (NLG) tasks, but high-quality linguistic resources remain scarce. In order to alleviate this gap in resources, this paper introduces *Aalamaram*, a treebank with rich linguistic annotations for the Tamil language. It is hitherto the largest publicly available Tamil treebank with almost 10,000 sentences from diverse sources and is annotated for the tasks of Part-of-speech (POS) tagging, Named Entity Recognition (NER), Morphological Parsing and Dependency Parsing. Close attention has also been paid to multi-word segmentation, especially in the context of Tamil clitics. Although the treebank is based largely on the Universal Dependencies (UD) specifications, significant effort has been made to adjust the annotation rules according to the idiosyncrasies and complexities of the Tamil language, thereby providing a valuable resource for linguistic research and NLP developments.

Keywords: Tamil Corpus, CoNLL-U, Annotation Guidelines, Tamil Treebank, Universal Dependencies

1. Introduction

Tamil, with a rich literary tradition spanning over two millennia, stands as one of the oldest surviving classical languages globally. Officially recognized by the Indian government as a classical language in 2004, Tamil holds significant cultural and historical importance, extending beyond being merely a means of communication (Keane, 2004). Boasting a global speaker base of approximately 89.7 million people[1], Tamil's influence is not only confined to its native regions such as India and Sri Lanka, but also extends to diaspora communities in countries like Singapore, Malaysia, Mauritius, Fiji, and South Africa[2].

However, despite the relatively large population that uses the language, the amount of data available for Natural Language Processing (NLP) in Tamil is arguably not commensurate, lagging behind major languages such as English, French, Spanish and Chinese. Although recent years have seen a growth in unannotated Tamil corpora (Kunchukuttan et al., 2020; Kakwani et al., 2020; Ramesh et al., 2021) as well as annotated data for certain benchmarking tasks in Natural Language Understanding (NLU) and Natural Language Generation (NLG) (Kakwani et al., 2020; Doddapaneni et al., 2023), datasets with rich linguistic annotations remain scarce.

Such annotated corpora, commonly known as treebanks, are important sources of data not just for linguistic research, but also for practical applications in NLP. Syntactic parse trees can be used directly in grammar checking (Li et al., 2022) and linguistic features engineered via syntactic parsers

[*] Co-first authors

[1] https://www.worlddata.info/languages/tamil.php

[2] https://www.britannica.com/topic/Tamil-language

can be used to enrich text representations and improve performance of models on downstream tasks such as machine translation (Deguchi et al., 2019; Bugliarello and Okazaki, 2020), machine reading comprehension (Zhang et al., 2020), and text summarization (Xu and Durrett, 2019; Huang et al., 2022). Moreover, although Large Language Models (LLMs) generally display strong performance in these aforementioned tasks, they still have room for improvement when it comes to understanding the correct morphosyntax of languages (Zhou et al., 2023), especially for low-resource languages such as Tamil (Leong et al., 2023), and preliminary research has shown that this gap can potentially be closed with treebanks as well (Yoshida et al., 2024). As such, it would be important to have treebanks built for the Tamil language as well in order to push the envelope of Tamil NLP systems.

As of now, there are two publicly available Tamil treebanks built under the Universal Dependencies (UD) framework (Nivre et al., 2016) – the Tamil Treebank (TTB) (Ramasamy and Žabokrtský, 2012) and the Modern Written Tamil Treebank (MWTT) (Krishnamurthy and Sarveswaran, 2021). Unfortunately, both treebanks are rather small, with a size of approximately 600 sentences each (see Table 1), which is not ideal for the training of end-to-end deep neural networks. Furthermore, these treebanks are also highly limited in data diversity, being drawn only from a single data source. This could reduce the effectiveness of models trained on them as they might not be able to generalize beyond the sentence structures and domains present in these treebanks. These treebanks also lack named entity annotations, which are important for information extraction applications.

As such, we propose *Aalamaram*[3], a large-scale treebank with almost 10,000 Tamil sentences annotated for parts-of-speech (POS), morphological features, named entities and dependency relations (see Figure 1). It is hitherto the largest publicly-available treebank for the Tamil language. *Aalamaram* is built from diverse data sources and significant efforts have been made to review and adjust the annotation rules from the UD framework and past Tamil treebanks in order to account for the idiosyncrasies and complexities of the Tamil language.

The rest of the paper is organized as follows: Section 2 presents related work. Section 3 describes the data curation process in detail. Section 4 dives into the annotation process and quality control cycle. Section 5 discusses certain linguistic phenomena that surfaced during the annotation process and which prompted reanalysis. Finally, we present our conclusions in Section 6 and put forward suggestions for future works.

2. Related Work

Although Tamil is a relatively low resource language, it is still classified as a class 3 language[4] according to Joshi et al.'s (2020) taxonomy, and this is possibly in part a result of the growth in raw Tamil text corpora for unsupervised pre-training in recent years, such as IndicNLP (Kunchukuttan et al., 2020) and IndicCorp (Kakwani et al., 2020). In addition, there have also been parallel efforts in building annotated datasets for certain tasks in NLU and NLG such as machine translation (Ram R and Devi, 2018; Siripragada et al., 2020; Ramesh et al., 2021), question answering and sentiment analysis (Doddapaneni et al., 2023).

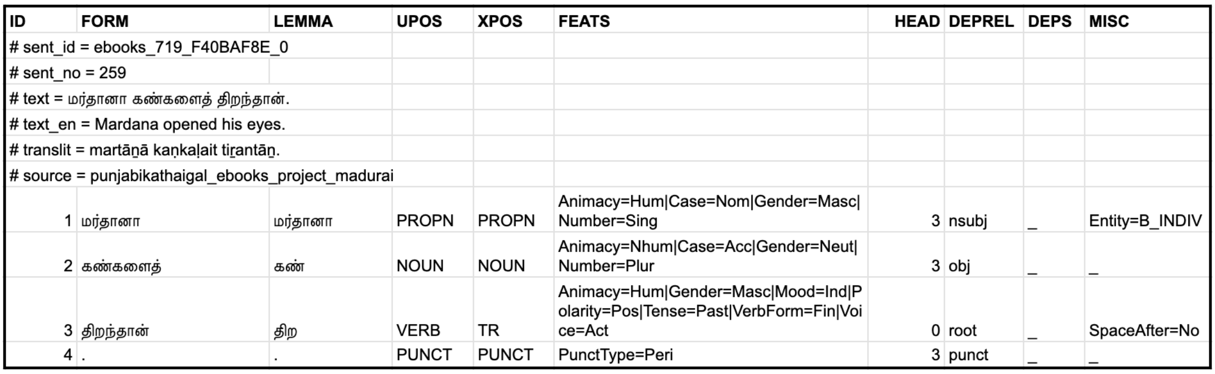
However, these datasets often lack the linguistic annotations that are essential for a granular syntactic and semantic analysis of Tamil texts. Such detailed analyses are vital in facilitating downstream NLP applications that require a nuanced understanding of the language. Currently, there have been a couple of efforts that looked at building such specialized corpora, tackling tasks such as POS tagging (Dhanalakshmi et al., 2009; Akilan and Naganathan, 2012; Chandra et al., 2014; Devi et al., 2016; Sarveswaran and Dias, 2021) and Named Entity Recognition (NER) (Pattabhi and Devi, 2013; Mhaske et al., 2023). However, there is a lack of a unified tag set for these linguistic annotations, which can make it difficult to harmonize and pool resources as well as to compare results across studies.

One promising work in unifying morphosyntactic annotations not just intra-linguistically but also cross-linguistically is the UD framework (Nivre et al., 2016). It aims to provide a linguistic representation conducive for morphosyntactic research, semantic interpretation, as well as practical NLP across diverse human languages (de Marneffe et al., 2021).

There have been to date two seminal works in applying the UD framework to the Tamil language, namely the Tamil Treebank (TTB) (Ramasamy and Žabokrtský, 2012) and the Modern

[3] *Aalamaram* (ஆலமரம்) is the Tamil word for the banyan tree, which is culturally significant to Tamilians. It is often featured in Tamil literature, folklore and proverbs, signifying its deep-rooted presence within the Tamil community. The use of the name *Aalamaram* is also a direct reference to the fact that the resource built is a treebank containing parse trees.

[4] Joshi et al. (2020) classified languages based on their existing resources into 6 categories, with class 3 languages being referred to as "rising stars" which have unsupervised pre-training data but lack labeled data collection.



ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
# sent_id = ebooks_719_F40BAF8E_0									
# sent_no = 259									
# text = மர்தானா கண்களைத் திறந்தான்.									
# text_en = Mardana opened his eyes.									
# translit = martāṉā kaṇkaḷait tiṟantāṉ.									
# source = punjabikathaigal_ebooks_project_madurai									
1	மர்தானா	மர்தானா	PROPN	PROPN	Animacy=Hum|Case=Nom|Gender=Masc|Number=Sing	3	nsubj	_	Entity=B_INDIV
2	கண்களைத்	கண்	NOUN	NOUN	Animacy=Nhum|Case=Acc|Gender=Neut|Number=Plur	3	obj	_	_
3	திறந்தான்	திற	VERB	TR	Animacy=Hum|Gender=Masc|Mood=Ind|Polarity=Pos|Tense=Past|VerbForm=Fin|Voice=Act	0	root	_	SpaceAfter=No
4	.	.	PUNCT	PUNCT	PunctType=Peri	3	punct	_	_

Figure 1: Example of an annotated sentence in *Aalamaram*

	TTB	MWTT	Aalamaram
Sentences	600	534	9567
Tokens	8635	2536	84253
Syntactic Words	9581	2584	95384
Multi-word Tokens	835	43	10211
Syntactic Word to Multi-word Token Ratio	2.13	2.12	2.09

Table 1: Comparison of Existing Tamil Treebanks with *Aalamaram*

Written Tamil Treebank (MWTT) (Krishnamurthy and Sarveswaran, 2021). TTB contains 600 sentences of news data and was initially annotated according to the Prague Dependency Treebank scheme (Hajič, 1998; Hajič et al., 2020) with 3 layers of annotations, including a morphological layer, surface syntax layer, and a tectogrammatical layer. It was then subsequently converted into the UD format. MWTT on the other hand contains 534 simple sentences sourced from Thomas Lehmann's reference grammar for the Tamil language "A Grammar of Modern Tamil" (Lehmann, 1993). MWTT was created with the intention of providing an error-free gold standard benchmark treebank for Tamil through the coverage of different sentence structures provided in the reference grammar, as it was observed that there were certain inconsistencies and errors in TTB that might have been a result of the automatic mapping from the Prague Dependency Treebank format to the UD format.

While both treebanks have been important resources given the dearth of morphosyntactically annotated datasets in Tamil, they are both relatively small and not ideal for the training of end-to-end neural networks. In fact, MWTT was also intended to be used only as a test dataset. Furthermore, they are both limited in the domains that are covered, with MWTT being drawn from a reference grammar and TTB being drawn from news only.

In addition, the highly agglutinative nature of Tamil (Lehmann, 1993; Krishnamurti, 2003; Annamalai and Steever, 2019) poses a challenge in determining the appropriate tokenization of Tamil words. A case in point would be the widespread occurrence of clitics which serve a gamut of semantic and pragmatic functions (Lehmann, 1993; Schiffman, 1999; Annamalai and Steever, 2019). These clitics are only marginally dealt with in the two existing Tamil UD treebanks, but a more in-depth treatment of the matter would be crucial in ensuring accurate analysis of Tamil texts. Both treebanks are also not annotated for named entities which are important in information extraction applications. As such, there is a need for a larger Tamil treebank with diverse data sources to support the training of deep neural networks, with named entity annotations to support NER applications, as well as a need for in-depth analysis of various linguistic phenomena in the Tamil language in order to arrive at a more accurate annotation. We therefore propose *Aalamaram* as a new treebank for the Tamil language in order to plug this gap.

3. Data Curation

As previous treebanks were relatively small and/or limited to a single source, we wanted to create a treebank that was larger in scale, with greater variety in data sources, and that also contained named entity annotations. Comparative statistical analysis of the Tamil treebanks is presented in Table 1, highlighting the growth in dataset scale in the proposed *Aalamaram* treebank. We also wanted the data to reflect real-world usage of Tamil, albeit with a focus on formal language for a start. This section describes the process of collecting and curating the data to arrive at the final set of sentences for annotation.

3.1. Data Sources

In order to enrich the diversity of texts in our dataset, we extracted data from a variety of sources:

- News - News articles written between 2021 and 2022 were scraped from Theekkathir[5], a Tamil newspaper operated by the Toiling Masses Welfare Trust Tamil Nadu. The data scraped primarily comprises formal news articles, with a predominant focus on political affairs. This data is available under a CC-BY-SA 4.0 IN license.
- Movie Reviews - Movie reviews were sourced from an existing dataset[6]. The language used in this source is not as formal as in the other sources.
- Wikipedia - Wikipedia articles were sourced from an existing dataset[7] and additional scraping of Wikipedia was done in order to enrich the representation of named entities in the dataset.
- Ebooks - Ebooks spanning publication dates from 1900 to 2021 were obtained from the Free Tamil Ebooks website[8]. These comprise mostly novels and are in the domain of fiction. These ebooks are mostly licensed under a CC-BY-NC-SA 4.0 license.
- Grammar books - Simple sentences were collected from Indian middle and high school Tamil grammar books, as they encompassed relatively simple examples that are well-crafted to demonstrate a variety of grammar rules. These sentences were only used in the initial phase for training the annotators, as well as for developing the annotation guidelines.

3.2. Data Filtering

After extracting all the data from the various sources, a series of data filtering steps were taken in order to obtain a subset that is suitable for linguistic annotation.

Although the goal for this initial work is to annotate approximately 10,000 sentences on sentence-level tasks, we initially filtered data on a paragraph level in order to obtain a set of paragraphs that could be used for paragraph-level or discourse annotations in the future. The final set of sentences were samples from this set of paragraphs. The following were the exclusion criteria that we set for removing data from the pool:

- Paragraph consists of more than 4 sentences
- 50% or more of the words in the paragraph are English
- High frequency of numerals are present in the paragraph
- Paragraphs begin with certain symbols such as , or !
- Sentences in the paragraph are shorter than 3 words or longer than 30 words

This allowed us to balance filtering out undesirable content and retaining useful data, with approximately 30% of the data being removed after these steps.

3.3. Sampling Strategy

The next step was to obtain a set of approximately 10,000 sentences from the pool of paragraphs from the filtering stage. We performed stratified random sampling by data source to obtain a corpus of 7,900 paragraphs with 30,000 sentences which can be used for future paragraph-level annotations. The target ratio of data sources (30% Wikipedia, 20% News, 20% Movie Reviews and 30% Ebooks) was decided through practical considerations of data availability as well as balance between sources.

The final set of sentences were filtered via another round of stratified random sampling with the same target ratios, with sentence segmentation performed using punctuation as boundaries. Upon inspection of the data, it was found that using punctuation for sentence segmentation may occasionally result in incomplete sentences. As such, we merged such split sentences back into a single sentence as far as possible and purged malformed ones that could not be salvaged from the dataset. This resulted in a final set of 9567 sentences available for linguistic annotation (see Table 2 for statistics).

4. Data Annotation

4.1. Annotation and Quality Control

The annotation process was divided into 3 main phases – Guideline Development Phase, Training

[5] https://theekkathir.in/

[6] https://www.kaggle.com/datasets/sudalairajkumar/tamil-nlp

[7] https://www.kaggle.com/datasets/disisbig/tamil-wikipedia-articles

[8] https://freetamilebooks.com/

Data Source	Sentences	Tokens	Syntactic Words	Multi-word Tokens	Proportion (Sentences)
News	1717	14959	17140	1954	17.95%
Movie Reviews	2191	22262	25054	2615	22.90%
Wikipedia	3098	29751	33319	3288	32.38%
Ebooks	2561	17281	19871	2354	26.77%
Total	9567	84253	95384	10211	100.00%

Table 2: Statistics of Various Data Sources in *Aalamaram*

Phase, and Annotation Phase. A total of 20 undergraduate and postgraduate students who are native speakers of Tamil and who are majoring in Data Science and Information Technology were recruited for this project. The quality control team involved 3 professors and 4 postgraduate students studying Data Science who also have Tamil as their native language and who have experience in NLP.

In the first phase, guidelines were developed with a top-down approach, using the UD guidelines as the main reference and drawing further inspiration from existing NER datasets (Sekine et al., 2002; Vijayakrishna and Sobha, 2008; Weischedel et al., 2011) and Tamil datasets with linguistic annotations. The guidelines were then further refined based on iterative linguistic analyses.

In the second phase, annotators were trained on the annotation guidelines using 200 sentences from grammar books as practice. The 20 annotators were divided into two teams of 10 (named Team 1 and Team 2). They were then further divided into 5 pairs each, one pair for each annotation task, namely POS, Lemma, Morphology, Dependency Relations, and NER. A careful learning and review process was put in place in which each member of a pair would review every annotation done by the other member. This allowed the annotators to reinforce their understanding of the guidelines and to surface challenges in regular discussions with quality controllers. Grammar book sentences were chosen for this phase as they are relatively more straightforward and helped to get annotators up to speed quickly without being bogged down unnecessarily by complicated cases. This phase also allowed us to update the guidelines based on feedback from the annotators' experiences. Inter-annotator agreement (IAA) scores based on Cohen's kappa score (Berry and Mielke, 1988) were also calculated at regular intervals to monitor the annotators' performance and the quality of the annotation.

Finally, following verification by the quality control team to ascertain the readiness of the annotators, determined through a combination of regular assessments and IAA scores, the annotators proceeded to work on the actual dataset consisting of 9567 sentences in the Annotation Phase. This phase was done without cross-reviews between members of each pair in order to speed up annotation. Team 1 and Team 2 were also not allowed to view each other's annotations to avoid inadvertent biases in annotation. 10% of the dataset selected at random was annotated by both Team 1 and 2 to allow for calculation of IAA scores. Simultaneously, this same set of sentences was also annotated by the quality control team and termed the "Gold" dataset. This allowed us to calculate the IAA between the two teams and the quality controllers in order to ascertain the accuracy of the annotations. The IAA reaches or exceeds 0.7 between both teams as well as between teams and the quality controllers (see Table 3), which indicates substantial agreement. We do not include the IAA for named entity annotation at the moment as reviews are still in progress. Furthermore, we also observed significant improvements in IAA between the initial and final stages of annotation (see Figure 2), suggesting that the quality control cycle was effective in improving the dataset quality over time. At the end of this phase, the data underwent a final quality check as well as automatic validation using the UD script[9]. This process resulted in some updates to the rules included in the UD for Tamil treebanks, showing the success of our large-scale treebank in expanding the variety of linguistic phenomena covered.

4.2. Annotation Tasks

4.2.1. Multi-word Segmentation

Given the highly agglutinative nature of Tamil, we decided to pay close attention to how words should be tokenized to best capture morpho-syntactic information in our dataset. We split auxiliary verbs and postpositions out as separate tokens, which is in line with existing work (Ramasamy and Žabokrtský, 2012; Krishnamurthy and Sarveswaran, 2021). Furthermore, we also split all clitics as listed in Lehmann (1993), which

[9] https://github.com/UniversalDependencies/tools/blob/master/validate.py

	UPOS	XPOS	HEAD	DEPREL
Team 1 vs Gold	0.8594	0.8185	0.7293	0.7003
Team 2 vs Gold	0.8748	0.8311	0.8081	0.7747
Team 1 vs Team 2	0.8342	0.7941	0.7275	0.6997

Table 3: Inter-annotator Agreement Scores for Full Dataset Annotation

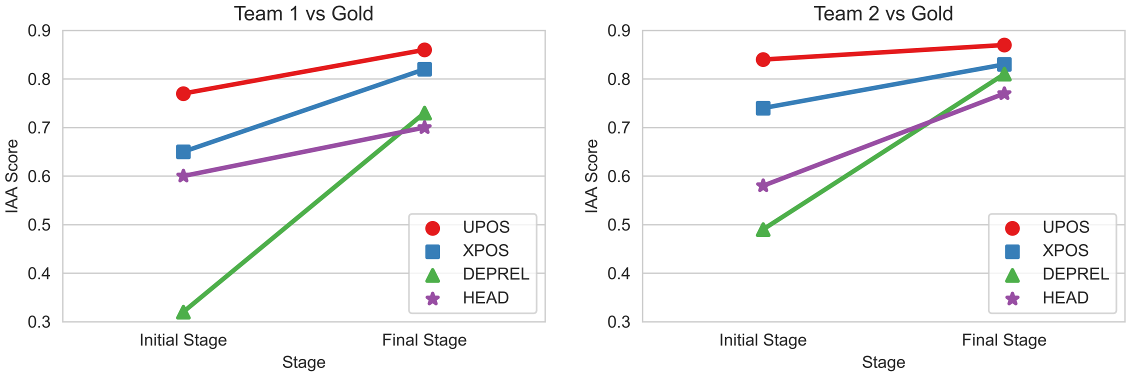


Figure 2: Improvement in Inter-annotator Agreement Scores

generally lack coverage in existing treebanks[10]. This allows us to better clarify the function of these clitics (see Section 5.2) in these sentences. On the other hand, we eschew the tokenization of case markers (as is done in TTB) and instead opt to acknowledge them under morphological feature annotations which is more in line with the UD annotation guidelines. We also do not split compound nouns.

4.2.2. POS Annotation

For POS annotations, we include both the Universal POS (UPOS) and more fine-grained language-specific (XPOS) tags. All 17 UPOS tags of the UD are used in *Aalamaram*, in contrast to TTB and MWTT which lack `SCONJ`, `INTJ` and `SYM`. This can be attributed to the scale and the coverage of *Aalamaram*. Certain words such as *eṉpatu*, and clitics such as *-um* were also re-analyzed in certain contexts as `SCONJ` (see Section 5.2), contributing to this difference.

4.2.3. Morphological Feature Annotation

The agglutinative nature of Tamil morphology makes the accurate analysis of morphological features crucial in NLP applications. As such, *Aalamaram* uses an expanded set of features compared to MWTT and TTB. One example of this expansion is in the annotation of the Animacy feature.

In MWTT and TTB, only the `Anim` label for animate nouns is used. However, nouns in Tamil have been analyzed in linguistic literature as being grouped along the axis of *rationality*[11] (Lehmann, 1993; Annamalai and Steever, 2019). Rational nouns include human-like entities such as humans, gods and demons, while irrational nouns can include both animate nouns like animals and babies as well as inanimate nouns. Rationality has a significant impact on grammar, such as in determining the inflection of nouns in certain grammatical cases or in subject-verb agreement. Although preliminary research has suggested that there can be intersections between rationality and animacy, with certain word inflections dependent on one but not the other[12], we leave exploration of this intersection to future work and tentatively use `Hum` and

[10] MWTT does not tokenize clitics and TTB only covers 4 clitics, namely *-um*, *-ē*, *-ēyē*, and *-āvatu*.

[11] This is sometimes referred to as [±human] (Krishnamurti, 2003), with rational nouns called உயர்திணை (*uyartiṇai*) and irrational nouns called அஃறிணை (*aẖṟiṇai*) in the Tolkāppiyam.

[12] For example, in the sentence *kumār ūr-ukkup pōṉāṉ* (Kumar went to a town), the inanimate noun *ūr* (town) takes the dative case marker *-ukku*. On the other hand, in a similar sentence like *kumār āppāv-iṭam pōṉāṉ* (Kumar went to father), the word *āppā* (father) has to take the locative case marker *-iṭam* instead due to it being an animate noun (Lehmann, 1993). This variation seems to be dependent on animacy and not on rationality, since the irrational animate noun *kuḻantai* (baby) takes the locative case marker *-iṭam* as well.

Nhum for the Animacy values, with Hum being used for rational nouns and Nhum for irrational nouns.

4.2.4. Dependency Relation Annotation

The dependency relations in *Aalamaram* were also annotated according to the UD guidelines, using 28 out of 37 relations, which is an expansion from the 22 used in MWTT and 25 in TTB. Significant linguistic inquiry was carried out in order to derive accurate dependency relations, especially due to the more extensive multi-word segmentation that was carried out. Some of these are explored in Section 5.

4.2.5. Named Entity Annotation

For named entities, we designed a hierarchical tagset with three levels of granularity, drawing inspiration from existing named entity hierarchies (Sekine et al., 2002; Vijayakrishna and Sobha, 2008) as well as the OntoNotes NER tagset (Weischedel et al., 2011). The first level comprises the standard ENAMEX, NUMEX and TIMEX labels, while the second and third levels comprise 14 and 35 fine-grained tags respectively. We also follow common conventions in employing the IOB2 tagging scheme.

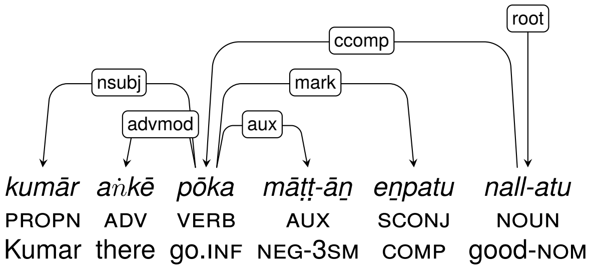
5. Discussion

This section discusses some of the linguistic phenomena in Tamil that surfaced through the annotation process and which prompted reanalysis.

5.1. *eṉpatu*

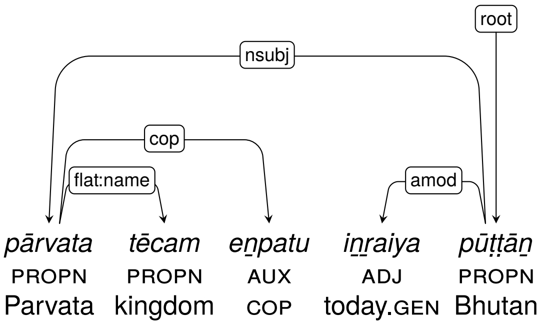
The word *eṉpatu*, which is the future verbal noun form of the verb *eṉ* (to say), has traditionally been analyzed as a complementizer (Lehmann, 1993), which would fall under the label of SCONJ (subordinating conjunction) under the UD framework, although past works in NLP have analyzed it as a particle (PART) instead (Akilan and Naganathan, 2012; Ramasamy and Žabokrtský, 2012). The annotation process also surfaced two different types of sentences containing *eṉpatu* which prompted a reanalysis of the function of *eṉpatu*.

Prima facie, the majority of sentences with *eṉpatu* seemed to involve its function as a complementizer, embedding a clause as a noun phrase (NP) that can occur in any NP position (Lehmann, 1993) (see Figure 3). While it has been proposed that *eṉpatu* in such a context can be analyzed as *eṉ-p-atu* (Lehmann, 1993) or even *eṉp-a-atu* (Butt et al., 2020), with the *-atu* suffix in both cases playing a nominalizing role, we find that more work needs to be done before this conclusion can be made and therefore opt to keep the entire word as a single token without splitting it into smaller morphemes. In such a context, we follow the UD guidelines and label *eṉpatu* as SCONJ with a dependency relation of *mark* given its complementizing function.


root
ccomp
nsubj
mark
advmod
aux
kumār aṅkē pōka māṭṭ-āṉ eṉpatu nall-atu
PROPN ADV VERB AUX SCONJ NOUN
Kumar there go.INF NEG-3SM COMP good-NOM

‘It is good that Kumar will not go there’

Figure 3: *Eṉpatu* with complementizer function


root
nsubj
cop
flat:name
amod
pārvata tēcam eṉpatu iṉṟaiya pūṭṭāṉ
PROPN PROPN AUX ADJ PROPN
Parvata kingdom COP today.GEN Bhutan

‘Parvata Kingdom is today’s Bhutan’

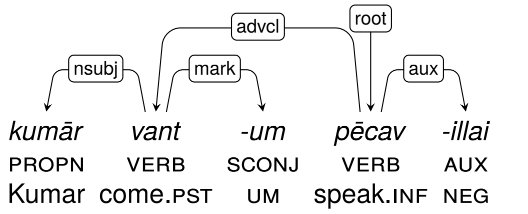
Figure 4: *Eṉpatu* with copula-like function

However, we found that there exists another group of sentences that do not seem to be featuring *eṉpatu* as a complementizer, but rather more like a copula (see Figure 4). As there is no clause with an inflected verb for *eṉpatu* to embed in such a context, it is challenging to analyze *eṉpatu* as a complementizer here. We leave potential reanalysis of this context to future work and opt to label *eṉpatu* as a copula in such contexts, which takes an AUX POS tag and *cop* dependency relation under the UD framework.

This reanalysis of *eṉpatu* as SCONJ and AUX not only clarifies the various functions of *eṉpatu* in different contexts, but is also in line with the recommendations of the UD guidelines to only use the PART label when no other label is possible.

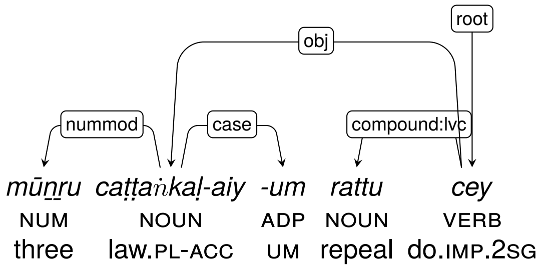
5.2. Clitics

Clitics abound in the Tamil language and serve a plethora of semantic and pragmatic functions (Lehmann, 1993; Schiffman, 1999; Annamalai and Steever, 2019). However, they have not been well studied in previous Tamil treebanking works and are often not treated as separate tokens in their own right. This presents problems in accu-


root
advcl
nsubj
mark
aux
kumār *vant* *-um* *pēcav* *-illai*
PROPN VERB SCONJ VERB AUX
Kumar come.PST UM speak.INF NEG

‘Although Kumar came, he did not speak’

Figure 5: *-um* used in a concessive sense


root
obj
nummod
case
compound:lvc
mūṉṟu *caṭṭaṅkaḷ-aiy* *-um* *rattu* *cey*
NUM NOUN ADP NOUN VERB
three law.PL-ACC UM repeal do.IMP.2SG

‘Repeal all three laws’

Figure 6: *-um* used in an all-inclusive sense

rately determining the dependency relations between words, conflating multiple syntactic and semantic functions in a single token. Therefore, as stated in Section 4.2.1, all clitics in *Aalamaram* were tokenized and rigorous analyses were done to determine their functions.

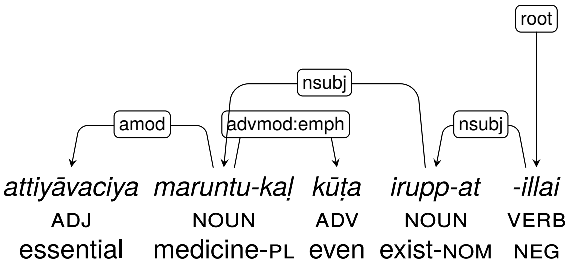
One example would be the particularly polysemous *-um* clitic which can have up to 5 functions based on the examples that we found in *Aalamaram*. In TTB, *-um* can take on a few different dependency relations such as *advmod:emph*, *cc* or *mark*, but the POS tag is always `PART`. In contrast, in *Aalamaram*, it can take on a POS tag of `CCONJ`, `SCONJ` (see Figure 5), `ADV`, `ADP` (see Figure 6) or `PART` depending on the context.

Other clitics that were also tokenized and analyzed include *-ā*, *-āvatu*, *-ām*, *-ē*, *-ō* and *-tāṉ*.

5.3. *illai*

The negative verb *illai* can express negation in both copulative and existential contexts (Lehmann, 1993). It has been suggested that the former should be labeled as `AUX` with a dependency relation of *cop*, while the latter should be labeled as `VERB` and should act as the head of the clause (Krishnamurthy and Sarveswaran, 2021). There were two other scenarios in which we found these rules to be insufficient for annotation.

The first scenario involves the use of *illai* as an auxiliary verb when used in the negative form of a main verb (see Figure 5). Such cases were not annotated in MWTT due to the lack of multi-word expansion for words ending in *illai*. A simple rule of thumb that we sought to implement was to treat *illai* as an `AUX` with a dependency relation of *aux* if it is not a standalone token, as the assumption was that the verb it is attached to should be the main verb.


root
nsubj
amod
advmod:emph
nsubj
attiyāvaciya *maruntu-kaḷ* *kūṭa* *irupp-at* *-illai*
ADJ NOUN ADV NOUN VERB
essential medicine-PL even exist-NOM NEG

‘Even essential medicines are unavailable’

Figure 7: *illai* as a main verb

However, the second scenario surfaced while implementing this rule as it was found that there are cases in which *illai* should be interpreted as the main verb when attached to a verb in the future verbal noun form (see Figure 7). While the linguistic arguments supporting this interpretation would require a more in-depth exploration, we tentatively suggest that *illai* be labelled as `VERB` in such cases.

6. Conclusion

In conclusion, we propose *Aalamaram*, the largest publicly-available dependency treebank for the Tamil language with a size of almost 10,000 sentences manually annotated for POS, morphological features, named entities and dependency relations, with close attention paid to multi-word segmentation. During the process of annotating the treebank, we also discovered various linguistic phenomena in Tamil that prompted reanalysis and adjustment of annotation rules. These include the analysis of clitics, the copula-like function of *eṉpatu*, and the interpretation of *illai* as a main verb or auxiliary. We hope that these discoveries and discussions will enable the field to get closer to a more accurate analysis of Tamil syntax, build more accurate parsers and improve Tamil NLP in general.

Moving forward, there remain certain aspects of the treebank that can be improved. Some possible improvements that can be explored are as follows:

1. More in-depth analyses of suffixes such as *-aana* and *-aaka* and whether multi-word tokenization is warranted for them

2. The use of Enhanced Dependencies[13] to handle linguistic phenomena such as ellipsis

[13] `https://universaldependencies.org/u/overview/enhanced-syntax.html`

3. Revisions to the Animacy feature to allow intersection of rationality and animacy

4. Further analysis of *illai* and *eṉpatu*

Future work would also include the training of tokenizers, POS taggers, named entity recognizers, morphological parsers and dependency parsers. This could allow us to explore the impact of various annotation decisions on model performance, such as the extensive segmentation of clitics and reanalysis of POS and dependency relations for them.

7. Acknowledgements

This work is supported by the National Research Foundation, Singapore under its AI Singapore Programme. The authors would like to thank the students of Indraprastha Institute of Information Technology, Delhi, India (IIIT-Delhi) and Thiagarajar College of Engineering, Madurai, India (TCE) who were involved in the project for their hard work in annotating the dataset. Additionally, we extend our thanks to Jian Gang Ngui for his invaluable assistance with the linguistic analysis.

8. Bibliographical References

R. Akilan and E. R. Naganathan. 2012. Pos Tagging for Classical Tamil Texts. *International Journal of Business Intelligent*, 1(2):15–17.

E. Annamalai and Sanford B. Steever. 2019. *The Dravidian Languages*, chapter Modern Tamil. Routledge.

Kenneth J. Berry and Paul W. Mielke. 1988. A Generalization of Cohen's Kappa Agreement Measure to Interval Measurement and Multiple Raters. *Educational and Psychological Measurement*, 48:921 – 933.

Emanuele Bugliarello and Naoaki Okazaki. 2020. Enhancing Machine Translation with Dependency-Aware Self-Attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1618–1627, Online. Association for Computational Linguistics.

Miriam Butt, S. Rajamathangi, and Kengatharaiyer Sarveswaran. 2020. Mixed Categories in Tamil via Complex Categories. In *Proceedings of the LFG'20 Conference*, pages 68–88, Stanford, CA. CSLI Publications.

Nitish Chandra, Sudhakar Kumawat, and Vinayak Srivastava. 2014. Various Tagsets for Indian Languages and Their Performance in Part of Speech Tagging. In *Proceedings of 5th IRF International Conference*, Chennai, India.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Hiroyuki Deguchi, Akihiro Tamura, and Takashi Ninomiya. 2019. Dependency-Based Self-Attention for Transformer NMT. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 239–246, Varna, Bulgaria. INCOMA Ltd.

Sobha Lalitha Devi, Sindhuja G., Gracy L., Padmapriya N., Gnanapriya A., and Parimala N.H. 2016. AUKBC Tamil Part-of-Speech Corpus (AUKBCTamilPOSCorpus2016v1). Chennai, India. Computational Linguistics Research Group, AU-KBC Research Centre.

V Dhanalakshmi, Anand Kumar, G Shivapratap, KP Soman, and S Rajendran. 2009. Tamil POS Tagging using Linear Programming. *International Journal of Recent Trends in Engineering*, 1(2):166–169.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards Leaving No Indic Language Behind: Building Monolingual Corpora, Benchmark and Models for Indic Languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Jan Hajič, Eduard Bejček, Jaroslava Hlavacova, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. Prague Dependency Treebank - Consolidated 1.0. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5208–5218, Marseille, France. European Language Resources Association.

Jan Hajič. 1998. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová*, pages 106–132.

Yen-Hao Huang, Hsiao-Yen Lan, and Yi-Shin Chen. 2022. Unsupervised Text Summarization of Long Documents using Dependency-based Noun Phrases and Contextual Order Arrangement. In *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)*, pages 15–24, Taipei, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Elinor Keane. 2004. Tamil. *Journal of the International Phonetic Association*, 34(1):111–116.

P. Krishnamurthy and K 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.

Bhadriraju Krishnamurti. 2003. *The Dravidian Languages*. Cambridge Language Surveys. Cambridge University Press.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul C., Avik Bhattacharyya, Mitesh Khapra, and Pratyush Kumar. 2020. AI4Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages.

Thomas Lehmann. 1993. *A Grammar of Modern Tamil*, second edition. Pondicherry Institute of Linguistics and Culture publication. Pondicherry Institute of Linguistics and Culture, Pondicherry.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. BHASA: A Holistic Southeast Asian Linguistic and Cultural Evaluation Suite for Large Language Models.

Zuchao Li, Kevin Parnow, and Hai Zhao. 2022. Incorporating rich syntax information in Grammatical Error Correction. *Information Processing Management*, 59(3):102891.

Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M. Khapra, Pratyush Kumar, Rudra Murthy, and Anoop Kunchukuttan. 2023. Naamapadam: A Large-Scale Named Entity Annotated Data for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10441–10456, Toronto, Canada. Association for Computational Linguistics.

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

R.K. Pattabhi and Sobha Lalitha Devi. 2013. NERIL: Named Entity Recognition for Indian Languages @ FIRE 2013–An Overview. In *Named-Entity Recognition Indian Languages FIRE 2013 Evaluation Track*, FIRE '13, New Delhi, India.

Vijay Sundar Ram R and Sobha Lalitha Devi. 2018. Overview of Verb Phrase Translation in Machine Translation: English to Tamil and Hindi to Tamil. In *Proceedings of the 10th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '18, page 6–10, New York, NY, USA. Association for Computing Machinery.

Loganathan Ramasamy and Zdeněk Žabokrtský. 2012. Prague Dependency Style Treebank for Tamil. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1888–1894, Istanbul, Turkey. European Language Resources Association (ELRA).

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2021. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages.

Kengatharaiyer Sarveswaran and Gihan Dias. 2021. Building a Part of Speech tagger for the Tamil Language. In *2021 International Conference on Asian Language Processing (IALP)*, pages 286–291. IEEE.

Harold F. Schiffman. 1999. *A Reference Grammar of Spoken Tamil*. Reference Grammars. Cambridge University Press.

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended Named Entity Hierarchy. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Shashank Siripragada, Jerin Philip, Vinay P. Namboodiri, and C V Jawahar. 2020. A Multilingual Parallel Corpora Collection Effort for Indian Languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3743–3751, Marseille, France. European Language Resources Association.

R Vijayakrishna and Lalitha Devi Sobha. 2008. Domain Focused Named Entity Recognizer for Tamil using Conditional Random Fields. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. OntoNotes: A Large Training Corpus for Enhanced Processing. In Joseph Olive, Caitlin Christianson, and John McCary, editors, *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*, pages 54–63. Springer New York, NY.

Jiacheng Xu and Greg Durrett. 2019. Neural Extractive Text Summarization with Syntactic Compression. 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 3292–3303, Hong Kong, China. Association for Computational Linguistics.

Ryo Yoshida, Taiga Someya, and Yohei Oseki. 2024. Tree-Planted Transformers: Large Language Models with Implicit Syntactic Supervision.

Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, and Rui Wang. 2020. SG-Net: Syntax-Guided Machine Reading Comprehension. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Houquan Zhou, Yang Hou, Zhenghua Li, Xuebin Wang, Zhefeng Wang, Xinyu Duan, and Min Zhang. 2023. How Well Do Large Language Models Understand Syntax? An Evaluation by Asking Natural Language Questions.