Towards Building a Modern Written Tamil Treebank

Parameswari Krishnamurthy
Centre for Applied Linguistics and Translation Studies
University of Hyderabad, India
pksh@uohyd.ac.in

Kengatharaiyer Sarveswaran
University of Moratuwa, Sri Lanka.
Department of Computer Science,
University of Jaffna, Sri Lanka.
sarves@univ.jfn.ac.lk

Abstract

In this paper, we describe the creation of a morphosyntactically annotated treebank for modern written Tamil following the Universal Dependencies (UD) framework to support the implementation and evaluation of Tamil dependency parsers. At present, this treebank consists of 534 sentences. This paper discusses unique constructions found in Tamil and explains sub-relations and language-specific relations introduced, apart from outlining the methodology. This carefully annotated treebank can also serve as the benchmark dataset to evaluate Tamil Natural Language Processing (NLP) tools. The treebank will be extended further to cover more complex constructions in Tamil, and annotations will be enriched by incorporating the Enhanced Universal Dependencies scheme.

1 Introduction

The paper presents a treebank for modern written Tamil following the Universal Dependencies (UD) framework called Modern Written Tamil Treebank (MWTT). The sentences in MWTT are extracted from Lehmann's *A Grammar of Modern Tamil* (Lehmann, 1989), which consists of various well-formed sentences from modern written Tamil covering different sentence structures. The first and the current release of the treebank has 534 sentences containing 2536 tokens.

Tamil is a Dravidian language spoken natively by more than 78 million people worldwide,[1] including in India, Sri Lanka, Malaysia, Singapore, and Mauritius. Despite its significant speaker population and historical time depth, Tamil is low-resourced from the perspective of Natural Language Processing (NLP) (Bhattacharyya et al., 2019). Although there have been enormous efforts in creating resources and building NLP applications for Tamil, most of them are not available for public use or obsolete/not maintained.

Neural-based approaches are the state-of-the-art when it comes to the development of NLP applications, including syntactic parsing. These approaches require a significant amount of annotated data for training. However, there are no morphosyntactically annotated data with acceptable quality and with a wide syntactic structural coverage available publicly to develop and evaluate applications. In this context, we have created the UD-based treebank carefully. This paper also gives an account of unique syntactic constructions in Tamil, which we encountered when analysing and tagging simple sentences.

2 Review of Literature

2.1 Tamil Treebanks

There are many syntactic annotation schemes that are being used to create treebanks, including PennTreebank (Marcus et al., 1993), Prague Dependency Treebank (Böhmová et al., 2001), AnnCorra (Bharati et al., 2006), and the UD (Nivre et al., 2016). There have been few attempts to create treebanks for Tamil using these schemes. However, except Loganathan's Tamil PDT (Ramasamy and Žabokrtský, 2012) and UD_Tamil_TTB,[2] others are not available for public use.

[1] http://www.languagesgulper.com/eng/Tamil.html
[2] https://github.com/UniversalDependencies/UD_Tamil-TTB/tree/master

Tamil_TTB is mapped from Tamil PDT using a script. We noticed several inconsistencies and errors in tokenisation and annotation in the Tamil_TTB treebank. For instance, some words are segmented incorrectly. Although in Tamil, nouns with dative case marking can be used to mark subjects, obliques and indirect objects, it is incorrectly used to mark objects at least in 37 instances, and indirect objects are wrongly marked as objects at least in eight instances. Further, there are inconsistencies with the usage of tags `nmod` and `obl` that are found widely in the treebank.

There is also a need to create an error-free gold standard treebank for Tamil, which can be used as a benchmark dataset, as so far, different researchers have used different datasets to validate the system they developed. The current attempt is to build such annotated dataset covering different sentence structures for modern written Tamil.

2.2 Universal Dependencies Framework

UD (Nivre et al., 2016) is a dependency framework, which proposes a morphosyntactic annotation scheme. This cross-linguistically consistent scheme has been developed by deriving existing standards on POS, morphology, and dependency annotations to facilitate multilingual research studies and parser development. The dependency relations are created between syntactic words; words that have more than one syntactic information are broken into separate tokens before the relations are established.

We identified that most of the newly created treebanks, even low-resource and morphosyntactically-rich languages, are annotated using the UD scheme; hence adopting it for Tamil is also beneficial. This allows us to create cross-lingual mapping with other languages, and to make use of tools and resources which are already built around UD.

2.3 Syntactic Parsers for Tamil

Tamil is morphologically rich and relatively free-order in nature. It has an (S)OV word order with left-branching. Several attempts have been made to develop syntactic parsers for Tamil using various formalisms. However, apart from *Thamizhi*UDp (Sarveswaran and Dias, 2020), and the other off-the-shelf parsers such as Stanza (Qi et al., 2020), UDPipe (Straka and Straková, 2017), and TranKit (Nguyen et al., 2021) others are not available publicly to use and build upon. All available parsers are implemented using state-of-the-art neural-based approaches. These approaches require more and more annotated data to improve the parsing accuracy. Further, it is also noticed that there is no well-curated benchmark dataset to evaluate and compare the accuracy of these parsers.

3 Data selection

We aim to create the UD annotated treebank consisting of different syntactic types of sentences to implement and evaluate syntactic parsers. Dataset without much noise and covering widely acceptable sentence structures in the modern written Tamil would be beneficial for such tasks. Though we initially put efforts in compiling datasets from real-time occurrences from news-papers, blogs and online platforms, they did not comprehensively cover all grammatical constructions which can be used as a representative dataset to implement and evaluate syntactic parsers. Hence, we have chosen sentences from Lehmann's *A Grammar of Modern Tamil* (Lehmann, 1989) to start with, as they cover different linguistic structures with exceptions and these sentences represent written Tamil which are even today widely accepted.

4 Methodology

The annotations of the treebank consist of POS, lemma, morphological, and dependency information in accordance with UD. We have used a step by step process given below to annotate the treebank:

1. The dataset is pre-processed and tokenised.
2. Multi-word tokens are identified and expanded. The multi-word expansion is an essential feature in UD, using which tokens are divided into multiple syntactic units; this is discussed in detail in Section 4.2.
3. The processed text are POS tagged using *Thamizhi*POSt (Sarveswaran and Dias, 2021).

4. Morphological tags are added to each token using Apertium Tamil morphological analyser (Parameswari, 2011).

5. Dependency relations are marked manually on the top of POS and morphological information.

4.1 Preprossessing and Tokenisation

We chose 534 simple sentences with mostly one clause as the first step. As the next step, we plan to extend the treebank by including complex sentences that comprise embedded clauses. We extracted the sentences and cleaned them manually. Unicode normalisation is done using the script we developed.[3] The sentences were then tokenised to separate symbols and special characters and converted to CoNLL-U annotation format.

4.2 Multi-word Token Expansion

In UD, the basic unit of annotation is syntactic words, not phonological or orthographic words.[4] When language is morphologically rich, it tends to add multiple grammatical pieces of information within a word that are morpho-syntactically relevant. There are instances where languages do add syntactic elements such as clitics, conjunctions, particles, compound verbs *etc.* within a word, which need to be split and provided with the token status for further processing. In MWTT, multi-word tokens are identified and given token status. For instance, Figure 1 shows the clitic *-um*, which expresses coordinating conjunction and it is identified as multi-word. Similarly, Figure 2 explicates the clitic *-ō* that functions as a complementiser; hence it is tokenised for its syntactic relation though morphologically manifested.

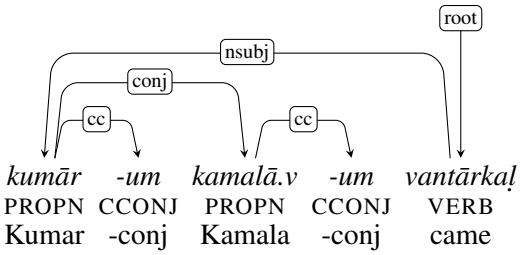


Figure 1: 'Kumar and Kamala came'

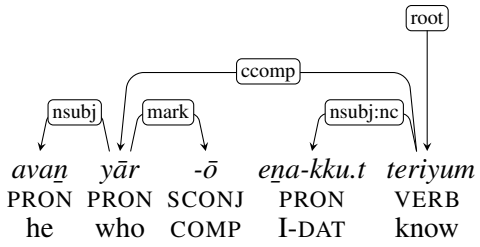


Figure 2: 'I know who is he'

Similarly, in compound verb (Verb (V1) + Verb (V2)) constructions, when V2s express aspect, mood, passive, causation and polarity, they are identified as multi-word tokens as in Figure 7. However, it is not split when the V2 is semantically bleached for its meaning and functions as an explicative verb. In the compound verb *ōṭip-pō* run.PART-go 'elope', the V2 *-pō* '*lit.* go' has partially lost its original meaning; hence it has not been split.

There are 43 multi-word tokens found in our treebank. On average, one multi-word token consists of 2.12 syntactic words.

4.3 POS Tagging

We used *Thamizhi*POSt (Sarveswaran and Dias, 2020), a contextual POS tagger to tag data with POS information. This POS tagger is trained on the text, which is not multi-word expanded. Therefore, most of the multi-word tokenised elements *i.e.* clitics, particles, compound verbs *etc* were not annotated correctly. Therefore, after the POS tagger output, we manually reviewed the POS tags.

Our treebank uses 14 POS tags out of 17 tags given in the UD POS scheme, see Table 1 for the POS tag inventory.

Table 1: POS tag frequencies of MWTT

ADJ	36	ADP	70	ADV	161	AUX	86	CCONJ	10	DET	57	NOUN	534
NUM	105	PART	2	PRON	171	PROPN	315	PUNCT	534	VERB	512	SCONJ	2

[3] https://github.com/sarves/thamizhi-preprocessor

[4] https://universaldependencies.org/u/overview/tokenization.html

Since the present version of MWTT consists only the simple sentences, the treebank does not contain any interjections, symbols and unknowns; therefore, INTJ, SYM and X were not utilised. PUNCT, NOUN, VERB and PROPN are the most frequent POS tags found with the frequency of 534(21%), 524(21%), 512(20%), and 315(12%), respectively.

4.4 Morphological Analysis

Tamil is known for its agglutinating morphology, where words are loaded with rich linguistic information. Words are morphologically analysed using Apertium Tamil Morphological Analyser (Parameswari, 2011), and then the output is mapped to UD features.[5] Nouns and pronouns are mainly analysed for their gender, number, person, case, politeness and rationality features. Adjectives are looked over for their gender, number and person details. Verbs and auxiliaries are analysed for gender, number, person, tense, aspect, mood, polarity, voice and verb form. However, since the morphological analysis is not contextual in nature, we have also reviewed the annotations.

4.5 Dependency Annotation

We annotated dependency relations according to UD schema. Around 22 relations are utilized to annotate the simple sentences out of 37 relations that are documented in UD. Apart from these main relations, although the treebank covers a simple and very limited number of syntactic constructions, it uses 17 sub-relation types that provide language-specific syntactic information separating by a colon (:) (see Table 2 for the top 5 sub-relations). All these annotations were done manually. Some of these syntactic analyses require in-depth linguistic inquiry. We have given an initial account of these constructions and issues in Section 5. This dataset is evaluated with Tamil UDPipe parser[6] and *Thamizhi*UDp. While UDPipe which is trained on Tamil_TTB treebank provides a Labeled Attachment Score (LAS) of 27.19, *Thamizhi*UDp which is trained using news data provides a LAS of 83.31.

5 Dependency Relations in Tamil

This section presents the discussion on predicates, subjects, oblique relations, compounds, coordinations, and other Tamil specific relations.

5.1 Predicates in Tamil

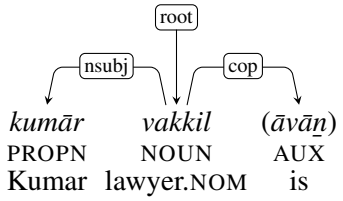
Sentences in contemporary written Tamil is constructed most commonly with verbal predicates. Complex predicates in Tamil consist of (i) verb+verb construction where a series of verbs can be added periphrastically to the first verb which is either in the form of verbal participle or infinitive forms to express aspect, mood, passive, causation, negative polarity and attitude (see Figure 7), (ii) noun+verb construction where a noun functions as the head and a verb as the light verb (see Figure 12 (cf. (Lehmann, 1989; Rajendran, 2004; Sarveswaran and Butt, 2019)).

However, in copular constructions, nominal predicates are identified as `root`. The copula verb *āku* '*lit.* to become' optionally occurs in nominal predicates. Figure 3 is an example of nominal predicate where the nominative case marked noun is identified as `root` and the copula verb is attached to it with the relation `cop`. The negative copula verb *illai* appears obligatorily to express constitution negation as in Figure 4. When the verb *illai* 'not' occurs as an existential negation, it is considered as *root* as seen in Figure 5. There are instances where the nominal predicate is dative-case marked to express benefaction as in Figure 6 and the copula is absent.

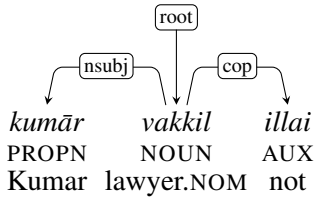
[5] https://universaldependencies.org/u/feat/index.html
[6] http://lindat.mff.cuni.cz/services/udpipe/

Table 2: Sub-relations or Language-specific relations used in MWTT

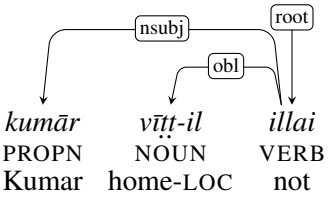
Relation	Description	Count
`nsubj:nc`	Non-canonical nominal subject	47
`obl:tmod`	Temporal modifier – oblique	45
`nmod:poss`	Nominal modifier – possessive	28
`compound:lvc`	Light verb	18
`obl:lmod`	Locative modifier – oblique	16



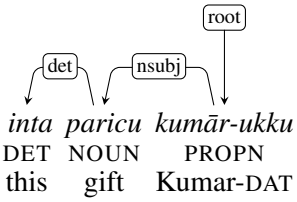
'Kumar is a lawyer'
Figure 3: Nominal Predicate



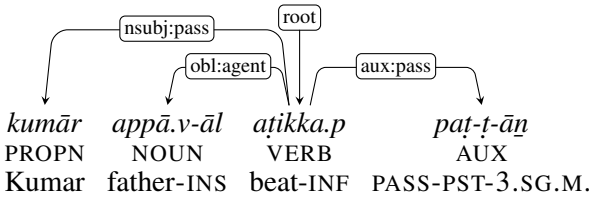
'Kumar is not a lawyer'
Figure 4: *illai* as Copula



'Kumar is not at home'
Figure 5: *illai* as Predicate



'This gift is for Kumar'
Figure 6: Dative as Predicate



'Kumar was beaten by (his) father'
Figure 7: Subject in Passive Construction

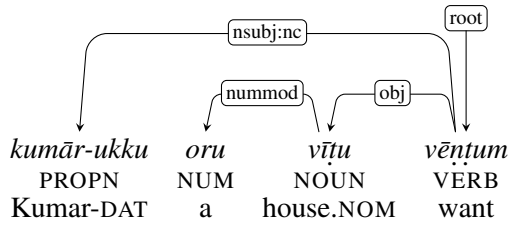
5.2 Subjects in Tamil

In Tamil, most commonly, a nominative case marked noun phrase functions as a subject and controls verb agreement. However, in the passive construction, the nominal subject with the nominative case marker (not a proto-agent) is identified as *nsubj:pass* which controls the verb agreement and the agent as *obl:agent* (see Figure 7).

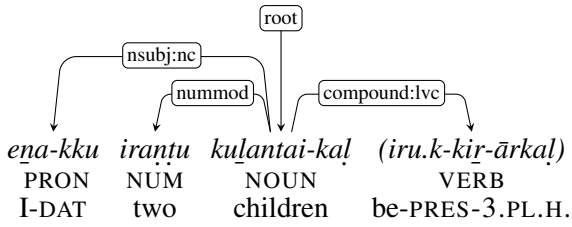
Subjects in Tamil are also realised with non-nominative markers such as the dative, the instrumental and the locative case markers. There are several discussions and diagnostic tests of Non-Nominative Subjects (NNS) (Siguresson, 2004; Subbārāo, 2012), which include NNSs that can occur as antecedents to anaphors and NNSs as controllers of PRO.

In NNS, the dative subject construction is the most widespread in Dravidian languages (Subbārāo, 2012) and are called experiencer subjects. Verma and Mohanan (1990) describes "in the so-called experiencer subject constructions in South Asian languages, the thematically prominent argument, which we expect to be a grammatical subject, is quite often an experiencer, and is marked with the case otherwise associated with indirect objects". In Tamil, stative predicates expressing the notion of mental, emotional and physical experience require the case-marking pattern of DAT-ACC (Lehmann, 1989; Pappuswamy, 2005). The tag *nsubj:nc* is used to mark non-canonical subjects. Dative subjects can also occur to express need or necessity, but the object is not explicitly marked for the accusative marker in Tamil as in Figure 8.

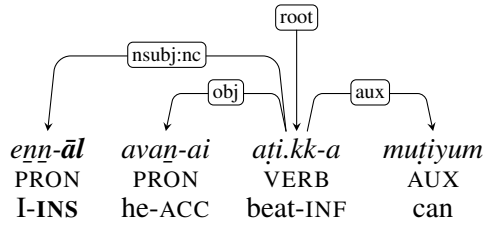
While expressing inalienable possession and kinship, the subject is marked with the dative in Tamil. The verb *iru* 'to be' is used as a possessive verb ('to have') and occurs optionally. In such constructions, the possessed noun is marked as `root` and the verb *iru* as a light verb i.e. `compound:lvc` as seen in Figure 9. The subject is marked with the locative case marker to show the temporary or alienable possession, and the existential verb *iru* 'to be' is used as the possessive verb (see Figure 11). When the predicate expresses capabilitative mood in Tamil, the subject is marked for the instrumental case marker, see Figure 10. It is also seen that the theme is marked for the accusative though the subject is NNS.



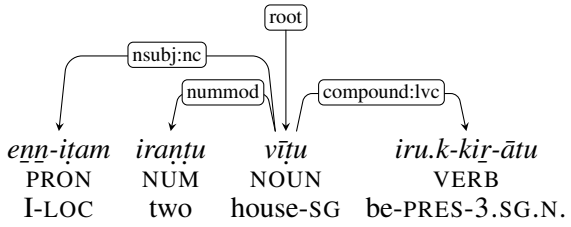
'Kumar wants a house'
Figure 8: Dative Subject-1



'I have two children'
Figure 9: Dative Subject-2



'I can beat him'
Figure 10: Instrumental Subject



'I have two houses'
Figure 11: Locative Subject

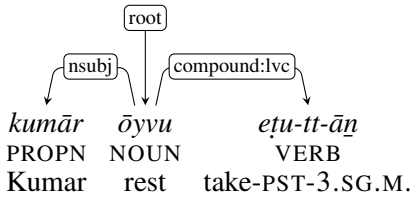
5.3 Oblique Cases

Non-core arguments are grouped under oblique cases in UD (Nivre et al., 2016). Language-specific oblique tags to differentiate locative (`obl:loc`), instrumental (`obl:inst`), ablative (`obl:abl`), sociative (`obl:soc`), place (`obl:pmod`) and temporal modifiers (`obl:tmod`), comparatives (`obl:cmpr`), agents in passive (`obl:agent`) are introduced in our treebank.

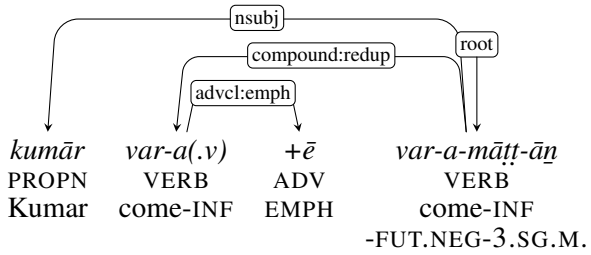
5.4 Compound

The relation `compound` is majorly marked for noun-noun compounds in UD. In Tamil, a language-specific tag `compound:lvc` is adopted for light-verb constructions where the noun occurs in juxta-position to verb and carries the semantic content. The same tag is also used in Telugu treebank (Rama and Vajjala, 2018), however in our case, the noun is marked as `root`, and the light-verb is marked as `compound:lvc` as in Figure 12 following the practice in UD standard, whereas they are seen in the reverse direction in Telugu.

Reduplication and echo-word formation are other linguistic processes that are commonly found in many South-Asian languages (Subbārāo, 2012). They provide emphasis or distributive meaning. In Tamil, verbs (see Figure 13), nouns, determiners, adjectives and adverbs can be reduplicated. They are marked with the relation `compound:redup`.



'Kumar took rest'
Figure 12: `compound:lvc`



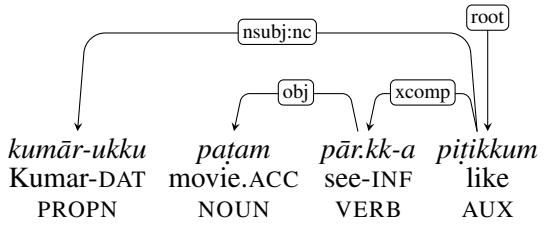
'Kumar won't come (with emphasis)'
Figure 13: `compound:redup`

5.5 Coordination

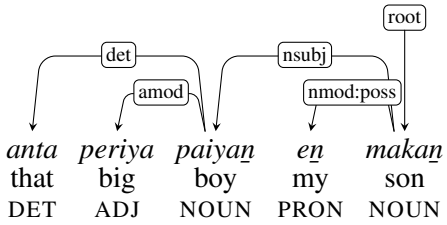
In Tamil, most commonly the clitic *-um* 'and' (see Figure 1), *-ō* 'or', *-āvatu* 'either.. or' are added as suffixes to each word or phrase or clause which are coordinated. The free morphemes *maṛṛum* 'and', *allatu* 'or', and *āṇāl* 'but' are also used, similar to English coordinators. The relation `conj` is used to conjoin them with the head-first approach in compliance with UD guidelines.

5.6 Other Relations

An open clausal complement (`xcomp` in UD) is found in Tamil as explicated in Figure 14, where the subject in the higher clause behaves as a subject to the subordinate predicate *pār* 'to see'. The relation `nmod:poss` is marked on possessive noun which is either realised in genitive case marker or oblique stem as in Figure 15. Auxiliaries are distinguished with relations `aux, aux:neg` and `aux:pass` in Tamil as negative and passive information are encoded as auxiliaries. Similarly, other relations such as `acl:relcl` is marked for relative clauses, `advcl` is marked for adverbial clauses and `advcl:cond` is for conditional adverbial clauses. Modifier relations such as `advmod`, `nummod`, `nmod` and `amod` are utilised in the treebank. To capture the emphasis to the meaning of any constituent, either the emphatic clitic *-ē* as a bound morpheme or the free morpheme *tāṉ* is used and identified as `advmod:emph`.



'Kumar likes to see movies'
Figure 14: The relation `xcomp`



'That big boy is my son'
Figure 15: The relation *nmod:poss*

6 Conclusion

In this paper, we have reported the creation of a Modern Written Tamil Treebank (MWTT) according to the Universal Dependencies framework.[7] We followed a hybrid approach and used an existing POS tagger and a morphological analyser to reduce manual annotation. We have also highlighted different syntactic constructions of simple sentences found in our corpus and how those are captured using the Universal Dependencies formalism. This treebank is useful as a benchmark dataset to evaluate syntactic parsers and other NLP tools such as POS taggers and morphological analysers.

As part of the future work, we will extend the resource by adding other complex syntactic constructions found in Lehmann's grammar book and other Tamil grammar books that we can access. Further, the Enhanced Universal Dependencies (EUD) scheme will also be incorporated to capture deep syntactic information.

Acknowledgements

We want to thank unknown reviewers for the valuable input, which shaped the final paper well. We extend our thanks to Keerthama B for involving in the annotation process. Further, we would like to thank Dan Zeaman for his continuous technical support in publishing MWTT.

References

Akshar Bharati, Rajeev Sangal, Dipti Misra Sharma, and Lakshmi Bai. 2006. Anncorra: Annotating corpora guidelines for POS and Chunk annotation for Indian languages. *LTRC-TR31*, pages 1–38.

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

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2001. The Prague Dependency Treebank: Three-Level Annotation Scenario. In Anne Abeillé, editor, *Treebanks: Building and Using Syntactically Annotated Corpora*, pages 103–127. Kluwer Academic Publishers.

Thomas Lehmann. 1989. *A Grammar of Modern Tamil*. Pondicherry Institute of Linguistics and Culture, India.

[7] https://github.com/UniversalDependencies/UD_Tamil-MWTT/tree/master

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Minh Van Nguyen, Viet Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A Light-Weight Transformer-based Toolkit for Multilingual Natural Language Processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, and Natalia Silveira. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666. European Language Resources Association.

Umarani Pappuswamy. 2005. Dative Subjects in Tamil: A Computational Analysis. *South Asian Language Review*, XV(2):40–62.

K Parameswari. 2011. An implementation of APERTIUM morphological analyzer and generator for Tamil. *Parsing in Indian Languages*, pages 41–44.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

S Rajendran. 2004. Strategies in the formation of compound nouns in Tamil. *Languages of India*, 4.

Taraka Rama and Sowmya Vajjala. 2018. A Dependency Treebank for Telugu. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pages 119–128, Prague, Czech Republic. Association for Computational Linguistics.

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) .

Kengatharaiyer Sarveswaran and Miriam Butt. 2019. Computational Challenges with Tamil Complex Predicates. In Miriam Butt, Tracy Holloway King, and Ida Toivonen, editors, *Proceedings of the LFG19 Conference, Australian National University*, pages 272–292, Stanford. CSLI Publications.

Kengatharaiyer Sarveswaran and Gihan Dias. 2020. ThamizhiUDp: A Dependency Parser for Tamil. In *Proceedings of the 17th International Conference on Natural Language Processing*, pages 200–207, Indian Institute of Technology Patna, India. NLP Association of India.

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

HA Sigurdsson. 2004. Icelandic non-nominative subjects. In Bhaskararao, P. and Subbarao, K.V., editor, *Typological Studies in Language*, chapter 7, pages 137–159. John Benjamins Publishing Company.

Milan Straka and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Kārumūri V Subbārāo. 2012. *South Asian Languages: A Syntactic Typology*. Cambridge University Press.

Manindra K Verma and KP Mohanan. 1990. Introduction to the Experiencer Subject Construction. In Manindra K Verma and KP Mohanan, editors, *Experiencer subjects in South Asian Languages*, chapter 1, pages 1–12. Center for the Study of Language and Information (CSLI), Stanford, CA.