Tamil Dependency Parser

Universität Konstanz

Dependency Parser

A **dependency parser** is a computer program that generates linguistic information, including dependency relations, for a sentence. This linguistic information-rich output is often referred to as a treebank.

Tamil language

- A classical language with over two millennia of literary tradition.
- Spoken by 80+ million people globally.
- Diglossic language. Regional variations in spoken Tamil.
- Low-resource language with limited computational resources and inadequate linguistic studies.
- Morphologically rich, flexible word order, Southern Dravidian language.

அழிந்துக்கொண்டிருந்தேன் - *aḻintukkoṇṭiruntēṉ*
(I) was being destroyed

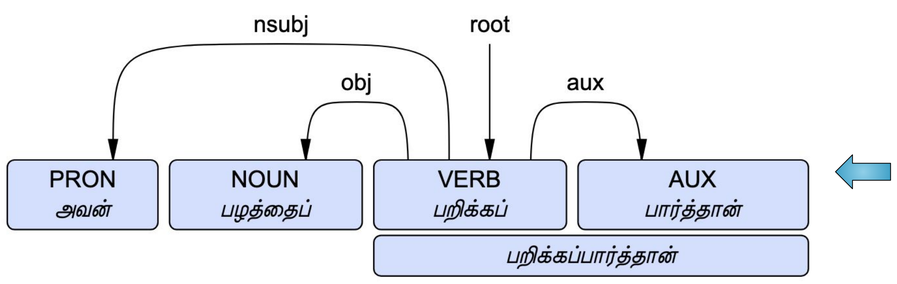
அழி	ந்து	கொண்டு	இரு	ந்த்	ஏன்
aḻi	*ntu*	*koṇṭu*	*iru*	*nt*	*ēn*
LEMMA destroy	TRANSITIVITY intransitive	ASPECT continuous	ASPECT continuous	TENSE past	PNG first person, singular

Universal Dependencies

- Used for 245 treebanks in 141 languages.
- Valuable for cross-linguistic analysis and computational model development.
- No large Tamil treebanks available.
- Uses CoNLL-U format for Part of Speech, Lemma, Morphology, and Dependency annotation with evolving guidelines.

அவன்	பழத்தைப்	பறிக்கப்பார்த்தான்.
avaṉ	*paḻattaip*	*paṟikkappārttāṉ*
He.NOM	Fruit.ACC	pick.try.PAST.3SgM

"He tried to pick the fruit."



Why Dependency Treebanks Matter

- **Part of Speech (POS), Morphology, and Syntax** are vital for understanding a language and its structure.
- **Treebank** is a computational resource containing sentences with the linguistic information mentioned above, as well as lemmas and named entities.
- Treebanks are crucial for **developing computer applications** like machine translation and question-answering systems, as well as for **linguistic analysis**.
- **Dependency treebanks** excel in handling complex word order and rich morphology for precise syntactic analysis.
- These are especially useful for languages like **Tamil**.

Methodology

Preprocessing, Multiwords

- Collected from online news sources and textbooks.
- Cleaned and Unicode-normalised text.
- Identified and tokenised multiwords.

Part of Speech tagger[1]

- Part of Speech (POS) tagger labels words with their grammatical category, such as nouns or verbs.
- Trained a tagger to tag text using the Universal POS.
- The Deep Learning tagger achieves 93.27% accuracy.

Morphological tagger[2,3,4]

- Morphological tagger labels words with their grammatical category, such as nouns or verbs.
- Conducted an in-depth Tamil morphology study.
- Built a Finite-State morphological analyzer.

Dependency parser[5]

- Multilingual learning was used to bootstrap annotation.
- Manually corrected dependency annotations.
- Trained a Deep Learning-based tagger, achieving a Labelled Attachment Score of 75.53%.

CoNLL-U scheme

```
#text = அவன் பழத்தைப் பறிக்கப்பார்த்தான்.
1     அவன்          அவன்    PRON   _   Case=Nom|Gender=Masc|Number=Sing|Person=3   3   nsubj   _ _
2     பழத்தைப்      பழம்    NOUN   _   Case=Acc|Gender=Neut|Number=Sing|Person=3   3   obj     _ _
3-4   பறிக்கப்பார்த்தான்    _      _   _                                           _   _       _ _
3     பறிக்கப்      பறி     VERB   _   VerbForm=Inf                               0   root    _ _
4     பார்த்தான்    பார்    AUX    _   Gender=Masc|Number=Sing|Person=3|Tense=Past 3   aux     _ _
5     .             .       PUNCT  _   PunctType=Peri                             3   punct   _ _
ID    Token         Lemma   POS    _   Morphological information                  Dependency
```


Selected References

1. Sarveswaran, K., & Dias, G. (2021). Building a Part of Speech tagger for the Tamil Language. In *2021 International Conference on Asian Language Processing (IALP)* (pp. 286-291). IEEE.
2. Sarveswaran, K., Dias, G., & Butt, M. (2021). Thamizhi Morph: A morphological parser for the Tamil language. *Machine Translation*, *35*(1), 37-70.
3. Sarveswaran, K., & Butt, M. (2020). Computational challenges with Tamil complex predicates. In *The 2019 Conference on Lexical Functional Grammar: LFG'19* (pp. 272-292).
4. Sarveswaran, K., Dias, G., & Butt, M. (2019). Using Meta-Morph Rules to develop Morphological Analysers: A case study concerning Tamil. *Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing*, 76–86. doi:10.18653/v1/W19-3111
5. Sarveswaran, K., & Dias, G. (2020). T*hamizhi*UDp: A Dependency Parser for Tamil. *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, 200–207.

Next steps:

How can the parser's performance be improved?
Can Lexical Functional Grammar be useful?

Kengatharaiyer Sarveswaran
University of Jaffna, Sri Lanka.
Herz Fellow, University of Konstanz, Germany.
sarves@univ.jfn.ac.lk

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
Baden-Württemberg
MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST
Zukunftskolleg

Tamil Dependency Parser

Universität Konstanz

Dependency Parser

A **dependency parser** is a computer program that generates linguistic information, including dependency relations, for a sentence. This linguistic information-rich output is often referred to as a treebank.

Tamil language

- A classical language with over two millennia of literary tradition.
- Spoken by 80+ million people globally.
- Diglossic language. Regional variations in spoken Tamil.
- Low-resource language with limited computational resources and inadequate linguistic studies.
- Morphologically rich, flexible word order, Southern Dravidian language.

அழிந்துக்கொண்டிருந்தேன் - *aḻintukkoṇṭiruntēṉ*
(I) was being destroyed

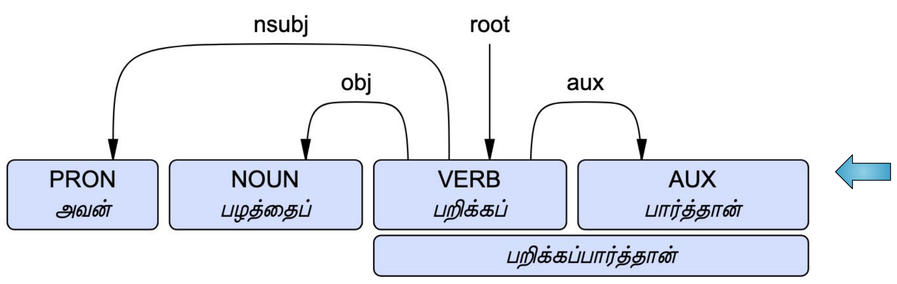
அழி	ந்து	கொண்டு	இரு	ந்த்	ஏன்
aḻi	*ntu*	*koṇṭu*	*iru*	*nt*	*ēn*
LEMMA destroy	TRANSITIVITY intransitive	ASPECT continuous	ASPECT continuous	TENSE past	PNG first person, singular

Universal Dependencies

- Used for 245 treebanks in 141 languages.
- Valuable for cross-linguistic analysis and computational model development.
- No large Tamil treebanks available.
- Uses CoNLL-U format for Part of Speech, Lemma, Morphology, and Dependency annotation with evolving guidelines.

அவன்	பழத்தைப்	பறிக்கப்பார்த்தான்.
avaṉ	*paḻattaip*	*paṟikkappārttāṉ*
He.NOM	Fruit.ACC	pick.try.PAST.3SgM

"He tried to pick the fruit."



Why Dependency Treebanks Matter

- **Part of Speech (POS), Morphology, and Syntax** are vital for understanding a language and its structure.
- **Treebank** is a computational resource containing sentences with the linguistic information mentioned above, as well as lemmas and named entities.
- Treebanks are crucial for **developing computer applications** like machine translation and question-answering systems, as well as for **linguistic analysis**.
- **Dependency treebanks** excel in handling complex word order and rich morphology for precise syntactic analysis.
- These are especially useful for languages like **Tamil**.

Methodology

Preprocessing, Multiwords

- Collected from online news sources and textbooks.
- Cleaned and Unicode-normalised text.
- Identified and tokenised multiwords.

Part of Speech tagger

- Part of Speech (POS) tagger labels words with their grammatical category, such as nouns or verbs.
- Trained a tagger to tag text using the Universal POS.

Morphological tagger

- Morphological tagger labels words with their grammatical category, such as nouns or verbs.
- Conducted an in-depth Tamil morphology study.
- Built a Finite-State morphological analyzer.

Syntactic parser

- Multilingual learning was used to bootstrap annotation.
- Manually corrected dependency annotations.
- Trained a Deep Learning-based tagger

CoNLL-U scheme

#text = அவன் பழத்தைப் பறிக்கப்பார்த்தான்.

ID	Token	Lemma	POS		Morphological information	Dependency		
1	அவன்	அவன்	PRON	_	Case=Nom|Gender=Masc|Number=Sing|Person=3	3	nsubj	_ _
2	பழத்தைப்	பழம்	NOUN	_	Case=Acc|Gender=Neut|Number=Sing|Person=3	3	obj	_ _
3-4	பறிக்கப்பார்த்தான்		_	_	_	_	_	_ _
3	பறிக்கப்	பறி	VERB	_	VerbForm=Inf	0	root	_ _
4	பார்த்தான்	பார்	AUX	_	Gender=Masc|Number=Sing|Person=3|Tense=Past	3	aux	_ _
5	.	.	PUNCT	_	PunctType=Peri	3	punct	_ _

Next steps:

How can the parser's performance be improved?
Can Lexical Functional Grammar be useful?

Selected References

1. Sarveswaran, K., & Dias, G. (2021). Building a Part of Speech tagger for the Tamil Language. In *2021 International Conference on Asian Language Processing (IALP)* (pp. 286-291). IEEE.
2. Sarveswaran, K., Dias, G., & Butt, M. (2021). Thamizhi Morph: A morphological parser for the Tamil language. *Machine Translation*, *35*(1), 37-70.
3. Sarveswaran, K., & Butt, M. (2020). Computational challenges with Tamil complex predicates. In *The 2019 Conference on Lexical Functional Grammar: LFG'19* (pp. 272-292).
4. Sarveswaran, K., Dias, G., & Butt, M. (2019). Using Meta-Morph Rules to develop Morphological Analysers: A case study concerning Tamil. *Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing*, 76–86. doi:10.18653/v1/W19-3111
5. Sarveswaran, K., & Dias, G. (2020). T*hamizhi*UDp: A Dependency Parser for Tamil. *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, 200–207.

Kengatharaiyer Sarveswaran
University of Jaffna, Sri Lanka.
Herz Fellow, University of Konstanz, Germany.
sarves@univ.jfn.ac.lk

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
Baden-Württemberg
MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST
Zukunftskolleg