A DEEP SYNTACTIC PARSER
FOR THE TAMIL LANGUAGE

Kengatharaiyer Sarveswaran

178097E

Degree of Doctor of Philosophy

Department of Computer Science & Engineering

University of Moratuwa
Sri Lanka

September 2022

A DEEP SYNTACTIC PARSER

FOR THE TAMIL LANGUAGE

Kengatharaiyer Sarveswaran

178097E

Thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Department of Computer Science & Engineering

University of Moratuwa
Sri Lanka

September 2022

DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: UOM Verified Signature Date: 20/09/2022

The above candidate has carried out research for the PhD thesis under our supervision.

Name: Professor Gihan Dias, University of Moratuwa, Sri Lanka.

Signature of the Supervisor: UOM Verified Signature

Name: Professor Miriam Butt, University of Konstanz, Germany.

Signature of the Supervisor: UOM Verified Signature Date: 20/09/2022

ABSTRACT

A Deep Syntactic Parser for the Tamil Language

Natural Language Processing (NLP) applications have become integral to human life. A syntactic parser is a vital linguistic tool that shows syntactic relations between the words in a sentence. These may then be mapped to a tree, a graph, or a formal structure. Syntactic parsers are helpful for building other NLP applications. In addition, they help linguists to understand a language better and perform cross-lingual linguistic analysis. A syntactic parser that performs a deeper analysis and captures argumentative, attributive and coordinative relations between the words of a given sentence is called a deep syntactic parser. Tamil is considered a low-resourced language in terms of tools, applications, and resources available for others to use and build NLP applications or carry out linguistic analyses. Not many resources, such as treebanks and annotated corpora, or linguistic analysis tools such as POS taggers or morphological analysers, are publicly available for Tamil. Available off-the-shelf language-agnostic syntactic parsers show comparatively low performance because of the rich morphosyntactic properties of Tamil. This study elaborates on how I developed the first grammar-driven parser for Tamil, which uses the Lexical-Functional Grammar formalism, and a state-of-the-art data-driven parser using the Universal Dependencies framework. I have also proposed an approach to evaluate a syntactic parser's syntactical coverage, experimented with transition-based and graph-based approaches, and for the first time, tried multi-lingual training to develop a data-driven parser for Tamil. A part of speech tagger, a morphological analyser cum generator, pre-processing tools, and treebanks are the other tools and resources I have developed to facilitate the development of the parsers. While all these tools give the current best score for their respective tasks, these resources are also available online for others to build upon. Moreover, the study also documents my contributions toward understanding different linguistic aspects of the Tamil language.

Keywords: Deep Syntactic Parser; Grammar-driven parser; Data-driven parser; Part of Speech tagger; Morphological Analyser

DEDICATION

அப்பா - அம்மா

appā - ammā

'Father - Mother'

for their unconditional love, support, and being the reason of who I am today.

பெரியப்பா - பெரியம்மா

periyappā - periyammā

'Uncle - Aunt'

for being my guardians when crossing the most important part of my life.

இயற்கை

iyaṟkai

'the great Nature'

(the god)

for always putting together and aligning things I required for the progressions of this study.

ACKNOWLEDGEMENTS

I am very grateful to the people who have supported me in various ways since the beginning of this study in 2018 to complete this study.

First, I must thank my two perfect supervisors, Professor Gihan Dias and Professor Miriam Butt. I would not have completed this study without their tremendous support. Their detailed and insightful comments have significantly improved my research work. I am additionally thankful to them for creating valuable opportunities to widen my knowledge and academic network.

I thank Professor Gihan Dias for encouraging me to start my PhD research and offering me the wonderful opportunity to be his doctoral student. He always knew the right moment when to lend me a word of encouragement and give me a push. I am grateful for that and for the valuable guidance he has provided to complete this study. Professor Gihan Dias also supported me with the funding to carry out research, attend conferences, and organise academic events.

I thank Professor Miriam Butt, my other supervisor, for helping me with the computational linguistics part of my study. She hosted me twice as a visiting researcher at the University of Konstanz and created opportunities to meet other leading scholars in this field of study. Professor Miriam also provided me with financial support to be able to attend conferences and supported me in co-organising workshops and a summer school. I am grateful for all these and the trust she kept in me.

I want to express my gratitude to the progress evaluation panel members, Professor Subathini Ramesh and Dr Charith Chitraranjan, and research coordinators Professor Sanath Jeyasena and Dr Shehan Fernando, for their continuous input and guidance in tailoring my research.

I am grateful to the following thesis evaluation panel members for their constructive comments and feedback to improve the thesis and plan my future work:

- Professor Jagath Premachandra - Professor at the University of Moratuwa, Sri Lanka.
- Professor Mary Dalrymple - Professor of Syntax at the University of Oxford, United Kingdom.
- Professor Sarmad Hussain - Professor of Computer Science, University of Engineering and Technology, Pakistan.

- Dr Uthayasanker Thayasivam - Senior lecturer in Computer Science at the University of Moratuwa, Sri Lanka.
- Dr Amal Shehan Perera - Senior lecturer in Computer Science at the University of Moratuwa, Sri Lanka.

I should thank my friend Dr Maris Camilleri for helping me with language editing and providing feedback on my write-up. I know, at times, she sacrificed her personal and family time to do these. I am very grateful to her in this respect.

I would like to thank my colleagues at the Department of Computer Science, the University of Jaffna for encouraging me to start the PhD research, and the administration of the University of Jaffna for providing me with the required permission and leave to carry out the research.

I am also thankful to the wonderful staff and researchers of the National Languages Processing centre, and the head and the staff of the Department of Computer Science and Engineering at the University of Moratuwa for their support, input, and the provision of all required resources. Moreover, I am grateful to them for making my time at the University of Moratuwa memorable and valuable. I would also like to thank the director and the staff Faculty of Graduate Studies for facilitating all administrative matters related to the registration and evaluation of my study.

I also thank my friends in the Computational Linguistics group at the University of Konstanz for making my stay at Konstanz productive and memorable. I would especially like to thank Jessica Zipf for her support in solving issues related to XLE.

I am thankful to Paul Meurer at the University of Bergen for providing the required permission and helping to set up the Tamil grammar on INESS. I am grateful to Dr Dan Zeman, Charles University, the Czech Republic, for helping me sort out technical issues and set up a Tamil Universal Dependencies treebank.

I would like to thank Professor Lauri Karttunen from Stanford University for introducing me Foma (a Finite-State compiler), which cleared one of my greatest obstacles in the development of the morphological analyser, and Dr Mans Hulden from the University of Colorado Boulder, for helping me sort out issues related to Foma.

I thank Dr Parameswari Krishnamurthy from the University of Hyderabad, for working with me to create a Tamil treebank that I then used as a part of my test set. I also would like to thank Professor Rajendran Sankaravelayuthan, Amrita Vishwa Vidyapeetham, Dr S. Rajamathangi, Jawaharlal Nehru University, and Ms Sujiththa Srikantharajah, University of Jaffna for helping me with reviewing linguistic annotations.

I owe a lot to my wife, Sharanyaa and daughters, Thenmozhi and Poongkuzhalii, for their patience and love. It would have been much easier for them if I had continued working at the University of Jaffna, but they encouraged me to go ahead with my studies. Our little daughter Nilavezhilii is the best thing that happened to us during

this challenging time. I know they made sacrifices to help me stay focused on my studies. I also thank my parents and brothers for supporting me in everything with selfless generosity.

காலத்தி னாற்செய்த நன்றி சிறிதெனினும்
ஞாலத்தின் மாணப் பெரிது. - திருக்குறள் (102)

kālatti ṉāṟceyta naṉṟi ciṟiteṉiṉum
ñālattiṉ māṇap peritu. - tirukkuṟaḷ (102)

“A favour conferred in the time of need,
though it be small (in itself),
is (in value) much larger than the world.”

Thank you!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	**Description**
1S	First person – Singular
1PLE	First person - Plural - Epicene
3SN	Third person – Singular – Neuter
3SM	Third person – Singular – Masculine
3SF	Third person – Singular – Feminine
ABL	Ablative
ACC	Accusative
ADJ	Adjective
ADJL	Adjectivalizer
ADV	Adverb
ADVL	Adverbializer
AGR	Agreement
AP	Adjectival Participle
ARG	Argument
ASCII	American Standard Code for Information Interchange
ASS	Associative
AUX	Auxiliary
AVM	Attribute-Value-Matrix
BEN	Benefactive
BILSTM	Bidirectional Long Short-Term Memory
BLEX	Bi-LEXical dependency score
CAR	Cardinal
CAUS	Cause Marker
COMP	Complementiser
COND	Conditional
CONJ	Conjunction
COP	Copula
CRF	Conditional Random Field

DAT	Dative
DEC	Declarative
DIST	Distal
DOM	Differential Object Marking
DUR	Durative
EMPH	Emphatic
EPI	Epicene
F	Feminine
FEATS	Features
FST	Finite-State Transducer
FUT	Future Tense
GEN	Genitive
GEND	Gender
HDT	Hamburg Dependency Treebank
HON	Honorific
HON	Double Honorific
HORT	Hortative
IMP	Imperative
INCL	Inclusive
INESS	Infrastructure for the Exploration of Syntax and Semantics
INF	Infinitive
INS	Instrumental
IRRAT	Irrational
INTJ	Interjection
IOBJ	Indirect Object
LOC	Locative
LAS	Labelled Attachment Score
LFG	Lexical Functional Gramma
LV	Light Verb
M	Masculine Analyzer
MA	Morphological Azad
MLAS	Morphology-aware Labeled Attachment Score
MWTT	Modern Written Tamil Treebank
N	Neuter
NEG	Negative
NER	Named Entity Recognizer
NLP	Natural Language Processing
NLTK	Natural Language ToolKit
NMLZ	Nominaliser

NOM	Nominative
NP	Noun Phrase
NTYPE	Noun Type
NUM	Number
OBJ	Object
OBL	Oblique
ORD	Ordinal
PART	Particle
PASS	Passive
PERM	Permissive
PERS	Person
PL	Plural
POS	Parts of Speech
POSS	Possessive
PP	Postposition Phrase
PRED	Predicate
PRES	Present tense
PROG	Progressive
PRON	Pronoun
PRS	Present Tense
PSP	Postposition
PST	Past Tense
QUOT	Quotative
RAT	Rational
REL	Relativiser
RNN	Rrecurrent Neural Network
SAN	Sandhi
SEM	Semantic
SER	Singular - Epicene - Rational
SG	Singular
SUBJ	Subject
SVC	Serial Verb Construction
SYM	Symbol
TB	Treebank
TNS-ASP	Tense-Aspect
TTB	Tamil TreeBank
UAS	Unlabelled Attachment Score
UD	The Universal Dependencies
UPOS	Universal Part of Speech

VP	Verbal Pharse
VPART	Adverbial Participle
VTYPE	Verb type
XCOMP	Non-finite clause argument
XLE	Xerox Linguistic Engine
XPOS	Language-specific Part of Speech

TRANSLITERATION SCHEMA

Vowels		Consonants	
அ	a	க்	k
ஆ	ā	ங்	ṅ
இ	i	ச்	c
ஈ	ī	ஞ்	ñ
உ	u	ட்	ṭ
ஊ	ū	ண்	ṇ
எ	e	த்	t
ஏ	ē	ந்	n
ஐ	ai	ப்	p
ஒ	o	ம்	m
ஓ	ō	ய்	y
ஔ	au	ர்	r
		ல்	l
		வ்	v
		ழ்	ḻ
		ள்	ḷ
		ற்	ṟ
		ன்	ṉ

Note: Composite characters are formed by adding consonants and vowels together. For instance, Tamil letter க is transliterated as *ka* as க = க் (k)+ அ(a) In this way there are 216 composite Tamil letters are formed by composing 18 consonants with 12 vowels, and the composite letters will be transliterated accordingly.