DEVELOPMENT OF A CORPUS AND A PARSER FOR WRITTEN MALAYSIAN TAMIL

ELANTTAMIL MARUTHAI

FACULTY OF LANGUAGES AND LINGUISTICS
UNIVERSITI MALAYA
KUALA LUMPUR

2022

DEVELOPMENT OF A CORPUS AND A PARSER FOR WRITTEN MALAYSIAN TAMIL

ELANTTAMIL MARUTHAI

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF LANGUAGES AND LINGUISTICS
UNIVERSITI MALAYA
KUALA LUMPUR

2022

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: **Elanttamil Maruthai**

Registration/Matric No: **17030063/3**

Name of Degree: **Doctor of Philosophy**

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

DEVELOPMENT OF A CORPUS AND A PARSER FOR WRITTEN MALAYSIAN TAMIL

Field of Study:

Corpus linguistics

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date: 31/3/2022

Subscribed and solemnly declared before,

Witness's Signature Date: 31/3/2022

Name:

Designation:

DEVELOPMENT OF A CORPUS AND A PARSER FOR WRITTEN MALAYSIAN TAMIL

ABSTRACT

Tamil is a classical language with ancient heritage, existing without any break or interruption in its long history. Over the years, the grammar and lexicon of Tamil have undergone changes. Some old linguistic features have disappeared while some new features have emerged. All these require serious investigation, especially given the current developments in corpus and computational linguistic research. It is therefore surprising that limited research has taken advantage of these technological advancements in studying Tamil in general and Malaysian Tamil in particular.

The present thesis aims to address this concern by developing the first corpus of Written Malaysian Tamil (WMTC). Based on this WMTC, the Tamil language in Malaysia can be analysed authentically. The thesis describes the creation and development of this WMTC, which consists of one million words, spanning eight different genres, with a total of 500 samples of text, each containing about 2000 words, comprising texts from periodicals, popular magazine, Internet, school textbooks, fiction, academy journal, to be spoken and unclassified category.

Since Tamil is a morphologically rich language, the present thesis further develops an automatic Tamil unified parser and POS tagger software. This thesis discusses the development of this software, which consists of a morphological parser, a tagger, an N-gram tool and a concordancer for the linguistic analysis of Tamil. Given the scope of research, this thesis focuses only on the morphological analysis of written Tamil as an illustration of how the software can be used to analyse the morphological aspects of written Malaysian Tamil and possibly other varieties of Tamil.

One innovation of this parser is the introduction of Tamil computational algorithm into the parser, which makes the analysis and processing of morphological features possible. 51 POS tags were developed for this research project. In addition, the noun and verb inflection charts explaining the computational morphotactics of Tamil words were developed along with lists of tokens, types and lemmas.

This thesis makes two major contributions to corpus and computational linguistic research: the creation and development of a corpus and a parser. All this is paving the way for future research in language technology, natural language processing, corpus development and computational linguistic research. This current research also has important implications for Tamil language pedagogy and language planning.

PEMBANGUNAN KORPUS DAN PARSER UNTUK PENULISAN BAHASA TAMIL MALAYSIA

ABSTRAK

Bahasa Tamil merupakan satu bahasa klasik yang memiliki warisan berzaman dan wujud secara berterusan tanpa putus atau sebarang gangguan dalam sejarah kewujudnya yang begitu lama. Nahu dan leksikon Bahasa Tamil telah melalui perubahan sepanjang tempoh ini. Beberapa ciri linguistik lama telah hilang dan ada pula beberapa ciri baru yang muncul. Perkara ini perlu disiasat dengan serius, terutamanya dengan adanya perkembangan semasa dalam bidang linguistik korpus dan komputasi. Adalah mengejutkan bahawa hanya sebilangan kajian yang terhad sahaja mengekploitasi kemajuan teknologi seperti ini dalam mendalami pengetahuan Bahasa Tamil umumnya dan Bahasa Tamil Malaysia, khususnya.

Tesis ini bertujuan menangani masalah ini melalui pembangunan korpus pertama Bahasa Tamil Malaysia bertulis atau Written Malaysian Tamil Corpus (WTMC). Berdasarkan WTMC ini, analisis autentik Bahasa Tamil yang digunakan di Malaysia boleh dilakukan. Tesis ini menerangkan pembentukan dan pembangunan WTMC ini yang mengandungi satu juta perkataan dan merangkumi lapan genre dari sejumlah 500 teks yang menganduing 2000 patah perkataan setiap satu. Teks ini termasuk dari terbitan berkala, majalah popular, Internet, buku teks sekolah, fiksyen, jurnal akademik hingga ke kategori perucapan dan kategori tidak diklasifikasi.

Disebabkan Tamil adalah Bahasa yang kaya dari segi morfologi, tesis ini membangunkan sebuah perisian penanda/tagger POS dan “unified parser” automatik. Pembangunan perisian ini yang mengandungi “parser” morfologi, satu penanda atau tagger, satu alat N-gram dan satu alat konkordansi khusus untuk analisis Bahasa Tamil juga akan dibincang dalam tesis ini. Memandangkan skop kajian ini, hanya analisis morfologi Bahasa Tamil bertulis yang menjadi tumpuan dalam tesis bagi meggambarkan potensi penggunaan perisian untuk menganalisis aspek morfologi Bahasa Tamil Malaysia serta variasi lain Bahasa Tamil.

Satu innovasi penting dalam “parser” ini dapat dilihat dari segi pengenalan algoritma komputasi Bahasa Tamil ke dalam parser tersebut. Ini menjadikan analisis dan pemprosesan fitur morfologi dapat dilakukan. Sebanyak 51 penanda POS juga dibangunkan untuk projek ini. Tambahan lagi, carta infleksi kata nama dan kata kerja yang menerangkan morfotaktik komputasi perkataan Tamil juga dibentuk berserta dengan senarai token, jenis token dan lemma.

Dua sumbangan utama dari tesis ini terhadap bidang kajian linguistik korpus dan komputasi termasuk: pembentukan dan pembangunan satu korpus dan satu parser. Kesemua ini membantu dalam penerokaan kajian dalam bidang teknologi Bahasa, pemprosesan Bahasa natural, pembangunan korpus, dan komputasi. Kajian ini juga mempunyai implikasi penting terhadap pedagogi dan perancangan Bahasa Tamil.

ACKNOWLEDGEMENTS

First and foremost, I am incredibly grateful to my esteemed supervisor, Dr Chau Meng Huat, for his invaluable advice, insightful comments and suggestions, continuous support, and patience during my PhD. study. His immense corpus linguistics knowledge and bountiful experience have encouraged me throughout my academic research. I would like to thank Dr. Suad (retired) for her guidance during the preliminary study.

I would like to thank my colleagues from the Tamil unit for their kindness, comprehension, help, and support that have made my study and work-life a wonderful time. My gratitude extends to the faculty of languages and linguistics for the opportunity to undertake my studies in corpus linguistics at the Universiti of Malaya.

I would like to thank Dr.N. Deivasundaram, Dr.A. Gopal and Dr K. Karunakaran for all their support. I wish to thank Mr Saravanan and Mr. Muhelan for being my backbone during software development. I would like to thank my secondary school teachers and university lecturers for their academic guidance. Vivekananda (PJ) Tamil school teachers (1978-1983) play a vital role in my educational journey. They shaped who I am today, and I would like to record a special thanks to them.

Finally, I would like to express my gratitude to my parents, brothers, sisters, nephews, and nieces. Without them, this would not have been possible. I also appreciate all the support I received from the rest of my family. I am grateful to my wife for her tremendous understanding and encouragement over the past years; it would be impossible for me to complete my study without her support and patience. Finally, I appreciate my friends for their encouragement and support throughout my studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

Eng. Abbr	**Eng. Exp.**
acc.	accusative
ADJ	Adjective
ADJ-N(PL-C)	Adjectival Noun (Plural-Common)
ADJ-N(PL-NH)	Adjectival Noun (Plural-Non-Human)
ADJ-N(SG-C)	Adjectival Noun (Singular-Common)
ADJ-N(SG-F)	Adjectival Noun (Singular-Feminine)
ADJ-N(SG-M)	Adjectival Noun (Singular-Masculine)
ADJ-N(SG-NH)	Adjectival Noun (Singular-Non-Human)
ADV	Adverb
AN	Appellative Noun
ANC	American National Corpus
Asp	Aspectual
Aux	Auxiliary
BMELC	Business and Management English Language Learner Corpus
BNC	British National Corpus
C	Case
CALES	Corpus-Based Archive of Learner English in Sarawak
CIIL	Central Institute of Indian Languages
CL	Clitic
CLAWS	Constituent Likelihood Automatic Word-tagging System
COBUILD	Collins-Birmingham University International Lexical Database
COCA	Corpus of Contemporary American English
COMEL	Corpus of Malaysian English

CONJ	Conjunctive Suffix
CS	Causative Suffix
DBP	Dewan Bahasa dan Pustaka
DPN	Demonstrative Pronoun
EMAS	English of Malaysian Schools Students
EMILLE	Enabling Minority Language Engineering
FV	Finite Verb
ICE-Malaysia	International Corpus of English-Malaysian Component
ICLE	The International Corpus of Learner English
INFITT	International Forum for Information Technology in Tamil
INT	Intensifier
INTJ, IJ	Interjection
IPN	Interrogative Pronoun
LCMC	Lancaster Corpus of Mandarin Chinese
LGSWE	Longman Grammar of Spoken and Written English
MACLE	Malaysian Corpus of Learner English
MAG	Morphological Analyzer and Generator
MALEX	MaLay LEXicon
MCWT	Malaysian Contemporary Written Tamil
MEC	Malaysian Corpus of English
MPGC	Malay Practical Grammar Corpus
MYCanCor	Malaysia Cantonese Corpus
N- H-PL	Noun-Human-Plural
N- NH-PL	Noun-Non-Human-Plural
N-C	Noun-Case
NEG	Negative

N-H-C	Noun-Human-Case
N-NH-C	Noun-Non-Human-Case
N-P	Noun-Postposition
NPL	Noun Plural
N-PL-C	Noun-Plural-Case
N-PL-P	Noun-Plural-Postposition
N-SG/PL-P	Noun-Singular/Plural-Postposition
OCR	Optical Character Recognizer
OPN	Oblique Pronoun
OPT	Optative
PAR-N	Participial Noun
PAR-N(PL-C)	Participial Noun (Plural-Common)
PAR-N(PL-NH)	Participial Noun (Plural-Non-Human)
PAR-N(SG-C)	Participial Noun (Singular-Common)
PAR-N(SG-F)	Participial Noun (Singular-Feminine)
PAR-N(SG-M)	Participial Noun (Singular-Masculine)
PAR-N(SG-NH)	Participial Noun (Singular-Non-Human)
Part.N	Participial Noun
PF	Prefix
PF-NEG	Negative Prefix
PL	Plural
PN	Pronoun
PN(F-PL)	Person Number (First person -Plural)
PN(F-SG)	Person Number (First person -Singular)
PN(S-SG)	Person Number (Second person -Singular)
PN(S-SG/PL)	Person Number (Second person -Singular/Plural)

PNG	Person Number Gender
PNG(T-PL)	Person Number Gender (Third person -Plural)
PNG(T-PL-NH)	Person Number Gender (Third person -Plural-Non-Human)
PNG(T-SG-CO)	Person Number Gender (Third person -Singular-Common)
PNG(T-SG-F)	Person Number Gender (Third person -Singular-Feminine)
PNG(T-SG-M)	Person Number Gender (Third person -Singular-Masculine)
PNG(T-SG-NH)	Person Number Gender (Third person -Singular-Non-Human)
POS	Part-Of-Speech
RP	Relative Participle
RP-DN	Defective Negative Relative Participle
RP-NEG	Negative Relative Participle
RPS	Relative Participle Suffix
RPS-DNEG	Defective Negative Relative Participle Suffix
RPS-NEG	Negative Relative Participle Suffix
SADV	Sentential Adverb
TDIL	Technological Development of Indian Languages
TNS-FU	Future Tense
TNS-PR	Present Tense
TNS-PT	Past Tense
USM	University Sains Malaysia
VA	Aspectual Auxiliary Verb
VF	Finite Verb
V-F-A	Appellative Finite Verb
V-F-D	Defective Finite Verb
V-IN	Infinitive Verb
VM	Modal Auxiliary Verb

VN	Verbal Noun
VN-FU	Future Verbal Noun
VN-PR	Present Verbal Noun
VN-PT	Past Verbal Noun
V-PAR	Verbal Particle
VP-CONC	Concessive Verbal Participle
VP-COND	Conditional Verbal Participle
VP-CONJ	Conjunctive Verbal Participle
VP-CONS	Consecutive Verbal Participle
VP-INF	Infinitive Verbal Participle
VP-NEG	Negative Verbal Participle
VP-SIM	Simple Verbal Participle
VP-SIMU	Simultaneous Verbal Participle
VS	Verb Stem
VV	Verb Voice
V-VP	Verbal Past Participle
WC	Word Class
WMT	Written Malaysian Tamil
XML	Extensible Markup Language

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.0 Introduction

This chapter introduces some background information of this research. It surveys such issues as the status of Tamil and the need for a corpus for Malaysian Tamil. The chapter also presents research objectives and research questions of this thesis. Finally, the scope and significance of the research are discussed.

1.1 Background to the Study

1.1.1 The Status of Tamil

The Tamil language has a history of more than two thousand years (Samuel, 1994; Selby, 2019; Steever, 2018; Zvelebil, 1960, 1973, 1992). It is one of the longest-surviving classical languages in the world (Mann et al., 2019; Zvelebil, 1974) and it is the only language of contemporary India that is recognizably continuous with a classical past (Pon, 1993). It has existed in the Tamil speech community without any break, and it has a vast literary and grammatical heritage (Thirumalai, 2004). In addition to its continuity throughout history, the Tamil language has incessantly been enriching its linguistic features, particularly its lexicon and grammar, to serve the current and emerging communicative needs of the Tamil speech community. It is an official language in Tamil Nadu, Sri Lanka, and Singapore (Shapiro & Schiffman, 2019). It is also one of the four main languages (Winskel, 2020) of instruction in the Malaysian schools and is a medium of instruction in 527 Tamil primary schools nationwide in Malaysia.

The Tamil language used in different countries varies because the language used by a speech community is dependent on various non-linguistic variables such as age, education, profession, domain, and medium. These factors lead to differences in language use patterns. The differences between the varied varieties of a language are best documented by the actual language use of the respective communities based on an authentic, representative, and well-balanced corpus. There is a need for a corpus for Malaysian Tamil.

1.1.2 Corpus Linguistics

A corpus is, as Sinclair (1991, p. 171) pointed out, “a collection of naturally occurring language chosen to characterize a variety of the language”. It is usually planned and designed for some linguistic purposes. The specific purpose of the design determines the selection of texts. A corpus in modern time contains a large collection of machine-readable texts stored in a computer database which could facilitate search and retrieval of these texts (Michael McCarthy & Ronald Carter, 2001; McEnery & Hardie, 2012b)

Corpus linguistics has become a major research method for linguistic studies (Hunston, 2002a; McEnery & Hardie, 2012b). Corpus linguistics does not only provide a large set of computerized linguistic data to be used as an empirical foundation for linguistic inquiries but also provides an alternative approach to the study of language. It allows linguists to study the human language based on the data representing its actual use in real-life contexts rather than based on human intuition of human language (Biber et al., 1998).

Thus, corpus linguistics enables us to study theories of language which are not possible before the development of large-scale computer-readable corpora (McEnery & Hardie, 2012b). For example, it allows a corpus-driven approach to the description of the lexical grammar of English (Hunston & Gill, 1998). It also allows investigations of patterns of language use ranging from colligations, collocations, and lexical bundles to phraseology (Cowie, 1998; Hunston, 2002b; Sinclair, 1991). Evidence from corpus investigations can refine and advance a range of theories of language.

1.1.2.1 The use of corpora

Apart from applications in research, corpus linguistics has also been applied to language teaching. It has been used to develop teaching materials such as word lists and reading materials (Allan, 2009; Nation, 2016; Sinclair, 2004) and to implement data-driven learning in language teaching (Chen, 2018; Hadley & Charles, 2017; Johns, 1991). While corpus linguistics has proven to be useful for linguistic studies, the corpora constructed and the relevant analyses based on these corpora are primarily in English (Biber et al., 1998; Davies, 2009), In contrast, well-designed corpora in other languages are still relatively rare.

As the field of corpus linguistics moves forward, there is a need to reflect on the resources available for linguistic studies and the methods and tools that have been used for such inquiries. Given the possible applications of corpora in research and teaching, there is a need to consider the construction of corpora in languages other than English. One such language is the Tamil language, which is widely used amongst the Tamil-speaking community in Malaysia, but which is surprisingly relatively less researched, particularly in studies based on authentic language data. The present research, as will be discussed in greater detail later, is a humble contribution to this direction.

1.2 Problem Statement

As one of the longest-surviving classical languages in the world, the Tamil language has been in use for over 2000 years (Hart, 2015; Hart & Heifetz, 2002). While the Tamil language has changed over time, the existing grammar descriptions about its rules of use largely rely on those prescribed centuries ago. Little is known about how the Tamil language is currently used, especially in countries like Malaysia, where the Tamil language is a minority language. Hence, there is a need to survey the actual use of Tamil in contemporary Malaysia. In this respect, an electronic and machine analyzable representative general corpus would be essential. As (Godfrey & Zampolli, 1997, p. 101) reminded us,

> Languages for which no adequate computer processing is being developed, risk gradually losing their place in the global information society, or even disappearing, together with the cultures they embody.

However, there is currently no computer programme that gives results of the tokenization of a Tamil corpus; nor is there a computer programme that can perform part-of-speech (POS) tagging for Malaysian Tamil: both are important for the processing of a corpus to render it usable for the study of Tamil. Unlike English corpus linguistics where corpus tools like WordSmith Tools and AntConc are widely accessible, there is no similar computer program in Tamil in Malaysia. Applications of corpus linguistics in the study of Tamil in this country have therefore been restricted, from linguistic description and Tamil language pedagogy to language technology.

It is high time that a corpus for Malaysian Tamil be built and a computer programme be developed to process the corpus. The current research seeks to to address both of these needs. Specifically, this thesis reports on efforts and initiatives to develop a written Malaysian Tamil corpus (WMTC) and a morphological parser with a POS tagger. Three important points to note here. First, as a starting point, a written corpus, rather than a spoken corpus, is built. As will be discussed in Chapter 8 later, builing a spoken Malaysian Tamil corpus will be an agenda for future research. Second, this thesis will focus on the development of a morphological parser and a POS tagger, given that Tamil is an agglutinative and inflectional language, with grammatical suffixes attached to the lexicon. The discussion on the development of this parser and the POS tagger will be presented in Chapter 5 of this thesis, along with a brief discussion of other tools of corpus analysis such as the concordancer and N-gram necessary to illustrate how the computer program was developed. Third, the algorithm chosen for the morphological parser and POS tagger is discussed and considered in Chapter 6 in this thesis.

1.3 Research Objectives

This thesis aims to achieve the following three objectives:

(1) To construct a written Malaysian Tamil corpus (i.e., WMTC).

(2) To develop a morphological parser and POS tagger to process the corpus; and

(3) To choose and design an algorithm for the morphological parser and the POS tagger.

1.4 Research Questions

Based on the above objectives, the current research aims to address the following questions:

1) What considerations were taken into account in the construction of WMTC?

2) What was involved in the development of a morphological parser and POS tagger?

3) How might a suitable algorithm be designed for developing the morphological parser and the POS tagger?

The first research question is addressed in Chapter 4; the second research question in Chapter 5 and the third question in Chapter 6.

1.5 Significance of the Study

A well-designed Tamil corpus has great value for the study of the Tamil language at three important levels: lexis, grammar, and discourse levels. A Tamil corpus can be used for the study of meanings, frequencies, and collocates of Tamil words as well as patterned use of Tamil language and distribution of various constructions. It is also useful to study non-linguistic factors (e.g., registers and genres) that affect choices between structural variants.

As far as the present research is concerned, a corpus is useful for studying the morphology of Tamil since Tamil is a morphologically rich language. Hence it is useful to focus on inflection and derivation of Tamil words, as explored in the current research in Chapter 5 when we consider the morphological parser and POS tagger. As noted earlier, the Tamil language has changed over time, and emergent morphological features are worthy of investigation. Thus, WMTC was created and discussed in this thesis.

The development of the parser and the POS tagger, together with a consideration of the algorithm chosen, will make a further contribution to language technology, language pedagogy, language planning, corpus linguistics, computational linguistics and relevant language research.

1.6 Scope of the present research project

The present research project restricts itself to Tamil computational morphology. Also, it is restricted to a description of a corpus project focusing on written Malaysian Tamil.

1.7 Organization of the Thesis

The thesis consists of eight chapters and is organized as follows:

Chapter 1	Introduction
Chapter 2	Review of literature
Chapter 3	Methodology
Chapter 4	Malaysian Tamil corpus development
Chapter 5	The development of morphological Parser and POS tagger
Chapter 6	A*lgorithms for the morphological Parser and POS tagger*
Chapter 7	Discussion
Chapter 8	Conclusion

Chapter 1 introduces the background of the thesis, states the objectives of the thesis, problem statement, points out the significance of the study and identifies the questions to address, scope and necessity of the study. It concludes with the outline of the whole thesis.

Chapter 2 reviews the literature related to corpus linguistics, and morphological parsing and POS tagging.

Chapter 3 presents information on the methodology of the thesis. It deals with the salient features of a Corpus and the structure of the morphological parser, POS tagger, and the steps followed to achieve the above.

Chapter 4 explains the design and the developmental process and the relevant considerations to develop the corpus for Written Malaysian Tamil, one of the primary results produced from the present thesis project.

Chapter 5 explains the software architecture and the design adopted in the development of a morphological parser and the POS tagger. These include data presentation, character encoding, development platform, programming language, and other tools related to the development of the morphological parser and POS tagger. An initial morphological parser was based on the inflectional morphology of Tamil as described in various modern Tamil grammars; and subsequently this was modified and enriched based on the result of the initial parser applied over the present constructed corpus. The illustration of process involved in morphological parsing and POS tagging were also provided.

Chapter 6 deals with the design of algorithms for the morphological parser and the POS tagger. This chapter discusses the various aspects of Tamil computational morphology which is the backbone of the software - morphological parser for Written Tamil, which is one of the objectives of the present research project. Both computational linguistic issues and linguistic issues are discussed here. The morphological structure involved in the construction of various word forms of Tamil is explained based on the Tamil inflectional process. It describes the morphological analysis of the Tamil word forms, which involve the various types of inflection - both noun declension and verb conjugation. The various grammatical suffixes, morpho-tactic and morph-phonemic rules involved in Tamil morphological inflection and the various POS types of Tamil word forms also are discussed in this chapter. Based on the construction of the present morphological parser, Noun and verb flow charts illustrating the Tamil morphological inflection process are provided in this chapter.

Chapter 7 discusses the issues and the solutions in developing the corpus and its analysis to understand the structure of Tamil word forms with the help of the morphological parser and POS tagger developed.

Chapter 8 the final chapter of the thesis, concludes with a summary of the findings from the whole research project and their implications for the understanding of modern Tamil Morphology. Also, it explains the utilities of the present research project in further linguistic research, computational linguistic research, language planning, curriculum development and language applications such as language teaching, lexicography and language technology tools.

1.8 Conclusion

In this introductory chapter, an overview of the research has been presented. The background of the thesis, its objectives, research questions, significance and scope of the research have been discussed. The next chapter reviews the literature relevant to the present research.

CHAPTER 2

REVIEW OF LITERATURE

2.1 Introduction

In this chapter, prior corpus research for various languages, their aims, designs, constructions and analytic tools, and the use of corpus data for language technology are reviewed. Since the present study also involves developing a morphological parser and a POS tagger, a comprehensive review of works on morphological parser and POS tagger, has been conducted. This chapter reviews the issues involved in the development of morphological parsing and POS tagging.

2.2 Corpus Linguistics

2.2.1 Definitions, descriptions, distinctive features and types of corpus linguistics

As suggested over three decades ago, a corpus is a collection of "naturally-occurring language chosen to characterize a variety of a language". It is usually planned and designed for specific linguistic purpose or purposes. According to Dash & Arulmozhi (2018), "it is to be designed for the faithful study of linguistic properties present in a language". The purpose of the design determines the selection of texts. A corpus in modern time contains a large collection of machine-readable texts stored in a computer database which could facilitate search and retrieval of these texts (M. McCarthy & Ronald Carter, 2001; McEnery & Hardie, 2012b).

A corpus is different from an electronic archive in that the texts stored in the corpus are kept in a specific manner that enables scholars to study its data, not only non-linearly but also quantitatively and qualitatively (Hunston, 2002a, p. 2). As per (Baker, 2006) they can use corpora "as a standard reference with which claims about language can be measured". The specific way the corpus data are stored allows easy access and enables complex calculations to be carried out on large number of texts leading to the discovery of linguistic patterns and frequency information that would otherwise take days or even months to uncover if done manually and may run counter to intuition.

Consequentially, Corpus Linguistics has emerged as a distinctive branch of applied linguistics in the past decades as it provides a solid empirical base upon which to formulate linguistic generalizations, explore variations, and test linguistic theories (Leech, 2007). It is certainly distinct from most other topics in linguistics as it is not directly about the study of any particular aspect of language. Rather, it focuses upon a set of procedures or methods for studying language (McEnery & Hardie, 2012b). Hence, it is considered by many as a methodology of linguistic study. Hoffmann (2008), for example, describes "corpus linguistics as the systematic study of linguistic phenomena using machine-readable collections of authentic language use. It is an essentially quantitative method, meaning that corpus linguists tend to count language features" with the assistance of computers as part of their analysis of linguistic features, nevertheless, analysis can also be done qualitatively.

Corpus linguists are typically interested in discovering general patterns or norms of language use rather than in establishing a mere collection of idiosyncrasies or peculiar features of language that speakers produce. The emphasis is on description rather than prescription, that is, corpus linguists aim to describe rather than establish or uphold rules about how language should or should not be used. According to Crawford and Csomay (2015, p. 5)

> corpus linguistics looks at how language is used in certain contexts, how it can vary from context to context, and describes language variation and use by looking at large amounts of texts that have been produced in similar circumstances.

Lindquist and Hans (2018, p. 1) share similar view that "corpus linguistics is not a branch of linguistics rather it is on par with the other branches since corpus does not tell you what is studied, but rather that a particular methodology is used". Certain researchers have isolated distinctive features of corpus linguistics. For example, (Biber et al., 1998) identifies four important characteristics of corpus linguistics as listed below:

(1) It analyses empirically the actual pattern of use in natural language texts.

(2) It utilizes a large and principled collection of natural texts.

(3) It makes extensive use of computers for analysis.

(4) It depends on both quantitative and qualitative analytical techniques.

In addition to the above four characteristics of a corpus, (Tognini-Bonelli, 2001) noted one more feature of corpus linguistics - it is read vertically. The texts in a corpus are not read horizontally from start to finish as with any news item in a newspaper. Rather, the texts comprise of a collection of different albeit related events and they are investigated as fragments. Many examples of a single feature of a language are seen in relation to one another at any one time. Hence, the corpus is not read horizontally but vertically.

Tognini-Bonelli also made a distinction between “corpus-based” research and “corpus-driven” research. In a corpus-based approach, corpus linguistic researchers are guided by former corpus findings or by specific concerns about language use. They have a clear notion of what linguistic feature they want to examine before they search the corpus.

Furthermore, corpora can be classified according to their purpose, whether general reference or specialized corpora (Gatto, 2014). British National Corpus (BNC), and COCA (Corpus of Contemporary American English) are some of the general-purpose reference corpus projects. They are developed to represent General English.

Unlike a general reference corpus, a specialized corpus aims instead at representing only a given variety or domain of language in use, such as medical discourse or academic discourse. It is generally smaller than a general-purpose corpus, and restrictions may apply not only to domain, but also to genre, time, and geographical variety.

Another crucial distinction is between synchronic and diachronic corpora, and between monolingual and multilingual corpora (Gatto, 2014; Malamatidou, 2017). In addition, distinction must be made between raw corpora and tagged corpora. Tagged corpora that are fully annotated allows a corpus to unleash its potential for sophisticated analysis (Gilquin, 2002).

Generally, a corpus is developed for certain research purposes, and based on its aims various types of corpora can be identified (Gatto, 2014). The first distinction is between general and specialized corpus. That is, it concerns whether the corpus is for general or specialized purpose. BNC and COCA are two general purpose reference corpora developed to represent General English. A specialized corpus, on the other hand, represents only a particular variety or domain of a language in use, for example, legal discourse, medical discourse or business discourse. Specialized corpus is of course smaller than a general-purpose corpus and restrictions may apply not only to domain, but also to genre, time, and geographical variety. The second distinction is between synchronic and diachronic corpora. And the third distinction is between monolingual and multilingual corpora.

2.2.2 Views on Corpus Linguistics: Methodology or Discipline

Despite a plethora of corpus research, there have been varying views of corpus linguistics. According to (Taylor, 2008), "corpus linguistics is a tool, a method, a methodology, a methodological approach, a discipline, a theory, a theoretical approach, a paradigm (theoretical or methodological), or a combination of these."

Leech (1999, p. 106) notes that:

> computer corpus linguistics defines not just a newly emerging methodology for studying language, but a new research enterprise, and in fact a new philosophical approach to the subject.

For Stubbs, (1993, pp. 23-24) "a corpus is not merely a tool of linguistic analysis but an important concept in linguistic theory". Teubert (2005) sees corpus linguistics as "a theoretical approach to the study of language". In Tognini-Bonelli 's view, corpus linguistics is "a pre-application methodology which possesses theoretical status" (2001, p. 1). The views of Mahlberg strengthens the opinion of Tognini-Bonelli. She points out that:

> advocates of corpus-driven approaches to the description of English claim that new descriptive tools are needed to account for the situation of real text, and ideas of theoretical frameworks to accommodate such tools have started to emerge. (2005, p. 370)

According to her, disagreement remains whether corpus linguistics is mainly a methodology, and whether there is a need for a theoretical framework. But Thompson and Hunston (2006, p. 8) argue that "at its most basic, corpus linguistics is a methodology that can be aligned to any theoretical approach to language".

McEnery et al. (2007, pp. 7-8) holds the view that corpus linguistics is a methodology. Similarly, Bowker and Pearson (2002, p. 9) consider it as "an approach or a methodology for studying language use". While scholars like (Teubert, 2005) and (Bauer & Aarts, 2000) feel strongly that it is a discipline. McCarthy (2001, p. 125) thinks that corpus linguistics is a "cutting edge change in terms of scientific techniques and methods".

This claim of scientific method is criticized by Chomsky (2004, p. 97) who writes, "The standard method of the sciences is not to accumulate huge masses of unanalysed data and to try to draw some generalization from them." But the opinion of corpus linguistics scholars like (McCarthy & Carter, 2004) and stands against his introspective linguistics view. Their stand could be compared with the statement of (Sinclair, 1991) that "one does not study all of botany by making artificial flowers".

2.2.3 The Potentials of Corpus Linguistics

Leon (2005, p. 36)says that corpus linguistics:

> covers various heterogeneous fields ranging from lexicography, descriptive linguistics, applied linguistics language teaching or Natural Language Processing to domains where corpora are needed because introspection cannot be used, such as studies of language variation, dialect, register and style, or diachronic studies.

Corpus-based analyses could also be used in the study of English grammar (Campoy et al., 2010). (Biber & Reppen, 2015; 2015) have used the Longman Grammar of Spoken and Written English (LGSWE) in his research related to grammar teaching. Learner corpora are directly related to language classroom.

According to McEnery et al. (2007, p. 65), “a learner corpus is a collection of the writing or speech of learners acquiring a second language (L2)’’. International Corpus of Learner English (Dagneaux et al., 1998; Granger, 2002) is one of the well-known learner corpora which is defined by Nesselhauf (2007, p. 40) as “a systematic computerized collection of texts produced by language learners”. Meng (2012, pp. 191-207) in his article has also elaborately dealt with the learner corpora and second language acquisition.

2.3 A survey of corpora projects

2.3.1 Corpora for English language

Numerous corpus projects have been developed in the English language in the past decades

2.3.1.1 Survey Corpus

One of the earliest projects was the Survey Corpus led by Randolph Quirk at the Survey of English Usage Centre (Ilson, 1982; Quirk, 1990). The Survey Corpus was originally compiled on paper, in the form of many thousands of slips, with detailed grammatical annotations. Following Quirk, John Sinclair initiated multiple corpus projects. He first launched a spoken corpus project at the University of Edinburgh between 1963 and 1964. Since then, he continuously contributed immensely to this field ranging from proposing the design principles for corpus construction to analyzing corpus from various linguistic points of views.

2.3.1.2 Brown Corpus

The first electronic corpus, the Brown Corpus, was developed at Brown University in the 1960s by Nelson Francis and Henry Kucera (Francis & Kucera, 1964). It contains about one million words from the works published in 1961 in America.

2.3.1.3 British National Corpus (BNC)

The Brown Corpus was soon succeeded by a larger electronic corpus, the BNC, which consists of 100 million words of British English from 1991-1995. BNC consists of both written (90%) and spoken (10%) British English. The written texts consist of extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books, popular fictions, published and unpublished letters, memoranda, and school and university essays. The spoken sub-corpus includes a large amount of unscripted informal conversation recorded by volunteers selected from different ages, regions, and social classes, together with language collected in different contexts ranging from formal business or government meetings to radio shows and phone-ins. The corpus is part-of-speech tagged using the Constituent Likelihood Automatic Word-tagging System (CLAWS) C5 tag set (Garside, 1995).

2.3.1.4 The Collins-Birmingham University International Lexical Database (COBUILD)

In 1980, in collaboration with Collins, a publishing company, Sinclair set up COBUILD as a research facility. It provided data, ideas, and analyses for Collins, to help them compile a new corpus-based dictionary, specifically, the Collins COBUILD dictionary published in 1987 (McEnery & Hardie, 2012b).

2.3.1.5 Collins Corpus

The Collins Corpus is a database of English with over 4.5 billion words or tokens. It contains written materials from Internet, newspapers, magazines, and books published around the world, and spoken materials from radio, TV, and everyday conversations. Extracted from the large-scale Collins corpus, the Bank of English is a subset of 650 million words from a carefully chosen selection of sources, to give a balanced and accurate reflection of English as it is used today. Analyses of the large-scale Collins corpus reveal how words are used, what they mean, which words are used together, and how often words are used. This information on frequency helps to decide which words to include in the dictionaries.

2.3.1.6 The American National Corpus (ANC)

ANC is a corpus of American English containing 22 million words of written and spoken data produced since 1990. The genres in the ANC include newer types of language data that have become available since the latter part of the twentieth century, such as web-based diaries (blogs), web pages, chats, emails, and rap music lyrics. Like the BNC, the ANC corpus is also part-of-speech tagged and encoded in Extensible Markup Language - XML (Lindquist et al., 2018).

2.3.1.7 The Corpus of Contemporary American English (COCA)

The Corpus of Contemporary American English is the first large and diverse corpus of American English (Davies, 2009). It was designed as a monitor corpus that can be used to track and study recent changes in the language (Davies, 2010). First released in 2008, it contained more than 385 million words from 1990 - 2008 (20 million words each year), balanced between spoken, fiction, popular magazines, newspapers, and academic journals. By 2017, the size of the corpus had reached more than 560 million words of text. The unique relational database architecture of COCA allows for a wide range of queries (Davies, 2005). Now the COCA has billions of words. An academic vocabulary list of the top 3,000 words of academic English is added in Jan 2022 (Davies & Gardner, 2013). Besides these corpora, other important corpus projects include Cambridge International Corpus and Lancaster-Leeds Treebank which is the first syntactically parsed corpus.

Apart from these major corpora in the English language, efforts have been devoted to corpus projects in other languages, for example, Russian National Corpus (Grishina, 2006) and Academica Sinica Balance Corpus of Modern Chinese (Huang & Chen, 2010) and Helsinki corpus of English for historical English (Rissanen et al., 1993) from 750 AD to 1700 AD.

2.3.2 Corpora for Indian languages

In India, the Central Institute of Indian Languages (CIIL) under the Ministry of Education of India and Technological Development of Indian Languages (TDIL) under the Ministry of Electronics and Information Technology have been working to develop corpora for twenty-four Indian languages including Tamil. Some corpora are monolingual while others are parallel corpora involving two languages or more. And they encompass both written and speech corpora. For Tamil, the corpus has around a million words. (https://www.ldcil.org/resourcesTextCorp.aspx)

Another major corpus project, EMILLE Corpus (Xiao et al., 2004) (Enabling Minority Language Engineering), was undertaken by Lancaster University in collaboration with the CIIL in Mysore, India. This project developed three types of corpora, monolingual, parallel, and annotated corpora, for fourteen South Indian languages. A total of 10 billion words were selected from texts in these languages.

Apart from the CIIL's corpora project for Tamil, some of the universities and institutions in Tamil Nadu have been involved in Tamil corpus studies. These include Anna University (Chennai), Madras Institute of Technology (MIT, Chennai), Amirtha University (Coimbatore), and Tamil Virtual Academy (Chennai).

Tamil corpus has also been developed in Singapore. The Ministry of Education of Singapore released it in the 12th International Tamil teacher's conference in Singapore. The corpus was accompanied by essential corpus tools for easy access and analysis.

So far, from the review of corpora that are available, it is clear that most of them belong to the English language. There are some corpus projects for languages in India, for example, the corpora that Lancaster University built in collaboration with local institutions there.

2.3.3 Corpora in Malaysia

The corpus studies in Malaysia are mainly concerned with the English and Malay languages. According to a survey by Siti et al., (2014), the corpus studies in Malaysia are mainly on English language use in Malaysia, Malaysian English learner language, Malaysian textbook content, Malay language and lexicography and Corpora development.

2.3.3.1 Malaysian Malay Corpora

The corpus approach made inroads in Malay language research with the development of the Malay language corpus by the Dewan Bahasa dan Pustaka (DBP) in the 1980s. The corpus was developed to facilitate systematic and objective analyses of the Malay language to enrich Malay dictionaries, grammar books, etc. The corpus, to date, is the largest Malay corpus in Malaysia.

In recent years, more Malay corpora have been developed to facilitate research on Malay language description and translation. For example, Malay Practical Grammar Corpus (MPGC) is an ongoing Malay corpus project that involves analysis of written texts in Malay from several major genres, mainly newspapers, magazines, and books. And it permits teachers and students to acquaint themselves with corpus analysis and its application in the classroom (Abdullah et al., 2021)

MALEX (MaLay LEXicon) "is an annotated lexicon designed as a relational database"(Don, 2010). The data for the project is a compilation of texts from novels (approximately 800,000 words), newspaper corpus (approximately 5 million words), academic texts (20,000 words), as well as speeches, containing 1.3 million words by the former Malaysian Prime Minister, Tun Mahathir Mohamad. However, the researcher considers this an archive rather than a corpus. This contains a spelling normalizer, a tag set, a list of lemmas, morphological derivations, and a pronouncing dictionary.

There exists an online Malay language lexical database based on a corpus of Malay textbooks for primary school, bilingual national schools and trilingual national-type schools in Malaysia (Lee & Low, 2011). (Joharry & Rahim, 2014) reported on the development of a 250,000-word English-Malay bilingual parallel corpus in the domain of agriculture and health but the current status of this project is unknown. Another Malaysian corpus in the making is The Development of the Malaysian Hansard Corpus: A Corpus of Parliamentary Debates 1959-2020 (Abdullah et al., 2021).

For legalese, there is the Translation Memory of legal texts and a Glossary of legal terminology Mahadi et al. along with a legal English-Malay parallel corpus of about 210,000 words and a Glossary of legal terminology.

Awal et al. (2011) investigated the linguistic features of the Malay preposition *untuk* (equivalent to ‘for’ in English) in a translation corpus. Their findings support the translation universals hypothesis which claims that translation language resembles the normative standard language of the original language.

The availability of corpora and the familiarity of the corpus approach have also increased work in Malay language description. The DBP corpus, for instance, has made lexical and grammatical studies on Malay more dynamic and empirical with the use of attested data and computational methods. (Rahim, 2005) study, for example, compared the connotations of the words, *perempuan* (female) and *wanita* (women).

Generally, corpus-related research in the Malay language in Malaysia has for the most part focused on issues concerning the development of Malay corpora and the description of the Malay language. The development of various Malay corpora reflects scholarly recognition of the significance of corpus linguistics methods in Malay language research.

2.3.3.2 Malaysian English Corpora

The corpora that have been developed for Malaysian English are as follows:

1. A Corpus-Based Archive of Learner English in Sarawak - CALES (Botley, 2007)
2. The Malaysian Corpus of Learner English - MACLE (Knowles et al., 2006)

3. The Business and Management English Language Learner Corpus - BMELC (Chaal, 2011)
4. The Malaysian Corpus of English -MEC (Kaur, 2010)
5. The Corpus of Malaysian English (COMEL) - a spoken corpus (Knowles et al., 2006)
6. The International Corpus of English-Malaysian Component (ICE-Malaysia) - developed at University Sains Malaysia (USM), a part of the global ICE project. (Rahim, 2014)
7. The English of Malaysian Schools Students - EMAS (Ang et al., 2011)

Despite these corpora, corpora of learner language in Malaysia are a rarity, the English of Malaysian School Students (EMAS) is a corpus worthy of attention particularly to language researchers

2.3.3.3 Chinese and Malaysian Chinese Corpora

Before delving into Malaysian Chinese corpora, Chinese corpora development in China is discussed first for a broad perspective.

Corpus linguistics had a modest beginning in the early 1920s in China. The first project involved counting Chinese language characters which was followed by The Applied Glossary of Modern Chinese in 1928 (Feng, 2006). However, these were not machine-readable. In 1979 China developed a machine-readable corpus called the Chinese Modern Literature Work Corpus, which led to the publication of The Dictionary of Modern Chinese Word Frequency. The trend continued resulting in the publication of corpora such as the High School Chinese Language Teaching Material Corpus and Modern Chinese Corpus in 1983.

The corpus, Lancaster Corpus of Mandarin Chinese (LCMC), is available for Mandarin Chinese (McEnery & Xiao, 2003). Though a number of Chinese corpora are available in China, there are understandably few Malaysian Chinese corpora available. The Malaysian Chinese corpus that is available is a spoken one and in Cantonese, a Chinese dialect. The Malaysia Cantonese Corpus - MYCanCor (Liesenfeld, 2018), is a collection of video recordings of spontaneous talk-in-interaction. However, the corpus is transcribed in CHAT format and presented in traditional Chinese characters (UTF8).

In conclusion, there are more Malaysian English corpora than Malay corpora in Malaysia while little effort has been devoted to the Malaysian Tamil language and Chinese language.

2.4 Corpus Design

2.4.1 Representativeness and balance

A major function of a general corpus is to use it as a database to study certain features of an entire language. A corpus is not simply a collection of texts but seeks to represent a language or some parts of a language. This means that any researchers who wish to conduct a study on a particular domain or a particular aspect of language use, can extract a representative corpus or sample from a general purpose corpus, and extrapolate from this corpus to answer their research questions Leech (2007, p. 135). The appropriate design for a corpus therefore depends upon what it is meant to represent (Biber, 1993). The representativeness of the corpus decides the kinds of research questions to be posed and answered as well as the generalizability of the research results.

Despite the widespread discussion of the idea of representativeness in corpus linguistics, the principle of representativeness has rarely been adopted in practice by corpus linguists (Leech, 2007; McEnery & Hardie, 2012b). This is probably because of the complexity of sampling as a corpus is representative only if it fully captures the variability of a language (McEnery & Hardie, 2012b).

Some rules of thumb toward building a representative corpus have been suggested by a few researchers. Hunston (2002a, p. 28) suggests “breaking down the whole corpus into component parts and including equal amounts of data from each of the parts”. The principle of including an equal number of materials from many different genres of relevance is referred to as “balance” (Weisser, 2016). A balanced corpus should cover a wide range of text categories that are considered to be representative of the variety of language being studied. The categories are sampled proportionally so as to offer a small-scale model that is manageable of the linguistic materials the corpus developer wants to investigate Atkins et al., (1992, p. 6).

The conventional practice used by developers of the early corpora is not to select samples all of the same size but to simplify data collection and avoid the difficulty of reconciling texts of different dimensions (Sinclair, 2005). The key to this work is to know the character of the ‘whole’ corpus.

From the above discussion, it can be concluded that ‘representativeness’ and ‘balance’ have much importance in the construction of a corpus. However, both are very difficult to achieve in the construction of corpora.

2.4.2 Authenticity

The linguistic data in a corpus should be authentic, i.e., they should be gathered from actual language in use. As (Gatto, 2014) asserts:

> Corpus linguistics is by definition an empirical approach based on the observation of authentic data. As such, it can well be considered as the most evident outcome of the resurgence in popularity of language studies grounded on real examples of language in use, rather than on introspection, enabled by the possibilities and affordances provided by computer storage.

2.4.3 Size

The question of the size of corpora has been central to corpus development, and there has been the overriding belief among many corpus creators that biggest is best. Many corpus linguists prefer a large corpus. One reason for this is that words are unevenly distributed in texts and that most words occur only once. Thus, to study the behavior of words in texts, a large number of occurrences should be available (O'Keeffe, 2012). For this purpose, it is always good to have a corpus that is as large as possible.

In practice, the work of collecting and compiling a corpus is constrained by time and resources. The work of corpus building cannot keep on with no end in sight. In view of this, the question that should be asked is whether a corpus is large enough for its specific purpose. There is no one fixed standard size for a corpus, and the size can vary depending on the nature of the corpus and its purpose. The estimate of a suitable size for a corpus is subject to considerations concerning other factors. The most crucial consideration should be the purpose of the corpus itself, as ultimately, it is what the corpus is for that should determine its size (Gatto, 2014; O'Keeffe, 2012)

Corpora look at the whole language and are primarily used for lexicographic purposes and because of that they are constructed to be as large as possible (Kennedy, 1998), Nevertheless, there are justifications for corpora of smaller sizes that examine specific language features, especially for language teaching. In fact, he feels that the quality of the data that corpus researchers work with is just as important as its size. Gatto (1995) also expresses the same opinion as Kennedy. She also emphasizes that there cannot be a standard size for corpora as the size depends on the nature of the corpus and its purpose.

2.4.4 Data markup and storage

Another concern is the storage of the corpus texts which have to be in a suitable computer-readable version. Basic processing includes digitalizing or scanning the printed texts of the chosen documents. The texts will then be converted from their existing electronic form (PDF. HTML, etc.) Into a suitable format, usually plain text. Machine-readability of the texts in the corpus is the default for contemporary linguistics.

When using online documents, it is also necessary to clean the text by separating the text itself from typical paratextual materials, that is, the non-informative parts made up of navigation links, advertisements, headers, and footers, or what is termed as 'boilerplate' by Gatto. Further, she says that some form of markup may be necessary to compensate for the loss of information determined by the reduction of texts, or text samples, to mere strings of characters (Gatto, 2014). Xiao (2004, p. 155) defines corpus markup as "a system of standard codes inserted into a document stored in electronic form to provide information about the text itself."

In this regard, (Weisser, 2016) points out that there are the different file formats within each online text that is found online and that the files are not always easy to process as the text is sometimes stored in a proprietary format that can only be processed by programmes designed to deal with them.

Explaining what needs to be done, (Crawford & Csomay, 2015) shared this:

> In order to take this superfluous information out of the text, you will need to convert any text that you collect into a text file (a file with the extension ".txt"). The ".txt" format removes all of the markup language found in many other file extensions and allows a software program such as AntConc to find textual patterns instead of other patterns related to format or font type.

The file format including character encoding of the final corpus texts or samples emphasises the fact that since the purpose of storing corpus texts or samples is finally to make them undergo various analytical processes by corpus access software such as concordance, N-gram, parsing etc. they should be stored in a neutral file format such as ".txt". This is common to all works on corpus linguistics.

As to the other information regarding the text samples such as publications details, author(s) details and other related facts, standard encoding schemes are required to represent them. According to McEnery et al. (2012), a corpus may have three types of annotation information: (1) metadata, (2) textual markup, and (3) linguistic annotation.

Metadata is information that tells something about the text itself. For example, in the case of written material, the metadata may tell you who wrote it, when it was published, and in what language it is written in. The metadata can be encoded in the corpus text or held in a separate document or database. Textual markup encodes information within the text other than the actual words. For example, where italics starts and ends in a print written text.

Together, metadata and textual markup that have been introduced into a corpus allow a range of research questions to be addressed. However, it is possible to go beyond merely recording features of a corpus text such as where italics or the speech of a certain speaker begins and ends.

Here it is to be mentioned that extratextual annotation plays an important role in corpus annotation to identify a text based on metadata information of a header file rather than referring to the actual texts of the text file (Dash, 2021).

Linguistic information within a corpus text can also be encoded in such a way that systematic and accurate recovery of the analysis can be done later. The corpus is thus said to be analytically or linguistically annotated. Annotation typically uses the same encoding conventions as textual markup.

In most corpus projects, the eXtensible Markup Language or XML is a standard which is used widely for corpus files. However, for the COCA project, the Structured Query Language (SQL) database software which is based on relational database approach is used (McEnery & Xiao, 2011).

2.5 Corpus Access Software

According to Hunston (2002a), a corpus by itself is nothing other than a store of used language and does not contain new information about the language. Observations of various language features require corpus access software to rearrange the data in the store and gives a new perspective on the familiar. Analyses of frequency, phraseology, collocation, colligation and lexical bundles may be performed with the assistance of corpus tools and computer programs (Scott, 2010). Results from such analyses reveal the distinctive features of the language data stored in the corpus. Some of the corpus tools and computer programmes developed for such corpus analysis purposes are AntConc (Anthony, 2014), Wordsmith Tools (Scott, 2016) and Wmatrix (Rayson, 2003) which will be reviewed in the subsections that follows.

2.5.1 AntConc Tools

AntConc is a concordance software developed by Laurence Anthony at Waseda University in Japan. This software can be used by the researchers for their own language corpora. With this, KWIC (keyword in context) searches can be carried out through the concordance lines. Lexical and grammatical analyses, like n-grams, collocates, and word lists in a particular text, can be carried out too. (Crawford & Csomay, 2015)

2.5.2 Wordsmith Tools

Wordsmith Tools is a package of corpus analysis tools developed by Mike Scot at Oxford University Press. It has been in the existence for over twenty-five years. The core areas of the software package are included in three core modules as follows:

(1) Concord is used to create concordances using all the hits from a search within a previously defined body text.

(2) Wordlist lists all words on word forms that are included in the selected corpus and the statistical data are different from the text corpus.

(3) Keyword creates a list of all words and word forms that significantly occurred rarely or frequently according to certain statistical criteria in the text corpus.

Each module offers a number of other features in relation to the text corpus or text being analyzed. Thus, for example, collocation and dispersion plots are computed with a concordance search. In addition, there are a number of additional modules that are useful for preparing, cleaning-up and formatting the text corpus. WordSmith Tools can be used in 80 different languages. The software is an internationally popular programme for works based on corpus-linguistic methodology. It is used by corpus researchers in assorted fields.

2.5.3 Wmatrix Tool

Wmatrix is a software tool for analysing and comparing a corpus, providing a web interface to the corpus annotation tools, and standard corpus linguistic tools such as frequency lists and concordances. In addition to these, the key grammatical categories and key semantic domains are provided by the keywords method.

2.6 Corpus Analysis Tools for Tamil

Regarding corpus analysis tools for Tamil, some academic institutions such as CIIL (Mysore, Karnataka), Anna University (Chennai, Tamil Nadu), Madras Institute of Technology (Chennai, Tamil Nadu), Amirtha University (Coimbatore, Tamil Nadu) have developed their own tools for their specific purposes. Some commercial enterprises such as NDS Lingsoft Solutions, Cre-A Publishers, Learnfun systems (all from Chennai, Tamil Nadu) have developed their own software tools for their commercial applications. However, these software tools are not available for public use.

2.7 Computational Linguistics and Language Technology

Linguistic data is the basis for computational linguistics and language technology. The studies of computational linguistics and language technology (Jurafsky, 2009; Mitkov, 2004) have been conducted for many western Romance languages. From word processor to automatic machine translation, language application software tools have been developed for many languages.

2.7.1 Corpora and language technology in Tamil

In the last twenty years, numerous studies have been conducted in the field of Tamil computational linguistics and language technology. In Tamil Nadu, contribution to this field comes from higher education institutes, research institutes, private companies and individuals. However, these researches have concentrated only on the Tamil used in Tamil Nadu (Renganathan, 2016). Relatively overlooked is the variety of Tamil language used outside of Tamil Nadu.

In Malaysia, research related to Tamil has been mainly focused on the development of fonts and keyboards for Tamil for various applications ranging from desktop to web applications. Meanwhile, the Malaysian Chapter of International Forum for Information Technology in Tamil (INFITT) has continuously been working towards the development of Tamil computing with respect to technological development and written Tamil in Malaysia.

Currently, existing corpus software are not yet able to make word lists from written texts in Tamil without substantial human interaction with the corpus. Educators and researchers who want to make word lists for Tamil language teaching have to spend considerable time interacting with a large number of written texts. It would be useful if computer programmes can retrieve from the corpus word lists with minimum human interaction. However, there is no such computer programme available for processing written Tamil.

Based on the success of research and development activities in Tamil computational linguistics, applications such as word processor (e.g., spellchecker and grammar checker), text to speech (TTS), automatic speech recognizer (ASR), text summarization, and automatic machine translation (AMT) can be developed for Tamil language processing. To help achieve these objectives, the present research project aims to do computational linguistic research of an exploratory nature for written Malaysia Tamil especially the morphological structure for developing a morphological parser and POS tagger.

2.8 Morphological Parsers

To understand what morphological parser is, one needs to know that a word can be divided into component morphemes. For example, the word *boxes* consists of *box* and -*es*. Parsing involves "taking an input and producing some sort of linguistic structure for it" (Jurafsky, 2009, p. 45). Morphological parser therefore involves "the constructing of a structural representation" of words that are broken down into morphemes (Jurafsky, 2009, p. 46)

According to them, morphological parsing is important for speech and language processing. For morphologically complex languages like German or Russian, it is crucial for online search as it enables automatic search for inflected forms of the base word as well as for part-of-speech tagging. Morphological parsing is also essential for constructing the massive online dictionaries that are required for comprehensive spell-checking. Lastly, it is also used for machine translation. To build a morphological parser for a language, Jurafsky et al. (2009) explains that the following are needed:

1. Lexicon: the list of stems and affixes, together with the basic information about them (whether a stem is a noun stem or a verb stem etc.)

2. Morphotactic: the model of morpheme ordering that explains which classes of morphemes can follow other classes of morphemes inside a word. For example, the fact that the English plural morpheme follows the noun rather than preceding it is a morphotactic fact.

3. Orthographic rules: these spelling rules are used to model the changes that occur in a word, usually when two morphemes combine (e.g., the *y* > *ie* spelling rule discussed above that changes *city* + *-s* to *cities* rather than *citys*).

Here, it should be noted that in Tamil, instead of orthographic rules, there are some Sandhi rules (morphophonemic rules) in Tamil inflection and derivation. For the morphological parsing the above said criteria are taken into account with respect to written Malaysian Tamil.

2.9 Parsing Approaches

The parsing of natural languages can be grouped into rule-based, statistical based and generalized. (Richardson, 1994). While the terms, ‘grammar-driven’ and ‘data-driven’, can be attributed to rule-based and statistical parsing techniques respectively. The different types of parsers are described below.

2.9.1 Rule Based Parser

Rule Rule Based Syntactic Parser: To get the relatively best parse tree for a sentence, a rule-based syntactic parser, developed based on a set of grammatical rules could be used. There are two such parsers available, one developed by Cocke - Kasami - Younger and the other developed by Earley (Jurafsky et al., 2000, p. 427 & 514).

Statistical Based Parser

Statistical Based Parser: As an alternative method to the rule-based one, there are some parsers available which are based on probabilistic statistics. These types of parsers work using statistical parsing algorithms based on the already correctly parsed sentences: whenever there are ambiguities, this type of parsers resolve them by the previous experience gained from working with statistical information.

2.9.2 Generalized parser

Generalized parser: As the backdrop of both rule-based and statistical parsing are identical, Melamed suggested a generalized parsing algorithm in 2005. In 2006, Goodman also proposed a general parsing algorithm.

2.10 Morphological Parser/Analyzer and Generator (MAG) Approaches

Regarding the types of morphological parsers, there are three types available: (1) corpus-driven (2) corpus-based (3) rule or grammar based. In the first type, purely based on the developed corpus, the morphological rules are framed for the parser to analyze the word forms. No previous available grammar is used in this type of parser. In the second type, the parser is developed based on the existing grammar; however, the parser is continuously being modified based on the new facts found in the developed corpus, till the accuracy is maximum. In the third type, no corpus is used: it is based on the existing grammar only. Based on the above types, parsers have been developed for many languages such as English, Arabic, Hindi etc.

Based on Regular Expression and Finite State Automata, Koskenniemi (1984) developed a Finite State Transducer (FST) for Finnish language. In this parser, there are two levels - the underlying level and the surface level. One level is obtained from the other level. This is not based derivation; but by matching one from the other. That is, “correspondence” matching is being used. Each and every morpheme of a word form at one level is matched for its corresponding morpheme in the other level. For this type of parser, all the knowledge needed for morphological parsing and generator - Lexicon, morpho tactics, and morphophonemic of a particular language - are used. The surface level is to describe word form as they occur in written text and the lexical level is to encode lexical units such as stem and suffixes.

The above mentioned FST method is used to develop a morphological parser for Arabic language by Xerox. Beesley (1998) and his team involved in this development of parser.

Another type of parser - a hybrid type of parser - using both the two-level morphology and Unification based formalism, is available for Basque language, which is highly inflectional one. This parser was developed by a highly Adurize I, Agirre E. et al., (2000).

2.10.1 Morphological Analyzers/Parsers and Generators for Indian Languages

Regarding the development of various morphological parsers in Indian languages, there are two main sources which are much useful and mentioned here: one is, the website of TDIL - Technology Development of Indian Languages - (http://tdil-dc.in), and the second one is a research article published by "International Journal of Computer Science & Engineering Technology".

According to the above-mentioned research article by Antony and Soman, there have been many attempts in the development of morphological analyser and generator (MAG) for Indian languages such as Kannada, Malayalam, Hindi, Punjabi, Bengali, Assamese, Bodo, and Oriya, but only a few are publicly available.

Language	Researchers	Year	Feature
Kannada	Vikram and Shalini	2007	Based on Finite State Automata, a morphological analyzer was developed for Kannada.
Kannada	University of Amrita	2008	Two kinds of morphological analyzer and generator (MAG) - one is, statistical method based, and the other is, rule-based one - were developed for Kannada.
Kannada	Engineering College (R.V) Bangalore	2010	Developed MAG using Trie data structure, handling up to around 3,700 root words and 88,000 inflected words.
	Hyderabad University	2011	Based on Network and Process Model, one morphological analyzer and generator was

			developed. It was claimed that it could handle all the inflectional processes involved in the morphology of Kannada. This tool includes all the affixes, the morphotactic rules and the morpho phonemic rules, in addition to a necessary lexicon.
	Antony & Team	2011	This syntactic parser is based on Kannada corpus, using statistical method. The method is based on the Penn Treebank based statistical syntactic parser. The parsing segments of word forms are displayed with their grammatical categories. Every word form is linguistically annotated with its POS. This statistical parser was trained using 1000 Kannada sentences, adapting Tree-bank basis.
	Sagar, Shobha and Ramakanth	2010	Using Context Free Grammar (CFG) and adapting both Top-Down and Bottom-Up parsing, this parser was developed. The developers finally claimed that a Top-Down parser is more preferable for parsing.
Malayalam	Saranya and team	2008	This morphological analyzer for Malayalam was developed using hybrid approach -

			involving both Paradigm and Suffix Stripping methods.
Malayalam	Jisha and colleagues	2011	To be used in the development of a MT system for Malayalam - Tamil, this MAG was developed. It consists of a bilingual lexicon and a table of all the suffixes involved in the inflection of both Malayalam and Tamil. Affix stripping of the source language Malayalam (morphological parsing) and affix joining in the target language Tamil (morphological generation) methods were used.
Hindi	Bajaj	2008	Bajaj made an attempt to use a semi-supervised machine learning method to enrich the available morphological parsers for Hindi.
Hindi	Meher Vijay Yeleti and team	2009	This parser was based on Dependency Parsing method. Though it was mainly based on grammar driven approach, it was enriched with a statistical based method in order to get high accuracy and speed.
Punjabi	IIIT, Hyderabad	2007	This is a Punjabi morph parser which could handle the Punjabi morphological inflection.

	Dr Mandeep Singh, Punjabi University	2008	Advanced Centre forTechnical Development of Punjabi Language developed a morphological analyser and a generator (MAG) for Punjabi language.
Bengali	Abu Zaher and his colleagues	2009	This parser for Bengali was based on Finite State Automata technology. This is an open-source software tool.
	Ghosh's & team	2009	This Parser was based on Dependency grammar. The developers used both statistical-based processing and a rule-based post-processing. ICON 2009 datasets were used to train the system.
Bangla, Hindi and Telugu	Joakim Nivre	2009	It was a transition-based dependency parser for three Indian languages. The performance of the system was tested with a test set of 150 sentences from each language. The parsers for Bangla and Hindi performed relatively better than the parser for Telugu, it was claimed.

Hindi, Bangla and Telugu	Daniel Zeman	2009	Based on the concept that the 'case' and 'vibhakti' are important features involved in parsing Hindi, Bangla and Telugu, this parser was developed. The uniqueness of this development is that it was a best combined parser for all the three languages.
	Sankar, Arnab, and Utpal	2009	This parser was based on constraint-based dependency parsing method to parse Bangla which is a free-word order language. Paninian Grammatical framework was adapted. To train the parser, a Treebank of 1000 annotated sentences was used. For evaluation process, a set of 150 Bangla sentences.
Assamese, Bengali, Bodo and Oriya	Mona Parakh and Team	2011	Developed a morphological parser for four Indian languages. This prototype model currently can handle inflectional suffixes.
Assamese	Navanath Saharia and Team	2011	Depended on a computational perspective rather than a linguistic perspective for Assamese. It can be used to parse any simple sentence with multiple nouns as well as both

			adjectives and adverbial clauses, the developers claimed.
Indian Languages	Akshar Bharati and Team	2009	This parser adapted the Dependency Framework. It was purely based on Paninian grammar. Also, it was claimed that it was a very simple parser for Indian Languages.

Table 2.1: Development of morphological analyser

2.11 Tamil morphological analyzer and generator

It is a well-known recognized fact that the Finite State Automata formalism plays an important role in the development of morphological parser and generator for many languages. Based on this formalism, the Finite State Transducer has been developed for languages to get both morphological parsing and morphological generation. The time taken for parsing and generation, the accuracy of the result as well as the storage space are very much optimal, especially for agglutinative and inflectional languages.

For Tamil, one of the agglutinative and inflectional languages, A.G. Menon developed a rule-based MAG for Tamil using the Finite State Transducer (Veerappan et al., 2011). He developed this tool with a well- organized lexicon and orthographic rules. This tool consists of a list of 50,000 nouns, around 3,000 verbs, and a relatively smaller list of adjectives.

Anand Kumar developed a morphological generator for Tamil. It was based on a suffix stripping algorithm with two modules; the first module handles the lemma/root part, and the second module handles the Morpho-lexical information. This system consists of the following information: morpho-lexical information file, suffix table, paradigm classification rules, and stemming rules. (Anand Kumar, Dhanalakshmi, Soman, et al., 2010)

He, with some other members of his team, also developed a morphological analyzer for Tamil language. This analyzer was based on corpus-based approach. He used SVM tool (Support Vector Machine) to train and test the generator. For training this tool, he developed a corpus consisting of 130,000 verb words and 70,000 noun words. The test tool contained 40000 verbs, and 30000 nouns taken from Amrita POS Tagged corpus (Anand Kumar, Dhanalakshmi, Rekha, et al., 2010).

Parameswari and her team, from the University of Hyderabad, adapted the Finite State Transducers algorithm for one-pass analysis and generation, and the Word and Paradigm based database (Parameshwari, 2011).

Menaka and her team, in 2010, developed a finite state automaton-based morphological generator. They conducted two experiments to evaluate the system for nouns and verbs using correct and wrong inputs; They established the fact that the finite-state-based morphological generator is well-suited for highly agglutinative and inflectional languages like Tamil (Antony & Soman, 2012; Menaka et al., 2010).

To use in the Machine Translation system for Tamil to Hindi, Rajendran developed one morphological analyser. This MT system was developed for the translation at word-level (Rajendran et al., 2003).

Another contribution from the University of Hyderabad was a research project as a part of doctoral research by Vaishnavi Ramasamy. She submitted a thesis on “A morphological generator for Tamil” (Ramaswamy, 2000) and “A morphological analyzer for Tamil” (Ramaswamy, 2003).

Ganesan, from the Central Institute of Indian Languages (CIIL), Mysore, developed a morphological analyzer to analyses the CIIL corpus. Based on the phonological, morphophonemic and morphotactic rules, he developed that analyzer (Rajendran, 2006).

2.12 Tamil morphological parsers

Some application software of Tamil like spellchecker needs first of all a robust morphological parser. From this perspective, N. Deiva Sundaram, for his Tamil Word Processor “Mentamizh - Tamil word processor” developed one parser for Tamil. It was purely a rule-based one (Sheshasaayee & Deepa, 2016).

AU- KBC Centre, Anna University of Technology, Chennai, has been involving in the basic research as well as the development of many applications’ software for Indian languages. Among them, one important contribution was a Tamil morphological parser. They developed this tool adapting FST machine like PC Kimmo (Anand Kumar, Dhanalakshmi, Rekha, et al., 2010).

Dhurai Pandi developed a morphological generator and parsing engine to analyze the verb patterns in modern Tamil (Duraipandi, 2006).

Resource Centre for Indian Language Technological Solutions, Anna University, Chennai, developed a morphological analyzer for Tamil. This analyzer could parse any Tamil word form into their various parts - root and affixes. Both Noun and verb inflection could be handled by this analyzer, it was claimed. For this analyzer, the developed one dictionary for Tamil consisting of twenty thousand words. These words were classified based on fiftenn grammatical categories. It had two modules: noun and verb analyzer - based on 125 rules.

Tamil has a rich tradition of Poetry from Sangam age. This poetry literary form is well structured one. The prosodical structures are bound to well defined formulae. Among them, one important type is Venpa. Bala Sundara Raman (2003) used Context Free Grammar (CFG) to analyze Venpa class of Tamil poetry. Push Down Automata parser was used to parse the CFG in the proposed system.

Selvam and colleagues developed hybrid language models , based on a part of speech tag set for Tamil language with more than 500 tags (Ajees & Idicula, 2020). About 326 Tamil sentences consisting of more than 5000 words formed the basis for this Phrase structured Treebank. The performance of the system was better than the grammar model, it was claimed.

Akilan, R and his team, Central Institute of Classical Tamil, Chennai, developed one morphological analyzer for Classical Tamil texts. In Sri Lanka, Kangatharaiyar Sarveswaran and others developed a Tamil morphological analyzer based of FST.

Various issues face by the researchers in this morphological domain of Tamil were The Part-of Speech system, called as POS, plays an important role in the morphological analysis of any language. In the analysis of any sentence in a language, first of all, the POS marking of every word in that sentence is much important. Both the individual segments or morphemes of a word and its syntactic role determine its POS. Depending on the purposes, different POS tag set are developed for languages. For the languages like English, Arabic and other European languages, there are many POS tag sets are available. But, for the Indian languages like Hindi, Bengali, Punjabi and South Indian languages, there are only a few tag sets are available. Nowadays, the machine translation systems need cross-linguistic POS tag set.

Hindi

For Hindi language, there are some POS tag sets available, which are different from one another because of the different approaches adapted in developing the tag set. Three different POS tagger systems were proposed: morphology driven, Maximum Entropy (ME) and Conditional Random Fields (CRF). There were two attempts for POS tagger development in 2008; both were based on Hidden Markov Model (HMM) approaches and proposed by Manish Shrivastava and Pushpak Bhattacharyya.

Nidhi Mishra and colleagues developed a POS Tag system for a Hindi Corpus in 2011. This system, by searching the tag pattern from the database, displays the tag of each Hindi word, for example, noun tag and adjective tag.

CIIL, Mysore, India proposed a tag set of 36 tags for Hindi language. This tag system was based on Penn tag set. It proposed a common tag set for all Indian languages which was much useful in doing cross-linguistic analysis.

IIIT, Hyderabad, developed a general standard tag set suitable for all Indian languages. The number of tags available in this tag set was 25. The 6th Workshop on Asian Language Resources, held in 2008, suggested three levels of tag sets were suggested, with the top-level consisting of obligatory 12 common categories for all Indian languages. The other levels contained the recommended tags and optional categories for verbs and participles.

2.12.1 Bengali

For Bengali language, a significant amount of work in the development of POS tag sets has been done, using different approaches. Two stochastic-based taggers were developed, in 2007, by Sandipan Dandapat, Sudeshna Sarkar, and Anupam Basu using HMM and ME approaches. A manually annotated corpus of about 40,000 words was used for both supervised HMM and ME models.

Ekbal Asif developed a POS tagger for Bengali. To train the tagger, a tagged word corpus of 72,341 was used. Ekbal Asif and Bandyopadhyay, using SVM algorithm, in 2008 developed another machine learning-based POS tagger. The entire training corpus was divided into two: one was a training corpus consisted of 57,341 and the other was a development one contained 15,000 words.

Hammad Ali in 2010, developed a POS tagger, based on unsupervised learning. For CDAC, Pune, Debasri Chakrabarti proposed a layered POS tagging in 2011. IIT-Kharagpur also, on its part, proposed a tag set for Bengali with 40 tags, and another with 51 tags, where 42 are general POS tags and the rest 9 for special symbols.

2.12.2 Punjabi

For Punjabi language, only one POS tag set was publicly available. Singh Mandeep, Lehal Gurpreet, and Sharma Shiv, in 2008, proposed a Punjabi POS tagger, adapting a rule-based approach. The fine-tuned tagset contains around 630 tags, consisting of all the tags for the various word classes, word-specific tags, and tags for punctuations.

2.12.3 Telugu

There are three notable POS tagger developments in Telugu, For Telugu, three significant POS taggers were developed, based on three different approaches: rule-based, transformation-based learning, and Maximum Entropy based. An annotated corpus of 12,000 words was constructed. To train the transformation-based learning and Maximum Entropy based POS tagger models, an annotated corpus of 12,000 words was constructed. Rama Sree and others, in 2007, improved the existing Telugu POS tagger. In 2008, Rama Sree team developed a Telugu tagset with additional tags to deal with inflectional languages.

2.12.4 Malayalam

For Malayalam, two separate corpus-based POS taggers were developed: one was based on a stochastic HMM based part of speech tagger, developed by Manju's team in 2009 and another one was based on machine learning approach, developed by Antony and his team. In the above first one, a tagged corpus of about 1,400 tokens was generated using a morphological analyzer and trained using the HMM algorithm.

Antony's team, in 2010, developed the second one which was based on a machine learning approach. Training, testing, and evaluations for this were performed with Support Vector Machine (SVM) algorithms. They also proposed a new AMRITA POS tag set. Using this tag set, a corpus of about 180,000 tagged words was used for training the system. This work led to a tag set for Malayalam, consisting of 29 tags.

2.12.5 Kannada

A POS tagger for Kannada language was developed by Antony's team, using SVM, with 30 tags. The approach adapted for this was statistical based one. For training and testing the accuracy of the tag set, a corpus size of fifty-four thousand words was used. In 2010, another tag set containing 29 tags was proposed by Vijayalakshmi F Patil. In 2010, Antony PJ and others developed a tag set for Kannada with 30 tags.

2.12.6 Tamil POS Tagger

Vasu Ranganathan, from USA, proposed a Tamil POS tagger which was based on a lexical phonological approach. He used an index method for morpho tactics. This tool can deal with both tagging and generation. Ganesan and his team, CIIL Mysore, developed a Corpus and tag set. With the tag set, he used dictionary and a morphological analyzer as well.

Selvam and Natarajan, in 2009, with the help of statistical methods like alignment, projection and induction, attempted a rule-based morphological analysis and POS tagging in Tamil. Dhana Lakshmi’s team, Amrita university, developed two POS taggers for Tamil in 2009. They used their own tag set for this purpose.

2.13 Conclusion

This chapter examined previous research on corpus projects, corpus analysis tools, morphological parsing, and POS tagging. Works done in corpus linguistics together with the tools developed have been reviewed. It has also been indicated that very few corpus studies were concerned with the Tamil language or with the development of tools for Tamil corpus analysis. In the next chapter, the methodology of this research project is discussed.

CHAPTER 3

METHODOLOGY

3.1 Introduction

The literature review in the chapter 2 reveals, there have been no corpora available for written Tamil of the Malaysian variety to date. The current corpus project represents a pioneering effort in this direction. This chapter will start with a restatement of the research aims and questions. It will then explain the methodology that will be used to answer the research questions. It is to be mentioned here that the first research question is dealt with in the chapter 4; the second research question in the chapter 5 and the third question in the chapter 6.

3.2 Restatement of the Research

This project has three aims. The first is to construct a corpus of written Malaysian Tamil. Second is to develop a morphological parser and a POS tagger for the processing of the corpus and third is to design the relevant algorithm for the morphological analysis of the written Malaysian Tamil along with the noun and verb inflection charts based on the morphological parser and the POS tagger. These can be used for corpus linguistic studies on written Malaysian Tamil specifically and for further research on language technology.

Based on the above three aims stated, the following three research questions are raised:

1) What considerations were taken into account in the construction of WMTC?

2) What was involved in the development of a morphological parser and POS tagger?

3) How might a suitable algorithm be designed for developing the morphological parser and the POS tagger?

The current thesis project does not aim to develop or describe a written grammar of Tamil but to analyze the morphological structure of words according to the computational purposes. To this end, a representative and authenticated corpus for written Malaysian Tamil is warranted.

For this, the practices from modern corpus linguistics have been examined to inform the methodology. From the design of the corpus to the final analysis of morphological parsing and POS tagging, the study followed the methodology adopted in development of the BNC corpus and COCA corpus. Moreover, in developing the tools for corpus construction and analysis, the methods of developing corpus access tools such as WordSmith and AntConc are considered in the current project. Though the methodology followed in BNC and COCA projects for English were studied, this research adopted those methodologies to the local conditions of Tamil community in Malaysia. In this context it is to be mentioned that Tamil community of Malaysia is only seven percent (https://www.malaysia.gov.my/portal/content/30114?language=my,2017); also, Tamil is not an official language in Malaysia. Students in Tamil schools learn Tamil up to standard six in Malaysia.

These three factors make their impact on the corpus construction for written Malaysian Tamil in size, number of domains, and number of media. However maximum effort has been taken to make this corpus a representative and balanced one.

3.3 The Corpus Construction

Since the 1960s, for Western languages, especially English, several corpus studies have been undertaken. During these studies, various rigorous features have been put in place to ensure that the corpora constructed are reliable and authentic.

Now the questions that are of interest here are:

1. How should a corpus for a language be constructed?
2. What should be its design?
3. What are the features to be adopted in the construction of a corpus?

In response to the above questions, it should be noted that it is not a simple task to build a well-designed corpus. However, if the task is to build a corpus representative of contemporary written Tamil of Malaysia and if the corpus is balanced concerning the domains of language use, the corpus design involves a series of considerations., For example, the types of texts to be included, the number of texts, the selection of particular texts, the selection of text samples from within texts, and the length of text samples (Biber, 1993).

The corpus mentioned above generally concentrates on the following aspects:

1. How to build an authenticated corpus for a language to represent its use and usage in contemporary time?
2. What are the various linguistic aspects to be studied in corpus analysis?
3. What are the linguistic tools to be developed for the above analysis?
4. How far could the corpus tools be used for corpus study?

3.4 Features followed in corpus design

From the pioneer works on corpus design such as BNC and COCA, the following features are arrived at for written Malaysian Tamil for the present research project.

3.4.1.1 Gathering or collection of linguistic materials

a) General-purpose corpus Special-purpose corpus: Based on the purpose of the research project, the nature of the corpus should be defined clearly. The first question that has to be answered is: What is the nature of the corpus - is it a general-purpose corpus or a special-purpose corpus? This WMTC is a general-purpose corpus.

b) Text sources: After deciding the purpose of the corpus, the next question is: what are the sources from which the proposed corpus will be built? Are they from various sources such as textbooks, literature, scientific texts, mass media or from just one particular source such as only materials from the mass media, literature or textbooks? For the WMTC project, the materials are taken from various sources, not from a particular source.

c) Domains: Next, the domain(s) of the corpus materials - education, politics, economics, and/or culture - to be covered should be decided. The language used in different domains may be different in their lexicon and sentence structure depending on the discourse type of the materials. For example, in the Science domain, different discourse types may be found, such as textbooks, popular books, or scientific articles. The corpus for the current project covers sources from virtually all the domains mentioned above. The details of the domains covered here are provided in the next chapter on corpus design.

3.4.1.2 Nature of the corpus

The next question to be answered is: what are the proportions of samples or words in each domain and each media? Here, the concepts of "representativeness" and "balanced" of the corpus have to be given due consideration.

3.4.1.2.1 Representativeness

Although a corpus is a collection of authentic machine-readable texts (Biber et al., 1998; McEnery & Hardie, 2012a), it is different from a random collection of texts or an archive of machine-readable data in that it is sampled by explicit criteria to be representative of, for example, a particular language, language variety, or text type (Atkins et al., 1992; Biber, 1993; Tognini-Bonelli, 2001). At this juncture, it is necessary to revisit the notion of representativeness and discuss the techniques (i.e., balance and sampling) to achieve representativeness in corpus construction. In corpus linguistics, representativeness refers to the extent to which a sample includes the full range of variability in a population (Biber, 1993). Any selection of texts is a sample. An evaluation of the representativeness of a sample is subject to a definition of the population that the sample is intended to represent. Thus, representativeness is a fluid notion closely related to the purpose of the corpus or a collection of texts sampled in a principled way.

A general corpus is intended to be representative so that the corpus can be used as the basis for generalizations concerning a language variety as a whole. The first step towards building a representative corpus is to have a full definition of population. The researcher needs to distinguish the boundaries for the language to be represented. In other words, the sample of texts should be able to include all the variability of the texts falling between the boundaries defined. In corpus design, variability can be considered from situational and linguistic perspectives. A corpus design can be evaluated for the extent to which it includes the following: 1) the range of text types in a language and 2) the range of linguistic distributions in a language.

3.4.1.2.2 Balanced corpus

A balanced corpus covers a wide range of text categories that are supposed to be representative of the language (variety) under consideration (Leech, 2007), However, there is no scientific measure for balance. The proportions of different kinds of texts it contains are best estimated based on informed and intuitive judgments. As Biber (1993) suggests the acceptable balance is determined by the intended use, i.e., the research questions.

To obtain a representative sample from a population, the first concern to address is to define the sampling unit and the boundaries of the population have to be defined. The population is the assembly of all sampling units, while the list of sampling units is referred to as the sampling frame. For the current project, the corpus has been constructed to be representative and balanced of written Malaysian Tamil. To achieve this, designing features of the BNC corpus and COCA corpus are adopted. The comparison of these with written Malaysian Tamil is given below.

Table 3.1: BNC, COCA & WMT comparison

Divisions	BNC	COCA	WMT
Period	1960 - 1994	1990 - till date	1980 to 2020
Written Text (proposition)	90%	80%	100%
Published & unpublished(misc.)	Included	Not included	Not included
Web Materials	Not included	Included	Included
Journals & Newspapers	Grouped under periodicals	(blogs & pages only)	(blogs, pages, emails & social media materials)
Fiction & nonfiction	Separated	Separated	Separated
Language	British English	Only fiction	Only fiction

It can be inferred from the above table that most of the features of BNC and COCA are being followed in WMTC.

3.4.2 Sampling Method

After defining the population and the sampling frame, the next step is to decide the sampling method for selecting texts that are representative of the population. The selection should not be based on one's subjective outlook or biased preference.

In this context, the following suggestions are given by some corpus linguists (Hunston, 2002b) McEnery & Hardy, 2012; (Gatto, 2014).

It is an established idea that corpus representativeness and balance are closely associated with sampling. The sampling is needed to achieve balance and representativeness. "A sample is assumed to be representative if what is found for the sample also holds for the general population" (Manning & Schutze, 1999, p. 119). The sampling frame secures a sample which, subject to limitations of size, will reproduce the characteristics of the population, especially those of immediate interest, as closely as possible. A representative sample from a population has to define the sampling unit and the boundaries of the population. A sampling unit may be a book, periodical, or newspaper, and the population may be the assembly of all sampling units. The list of sampling units is a sampling frame (McEnery et al., 2006).

There are two sampling methods of interest for corpus development: (1) simple random sampling and (2) stratified random sampling (Biber, 1993). Simple random sampling is a basic sampling method wherein all sampling units within the sampling frame are numbered, and the sample is chosen using random numbers. While simple random sampling, if implemented, is expected to produce the most representative sample. It may exclude relatively rare linguistic items in the population as the chance of an item being chosen correlates positively with its frequency in the population even though these rare linguistic items are of interest to researchers (McEnery & Xiao, 2007). One solution to this problem is stratified random sampling, which first divides the whole population into relatively homogeneous groups called strata and samples each stratum randomly. The current project has adopted the stratified random sampling method.

3.4.2.1 Sample size and number of samples

There are two more decisions to be taken in selecting the samples: (1) sample size and (2) number of samples.

a) **Sample size:** The question to consider with written language is whether full texts or whole documents should be sampled or texts chunks. If text chunks are to be sampled, should the text initial, middle, or end chunks be sampled? Generally, itis better to sample text segments unless the intention of the corpus is to study textual organization or related features. This is because to use full texts, permission has to be sought from the copyright holders. According to Biber (1993), “frequent linguistic features are quite stable in their distributions, and hence short text chunks of about 2,000 running words are usually sufficient to study such features, while rare features are more varied in their distribution and thus require larger samples. In choosing samples to be included in a corpus, however, “attention must also be paid to ensure that text initial, middle and end samples are balanced.” (McEnery & Xiao, 2007, p. 20).

b) **The number of samples**: sampling issue that relates to stratified sampling specifically is the number of samples and proportion for each text category. For the corpus to be considered representative, the numbers of samples across the different text categories should be proportionate to their level of frequencies and/or weights viz a viz the target population Here, the quantum of the sources and domains in the total corpus play a crucial role.

c) Nevertheless, it should be noted that defining a target population and deciding on the proportions are difficult to determine objectively (Hunston, 2002b). Furthermore, the criteria used to classify texts into different categories or genres are often dependent on intuitions. As McEnery et al. (2007, p. 20) advise “the representativeness of a corpus... should be viewed as a statement of belief rather than fact.”

The present research project consists of 500 samples, and each sample contains about 2000 words/word forms/tokens. In this context it should be mentioned here that the size of Brown corpus which has 500 samples each of which contains 2000 words with 1million words in total is being followed for WMTC. The words in the samples of WMTC are presented in plain text format. The total number of word forms obtained for this present thesis project is around one million, that is, 10,03,298.

3.4.2.2 Metadata

Based on the corpus design features described above, the data entry may include every sample text with necessary metadata, providing information on the various linguistic, sociolinguistic, and other linguistic features. This type of metadata may help other researchers use the corpus for their own research goals or purposes.

In the present project, for each sample, all the necessary metadata have been provided in the sample entry form developed especially for this purpose, so that the data can be saved in the XML format. The preview and the samples entry form containing the metadata are provided below:

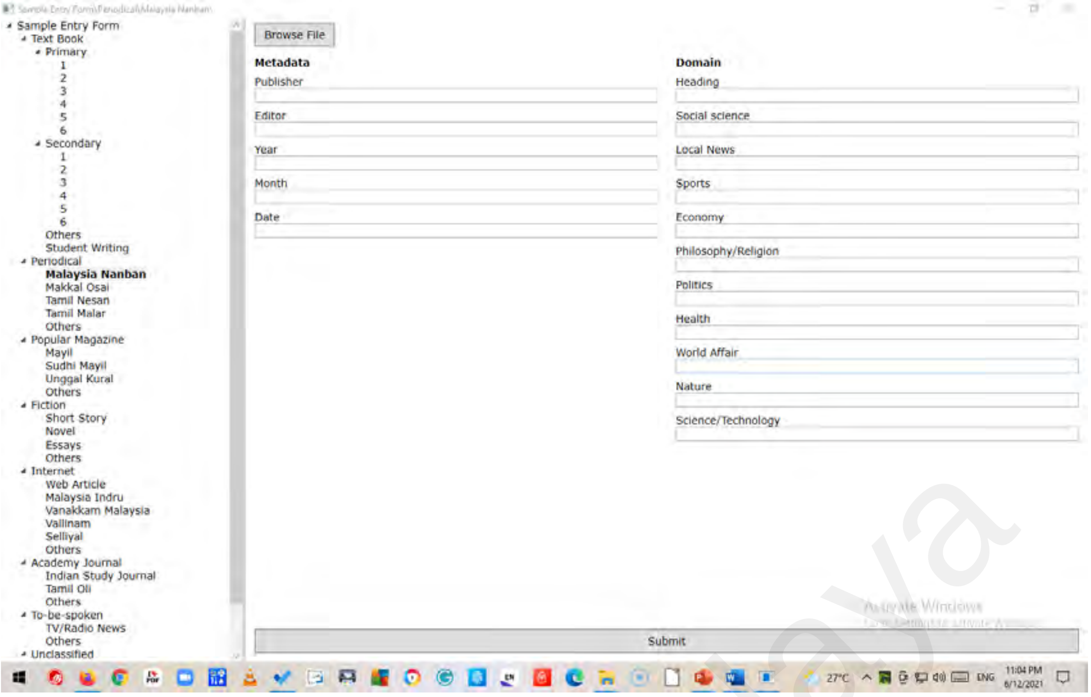


Figure 3.1: Sample entry form for periodicals

The preview and the samples entry form magazine are provided below:

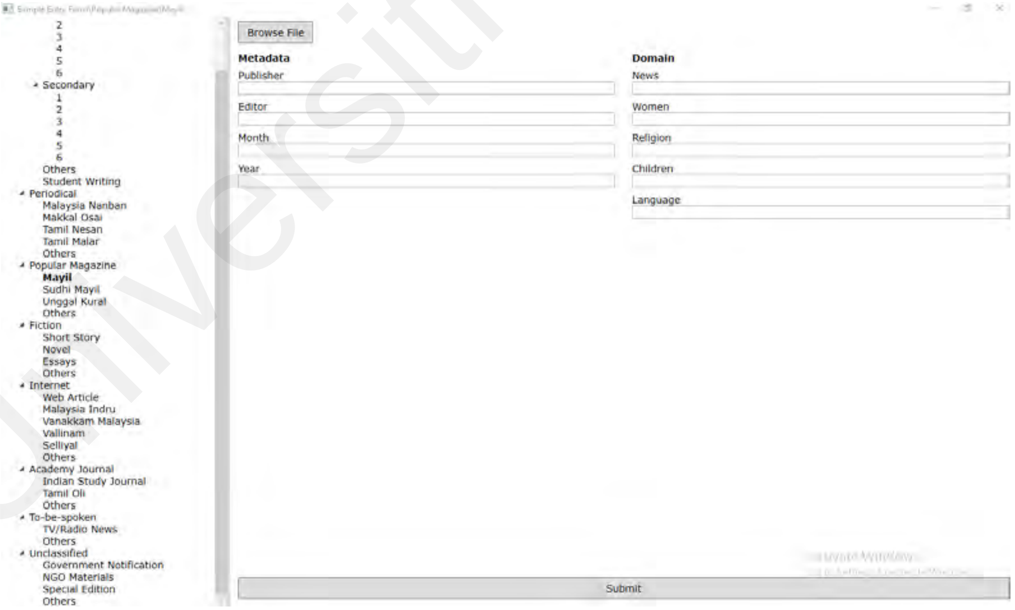


Figure 3.2: Sample entry form for magazines

The preview and the samples entry form textbooks are provided below:

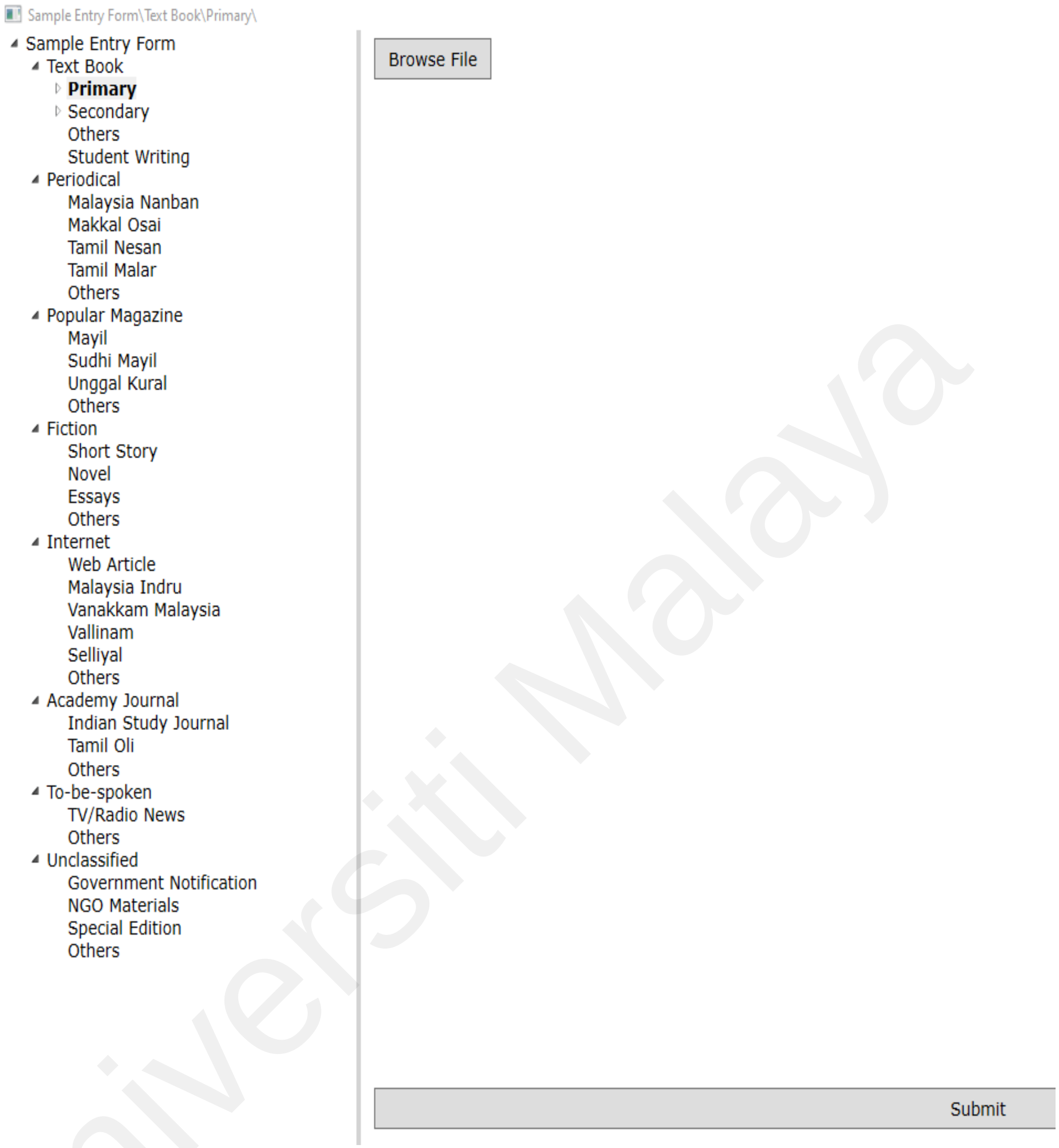
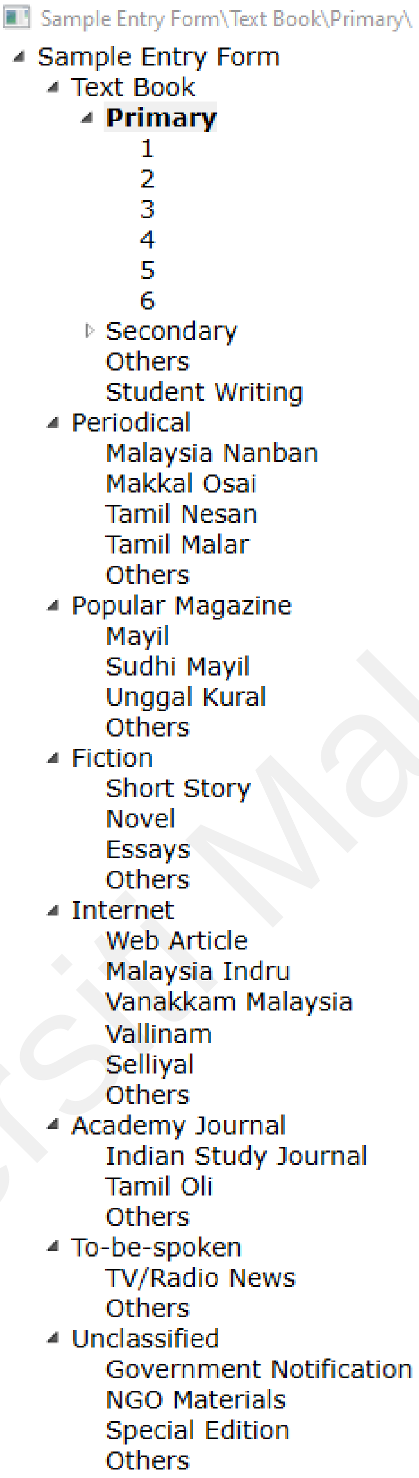


Figure 3.3: Sample entry form for textbooks



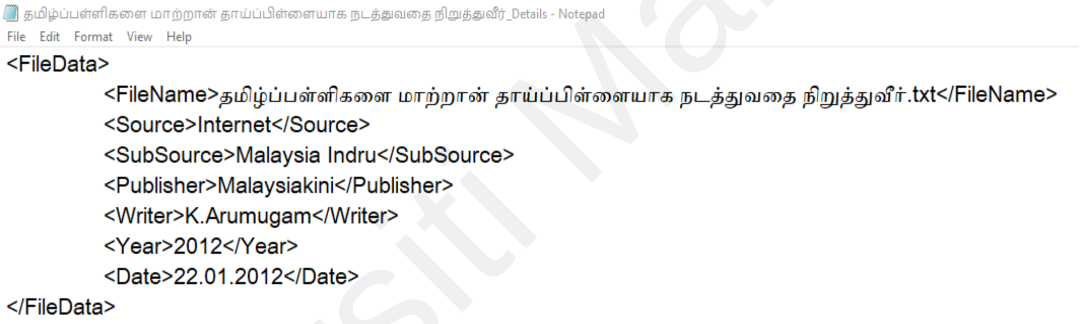
Sample Entry Form\Text Book\Primary\

- Sample Entry Form
  - Text Book
    - **Primary**
      - 1
      - 2
      - 3
      - 4
      - 5
      - 6
    - Secondary
    - Others
    - Student Writing
  - Periodical
    - Malaysia Nanban
    - Makkal Osai
    - Tamil Nesan
    - Tamil Malar
    - Others
  - Popular Magazine
    - Mayil
    - Sudhi Mayil
    - Unggal Kural
    - Others
  - Fiction
    - Short Story
    - Novel
    - Essays
    - Others
  - Internet
    - Web Article
    - Malaysia Indru
    - Vanakkam Malaysia
    - Vallinam
    - Selliyal
    - Others
  - Academy Journal
    - Indian Study Journal
    - Tamil Oli
    - Others
  - To-be-spoken
    - TV/Radio News
    - Others
  - Unclassified
    - Government Notification
    - NGO Materials
    - Special Edition
    - Others

Figure 3.4: Textbooks options of the standards

Figures 3.1- 3.4 show the metadata containing the sources such as textbooks, periodicals, popular magazine, fiction, internet, to-be-spoken, unclassified along with the details about the author, publisher, month, year and date; the domains such as heading, social science, local news, sports, economy, philosophy, politics, health, world affair, nature, science and technology; Also figures 3.3 and 3.4 depict the options of the Malaysian school standards, that is, primary (standards 1-6) and secondary (forms 1 to 6).

Figure 3.5 shows the final XML format of the metadata featuring the file name, source, sub source, publishers, author, year and date.


தமிழ்ப்பள்ளிகளை மாற்றான் தாய்ப்பிள்ளையாக நடத்துவதை நிறுத்துவீர்_Details - Notepad
File
Edit
Format
View
Help

```
<FileData>
        <FileName>தமிழ்ப்பள்ளிகளை மாற்றான் தாய்ப்பிள்ளையாக நடத்துவதை நிறுத்துவீர்.txt</FileName>
        <Source>Internet</Source>
        <SubSource>Malaysia Indru</SubSource>
        <Publisher>Malaysiakini</Publisher>
        <Writer>K.Arumugam</Writer>
        <Year>2012</Year>
        <Date>22.01.2012</Date>
</FileData>
```

Figure 3.5: XML file data

3.4.2.3 Conversion into machine-readable texts

The selected materials for the proposed corpus must be machine-readable electronic texts. The selected materials may be in three forms: (1) web or digital materials, (2) printed materials, (3) handwritten materials. Since the materials in the corpus should be machine-readable texts, the above-mentioned printed materials should be converted to digital forms. It can be done in two ways: (1) manual keying and (2) scanning. Electronic scanners may be used to convert the printed materials into computer image files which may be converted into machine-readable texts using Optical Character Recognizer (OCR) software available for the particular language.

For this project, all the means mentioned above are used to get the sample texts in the necessary plain text format, avoiding the proprietary software format commands. Here, it is worth noting that the Google OCR software has been essential in converting scanned images of the printed materials selected for this project.

3.5 Cleaning of the Corpus

The linguistic analysis tools such as morphological parser and syntactic parser developed for the corpus analysis could work only with the correct spelt words and correct grammatical sentences.

The texts in the selected samples may not be perfect in their spelling and grammar because they are selected from various sources, including web materials. So, all the samples selected should be checked for the correct spelling. Since the corpus may contain millions of words in size, it is impossible to do this spell-checking process manually. The automatic spellchecker available for the specific language may be used for this purpose.

Tools for corpus cleaning are language specific. For the present project, Tamil spellchecker proofing tool in the software 'Mentamizh' (Appendix A) is used to correct and clean the texts.

3.5.1 Normalization

In the corpus analysis, first, the token and type selections are the most important. In token selection, there may be some problems in deciding whether two are single tokens or one word consists of two tokens. For example, is the English utterance “according to” two tokens or a single token? Likewise, is the word “won’t” a single token or two tokens (“will” and “not”) has to be decided. This kind of normalization should be done for type selection for further corpus analysis.

Also, there may be some problems in identifying whether two tokens belong to a single type or two different types. For example, in the two sentences, “He is walking” and “Walking is good”, “walking” in the first sentence is a verbal participle whereas it is gerundial in the second one. They are two different types, though in the phonological form they are similar. This kind of normalization could be done with the help of a concordance tool with manual intervention.

For the current project, the issues mentioned above for English have also been faced. When such issues occurred, morphological parsing, POS tagging, and concordance were all used. This is explained in Chapters 5 and 6.

3.6 Tokens and Types

Once the corpus samples are ready for analysis through the computer, the tokenizer programme may be called for to pick up all the tokens in the samples. There may be repetitions of tokens and some ambiguous tokens. Here, it should be pointed out that before sending the sample text for further linguistic analysis, these repetitions and ambiguities should be resolved. The tokens with the same meaning and grammatical category are grouped into a single “type”. Here, the human intervention with the help of the “concordance” programme would do this job. Based on the “linguistic context” of the particular token, the ‘type’ selection process is completed. Based on the tokenization task, the types/word forms are selected for further morphological analysis and POS tagging in the project.

3.6.1 Development of Morphological parser and POS tagger for written Malaysian Tamil

After the type of selection process is completed, all the types are sent for morphological parsing, and POS tagging with the help of the morphological parser developed based on the available Tamil grammars. No on-the-shelf morphological parser is available for the above process of parsing and tagging. While the method adopted for the present morphological parser is elaborated in Chapter 6, the salient features are first briefly presented here.

The four components behind the morphological parser are (1) Tamil Lexicon, (2) Grammatical affixes, (3) Morphotactics, and (4) Morphophonemics. A detailed flowchart containing the word structure prepared for this morphological parser can be found in Chapter 6. The current morphological parser is based on the left to right parsing order.

3.6.2 Initial Morphological parser and POS tagger

At the outset, it should be noted that the kinds of POS and its numbers are mostly language-specific though there are some common POS in all languages. A detailed discussion on Tamil POS is presented in the chapter 6 on Tamil morphological parsing and POS tagging. All the POS tags are based on Computational Tamil morphology and will be explained accordingly.

For the present project, initially, a morphological parser and POS tagger are constructed, based on existing Contemporary Written Tamil grammars such as that authored by Malaysia S. Seeni Nainaa Mohammed, Singapore Siddhartha, Prof. S. Agesthialingom, Prof. Porko, and Prof. K. Karunakaran (Tamil Nadu). Of course, the ancient grammar treatises Tolkappiyam and Nannul are the basis for the above-mentioned grammatical books.

Here, it should be pointed out that contemporary written Tamil is a continuation of the previous written Tamil. Also, there is no fundamental difference in the grammatical structures of the traditional and modern Tamil. Some grammatical features of the old Tamil have disappeared while, some new grammatical features have emerged. Likewise, there are changes in the Tamil lexicon. These changes are necessary to develop a morphological parser and POS tagger for written Malaysian Tamil, which is one of the objectives of the current project.

All the types found in the constructed corpus are sent to the above morphological parser and POS tagger. However, it is found that many words could not be handled by this processor because of the presence of new lexicons, grammatical suffixes, and new morphotactic rules, in addition to some new morphophonological rules.

The following figure (Figure 3.6) shows the morphological parsing process. As illustrated, the words are first input into the morphological parser. Then, their roots are checked and once this is completed, they are sent for affix stripping. After the affixes have been stripped, they are checked to ascertain if they have accomplished morphophonemic rules and then they are sent for tagging and the final output with the parsed word is obtained.

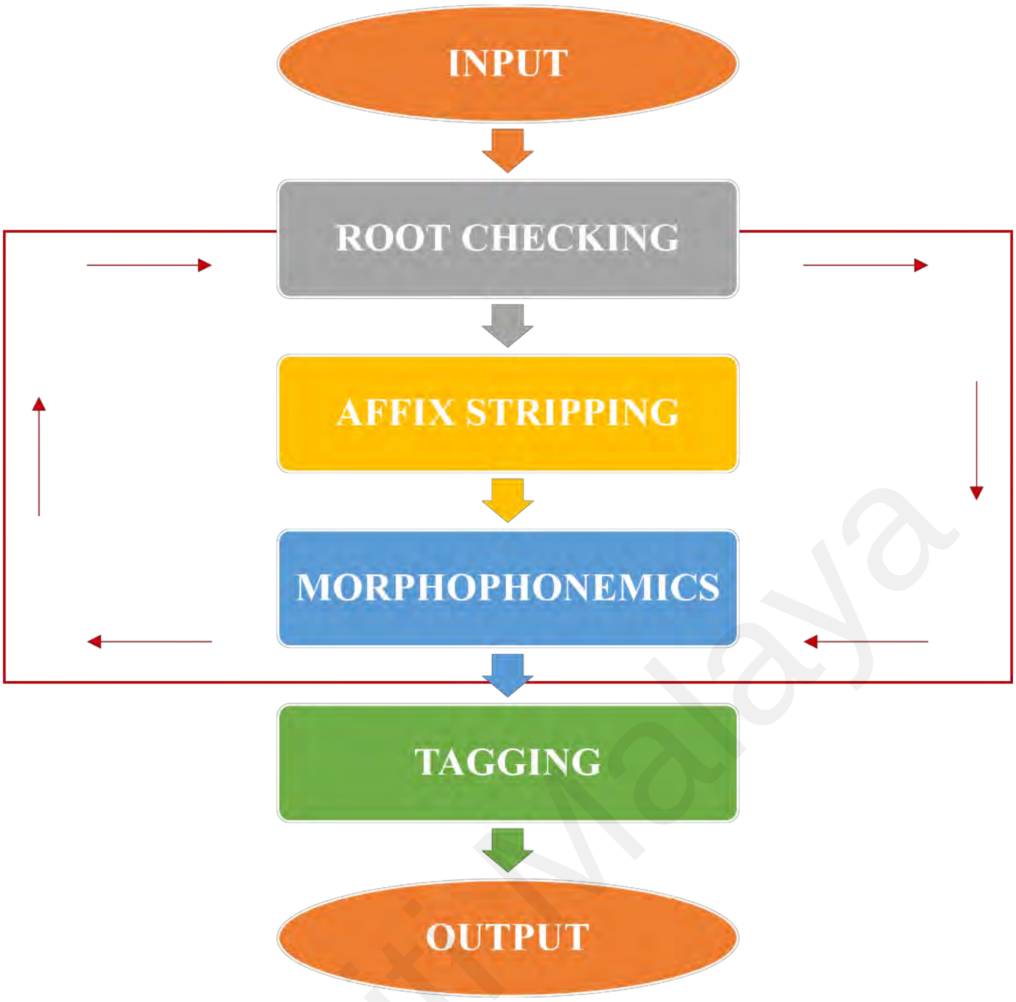


Figure 3.6: The flow chart of the morphological parsing process

3.7 Enrichment of the parser and POS tagger

Based on the parsing and tagging described, the new developments in contemporary written Tamil are taken care of. By accommodating those new grammatical features as well as the new lexicons, the morphological parser and POS tagger have been enriched and capable of handling all the types found in the corpus samples.

After the enrichment of the morphological parser, all the types again are sent for parsing and tagging. Based on the 'Precision and Recall' evaluation techniques, the efficiency of the new parser is enhanced to accommodate all the morphological features of written Malaysian Tamil. After getting the parsed output with the POS tagging, all the tokens/word forms are morphologically annotated.

The schematic representation of final Morphological Processing is illustrated in Figure 3.7 below.

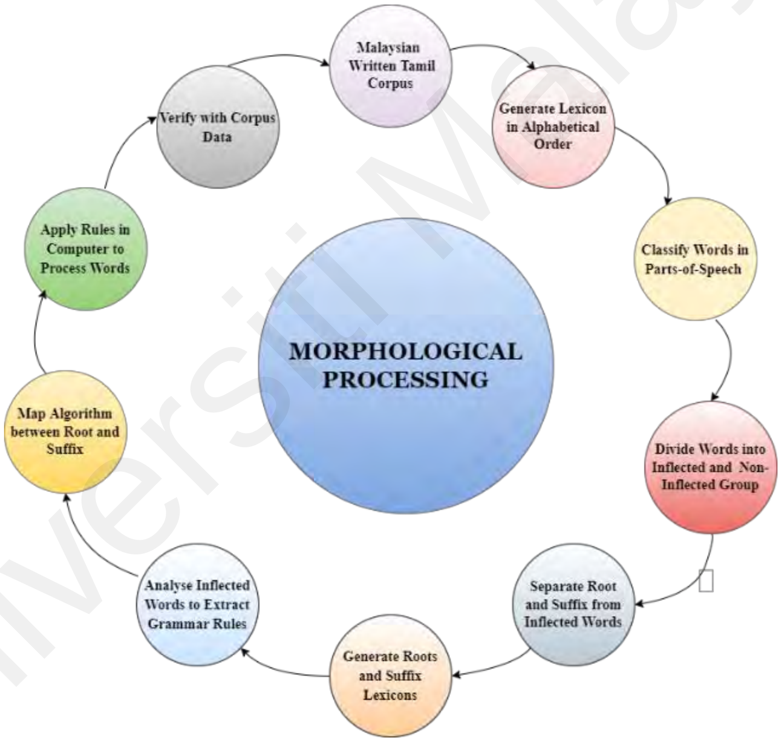

Malaysian Written Tamil Corpus
Generate Lexicon in Alphabetical Order
Classify Words in Parts-of-Speech
Divide Words into Inflected and Non-Inflected Group
Separate Root and Suffix from Inflected Words
Generate Roots and Suffix Lexicons
Analyse Inflected Words to Extract Grammar Rules
Map Algorithm between Root and Suffix
Apply Rules in Computer to Process Words
Verify with Corpus Data
MORPHOLOGICAL PROCESSING

Figure 3.7: Morphological Processing

The samples are input into a morphological parser. Then the words are divided into inflected and non-inflected word forms. The non-inflected word forms do not undergo any further processing since they are not ambiguous, the inflected word forms are parsed into their roots and suffixes and are sent to morphotactic and morphophonemic rule units for verification and in turn, to the tagging process, resulting in the well-parsed output. The backdrop of the developing the morphological parser and POS tagger is, of course, the Tamil computational morphological algorithm which is explained in detail in the chapter 6.

3.8 Conclusion

This chapter has discussed the methodology used to address the three objectives of this thesis. It has described the methodology adopted to develop the corpus and then explained how the morphological parser and the POS tagger along with other analysis tools for the corpus. The next chapter will present a discussion of the considerations involved in the development of WMTC.

CHAPTER 4

MALAYSIAN TAMIL CORPUS DEVELOPMENT

4.1 Introduction

This chapter presents information about the corpus that has been constructed based on the methods described in the preceding chapter. There are four stages in corpus compilation: corpus design, text collection, text encoding and mark-up, and storage (Adolphs, 2008; Kennedy, 1998). This chapter will consider these four stages. It discusses the features of corpus construction, that is, size, representativeness, balance, and sampling with a particular reference to BNC. Then it describes how these design features are adapted to build the WMTC and what considerations have been taken into account in the construction of WMTC.

4.2 Nature of the Corpus

As already explained in the previous chapters, the first of the three objectives of this research project is to develop a written Malaysian Tamil corpus. The corpus contains samples of texts produced in the last forty years.

According to Kennedy (1998), a corpus is “systematic, planned and structure compilation of text” and a database “designed and structured specifically to be used for linguistic description and analysis”.

Since corpus linguists aim at studying the various linguistic patterns of a language being used in different contexts in that particular speech community, they strive to build some corpora from the authentic texts of that language.

It is a known fact that the language use of the speech community is dependent on various non-linguistic factors such as age, education, profession, and domain and medium. These factors lead to differences found in the language patterns of the language of that speech community. Because of this relation between language structure and language use, in the language analysis, the authentic texts get their due importance in the construction of the corpus for that language. In constructing a general corpus of a particular language, corpus linguists give much importance to the size, representativeness, and balance of the corpus (Biber, 1993).

These three are the external criteria followed in the construction of any corpus. The design of WMTC has drawn on the designs of two significant corpus projects - BNC and COCA. The design features of these corpus projects have been adapted to construct the WMTC.

4.2.1 BNC and COCA projects

As it is noted, the BNC and COCA are two important corpora in the field of corpus linguistics. The BNC is a reference corpus of late 20th century British English, which consists of 100 million words of British English from 1991-1995.

COCA contains more than one billion words of data, including 20 million words each year from 1990-2019. The COCA is one of the world's widely used corpora. In early 2020, the scope and size, and features of COCA were expanded to make it even more useful for researchers, teachers, and learners with the same genre balance year by year. Thus, COCA is the monitor corpus of English that is large, recent and has a wide range of genres.

4.2.2 Tamil Language Corpora

The Tamil language used in different countries varies because the language use of the speech community is dependent on various non-linguistic variables such as age, education, profession, domain, and medium. These factors lead to differences found in the language patterns of the language of that speech community. The differences between varieties of language are best documented by the authenticated actual language use of the respective communities.

For Tamil used in Tamil Nadu, some corpus projects have been undertaken by some of the universities and research institutions such as Tamil Virtual Academy (Chennai), AU-KBC Centre of Anna University (Chennai), and Central Institute of Indian Languages (Mysore). For Singapore Written Tamil, the Ministry of Education (Singapore) has built a corpus. In Sri Lanka, Moratuwa University (Colombo) has been involved in building a corpus for Written Tamil.

However, no corpus is available for written Malaysian Tamil. So, there is a need to construct a corpus for contemporary written Tamil of Malaysia that comprises natural language use or authenticated texts produced by Tamil language users in their real-life contexts in Malaysia. In view of this, this research project is undertaken to accomplish this task.

4.3 Corpus Construction

In this section, the details of the construction of the WMTC are discussed. As noted earlier, the following are the four stages to compile the present WMTC:

1. corpus design
2. text collection
3. text encoding and mark-up
4. storage

4.3.1 Design of the Corpus

The size of the present corpus and the selection of samples are meticulously planned in order to make this WMTC representative of written Malaysian Tamil and a balanced one. For this, mostly the features of BNC have been adopted. As McEnery et al., (2012b, p. 250) observed that:

> If a corpus is to be claimed as a representative of a particular language, it should contain all the types of text, in the correct proportions, that are needed to make the contents of the corpus an accurate reflection of the whole of the language or variety that it samples.

In the construction of the WMTC, due attention was paid to make the corpus a representative and a balanced one.

4.4 WMTC

In the present project, corpus design is developed based on the features adopted in the BNC and COCA projects.

However, the WMTC is developed by the present researcher exclusively. Also, it is the first one of its kind for Malaysian Tamil. In comparison with the above two corpora projects, in size and time, the present project is clearly a small one. However, it incorporates the relevant features of the above two mega projects.

Here it is to be mentioned that the population of Malaysia is 32 million in which the Tamil population is only 2.6 million only (Department of Statistics, 2020). The number of journals, textbooks, periodicals, fictions published in Malaysia are very much limited in comparison with that of the state of Tamil Nadu. There are only four daily newspapers, and about ten are weekly and monthly journals in Malaysia.

In schools, only up to the primary school level, the Tamils have Tamil medium education for subjects. After that, only as a language subject, Tamil prevails. In university education, only Tamil and Tamil Linguistics subjects are in the Tamil medium.

It should be noted that in contrast to BNC and COCA, the written Malaysian Tamil exclusively contains samples from written Tamil works. Overall, the construction of the written Malaysian Tamil has drawn on the features of the design of BNC taken into consideration for the present written Malaysian Tamil. The size of the written Malaysian Tamil is 1 million words. It consists of 500 samples and each sample contains approximately 2,000 words. That is, not all the samples have the same number of words.

The total number of words compiled for this present thesis project is around one million, that is 10,03,298.

4.4.1 Features for corpus design adopted

First, the present project pays its attention to the gathering or collection of language materials to be included in the present corpus.

The present project is designed in the light of BNC and COCA projects, as a general-purpose corpus one and not a special corpus one. So, the language samples are collected almost from all the communicative domains of Tamil speech community in Malaysia. Regarding media, both BNC and COCA corpora have also included a spoken component. But this present project, as noted in Chapter 1, does not contain spoken Tamil sample. Other than this, all other aspects such as purpose, time, sources or media of data, the domains and the proportion of the media, contained in the BNC and COCA projects are consulted during the collection of samples.

The BNC project, as a synchronic one, contains 1 billion orthographic words whereas the COCA, as a monitor corpus, contains 5 billion orthographic words in total. The present WMTC contains 1 million orthographic words only, restricted to 40 years (from 1980 to 2020). Given the size of WMTC that is similar to the Brown Corpus, the same sampling size, that is, 500 sample texts with 2000 words in each one of them is adopted.

4.5 Representativeness, Balance, and sampling

4.5.1 Representativeness

As it is well-known, a corpus should be a collection of well-organized samples to achieve the desired research results. It should be well represented and balanced one of the particular languages for which the corpus is constructed. In other words, balanced and sampling must be considered to have representativeness.

A corpus is known to be a collection of machine-readable authentic texts (Biber et al., 1998; McEnery & Hardie, 2012a). However, a corpus is different from a random collection of texts or an archive of machine-readable data; it is sampled by explicit criteria to be representative of, for example, a particular language, language variety or text type (Atkins et al., 1992; Biber, 1993; Tognini-Bonelli, 2001). Before moving on to the next section on the construction of the contemporary written Tamil corpus, first it would be necessary to revisit the notion of representativeness and discuss the techniques (i.e., balance and sampling) to achieve representativeness in corpus construction.

In corpus linguistics, representativeness refers to the extent to which a sample includes the full range of variability in a population (Biber, 1993). Any selection of texts is a sample. An evaluation of the representativeness of a sample is subject to a definition of the population that the sample is intended to represent. Thus, representativeness is a fluid notion closely related to the purpose of the corpus or a collection of texts sampled in a principled way.

A corpus could be thought of as a "representative of the language variety it is supposed to represent if the findings based on its contents can be generalized to that language variety" (Leech, 1992). The representativeness of most corpora is to a great extent determined by two factors:

1. Balanced - the range of genres included in a corpus
2. Sampling - how the text chunks for each genre are selected

For Berber-Sardinha (1996), a representative corpus includes the majority of the types in the language as found in a comprehensive dictionary. According to Biber (1993, p. 256) "the compilation of a representative corpus should proceed in a cyclical fashion". In addition to text selection criteria, Hunston (2002a) suggests that another aspect of representativeness is change over time. The relevance of permanence in corpus design actually depends on how one views a corpus.

In this context, it is apt to mention Biber's (1993) concept of representativeness. For him, a corpus is representative, only if it fully captures the variability of a language. However, McEnery et al. (2012b) have said that the above is yet to be adopted in practice.

It is to be noted that a linguistic corpus is not a random compilation of a large number of texts, but the individual pieces of language need to be selected so that they fulfil a particular function, that is, they can be regarded as representative of the whole - an entire language or a specific variety or subset of it like academic journal articles.

In other words, although a corpus is only a subset of what it is supposed to represent, its function is to mirror the whole in such a way that linguists can use it to say something about the language variety that was sampled; observations on the basis of corpus data are generalized back to a whole from which a corpus was initially selected.

It is important to note that not every corpus is suitable for all types of linguistic analysis. It is therefore very important to know the characteristics of a corpus before using it. The interpretation and application of the research are very much dependent on this aspect (Hoffmann, 2008).

4.5.1.1 The representativeness of general and specialized corpora

There are two broad types of corpora in terms of the range of text categories represented in the corpus: general and specialized corpora. General corpora typically serve as a basis for an overall description of a language or language variety. The British National Corpus for example, is supposed to represent modern British English as a whole. In contrast, specialized corpora tend to be domain or genre specific. The representativeness of a general corpus depends heavily on sampling from a broad range of genres.

A general corpus is intended to be representative so that the corpus can be used as the basis for generalizations concerning a language variety as a whole. The first step towards building a representative corpus is to have a full definition of population. The researcher needs to distinguish the boundaries for the language to be represented. In other words, the sample of texts should be able to include all the variability of the texts falling between the boundaries defined. In corpus design, variability can be considered from situational and from linguistic perspectives. A corpus design can be evaluated for the extent to which it includes: 1) the range of text types in a language, and 2) the range of linguistic distributions in a language.

4.5.2 Balanced corpus

The representativeness of a general corpus depends on how balanced the corpus is, that is, the range of text categories included in the corpus. The acceptable balance of a corpus is dependent upon its intended uses.

A balanced corpus usually covers a wide range of text categories which are supposed to be representative of the language or language variety. The text categories must be sampled proportionally. That is, "…it offers a manageably small-scale model of the linguistic material which the corpus builders wish to study" (Atkins et al., 1992, p. 6).

According to Sharoff (2003), the "BNC is generally accepted as being a balanced corpus. The BNC model has been followed in the construction of a number of corpora like American National Corpus, Korean National Corpus, Polish National Corpus and Russian Reference Corpus". That is why the present research project also followed the BNC model.

Since the BNC is designed to represent the whole of contemporary British English, the sole aim of using the above text selection criteria was to have a balanced selection under each text category (Aston, 1998).

Balance appears to be a more important issue for a static sample corpus than for a dynamic monitor corpus. As corpora of the latter type are updated frequently, it is usually "impossible to maintain a corpus that also includes text of many different types, as some of them are just too expensive or time consuming to collect on a regular basis" (Hunston (Hunston, 2002a, pp. 30-31).

Like corpus representativeness, balance is an important issue for corpus creators, corpus users and readers of corpus-based studies alike.

However, there is no scientific measure for balance. The proportions of different kinds of texts it contains are best estimated based on informed and intuitive judgements. The acceptable balance is determined by the intended use, i.e., the research questions. To obtain a representative sample from a population, the first concern to be addressed is to define the sampling unit and the boundaries of the population. The population is the assembly of all sampling units while the list of sampling units is referred to as sampling frame. Based on the above discussion, the sampling frame for the present WMTC is also predetermined as in BNC.

Regarding this, Biber (1993, pp. 7-8) noted the following:

> Some of the first considerations in constructing a corpus concern the overall design: for example, the kinds of texts included, the number of texts, the selection of particular texts, the selection of text samples from within texts, and the length of text samples. Each of these involves a sampling decision, either conscious or not.

McEnery et al. (2007) commented that in order to have a representative sample from a population, the prime concern to be addressed is to define the sampling unit and the boundaries of the population. For instance, it can be said that a sampling unit of written text may be a book, periodical or newspaper. The population is the assembly of all sampling units while the list of sampling units is referred to as sampling frame.

McEnery et al. (2012b, p. 250) further added that:

> A sampling frame is a definition, or set of instructions, for the samples to be included in a corpus. A sampling frame specifies how samples are to be chosen from the population of text, what types of texts are to be chosen, the time they come from and other such frames. The number and length of the samples may also be specified.

In Biber's (1993) view, a sampling frame is an operational definition of the population, an itemized listing of population members from which a representative sample can be selected.

4.5.3 Sampling Method

After defining the population and the sampling frame, the next step is to decide the sampling method for selection of texts that are representative of the population. **"**Corpus representativeness and balance are closely associated with sampling. Given that we cannot exhaustively describe natural language, we need to sample it to achieve a balance and representativeness which match our research question". According to Manning & Schutze (1999, p. 119) "sample is assumed to be representative if what we find for the sample also holds for the general population".

In order to obtain a representative sample from a population, the first concern to be addressed is to define the sampling unit and the boundaries of the population. For Malaysian written text, the book, the periodical or the newspaper is the sampling unit. The population is the assembly of all sampling units while the list of sampling units is referred to as a sampling frame.

4.5.4 Sampling Techniques

There are different sampling techniques followed to choose a sample which must be representative of the population. There are two sampling techniques for corpus development: simple random sampling and stratified random sampling.

A simple random sampling is a basic sampling technique. In this technique, the sampling units within the sampling frame are numbered and the sample is chosen randomly from these numbers. However, a simple random sampling may generate a sample that does not include relatively rare items in the population which may be crucial for researchers (McEnery et al., 2006).

One solution to this problem is stratified random sampling. Here in this sampling the whole population is divided into homogeneous groups and each stratum is sampled at random. Biber (1993) observes that a stratified sample is never less representative than a simple random sample. Also, even a coverage of full texts as sample may not sometimes bring out "the peculiarity of an individual style or topic may occasionally show through into the generalities" (Sinclair & Sinclair, 1991, p. 19).

According to Biber (1993), "frequent linguistic features are quite stable in their distributions and hence short text chunks, around 2,000 words are usually sufficient for the study of such features while rare features are more varied in their distribution and thus require larger samples".

In selecting samples to be included in a corpus, however, attention must also be paid to ensure that text initial, middle, and end samples are balanced. WMTC followed Biber's suggestion and is, as noted earlier, similar to the sampling design of the Brown Corpus.

Out of 1 million WMTC words under each category the number of words selected is 10-20%, which is 2000 words in each 500 samples. The above discussion suggests that in constructing a balanced and representative corpus, stratified random sampling is preferred over simple random sampling. For the present, for WMTC the Stratified Random Sampling method is adopted.

The following 8 types are considered as different Strata and samples are drawn from each strata: Textbook, Periodical, Popular Magazine, Fiction, Internet, Academy Journal, To-be-spoken, Unclassified. Here the following are some of external criteria adopted in choosing the texts and selecting the samples: Time Period, Population, Members of the Population, Selection of texts, and Selection of Samples.

4.6 Corpus Design -Three independent criteria in selecting samples

It is to be mentioned that if a corpus is claimed to be balanced one:

> The relative sizes of each of its subsections have been chosen with the aim of adequately representing the range of language that exists in the population of texts being sampled. (McCarthy & McCarten, 2012, p. 250)

To make the present Corpus as a 'representative, balanced' one, the following 3 independent criteria are adopted in the selection of samples (Hoffmann & Rayson, 2008).

(a) Time Period

(b) Domain

(c) Medium

Time is concerned with the publication date of the text materials. Domain is concerned with the subject field or broad topic area and Medium is concerned with the type of application in which they appeared. With some corpora, after the construction of corpus, some more descriptive features such as age, sex, domicile and type of authors, age and sex of targeted audience are being added to the corpus. With the BNC, independent work was carried out to categorize the material further according to genre (Lee, 2001).

4.7 Corpus Design of written Malaysian Tamil

Based on the above, for the present corpus, every effort has been taken to contain sample texts which could represent written Malaysia Tamil, including its language variability attested especially at the morphological level due to non-linguistic variables such as contents of the samples - imaginative or informative and medium of the samples - books or periodicals.

4.7.1 Domain, Time Period and Medium for WMTC

To make the WMTC as a ‘representative, balanced’ one, the following three independent criteria - time, domain and medium. -: as already mentioned, are adopted in the selection of samples (Hoffmann, 2008).

4.7.1.1 Time Period

The BNC project covered the years from 1974 to 1993. The COCA project as a monitor one, started from 1990 and has been enriched every year. The present WMTC project - a synchronic corpus - covered 40 years, spanning the period between 1980 and 2020. The years covered with their percentage: 1980 - 2000 (20%); 2001 - 2010 (30%); 2011 - 2020 (50%).

4.7.1.2 Domain

The domains selected for this sample are the following: Imaginative (20%), Science and Technology (10%), World affairs (10%), Family and Society (20%), Commerce and Finance (10%), Arts (15%), Belief and Thought (5%), Leisure (10%).

The BNC project considered nine domains: (1) Imaginative (2) Natural & pure Science (3) Applied Science (4) Social Science (5) World affairs (6) Commerce and Finance (7) Arts (8) Beliefs and thought (9) Leisure whereas the COCA project is also consisted of almost all the above domains. However, the domains depend upon the medium selected. For example, the domains covered under magazines may be different from the domains of the academic journals.

The present WMTC project covers eight domains: (1) Imaginative, (2) Arts, (3) Beliefs and thought, (4) Science and Technology, (5) World Affairs, (6) Commerce and Finance, (7) Family and Society, (8) Leisure.

4.7.1.3 Medium

The written text types that are used in the context of Malaysia are listed as follows:

i. Textbook
ii. Periodical
iii. Popular magazine
iv. Fiction
v. Internet
vi. Academy journal
vii. To-be-spoken
viii. Unclassified.

The media selected are the following: Textbook (10%), Periodical (20%), Popular Magazine (20%), Fiction (10%), Internet (10%), Academy Journal (10%), To-be-spoken (10%), Unclassified (10%).

The BNC project covered five media. They are: (1) Book (2) Periodical (3) Misc. published (4) Misc. unpublished (5) To-be-spoken. The proportion among the media are not same in BNC.

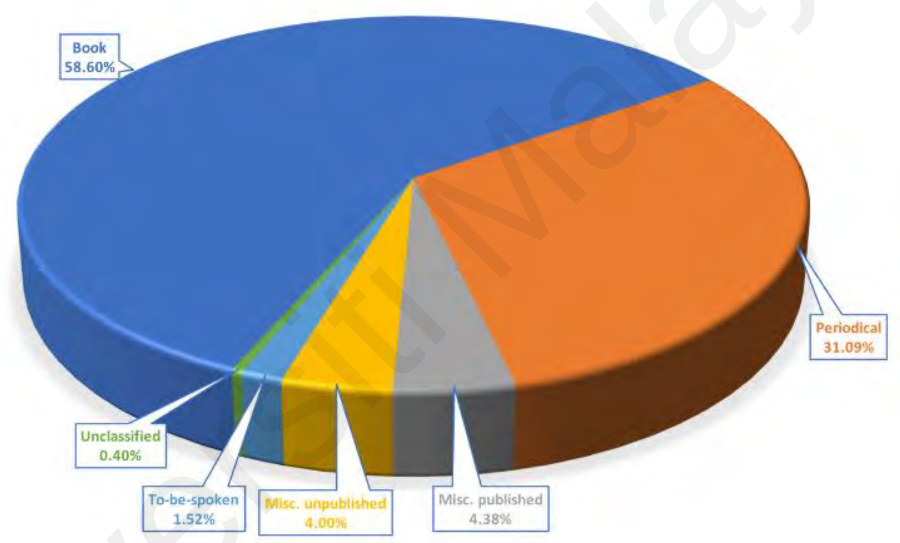


Figure 4.1: BNC - Proportion of Medium

The COCA project covered (1) Spoken (2) Fiction (3) Popular magazines (4) Newspaper (5) Academic (6) Web (7) Blog (8) TV/ Radio. The COCA project gave equal proportion to all the media.

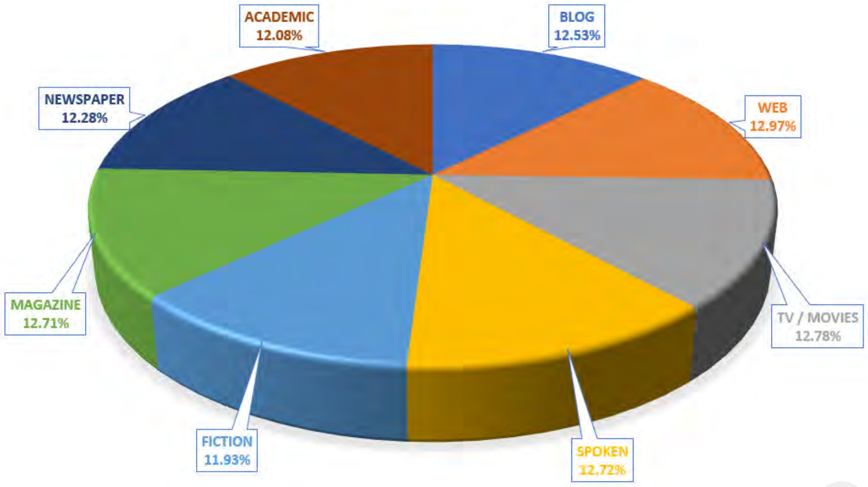


Figure 4.2 COCA - Proportion of Medium

The present project WMTC covers (1) Textbook, (2) Periodical (3) Popular magazines (4) Fiction (5) Internet (6) Academic journals (7) To-be-spoken (8) Unclassified. The proportion among the media is almost equal. All these text types are sampled for the construction of the WMTC.

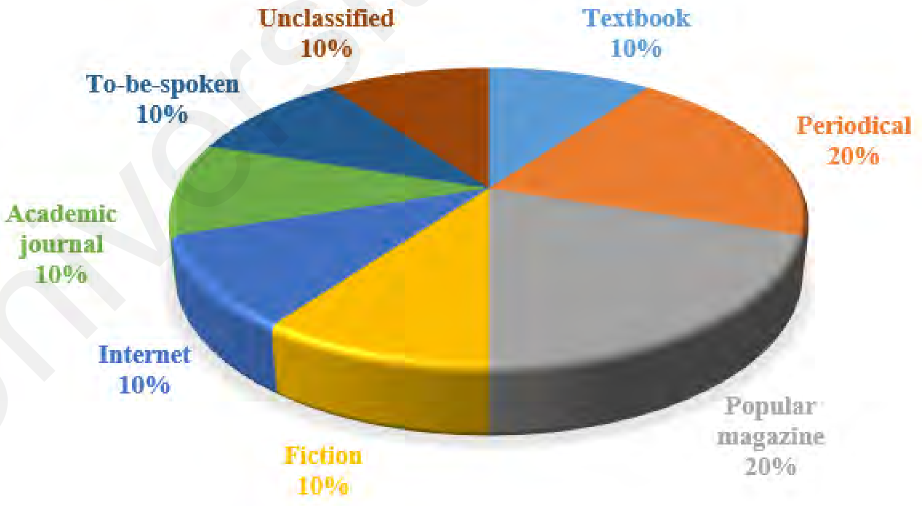


Figure 4.3 WMTC - Proportion of Medium

All the media from which this collection is done are Written and Web/blogs only. Wherever necessary, in accordance with Malaysian regulations concerning copyright, consent was obtained from the parties holding the copyright on the relevant data.

4.7.2 Similarities and differences between BNC and WMTC

Table 4.1: Similarities and differences between BNC and WMTC

	British National Corpus	Written Malaysian Tamil Corpus
Domain	Imaginative (19%)	Imaginative (20%)
	Arts (7%)	Arts (15%)
	Belief and thought (3%)	Belief and Thought (5%)
	Commerce/Finance (8%)	Commerce and Finance (10%)
	Leisure (14%)	Leisure (10%)
	Natural/pure science (4%)	Science and Technology (10%)
	Applied Science (8%)	World affairs (10%)
	Social science (16%)	Family and Society (20%)
	World affairs (20%)	
	Unclassified (1%)	
Date	1960-1974 (2.26%)	1980 - 2000 (20%)
	1975-1993 (89.23%)	2001 - 2010 (30%)
	Unclassified (8.49%)	2011 - 2020 (50%)
Medium	Book (58.58%)	Textbook (10%)
	Periodical (31.08%)	Periodical (20%)
	Misc. published (4.38%)	Popular magazine (20%)
	Misc. unpublished (4.00%)	Fiction (10%)
	To-be-spoken (1.52%)	Internet (10%)
	Unclassified (0.40%)	Academy journal (10%)
		To-be-spoken (10%)
		Unclassified (10%)

Here, it is to be noted that the BNC has not taken Internet materials as part of its design. However, the COCA has included the internet web materials - the blogs only. As the present WMTC represents the written Malaysian Tamil, it has included web materials such as blogs, emails, and other social media.

It should be noted that, in contrast to BNC, the WMTC exclusively contains samples from the written Tamil works. Overall, the construction of the WMTC has drawn on the features of the design of BNC (Table 4.1). The design of the COCA is also taken into consideration for the present written Malaysian Tamil. Though COCA followed the features of BNC design, there are some differences between these corpora. WMTC followed the features of both these corpora but with some differences which are pointed out as below:

a) BNC comprises texts from 1960 to 1994 and COCA texts from 1990 onwards. WMTC from 1980 to 2020.

b) BNC and COCA have almost similar proportions of the written texts. That is, the written texts are 90% in BNC and 80% in COCA. But the written texts are 100% in the present WMTC.

c) BNC includes miscellaneous published and miscellaneous unpublished (Burnard, 2000).COCA and WMTC did not include them.

d) COCA includes web blogs and webpages only, whereas WMTC in addition to blogs and web pages includes emails and other social media materials also. BNC didn't include this, maybe, because the time period was taken before the advent of these media.

e) COCA and WMTC grouped journals and newspapers separately, whereas the BNC groups them under periodicals.

f) BNC deals fiction nonfiction books separately but COCA and WMTC deals only fiction books.

g) In the Domains BNC and WMTC cover the following fields: applied science, arts, belief and thought, commerce and finance, imaginative, leisure, natural and pure science, social science, and world affairs.

h) BNC and COCA are meant for British English and American English respectively. WMTC is meant for Malaysian Tamil only.

i) According to Ide (2003) it is hard to ensure that an author whether he/she is a native or non-native. But WMTC ensures the author as a native Malaysian Tamil.

j) Annotation also varies in these 3 corpora. The annotation of the WMTC is different from the BNC and COCA. Unlike BNC and COCA whose annotation is based on English grammar, WMTC bases its annotation on Tamil grammar in general and morphological structure in particular.

4.7.3 Sampling method

The stratified sampling method is adopted for the development of the corpus. The stratified sampling considers the following eight types of texts as the strata: textbook, periodical, popular magazine, fiction, internet, academy journal, to-be-spoken, and unclassified. As noted earlier, the stratified sampling method requires a definition of population and a sampling frame. In the current research, the target population comprises the above eight genres. Sub-groups of the genres are listed in Table 4.2. Each of the sampling unit in the sub-groups is sampled using random techniques.

Table 4.2: The divisions of the metadata

Text Book	Primary 1,2,3,4,5,6, Secondary 1,2,3,4,5, Student Writing and Others
Periodical	Malaysia Nanban, Makkal Osai, Tamil Nesan, Tamil Malar, and Others
Popular Magazine	Mayil, Sudhi Mayil, Unggal Kural, and Others
Fiction	Short Story, Novel, Essays, and Others
Internet	Web article, Malaysia Indru, Vanakkam Malaysia, Selliyal, and Others
Academy Journal	Indian Study Journal, Tamil Oli, and Others
To-Be Spoken	TV/Radio news, and Scripts
Unclassified	Government Notification, NGO Materials, Special edition, and Others

4.7.4 Corpus Size

Determining the sample size is quite challenging in sampling procedure. Whether to sample full texts or part of texts? Which part of texts to be sampled - whether initial, middle or end parts? In the present project, only the parts of the text are taken. It is given much importance to the coherence of the sample texts. That is, the selected sample should have expressed conceptual coherence with their necessary metadata because language development and language use are inseparable (Chau, 2015; Tyler, 2010).

Going through these issues, however, it is decided to have the sample size for the present corpus consisting of 1 million. They are accommodated in almost 500 Texts or Files, each comprising more or less 2000 words, as noted earlier. That is, two of the sample size may be of some hundred words whereas others some thousands, depending on the domains and the genres they represent. That means, two samples from advertisements and announcements, contain less words because of their discourse types. In contrast, some samples from textbooks or creative literary works have longer size of words, to accomplish the discourse accordingly. Only two (0.4%) out of 500 samples comprising of different texts (each containing 100-300 words) in the same domain were combined to make 2000 words accordingly.

It should be noted here that these samples are exclusively taken from the Written Tamil Works. Here, it is to be noted, that the BNC has not taken Internet materials as part of its design; however, the COCA has included the internet web materials - the blogs only. The present WMTC has included web materials such as blogs, emails and other social media.

Ideally, a corpus may comprise as many texts as possible. However, compiling a machine-readable corpus can be very costly and time-consuming and the accuracy of any transcription and scanning is a primary consideration (Rayson, 2015). The cost of corpus compilation increases with the size of corpus. For practical reasons, there is a need to seek a balance between the number of texts and the length of text. The desired text length is contingent on linguistic distributions. “Common linguistic features are distributed in a quite stable fashion within texts and can thus be reliably represented by relatively short text segments while rare linguistic features show much more distributional variation within texts and thus require longer sample texts for reliable representation” (Biber, 1993).

In other words, sample texts should be long enough to reliably represent the distributions of linguistic features (Sinclair, 1996). The present corpus, as noted earlier, follows the Brown Corpus in deciding the length of text samples to be included in the corpus. Texts of the Brown corpus range between 1,000 to 2,000 words. In the sampling of Tamil written texts, 2,000 words were extracted from each of the text samples.

4.8 Gathering, computerizing, and organizing the written texts

This section discusses a next practical challenge in developing WMTC, which is to obtain the sample texts. Two major questions arise in the process: How to deal with copywrite issues after data gathering? How should the text be entered, stored, arranged and catalogued once they have been obtained? (Nelson, 2010). As for the former issue, most of the data collected are publicly available, for example, data from newspapers, journals, magazines, and websites. In accordance with Malaysian regulations concerning copyright, consent was obtained from the parties holding the copyright on the relevant data.

The text samples come from various sources, for example, Internet, scripts for speeches and printed books. In preparing data for entry into the corpus, data in electronic format were adapted and hard copies of books and magazines scanned and converted into electronic format with the assistance of OCR software. The OCR output were manually checked, and mistakes made by the software were manually corrected. All the sample texts gathered from the above were first stored in a NotePad File. The format commands such as Bold, Italics which are dependent on the specific proprietary software were stripped.

4.8.1 Text Collection

Identification of texts from each category: Regarding the content of (front and back matters) these sections represent meta-data*,* i.e., additional data about the text, but does not form part of the original text. An example of the meta-data that you encounter in everyday life would be the imprint page inside the front matter of a book.

Texts from each category are identified based on the following:

1. In deciding the domain of the printed sources, the present research project depended upon their title, preface, and content page. With some books such as Science and Technology, the materials printed on the back cover of them on science and Technology helped in this task.
2. The abstract given in the beginning also helped to identify the domain.
3. Also, the keywords given for the respective works also are considered for this task.
4. Regarding other media, the materials were carefully studied and the domains to which they belong were decided accordingly.

4.8.2 Metadata

During the collection of text materials for the present project, the particulars about the text, the extra-text details such as the author, year, domain, and media of the materials are properly entered in the entry form (Figure 4.4).

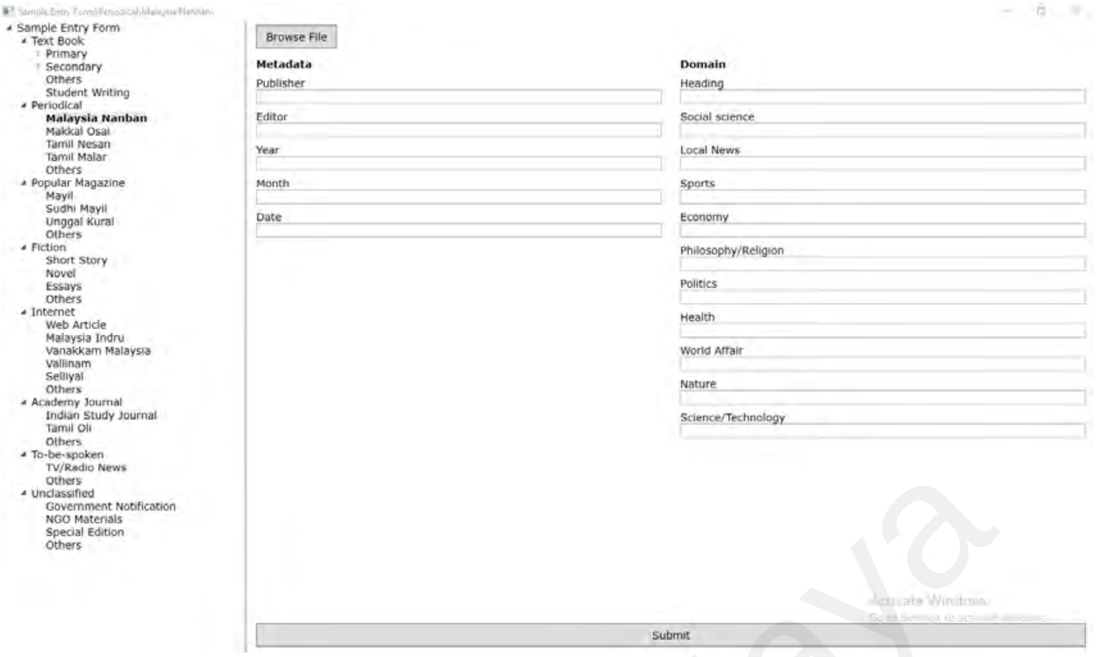


Figure 4.4: Entry Form

For the metadata, the view of Weisser (2016, p. 34) is that “the front and back matter” may be helpful. He pointed out that the “imprint page inside the front matter of a book contains the title of the book, the author, the publisher, edition, year of publication, its ISBN, the typeface and size wherein it is set in, and other types of information that are mainly independent of the content perse and the table of contents also forms the front matter.”

Another type of meta-information is represented by an index in the back matter of a scholarly book, where it serves as a kind of navigational aid in accessing individual parts of the book, and where it is clearly linked to the content itself. Also, regarding the storing of this meta-data, it is often stored in either externally in a different file or database, or inside the document itself.

4.8.2.1 Genre

They belonged to anyone of the genres such as textbooks, periodicals, popular magazines, fictions, internet materials, academic journals, to-be-spoken texts or unclassified items. They were further generalized and thus broadly classified as: Tamil textbooks (primary, secondary school textbooks), periodicals, popular magazines, internet materials, academic journals, to-be-spoken texts, literature (fictions) and the unclassified items.

Sub classification of Genres: Under each genre, the following required query items are included:

a. **Textbooks:**
School type - primary school; secondary school; students' writing, title of the book, language level (standard), topic, year, publisher

b. **Periodicals**
Type - name, year, month, date, editor, publisher

c. **Magazines**
Type - name, year, month, date, editor, publisher

d. **Fiction:**
Literary genre - story book, novel, short story, drama, author, year, publisher

e. **Internet:**
Website - name, year, month, date, editor

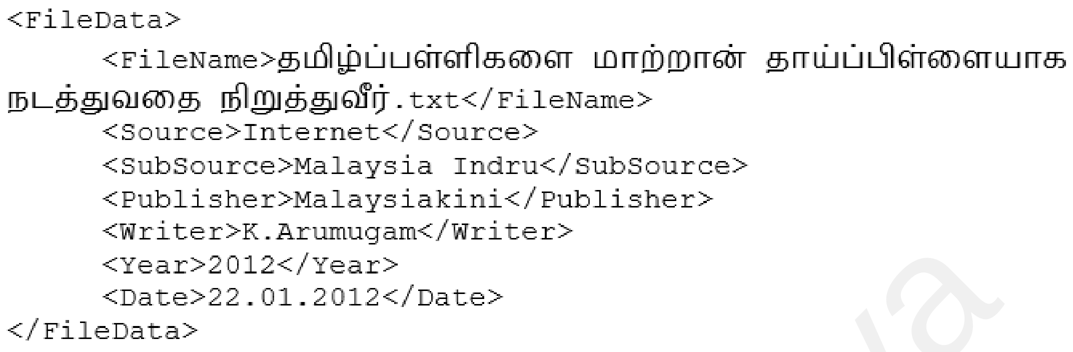
f. **Academy Journal:**
Type - name, volume, year, edition, editor, publisher

g. **To be spoken:**
Editor, date, month,year, media

h. **Unclassified:**
Type, publisher, year, month, and date

4.8.2.2 XML File format for storing the metadata

In the present project, this meta-data is stored in header of XML files.



```
<FileData>
      <FileName>தமிழ்ப்பள்ளிகளை மாற்றான் தாய்ப்பிள்ளையாக
நடத்துவதை நிறுத்துவீர்.txt</FileName>
      <Source>Internet</Source>
      <SubSource>Malaysia Indru</SubSource>
      <Publisher>Malaysiakini</Publisher>
      <Writer>K.Arumugam</Writer>
      <Year>2012</Year>
      <Date>22.01.2012</Date>
</FileData>
```

Figure 4.5: XML Data

4.9 Specific issues that come across while designing the WMTC

4.9.1 Data Collection

The data collection has several hurdles like cost, copyright, non-availability of old documents in digitized form, keying large texts, and issues in scanning. Among these, the present work encountered many problems with scanning.

4.9.2 Graphics issues

In many instances the graphics are intermingled with the texts. Due to this, the contents of the graphics contained in the texts are not scannable. Its required re-work on the text by manually keying the text.

4.9.3 The problem of Encoding

The problem of encoding of written Tamil texts is the foremost hindrance which requires much edition. Since the language technology reached the Tamil language only in the 90’s, most of the texts collected in the 80’s was not contained in the digitized form and hence they have to be manually keyed. However, this required a lot of proof reading. Though some of the materials are digitized in 90's, they were in old non-Unicode encoding formats such as TAM, TAB, TSCII based on ASCII. These formats had to be converted to Unicode encoding format. During this process, many important words were lost, misspelled, distorted or unknown characters cropped in. In some of the cases all the three formats were mixed together in the texts which lead to many ungrammatical errors.

4.9.4 Graphemic problem

For scanning, a google OCR software was used, since no other effective and full-fledged OCRs were not available at the time of data collection. However even the google OCR is not 100% accurate. There was a need to clean the texts. Some graphemes were misrepresented which needed immediate attention. For example, Tamil words **அ** as **ஆ**, **ல** as **வ**, **தூ** as **துர**.

4.9.5 Switching of Spoken Variety

Sometimes spoken variety that intervened had to be removed. This resulted in non-coherence of the texts or reduction of sample size, which had to be managed.

4.9.6 The problems with External Storage

Sometimes CDs may be available that had to be transferred which required much time and care. While copying the CDs, some problems arose as some of them were old and damaged. Though the e-mails and blogs were personal, due permission from the concerned persons were obtained.

4.9.7 Web sources

The web sources are not that straightforward. Sometimes the needed text materials may have to be searched like in a ‘whirlpool,’, and they would be in different formats. However, a web document is removed from its original context. For example, multimedia texts are stripped of their multiple contents and reduced to plain word documents (txt. or doc.).

4.10 Steps followed in Formatting

Once a text was chosen for a corpus, and the location of a copy in some usable format was determined, the following steps were undertaken:

1. A **copy** of the text made, in the format received, for later reference.
2. The text was saved in **plain text** format in XML.
3. An **identification** of the text was provided at the very beginning.
4. The **metadata** were coded in a mark-up format which allowed the metadata to be hidden when the corpus is searched.
5. A **pre-processing** (cleaning and normalization) of the text was carried out by the software such as spell checker, concordance, N-gram to undergo further analysis.

4.11 Text encoding, mark-up, and storage

All the sample texts gathered from the above are first stored in a Note Pad file. Since all the data are stored in the XML format, caution was taken to have plain texts by removing the format commands such as Bold, Italics which are dependent on the specific proprietary software.

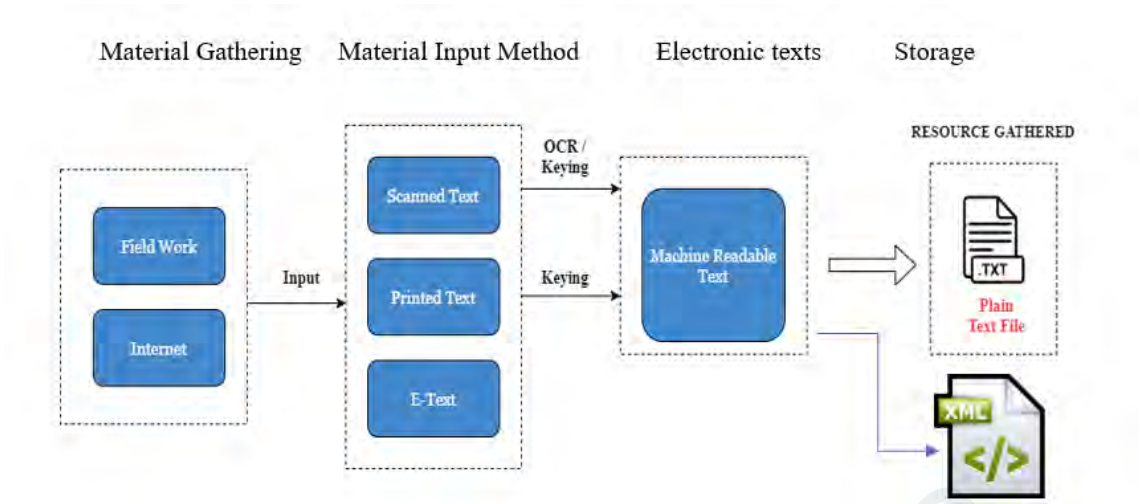


Figure 4.6: Text encoding, mark-up, and storage

1. Material gathering

2. Materials Input method

3. Data Collection and storage process.

4. Electronic text with metadata.

Every sample consists of two parts:

1) metadata in the XML format

2) sample text in the plain text format

4.12 Cleaning of the corpus

Once the samples were collected, the next task done was to check for the errors such as mistakes in spelling and grammar in the samples. For this task, the Tamil Spell-checker and Sandhi checker tools available in the commercial Tamil word processor 'Mentamizh' were used.

Though the above-mentioned word processor was used, since it was based on Tamil Nadu written Tamil, even some of the written Malaysian Tamil words which were correct were shown as error words. So, to decide and correct the mistakes found in the samples some manual work was done. The same problem was encountered with Malay and Chinese words transliterated into written Malaysian Tamil such as titles like *Dato, Datosri, Tun,* place names like *Selangor, Perak, Pahang*, food names like *Nasi Lemak*, *Roti Canai, Bak Kut Teh* and proper names like *Najib, Mahathir, Samy Vellu, Kit Siang* as well. Here it is to be mentioned that the above two types of words were included in the final lexicon of the morphological parser.

Regarding the scanned materials, as mentioned earlier, the Google OCR tool was used to convert the scanned images into texts. However, there were some spelling errors due to the corruption in the image of the original texts, due to the age and the ink used of the papers scanned, and the efficiency of the particular scanner machine used. These errors could not be perfectly corrected using the above-mentioned word processor. This correction task was manually completed.

4.13 Morphophonemic (Sandhi) problem in corpus cleaning:

In Tamil, the adoption of Sandhi rules is very important. In many contexts, Sandhi is very helpful for the process of disambiguation. There are two kinds of Sandhi (morphophonemic) rules. One is, Internal Sandhi which is present within compounds, between word and suffixes, and between suffixes. Another one is External Sandhi which is present between individual words.

(a) Internal Sandhi:

1. “yaanaiththantham”

   “yaanai + th + thantham”

   ‘elephant + sandhi + tusk’ > ‘the tusk of an elephant’

2. “therukkaL”

   “theru + k + kaL”

   ‘street + sandhi + plural suffix’ > ‘streets’

3. “avanaippaRRiththaan”

   “avan + ai + p + paRRi + th + than”

   ‘he + acc.case + sandhi + about + sandhi + only’ > ‘(it is) only about him’

(b) External Sandhi:

1. “avan kataikkup poonaan”

   “avan + kataikku + p + poonaan”

   ‘he + to shop + sandhi + went’ > ‘he went to the shop’

   There are well-defined sandhi rules in Tamil. These rules should not be ignored. Otherwise, the words or sentences would become ungrammatical or sometimes become ambiguous if the sandhi principles are not adopted.

4.14 Inflectional increment / empty morph (Caariyai)

In Tamil, with some words with case suffixes, there should be some inflectional increments between the root words and the cases. These rules are also well-defined in Tamil grammar.

1. “maraththai” (“mara(m) + thth + ai”)

   ‘tree + inflectional increment + case suffix’ > ‘tree (acc.)’

2. “viittai” (“viitu + (t) + case suffix”)

   ‘house + inflectional increment + case suffix’ > ‘the house (acc.)’

To correct the mistakes in the text samples for the above Sandhi and inflectional increment addition, the above-mentioned Tamil word processor “Mentamizh” is used.

4.15 Storage of Text Samples

After the samples are selected, the metadata part of the sample is stored in an XML file. The other part, that is, the text samples are stored in plain text format. The metadata of sample text are stored in XML format because it is ideally suited to represent the meta textual categories (Hoffmann et al., 2008). Necessary headers were created to accommodate the meta-data. This process is digramatically represented as follows:

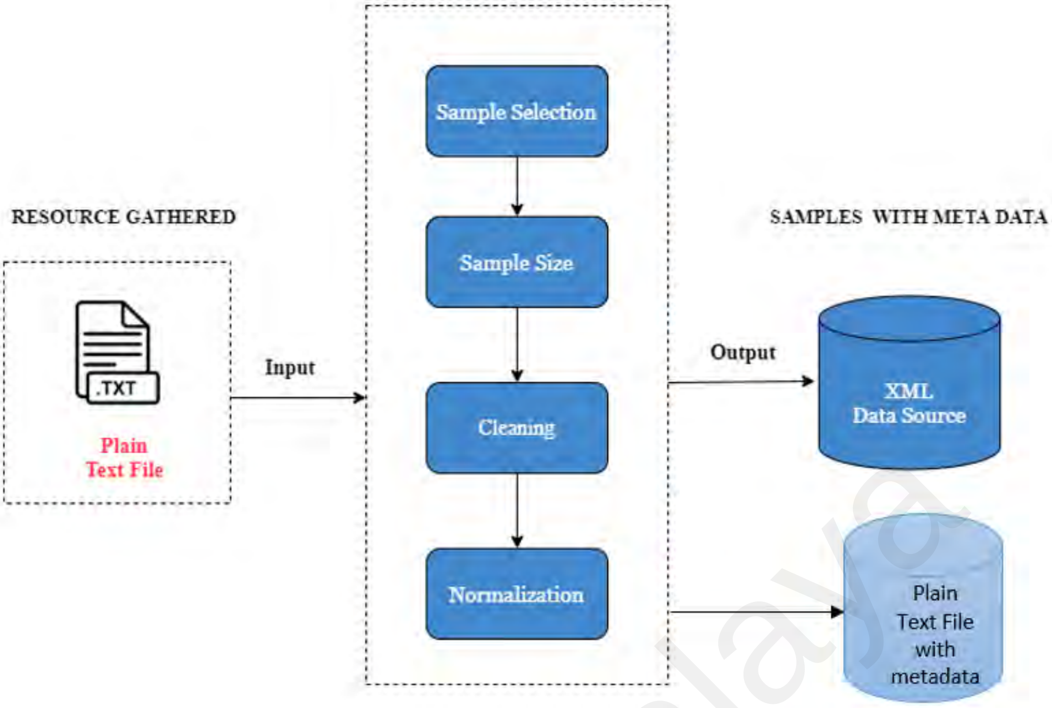


Figure 4.7: Storage of Text Samples

4.16 Corpus annotation

Another important aspect in corpus projects is the annotation provided to the sample texts. Corpus by itself cannot do anything by itself as it is only a store of used language. That storage can be rearranged with a corpus access software, so that observations of various kinds can be made. If a corpus represents, very roughly and partially, a language user's linguistic experience, the access software re-orders that experience so that it can be examined in ways that are usually impossible. A corpus does not contain new information about language, but the software offers us a new perspective on the familiar.

In three ways the data can be processed with some readily available software packages: 1. showing frequency, 2. phraseology, and 3. collection (Hunston, 2008). The samples should be analyzed with suitable software tools such as parsers, concordancer, collocation etc., to get the information about that language. With this morphological, syntactical, and other non-linguistic aspects, the samples should be annotated.

Regarding types of corpus annotation, McEnery and Hardie (2012) suggested that three types of information, that is, metadata, textual mark-up, and linguistic annotation are contained within them, helping in the investigation of the data in the corpus. “Metadata is information that tells something about the text. In the case of written material, for example, the metadata may tell who wrote it, when it was published, and what language it is written in. The metadata can be encoded in the corpus text or held in a separate document or database.”

“Textual markup encodes information within the text other than the actual words. For example, in a printed, written text, textual mark-up would typically be used to represent the formatting of the text, such as where italics start and end. Linguistic information can also encode within a corpus text in such a way that we can systematically and accurately recover that analysis later.” The information given by the metadata is then available to users of the corpus and is very useful in helping to interpret and explain the findings (Cheng, 2011).

However, when a corpus includes linguistic annotations, one must note what can and cannot be said about those annotations. It is to be noted that the corpus does not contain any new information. That is, the explicit information could be made which is there implicitly in the data. In other words, identifying a word as a noun does not mean that we transform it into a noun in so doing. In corpus annotation, nothing is created nor transformed but only the process of labelling is taking place. According to McEnery and Hardie (2012b, p. 34) by annotation "the corpus is enriched, from the point of view of a program or user, but the corpus has not added any new information to it".

Two things are to be emphasized in corpus annotation: one is 'annotation scheme'; the other is 'format' given to the annotation. Regarding the first, it can be said that the annotation scheme is a critical part of any corpus annotation project, irrespective of its type and scale. During the annotation process, it would contain explicit and complete information on the linguistic categories to be differentiated, which depend not only on the type of annotation at issue but also on the degree of specificity that is desired or required given the purpose of the annotation.

Also, the annotation scheme is a set of labels, designed to denote the linguistic categories, with a one-to-one correspondence between them and the categories. Moreover, these labels must be concise and meaningful. Also, whenever distinctions are made between subcategories of a general category, the labelling system should be designed so as to find the commonalities among the subcategories.

The annotation scheme should be comprised of some guidelines, explaining how different linguistic units in the corpus are to be designed to the linguistic categories found in the scheme and annotated with appropriate labels accordingly.

Regarding the format, which is used in the annotation, it could be said that this has to do with how the labels are to be applied to the appropriate linguistic units in the raw corpus. Regardless of the format adopted, it is to be noted that the annotations are easily separable from the raw corpus. In other words, it should be possible for one not only examine the raw texts and their annotations at the same time but also separate the texts from their annotations and examine them individually (Lu, 2014).

In the present research project, every sample has two annotations: (1) Non-linguistic annotation (i.e., metadata), (2) Linguistic annotation (i.e., POS tagging).

4.17 Conclusion

This chapter has provided information on the WMTC, such as its size, composition, and the major decisions regarding its construction. The compiled corpus contains 1 million 3298 words, with 500 sample texts, and approximately 2,000 words in each sample text. For the corpus to be accessible, error correction and normalization of the texts have been carried out. With careful sampling procedures and proper storage, the corpus is expected to be representative and balanced and useful for future investigations of actual language use of written Malaysian Tamil.

The development of the morphological parser and POS tagger and the design of algorithms for the same are discussed in the following 5th and 6th chapter, respectively.

CHAPTER 5

THE DEVELOPMENT OF MORPHOLOGICAL PARSER AND POS TAGGER

5.1 Introduction

The previous chapter discussed the development of the corpus, which is the first research objective of the present research project. In this chapter, the second objective of the thesis: the development of a morphological parser and a POS tagger is being addressed. This chapter discusses system architecture and software design with illustration, and briefly considers the algorithm of the software before it is discussed in greater detail in Chapter 6.

5.2 Morphological parser and POS tagger

The second objective of the current research project is the development of a morphological parser and a POS tagger. It should be noted though that two important corpus tools, a concordancer and N-gram along with the morphological parser and the POS tagger, have been developed and integrated into this software programme. These two corpus tools will be briefly considered in the relevant section.

It is important to point out that the development of the morphological parser and the POS tagger is informed by the following four major language components:

(1) lexicon (root words),

(2) grammatical affixes, postpositions, particles and clitics (grammatical features),

(3) morphotactic rules (ordering of morphemes), and

(4) morphophonemic rules (changes within words and between words) (Evangeline & Shyamala, 2020; Jayan et al., 2011; Lushanthan et al., 2014; Sarveswaran et al., 2021). Further details about these four components are presented in Chapter 6. These major language components are represented as follows:

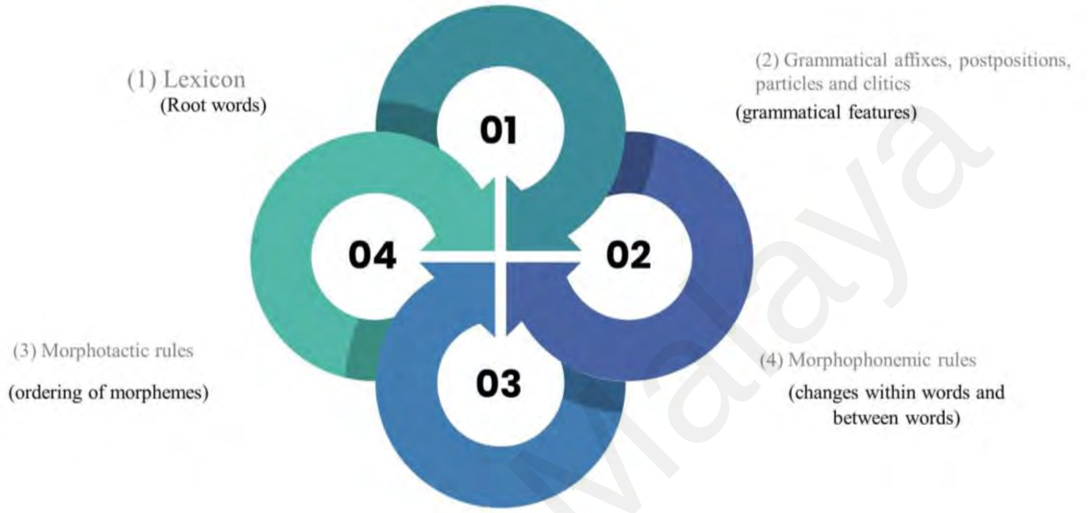


Figure 5.1: Major Language Components

While making use of above-said four major components for the development of the morphological parser and POS tagger, the corpus-based approach has been implemented as opposed to the corpus-driven approach (Tognini-Bonelli, 2001). In the corpus-based approach, some available grammars preferred by the researcher could be chosen to study the new linguistic features leading to the modification of the applied grammar chosen. That is, the modified grammar is not a new one, but a revised version based on the corpus data. In a corpus-driven approach, the researcher does not choose any available grammar for the study, but with his/her trusting the data and letting the data drive the discovery process (Sinclair, 1996), a new description of the grammar would be constructed; that is, a new theory would be derived or constructed from the corpus itself. In this study, a corpus-based approach has been adopted to develop the morphological parser and the POS tagger.

For modern written Tamil, there are some grammars written by some eminent scholars. Though they have shown some new linguistic features emerged in modern Tamil, they are not based on a well-constructed electronic corpus of Tamil. The main reason may be due to the non-availability of advanced computing technology to build a big corpus or to construct the necessary corpus software. This lacking is reflected in their study, leading to the challenges in finding or identifying all the new linguistic features found in the modern Tamil. The present research project contributes to the field by developing a corpus of Malaysian Tamil and relevant corpus analysis tools based on emerging computing technology, which addresses the gap found in the earlier studies. For some analysis such as frequency studies of the words and the grammatical features, some descriptive statistical analyses have been made in the present project.

5.3 System Architecture

After surveying a range of computational Tamil morphological tools and designs (Deivasundaram & Gopal, 2003), the following system architecture was arrived at. The following figure (Figure 5.2) illustrates the whole system architecture of the corpus-based morphological parser and POS tagger developed in the present research project.

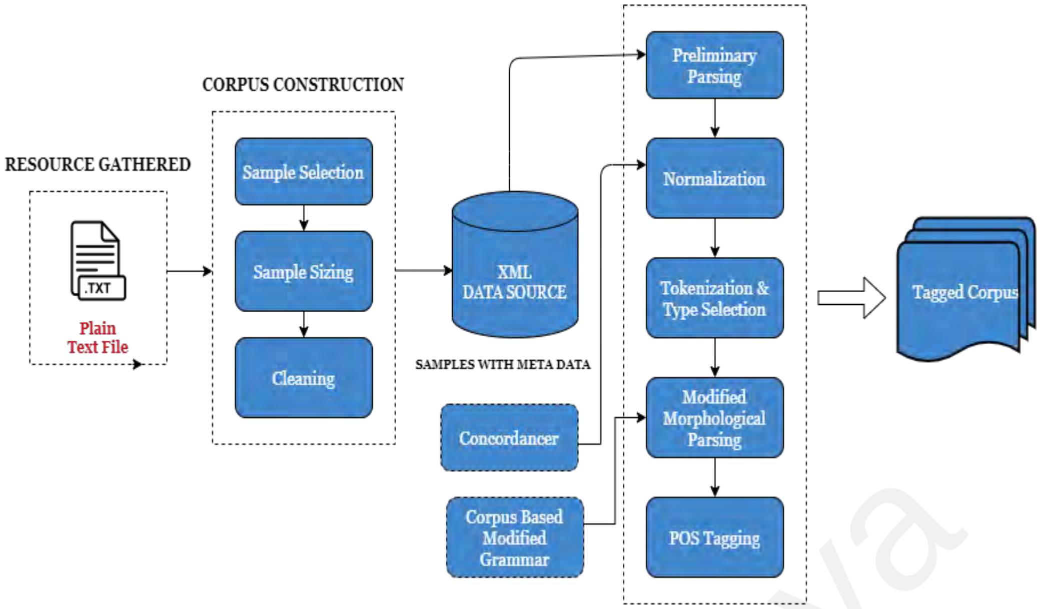


Figure 5.2: System architecture of morphological parser and POS tagger

The above diagrammatic representation describes the following processes involved in the present project:

1. Plain text File: Materials collected for this project were initially available in different format commands which are part of different proprietary software. For this project, all the proprietary format commands were removed, and final sample texts were stored as plain text files.
2. Sample Selection: As explained in the previous chapter, necessary samples were selected in a way to make this corpus a representative and balanced one. The necessary metadata for each sample was stored in an XML file.
3. Sample Size: 500 samples. Each sample consists of about 2000 words. In total, the present corpus consists of 1million words/tokens.
4. Cleaning: The selected samples were processed for spelling, sandhi, and other language errors with the help of the software 'Mentamizh.'
5. Data Source: Every sample consists of two parts- metadata in XML format and a plain text file of the sample content.
6. Morphological Parsing and POS tagging: After the texts were cleaned, the tokens were parsed, and POS tagged.
7. Concordancer: For disambiguation of POS ambiguities, a concordance program was developed.

8. Normalization: After the above process, the sample texts were normalized to the finalized token and type selection.
9. Modified Morphological Parsing: Based on the previous parsing and POS tagging, the missing morphological aspects and POS categories were found and included in the final morphological parser and POS tagger for further parsing and tagging.
10. Tagged Corpus: In the last step, the normalized texts were fed into the modified parser and POS tagger to get the final tagged corpus.

5.4 The necessity of linguistic knowledge for designing the software

The morphological parser should be able to segment the input word forms correctly. It should be given the necessary linguistic knowledge for correct segmentation: where to segment a word form and how many segments to be parsed, according to the Tamil morphology. The development of parser was informed by an algorithm designed with the help of the Tamil morphology. The algorithm used is discussed in detail in the next chapter. The knowledge required for designing software and its illustrations is outlined in this chapter.

To step into the system architecture, the knowledge about Tamil orthography, phonemes, graphemic encoding, parsing direction and Tamil computational morphological segmentation is needed (Balakrishnan, 2002; Deivasundaram, 2021; Deivasundaram & Gopal, 2003).

5.5 Tamil orthographic Words

In Tamil orthographic words, vowel graphemes could occur only initially. In other places, the vowels join with the consonants becoming syllables. For this process, there are allographs for all the vowel graphemes. All the individual vowel and consonant phonemes have separate graphic forms - that is, scripts. In addition to these, there are syllabic graphic forms that represent the syllables - that is, the consonant with the addition of vowel.

In Tamil, there are separate graphemes-scripts for all the 12 vowel phonemes, 18 consonant phonemes, and one archiphoneme (a dependent vowel phoneme). In addition, for 216 syllables (12 vowels times 18 consonants). In total, there are 247 graphemes (12 + 1 + 18 + 216). In addition to the Tamil consonants, there are five more grantha (borrowed from the Sanskrit language) phonemes with their respective graphemes existing in Tamil. These also combine with the Tamil vowels forming syllables (60). One more grantha ‘sri’, which is an independent one, will not join with any vowel. Therefore, in modern written Tamil, the total graphemes are 282 (216 + 5 + 60 +1).

Morphology is the “study of word structure, and words are the interface between phonology, syntax and semantics” (spencer, 2005). As it is well known, the minimum meaningful unit is a morpheme. To analyse words into different morphemes in Tamil, the orthographic form of the Tamil words has to be segmented into individual consonants and vowels (Anand Kumar, Dhanalakshmi, Rekha, et al., 2010).

Instead of syllabic scripts, they have to be represented as consonants and vowels. For example, the syllabic script ”கி” (ki) would be represented as ”க் + இ” (k + i). This task is very important because if a lexicon or suffix ends with a consonant and the following suffix starts with a vowel, they would be represented in a single syllabic script. Then only the individual morphs could be identified by the software.

Example: “aaciriyarukku” (“aaciriyar-ukku”) (‘teacher - dative case suffix’).

When the above two morphemes combine to form a single word form, the final phoneme “r” (‘ர்’) in “aaciriyar” would combine with the initial phoneme “u” (‘உ’) of the following, resulting in one syllable and represented in the syllabic script “ru” (‘ரு’). However, for the software to parse morphologicaly, they have to be split into separate phonemes: “r” and “u”.

5.5.1 Individual phonemes

As a first task, for morphological parsing of Tamil word forms, all the words are represented as individual phonemes - consonants and vowels, not as syllables (Balasubramanian, 1980). This is the basic requirement which is one of the tasks in the construction of the present software.

5.5.2 Graphemic Encoding

All the Tamil graphemes are encoded into UNICODE (Appendix B). It is UTF-8 (Tamil: U+0B80-U+0BFF) in this project.

5.6 Parsing

An important task in Computational Linguistics is the parsing of sentences (Ramasamy & Žabokrtský, 2011). A sentence is parsed to capture its syntactic structure. For this, the words in that sentence have been analyzed along with the linguistic annotation. All the words are POS tagged such as Noun, Verb, Adjective and Adverb. These categories are further subcategorized. Parsing is a standard technique used in the field of Natural

Language Processing (NLP). If a morphological parser is to be developed, a word must be parsed into its various parts with linguistic annotation (Rajendran, 2006).

In English, this information is contained in the lexicon with a list of all word forms and the part-of-speech (POS) along with inflectional information such as the number and tense since it has a simple inflectional system. Hence, they are accommodated in a database. To cite, the English nouns have merely two inflections (singular and plural) and English verbs usually have five forms (verb root, present tense, including -Ø and -s forms, past tense, past participle, and verbal participle). There are also some irregular English verb forms, for example, *speak* and *go*, which have irregular derivations (*SPEAK*: *speak, speaks, spoke, spoken, speaking; GO: go, goes, went, gone, and going*).

In contrast, an exhaustive lexical database for the Tamil language is not quite possible as in English since it has plenty of inflected forms for nouns and verbs (Annamalai & Steever, 2015). A morphological parser is a must for the Tamil language. Any word could be processed using the morphological parser and the POS tagger.

5.6.1 Parsing direction: Right to Left or Left to Right

There are two ways to be adopted in morphological parsing: one is Right to Left; another is Left to Right. (Hudson & Buijs, 1995). In both methods, the parser would match the whole word form in the lexicon. If it is found, there is no necessity for any further parsing process. For example, when the parser analyses the word “ammaa” ‘mother,’ it could find this word in the root lexicon. There is no need for any parsing.

If the input word is not found in the lexicon, there is a need for parsing or segmenting the input word form. In the first way - Right to Left - the parsing program starts from the endings of the inflected word form. Here, the suffixes are one by one identified till reaching the lexical root or stem.

In the second way - Left to Right - the parsing starts from the beginning of the word form. This is the method adopted for the present morphological parser software. Here, the attempt is to find the lexical root or stem first and then identify the various suffixes attached to this lexical root or stem. The root searching continues till some phonemic sequences can be identified with a lexicon available in the Lexicon database.

Example: “uurukku” > “uur + ukku” = ‘village + to’ > ‘to village’

Table 5.1: Word Form

“uurukku” ஊருக்கு ‘to village’	
uur ஊர்	ukku உ க் க் உ
Village	to

Here, the above word form “uurukku” consists of two morphemes: one is the root lexicon “uur” and the other one is the dative case suffix “ukku”. The parser which is searching for the phonemic sequence which could be found in the lexicon here finds “uur” as the root lexicon. Then, in its further searching, it could not find any root lexicon till the end. But it could identify the phonemic sequence “ukku,” a case suffix in the affix database.

The schematic representation of parsing process (Left to Right / Front to Back) of noun word form and verb word form:

Noun:

‘peNnkaLukkaakamattumthaaNaataa?’ “is it for ladies only?”

‘peNn - kaL - uk - ku - aaka - mattum - thaaN - aa - ataa’

Table 5.2 Parsing Process (Noun)

Grammatical Category	**Word / suffix**	**சொல் / விகுதி**	**Meaning**	**இலக்கண வகைப்பாடு**	**Number of morphemes**
Noun stem	peNn	பெண்	‘woman’	பெயர்ச் சொல்	1
Number	kaL	கள்	Plural suffix	பன்மை விகுதி	2
Filler	uk	உக்	Stem formative	சாரியை	3
Case	K(u)	கு	Dative case suffix	வேற்றுமை 4	4
PP	aaka	ஆக	‘for’ post-position	பின்னொட்டு	5
Cl.1.	mattum	மட்டும்	‘alone’	மிதவை ஒட்டு 1	6
Cl.2.	thaaN	தான்	‘only’	மிதவை ஒட்டு 2	7
Cl.3.	aa	ஆ	question	மிதவை ஒட்டு 3	8
Addressive suffix	(a)taa	அடா	vocative	விளி விகுதி	9

Varius stages of affix striping in noun word form:

பெண்களுக்காகமட்டும்தானாடா **“peNnkaLukkaakamattumthaaNaataa”**

Table 5.3 Affix Striping In Noun

Input	**Root/Stem**	**Plural**	**Case**	**P.Pos**	**Clitics 1**	**Clitics 2**	**Clitics 3**	**Addresive suffix**
பெண்	1	0	0	0	0	0	0	0
பெண்கள்	1	1	0	0	0	0	0	0
பெண்களுக்கு	1	1	1	0	0	0	0	0
பெண்களுக்காக	1	1	1	1	0	0	0	0
பெண்களுக்காகமட்டும்	1	1	1	1	1	0	0	0
பெண்களுக்காகமட்டும்தான்	1	1	1	1	1	1	0	0
பெண்களுக்காகமட்டும்தானா	1	1	1	1	1	1	1	0
பெண்களுக்காகமட்டும்தானாடா	1	1	1	1	1	1	1	1

Verb:

‘eluthikkaattavaikkapparththaanaataa?’ “did he try to cause (somebody) to demo writing?”

‘eluthi - kaatta - vaikkap - paar - thth - aan - aa -ataa’

Table 5.4 Parsing Process (Verb)

Grammatical Category	Word/ suffix	சொல் / விகுதி	இலக்கண வகைப்பாடு	Number of Morphemes
Verb stem	Eluthi(k)	எழுது ‘write’	வினை	1
Past participle	i	இ	‘செய்து’ விகுதி	2
Asp. Aux	kaatta	காட்டு	வினைக்கூறு	3
Inf. Marker	a	அ	‘செய’ விகுதி	4
Voice. Aux	Vaikka(p)	வை	வினைப் பாங்கு	5
Inf. Marker	(kk) a	(க்க்)அ	‘செய’ விகுதி	6
Mod. Aux	paar	பார்	வினை நோக்கு	7
Tense	thth	த்த்	கால விகுதி	8
PNG	aaN	ஆன்	தி-எ-பா விகுதி	9
Cl.1.	aa	ஆ	மி.ஒ 1	10
Cl.2.	(a)taa	அடா	விளி விகுதி	11

Note:

1) ‘eluthi ‘itself consists of main verb + past participle suffix (‘eluthu + I’)
2) ‘kaatta’ consists of Aspectual auxiliary verb ‘kaattu’ + infinitive verbal particple suffix (‘kaattu+a’)
3) ‘vaikka’ consists of Voice auxiliary verb ‘vai’ + infinitive verbal participle suffix (‘vai+kk+a’); here the segment ‘kk’ is stem formative.
4) In Tamil, the verb (either main or auxiliary verb) should be in the past participle form, if it is followed by an Auspectual auxiliary verb. It is obligatory.
5) The verb (either main or auxiliary verb) should be in the infinitive verbal participle form if it is followed either by Voice auxiliary or Modal auxiliary verbs. It is obligatory.
6) Some root verbs will take increments before it (‘the stem’) is affixed with the tense suffixes. In the above example, the Voice auxiliary ‘vai’ takes the increments ‘kk’ to form the stem for further affixation.

So, the above-mentioned verb example consists of 8+3 = 11 morphemes in total.

Varius stages of affix striping in verb word form:

எழுதிக்காட்ட வைக்கப்பார்த்தானாடா “**Eluthikkaattavaikkapparththaanaataa”**

‘eluthi - kaatta - vaikkap - paar - thth - aan - aa -ataa’

Table 5.5 Affix Striping In Verb

Input	Root/Stem	Past participle	Aspectual	Infinitive marker	Voice auxilary	modal	Infinitve marker	tense	PNG	Clitics	Clitics
எழுது	1	0	0	0	0	0	0	0	0	0	0
எழுதி	1	1	0	0	0	0	0	0	0	0	0
எழுதிக்காட்டு	1	1	1	0	0	0	0	0	0	0	0
எழுதிக்காட்ட	1	1	1	1	0	0	0	0	0	0	0
எழுதிக்காட்ட வை	1	1	1	1	1	0	0	0	0	0	0
எழுதிகாட்டவைக்க	1	1	1	1	1	1	0	0	0	0	0
எழுதிக்காட்ட வைக்கப்பார்	1	1	1	1	1	1	1	0	0	0	0
எழுதிக்காட்ட வைக்கப்பார்த்த்	1	1	1	1	1	1	1	1	0	0	0
எழுதிக்காட்ட வைக்கப்பார்த்தான்	1	1	1	1	1	1	1	1	1	0	0
எழுதிக்காட்ட வைக்கப்பார்த்தானா	1	1	1	1	1	1	1	1	1	1	0
எழுதிக்காட்ட வைக்கப்பார்த்தானாடா	1	1	1	1	1	1	1	1	1	1	1

5.6.1.1 The Lexicon within another root lexicon

In some cases, there may be some ambiguity because of the possibility of identifying more than one root lexicon. That is, one lexicon may be a part of another root lexicon.

Example:” aaciriyarai” > (1) “aaci ‘blessing’ + riyarai” ‘*’ or

(2) “aaciriyarai” ‘teacher (acc.)’

* - it means the particular form is not grammatical or acceptable.

In (1), the phonemic sequence “aaci” could be found in the lexicon. In (2), another phonemic sequence except the final one - “aaciriyar” - could also be found in the lexicon. However, in (1), the next segment, “riyarai” is a meaningless segment; It is neither a root lexicon nor some other inflectional suffixes which could occur with nouns. So, this possibility could be ignored.

But in (2), after segmenting the form “aaciriyar”, the remaining final phoneme “ai” could be found as a case suffix in the suffix database, which could follow a noun. So, the (2) “aaciriyarai” could be the correct word form since it could be handled by the parser.

5.6.1.2 One segmentation for different legitimate forms

In some cases, there may be more than one possible segmentation, and both of them may be legitimate forms.

Example: “patiththavarai” ‘the one who studied (acc.)’ or ‘till where it is read.’

Both possibilities are acceptable.

(1) “pati - thth - a - var - ai” ‘study - past tense - relative participle marker - number - gender suffix- accusative case suffix’

(2) “pati -thth - a - varai” study - past tense suffix - relative participle marker - verbal particle’

Here, the above segmentations are legitimate ones. The expected parsing result - one among these two possible segmentations - could be decided only by its linguistic contexts for which the concordance tool may help.

5.6.1.3 Homophonous forms

Another problem in deciding the segmentation is the homophonous forms of the lexicons.

Example: (1) “malarai” ‘flower (acc.)’ (2) “malarkiRathu” ‘flowers - it’In the above both, the first segmented part is “malar.” This form may be a noun or a verb root. Here, the inflection process would help. If it is a noun, it could be inflected either for plural or for case. On the other hand, if it is a verb, it could be inflected either for tense or other verbal participles.

In the above examples, the “malar” in (1) is followed by the case suffix - “ai” whereas in (2) is followed by a tense marker, from this, it could be decided that “malar” in (1) is a noun and in (2) is a verb. Thus, here, for correct parsing, this inflectional process helps the parser.

5.6.1.4 Knowledge of Morphophomics rules in segmentation

There is one more problem in the segmentation of a word form for the parser. Here, the parser could solve this with the knowledge of the morphophonemic rules. For example, the word form “veelai” in the following word forms are segmented in two ways.

(1) “avar kataiyil veelai paarkiRaar” ‘He is working in a shop’

(2) “avar cilaiyil irukkiRa veelaip paarkiRaar” ‘He sees the spear in the statue.

In (1), “veelai” is a single root lexical noun, whereas in (2), “veelaip” has two segments - “veel” ‘spear’ + “ai” ‘acc.case’. The final “p” phoneme occurs to satisfy a morphophonemic rule of Tamil. The parser could parse the above forms into correct segments using this knowledge of morphophonemics. In some instances, there will not be any morphophonemic presence.

Example: “avar veelai ethirpaarkkiRaar” ‘He expects a job’ or ‘He expects a spear’.

In the above sentence, with the word form “veelai” there is no sandhi presence. So, here, we could not expect any help from the morphophonemic rules; instead, only the concordance program would help to solve this ambiguity. The sandhi or morphophonemic changes involve various levels of linguistic analysis - Tamil phonology, morphology, syntax and semantics (Rajan et al., 2012).

5.6.1.5 Knowledge of Empty Morphs (caariyai) in segmentation

There is one more factor to be mentioned here - that is, “caariyai”, an empty morph. Nouns and verbs, when they are inflected, in some places take some empty morphs. For example, the noun “maram” ‘tree’ is inflected for a case, it takes first an empty morph “aththu” before the occurrence of the case suffix. Some empty morphs are obligatory, and some are optional. This knowledge of the empty morphs is provided to the morphological parser to be used in the segmentation process.

5.7 Development of Morphological Parser and POS tagger

As noted earlier, the algorithm based on Tamil computational linguistic knowledge is described in the next chapter, on which the present morphological parser and POS tagger were constructed.

5.7.1 Database of Tamil lexicons

The analysis of Tamil morphology helped to get a list of various Tamil affixes to be accommodated in the Tamil morphological parser. Also, it was helpful to arrive at all the morphotactic and morphophonemic rules of Tamil. Then it was necessary to build a Tamil lexicon with which this parser could be developed. Initially, a database of Tamil lexicons based on the existing resources was built. In this database, the gathered lexicons or root words were placed with their due grammatical/POS categories.

5.7.1.1 Verb Roots

The verb roots were classified into 13 types based on their conjugation class drawing on the Tamil Lexicon published by the University of Madras in 1936. The list of the classification of Tamil verbs is attached in the Appendix C.

5.7.1.2 Noun Forms

With nouns, it was necessary to mark whether they are human or non-human. Because some case inflections, such as Locative and Ablative, the human nouns behave differently from the non-human ones.

5.7.2 Checking of tokens in the corpus

Here, the tokens are identified just by the space between words. But they are not normalized ones. The first reason is that the number of orthographic words is huge. Manually it is not possible to normalize the texts for arriving at the tokens and the types. It is not possible to identify exactly whether a word is a separate lexeme or grammatical words such as postpositions, verbal particles, and auxiliaries. A check list containing postpositions and auxiliaries, (Table 5.6 - 5.13) which could occur as separate lexemes or as grammatical suffixes, was prepared. With the help of this checklist, it was able to find out whether a particular form was a word or a suffix.

Table 5.6: Check List - Postpositions - Accusative Postpositions

	Postposition	meaning	Example
1	vita/vitavum	than	avaNaivita ‘than him’
2	poola/pool/pooNRu	like	avaNaippoola ‘like him’
3	koNntu	with	kaththiyaikkoNntu ’with a knife’
4	nooki	towards	eNNainookki ‘towards me’
5	paRRi	about	eNNaippaRRi ‘about me’
6	kuRiththu	about	eNNaikkuRiththu ‘about me’
7	cuRRi/cuRRilum	around	eNNaiccuRRi ‘around me’
8	vittu	from	viittaivittu ‘from the house’
9	thavira	except	avaNaiththavira ‘except him’
10	muNNittu	on account of	theerthalaimuNNittu ‘on account of the election’
11	veeNnti	on account of /for the sake of	avaNaiveeNnti ‘for the sake of him’
12	otti	on the lines of	avar karuththaiyotti ‘on the lines of his idea’
13	poRuththu	for the sake of	unkkaLaipoRuththu ‘for your sake’
14	poRuththavarai	as far as	eNNaippoRuththavarai ‘as far as I am concerned’

Table 5.7: Check List - Postpositions - Dative postposition

	Postposition	Meaning	Example
1	aaka	for	uNakkaaka ‘for you’
2	eNRu	for	uNekkeNRu ‘for you’
3	muN/muNNee/muNpu/ muNNaal/muNNaalee	before	uNakkumuN ‘before you’
4	piN/piNpu/piNNaal/ piNNaalee/piNNee	after, behind	eNakkuppiN ‘behind me’
5	uL/uLLee	inside	pettikkuL ‘into the box’
6	itaiyee/itaiyil	between, among	namakkitaiyee ‘among us’
7	natuvee/natuvil	between, among	namakkunatuvee ‘among us’
8	maththiyil	amidst	unkkaLukkumaththiyil ‘amidst you’
9	veLiyee	outside	viittukkuveLiyee ‘outside the house’
10	meel	above	viittukkumeel ‘above the house’
11	kiiz	under	meecaikkukkiiz ‘under the table’
12	ethiril/ethiree	opposite	viittukkethiril ‘opposite to the house’
13	pakkaththil	beside	eNakkuppakkaththil ‘beside me’
14	arukil/arukee	near	uNakkarukil ‘near you’
15	pathil/ pathilaaka	instead of	eNakkuppathil ‘instead of me’
16	maaRaaka	against	uNakkumaaRaaka ‘against you’
17	neeraaka	in front of	eNakkuneeraaka ‘in front of me’
18	uriya	belonging to	eNakkuriya ‘belonging to me’
19	uLLa	belonging to	eNakkuLLa ‘belonging to me’
20	thakuntha	suitable to	uNakkuthakuntha ‘suitable to you’

Table 5.8: Check List - Postpositions - Genitive Postpositions

	Postposition	Meaning	Example
1	miithu/miithil	on	enmiithu ‘on /with me’
2	meel/melee	on	enmeel ‘on / with me’
3	vaziyaaka	through	thalaivarvaziyaaka ‘through the Chairman’
4	vaayilaaka	through	thalaivarvaayilaaka ‘through the Chairman’
5	peeril	on	eNpeeril ‘on/with me’
6	muulam/muulamaaka	through	thalaivarmuulamaaka ‘through the Chairman’
7	poruttu	for the sake of	eNporuttu ‘for the sake of me’

Table 5.9: Check List - Postpositions - Locative postposition

	Postposition	Meaning	Example
1	“irunthu”	‘from’	“viittilirunthu” ‘from the house’

Table 5.10: Check List - Postpositions - Plain postpositions

	Postposition	Meaning	Example
1	utaN	with	eNNutaN ‘with me’
2	kuuta	with	eNkuuta ‘with me’
3	utaiya	of	kaNnNnaNutaiya ‘of Kannan’
4	vacam	on/with	eNvacam ‘with me’
5	itam	on/with	eNNitam ‘on/with me’
6	varai	till/upto	paththuvarai ‘upto ten’
7	aaka	for	paththunaaLaaka ‘for the past ten days’
8	aaka	as	aracaraaka ‘as king’
9	thooRum	at/every	viituthooRum ‘at every house’
10	aara	full of	kaNnNnaara ‘eyeful’

Table 5.11: Check List - Auxiliaries - Aspectuals

	Aspect	Auxiliary	Literal meaning
1	Progressive	koNntiru	keep and be
2	Perfective	iru	be
3	Definitive	vitu	leave
4	Trial	paar	see
5	Demonstrational	kaattu	show
6	Reflexive	koL	receive
7	Reciprocal	koL	receive
8	Accidental	poo	go
9	Contemptive	tolai	get lost
10	Preservative	vai	keep, preserve
11	Benefactive	aruL	sanction with mercy
12	Habitual	vaa	come

Table 5.12: Check List - Auxiliaries - Modals

	Modal	Auxiliary	Literal meaning
1	Inceptive	poo	go
2	Attemptive	paar	see
3	Probablitative	kuutu	join
4	Obligatory	veeNntu	need
5	Future negative	maattu	(denial)
6	Factive negative	illai	no, not

Table 5.13: Check List - Auxiliaries - Voices

	Voice	Auxiliary	Literal meaning
1	Passive	patu	"kollappatu" 'be killed'
2	Causative	vai	"kollavai" 'make-kill'

For example, the orthographic word "paRRi" may be a separate lexeme or a postposition in a sample text. If it is a postposition, it should have been added with the respective noun. It should not occur as an independent word. However, if it is a separate lexeme inflected for verbal participle, it should be written separately.

In the following two sentences,

(1) "avaN unkaLaippaRRip peeciNaaN" 'He talked about you'

(2) "avaN unkaL kaikaLaip paRRi izuththaaN" 'By holding your hand, he dragged,'

"paRRi" 'about' is a postposition in (1) added with the case inflected noun "unkkaLai" without any space between them; but in (2), since "paRRi" 'having held' is the verbal participle form of the verb lexeme "paRRu", it is written with a space from the case inflected noun "unkkaLai". This is the correct usage of the orthographic word "paRRi".

Suppose, by mistake, if the author of the particular text (though intended to use it as a postposition) leaves a space between the respective case inflected noun “unkkaLaip” and “paRRi,” the computer program would consider it as a separate token. That is, instead of taking “unkkaLaippaRRi”, as a single token, the mistake was done in the text would lead the computer program to consider this postposition as a separate token. It would also lead to incorrect POS tagging. “PaRRi” would be tagged as a verbal participle of the lexeme “paRRu” ‘to hold.’

5.7.3 Tokenization

For this, the sample texts were sent to the developed morphological parser where a check list of such susceptible auxiliary verbs and postpositions was provided, which in turn filtered them all. After this process, using the concordance tool, these were classified either as individual tokens or part of some other tokens. To finalize the tokens in the collected sample texts, some normalization process should be done which could be carried out only after parsing.

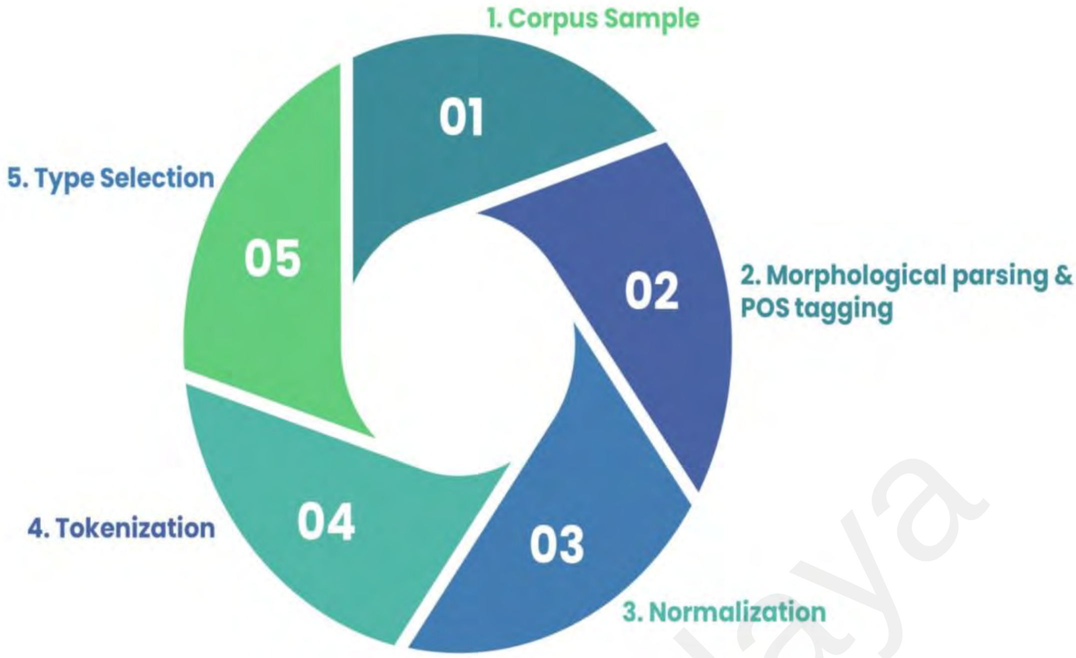


Figure 5.3 Token/type selection Process

5.8 Tokenization process

The output (tokens) from the morphological parser is represented thus:

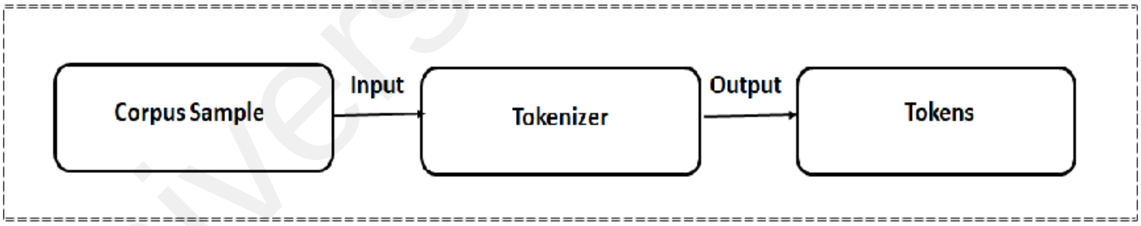


Figure 5.4: Tokenization Process

5.8.1 Parsed and unparsed words

The output from the morphological parser could be analysed into three parts.

(1) The first part is related to the unparsed words. A few thousand words could not be parsed by the parser because of two reasons: the root words or lexicons may not be found in the lexicon of the parser; or the morphotactics or morphophonemic rules present in the

parser may not be adequate to parse these words (Evangeline & Shyamala, 2020; Saravanan, 1999).

(2) The second part contains the words which could be parsed but having ambiguous POS tags.

(3) The third and the final part contains the words which are rightly parsed and rightly POS tagged. The issues involved in the above three parts are discussed below.

5.8.2 Wrong input forms

This morphological parser could not handle the orthographic words which are neither lexicons nor grammatical suffixes in Tamil. For example, the phonological form “maaka” is neither a lexicon nor a grammatical suffix. So, it comes under an unhandled words list. Likewise, some spoken Tamil words such as “pooRathu”, “pooRoom” “vaankkaNum” “mutincci” could not be handled by the parser. The orthographic word “vilakkeel” which is a pure literary word also could not be handled. “kaayaippaRiththu” “capaikkuvarukiRaar” are forms that are also under this category. When we visit the marked red words which could not be handled by the parser, we get many words such as above. But any Tamil speaker could understand the reason behind these unhandled forms.

Here it is to be noted that this is not due to any mistake in the parser, which is meant for written Tamil, but because of the wrong input forms. The solution for this issue is, either they could be corrected manually (“pooRathu” > “pookiRathu” ‘goes-it; “pooRoom” > “pookiRoom” ‘go-we’; “vaankkaNnum” > “vaankkaveeNntum” ’should be bought’; “mutiincci” > “mutinthu” ‘having completed’) or could be left out from the corpus data. Regarding the orthographic word “kaayaippaRiththu” “capaikkuvarukiRaar,” they could be understood by the Tamil speaker that they are part of Tamil. But they should be split into two parts - “kaayaip paRiththu” ‘the vegetable (acc.) having plucked’ > ‘having plucked the vegetable,’ “capaikku varukiRaar” ‘the assembly (dative) comes-he > ‘comes-he to the assembly’; then only the morphological parser could handle them. Therefore, all these orthographic words do not necessitate to modify the morphological parser.

5.8.3 Similar phonological forms (root level) but different POS categories

Examples: Noun and Verb:

“malar” (1) ‘flower’ (noun); (2) ‘blossom’ (verb)

“aatu” (1) ‘goat’ (noun); (2) ‘dance’ (verb)

“naatu” (1) ‘country’ (noun); (2) ‘seek’ (verb)

5.8.4 Root-level lexicon and inflected one

The orthographic word “veelai” can be a single root level lexicon for the noun root ‘job’; at the same time, it could be an inflected one which could be parsed into two parts: “veel + ai” ‘spear (acc.).’ This ambiguous tagging problem could be solved only with the help of linguistic context.

1. “avaNukkuk kataiyil veelai illai” ‘he has no job in the shop’
2. “avaN veelaith thozuthaaN” ‘he worshiped the spear’

5.8.5 Inflected words but with different POS categories.

1. Finite Verb (FV) and Relative Participle (RP):

   "varum" (1) 'will come' (FV)

   (2) 'that which will come' (RP)

   "tharum" (1) 'will give' (FV)

   (2) 'that which will give' (RP)

   "ootum" (1) 'will run' (FV)

   (2) 'that which will run' (RP)

2. Finite Verb (FV), Participial noun (Part. N) and Verbal noun (VN):

   "vanthathu" (1) 'came' (FV)

   (2) 'the one which came' (Part.N)

   (3) 'coming' (VN)

   "patiththathu" (1) 'studied -it' (FV)

   (2) 'the one which studied' (Part. N)

   (3) 'studying' (VN)

5.8.6 Lexical and grammatical words.

In the following orthographic words, the second part of each has ambiguous POS tags because of their ambiguity. They are either lexicons or grammatical suffixes.

Table 5.13: Orthographic Words

“notiththup **pooNathu**”	(“poo” ‘go’ / ‘Aspectual auxiliary’)
“thaNNaip **poola**”	(“poola” ‘like’ / ‘postposition’)
“kaattiN **uLLee**”	(“uLLee” ‘inside’ / ‘postposition’)
“kaaNnpaNavaRRaip **paRRi**”	(“paRRi” ‘having held’ / ‘postposition’)
“peyarkaLukku **eeRpa”**	(‘eeRpa” ‘to accept’ / ‘postposition’)
“aaNntukaLukku **muN**”	(‘muN” ‘before’ / ‘postposition’)
“aLavu **paRRiya**”	(“paRRiya” ‘held’ / ‘postposition’)
“ceythu **koNntaNar**”	(“koNntaNar” ‘had-they’ / ‘Aspectual auxiliary’)
“pooy **vitu**”	(“vitu” ‘give up’ / ‘Aspectual auxiliary’)
“thirutik **koNntu”**	(“koNntu” ‘having had’ / ‘Aspectual auxiliary’)
“etuththuk **koLvaaN**”	(“koLvaan” ‘will have-he’ / ‘Aspectaul auxiliary’)
“muttai ittu **vanthathu”**	(“vanthathu” ‘came-it’ / ‘Aspectual auxiliary’)
“vaLarththu **vanthaaN**”	(“vanthaaN” ‘came-he’ / ‘Aspectual auxiliary’)

Similarly, in English, some word forms like “have”, “can”, “be” ,“do” are either lexicons or grammatical words. Only the linguistic contexts help to disambiguate them. In the sentence “I have a book”, “have” is a lexicon; in “I have gone”, “have” is a grammatical word.

In the history of the Tamil language, some lexicons had been grammaticalised. However, after this grammaticalisation process, they continue to be the lexicons also. However, the grammaticalised word forms, when do the grammatical function, they could not occur as a free morpheme or word in Tamil. This is the criteria to decide whether they are lexicons or grammatical words. Depending upon the previous linguistic contexts, it could be decided whether they are lexicons or grammatical words.

For example, in the words “kaaNnpaNavaRRaip paRRi,” the second word may be a lexicon (as a verbal participle form of the main verb “paRRu” ‘to hold’) or a postposition (‘about’). If it is a separate lexicon, it should occur independently (as “kaaNnpaNavaRRaip paRRi”; ‘seen-they (acc.) having held’ > ‘having held the seen-they (acc.)’ whereas if it is a postposition which could follow an accusative case phrase, it should be added with the casal phrase as “kaaNnpaNavaRRaippaRRi” ‘about seen-they (acc.)’ without any space. The reason is, in Tamil, no grammatical word could occur independently. Since the above-mentioned example “paRRi” occurred separately, the parser tagged this as a verbal participle though in fact, it is a postposition. This is a wrong tagging. This problem could be solved only with the help of concordancer.

5.8.7 Segmentation problem

Another problem is how the parser segments an input. For example, with the orthographic word “patiththavarai” in the following sentences, the parser would segment this in two ways:

(1) “nii patiththavarai poothum” ‘It is enough upto what you read’

(2) “nii patiththavaraip paarththaayaa?” ‘Did you see the one who read?’

“patiththa - varai” > “patiththavarai” ‘upto one read’

“patiththavar -ai” > “patiththavarai” ‘the one who read (acc.)’

That is, the orthographic input word gives a place for two ways of segmentation which leads to providing two POS tags to this orthographic word. Only with the help of the concordance program, this problem could be solved.

5.8.8 Initial Parsing and POS Tagging

All the above problems with the segmentation of the input words and disambiguation of POS tagging of the input words could be solved with the concordance programme. This normalisation task is very much important with the sample texts, and in the present research project, this task is done with the initial parsing and POS tagging of the developed morphological parser.

With the words which have lexical POS as well as the grammatical POS (such as Postpositions and Auxiliary verbs), once the right POS is selected, the next task is normalization. If a word has a grammatical POS, it should be joined with its preceding lexical word because, in Tamil, as it is earlier explained, grammatical words could not occur independently.

Example: “avaN inkku vanthu, eNNaip paarththaaN” ‘He came here and saw me’

“avaN athaic ceythupaarththaaN” ‘He tried to do that’.

Here, the orthographic word “paarththaaN” in the first example is an inflected lexical word; in the second example, it is an aspectual auxiliary verb that is preceded by a verbal participle form of a lexical verb. In the second example, it should be combined with the main lexical verb. By this, now they will become a single token, whereas, in the first example, it is a separate token.

Likewise, with the inflected words, “vanthathu,” “patiththathu,” after selecting the correct POS tag for a particular form in a particular sentence with the help of the concordance program, it is possible to decide whether they are three different types or a single type.

5.8.9 Completed Tasks

At this stage, all the following tasks are completed:

1. Finalization of tokens
2. Finalization of types with POS tags.
3. Identification of new grammatical suffixes with proper tags.
4. Identification of new morphotactic rules, if any
5. Changes in morphophonemic rules, if any
6. Enrichment of Tamil database/ lexicon

5.8.10 Enhanced Parser

Based on the above tasks, now the morphological parser is further updated. Now the finalized types could be sent for morphological parsing and POS tagging. The samples could be sent to the morphological parser. After this process, the word forms were parsed as either lexical roots or stems with the help of the enhanced morphological parser. That is, the samples were fed at regular intervals, and the resultant residual issues were remedied by modifying the lexicon and suffixes lists; morphotactics; and morphophonemic rules accordingly till the maximum level was reached.

The following diagram shows the morphological parsing of a noun (“malarkaL” ‘flowers’) and a verb (“patiththaaN” ‘He read’).

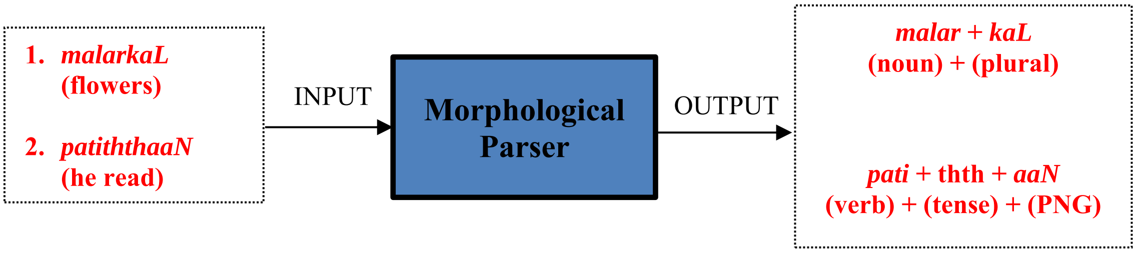


Figure 5.5: Morphological Parsing

5.8.11 POS Tagger

Once the word forms are segmented into lexical roots or stems, the next task to address is to tag every word form with its POS (Parts-of-Speech). This is done by the POS tagger program which is based on the output segments with their respective categories of every word form.

To decide the word forms or types using the Morphological Parser or to finalize the POS of a Tamil word form or type, the knowledge of computational morphological algorithm of Tamil was provided to the morphological parser and POS tagger tools/programs.

The knowledge of Tamil morphology was fed to the morphological parser program. So, once a word form was input into the parser it had been parsed into its respective segments.

The knowledge of Tamil inflectional suffixes was provided to the POS tagger programme. Once a word form was segmented into their segments - morphs -, the POS tagger could tag it for its legitimate POS.

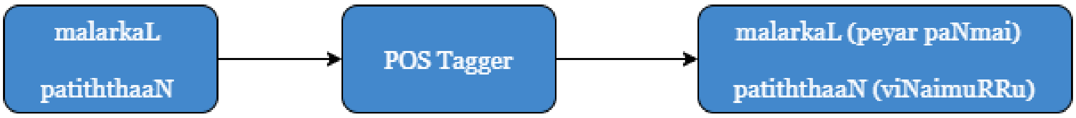


Figure 5.6: POS Tagging

Here the tagged Tamil word classes are: malarkaL (peyar paNmai) ‘Flowers (noun-plural: NPL)’; patiththaaN (viNaimuRRu) ‘Read-he (finite-verb: FV)’

1. Flowers/ flowers (noun-plural: NPL)
2. Read-he/read-he (finite-verb: FV)

5.8.12 Disambiguation Tool

Some word forms may be given more than one tag. This challenge of ambiguities is addressed based on the linguistic context using an N-gram analyser and concordancer. The process involved in software system architecture for morphological parsing and POS tagging was discussed earlier, and the enhanced process is represented here:

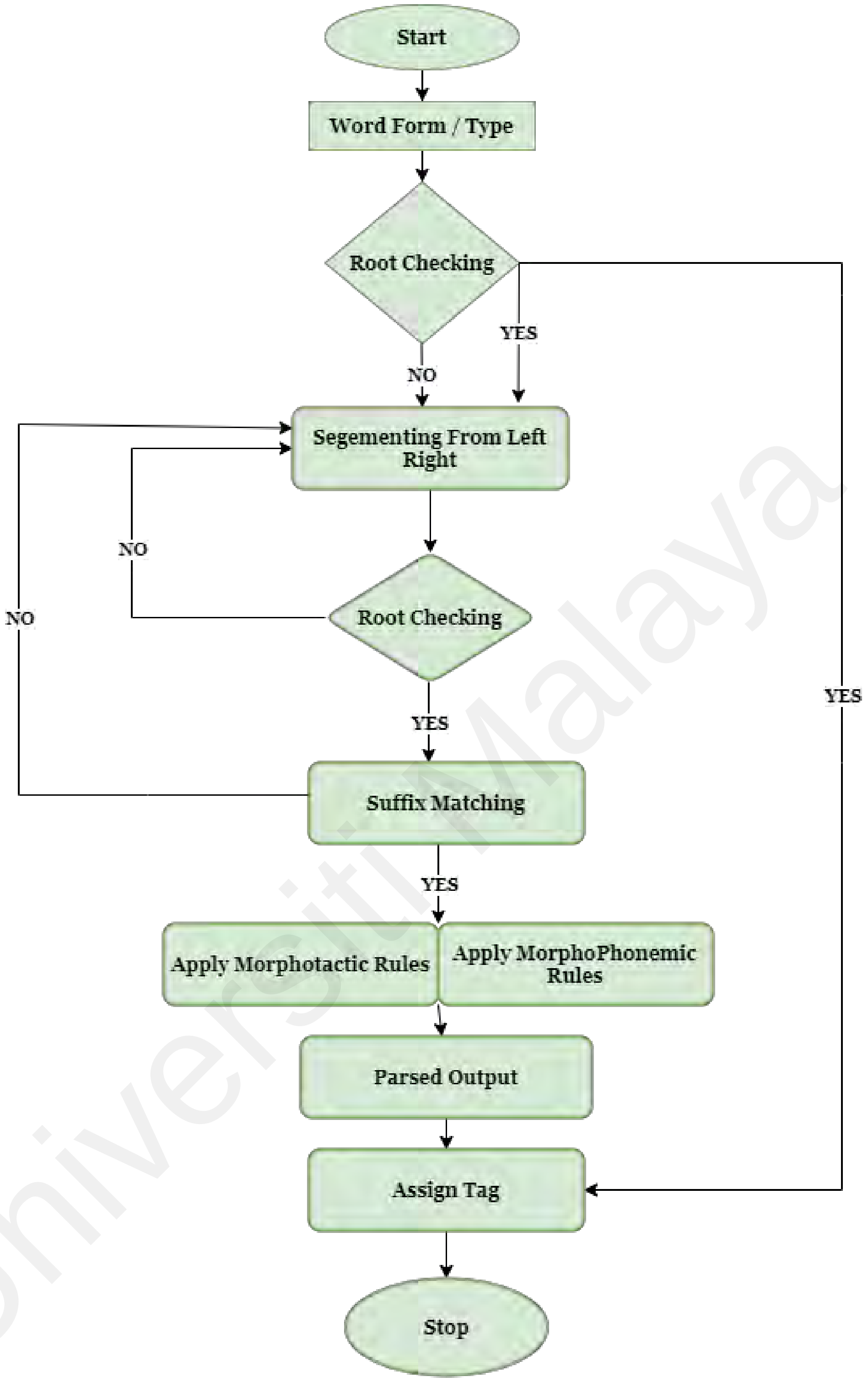


Figure 5.7: Software system architecture for morphological parsing

5.9 Software Design

The present software was developed using .net framework 3.5 platform. The programming language is C#. The corpus metadata is stored in XML file format, and sample text is stored in plain text file format. The software tools found in the software were developed using a developer tool - DevExpress. The results are displayed via the database. The encoding of Tamil characters is UTF-8 (Unicode). The software developed for this research is a desktop application. The software is compatible with the Microsoft Window platform. Also, some of the editing tools available in MS Word, MS Excel, such as sorting, duplicate removal, find and replace are used in this project.

5.10 Developer Tools

a. DevExpress

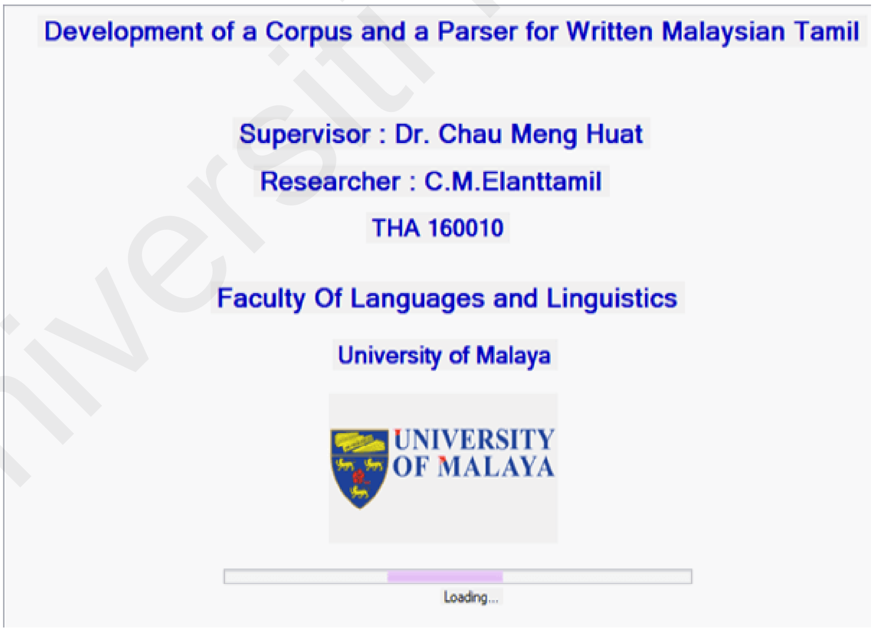
"DevExpress Universal is a complete software development package for .NET developers. It helps to build applications for Windows, Web, mobile and tablet. DevExpress universal subscription includes source code for all WinForms, DevExtreme HTML5 Widgets, ASP.NET, WPF, Dashboard and Windows 10 Apps controls. DevExpress - is a custom third-party provider of .NET controls. They customize the. NET controls by making it more attractive and more flexible than inbuilt." (https://www.devexpress.com)

b.NET and C#:

“.NET is designed to provide a new environment within which we can develop almost any application to run on Windows, whereas C# is a programming language that has been designed specifically to work with .NET. Using C# one can write a dynamic Web page, an XML Web service, a component of a distributed application, a database access component, a classic Windows desktop application, or even a new smart client application that allows for online / offline capabilities.

Both the .NET Framework and C# are entirely based on object-oriented principles right from the start. C# is an object-oriented language intended for use with .NET.” (Nagel et al., 2010)

5.11 Software Design



Development of a Corpus and a Parser for Written Malaysian Tamil

Supervisor : Dr. Chau Meng Huat

Researcher : C.M.Elanttamil

THA 160010

Faculty Of Languages and Linguistics

University of Malaya

Figure 5.8: Index Page

5.11.1 Main Window

The Main Window of the present software consists of 5 Menus: (1) File, (2) Home, (3) Corpus Construction, (4) Corpus Analysis, and (5) Morphological Analysis. These 5 Main Windows have Child Windows accordingly.

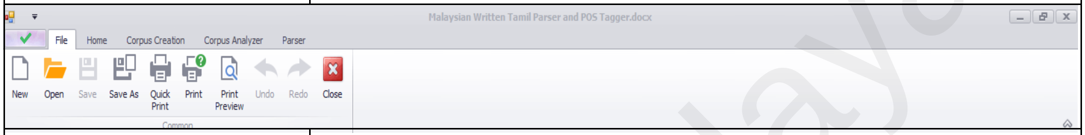
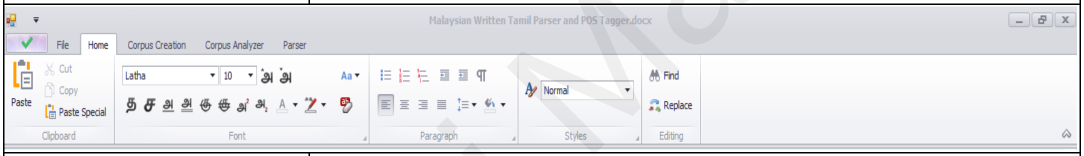
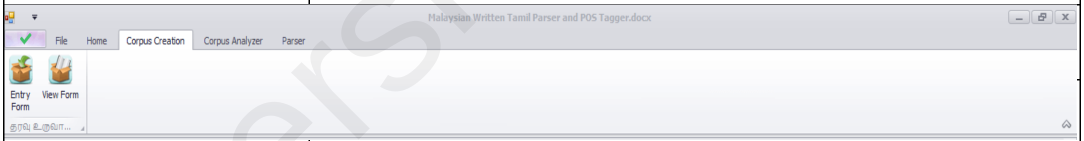
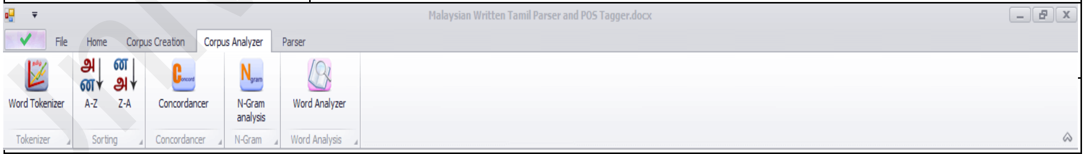
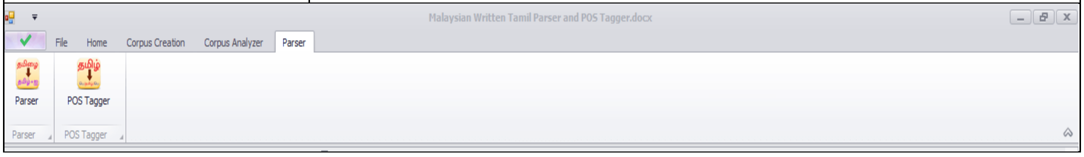
<table>
<tr><th>Menu</th><th>Child Windows</th></tr>
<tr><td>File</td><td>New, Open, Save, Save as, Quick print, Print, Preview, Undo, Redo, Exit</td></tr>
<tr><td colspan="2">
</td></tr>
<tr><td>Home</td><td>Clipboard, Font, Paragraph, Style, Editing</td></tr>
<tr><td colspan="2">
</td></tr>
<tr><td>Corpus Creation</td><td>Entry Form and View Form</td></tr>
<tr><td colspan="2">
</td></tr>
<tr><td>Corpus Analyser</td><td>Analyzer</td></tr>
<tr><td colspan="2">
</td></tr>
<tr><td>Parser</td><td>Parser, POS Tagger</td></tr>
<tr><td colspan="2">
</td></tr>
</table>

Figure 5.9: Main Window Menus

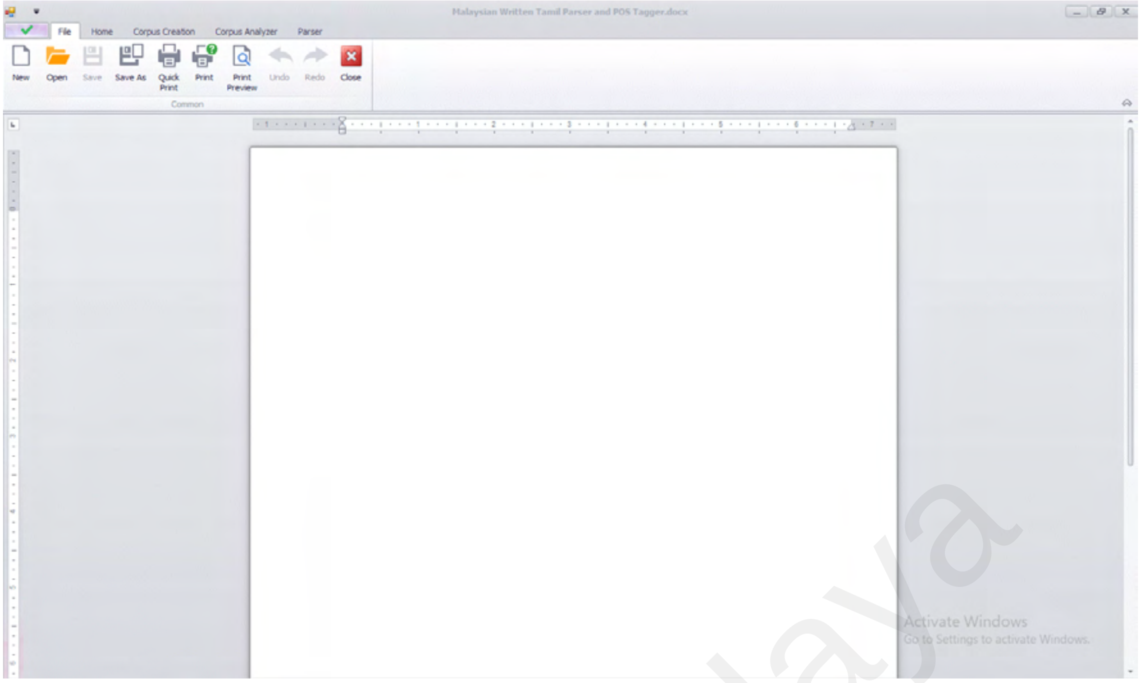
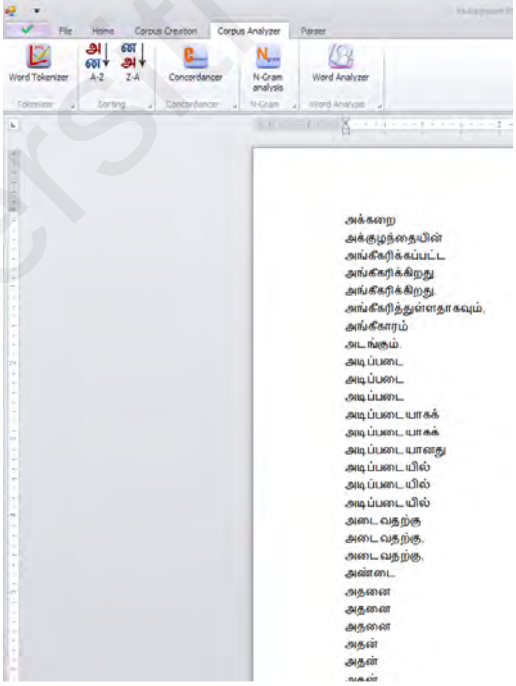


Figure 5.10: Main Window

Sorting tools (Ascending - Descending) are available in this software:



அக்கறை
அக்குழந்தையின்
அங்கீகரிக்கப்பட்ட
அங்கீகரிக்கிறது
அங்கீகரிக்கிறது.
அங்கீகரித்துள்ளதாகவும்,
அங்கீகாரம்
அடங்கும்.
அடிப்படை
அடிப்படை
அடிப்படை
அடிப்படையாகக்
அடிப்படையாகக்
அடிப்படையானது
அடிப்படையில்
அடிப்படையில்
அடிப்படையில்
அடைவதற்கு
அடைவதற்கு.
அடைவதற்கு.
அண்டை
அதனை
அதனை
அதனை
அதன்
அதன்

Figure 5.11: Ascending - Descending

5.11.2 Input Form and Review Form:

This menu consists of two Child Windows - Input Form and Review Form. The first ‘Input Form’ does the task of storing the sample texts with the necessary meta data in XML file format. The second ‘Review Form’ is useful to retrieve a particular sample for reviewing.

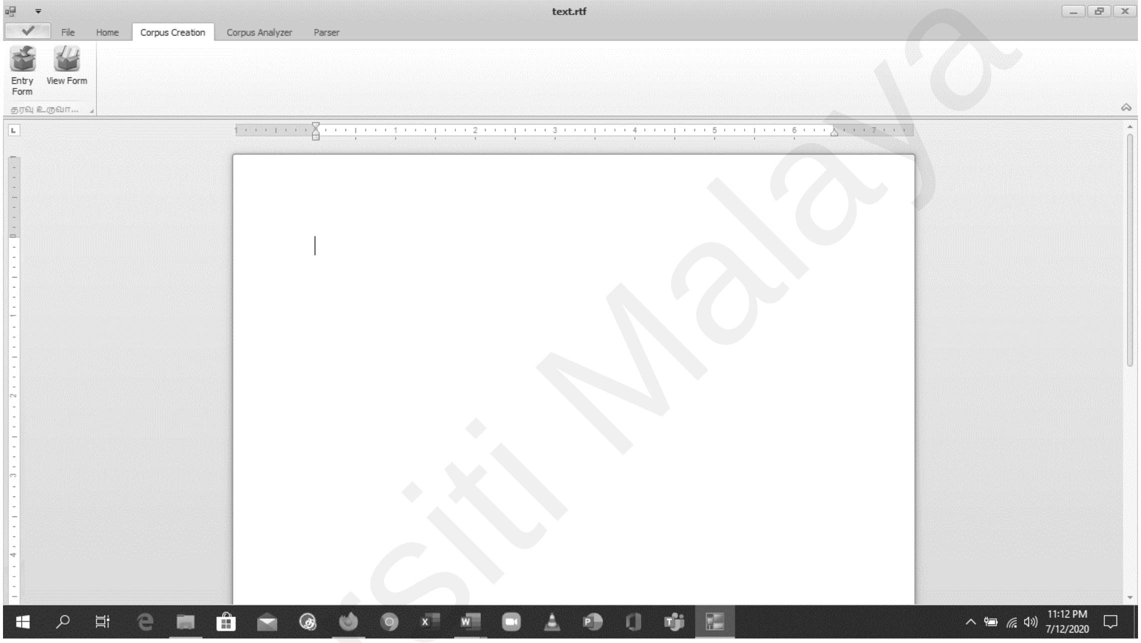


Figure 5.12: Input and Review

5.11.3 Input Form

This form consists of two Child Windows: (1) meta data forms and (2) sample text input Child Window. Then the user has to go to the second option sub-menu, whether the text sample is same type or mixed one. The third option button requires the user to specify whether the sample text is a monologue or a part of an interaction among persons. The fourth one asks the user to specify whether the sample text is a typed one (keying) or copied from some other source. The next fifth one asks the user regarding the domain of the particular sample text. The next Child Window is for choosing the written or spoken variety. Though the present research project is concerned only with Written Tamil, here provision for storing spoken variety also.

Once the user chooses a particular variety - written or spoken - this input tool requires information regarding the type of text: written - printed, manuscript, palm leaves or digital texts; spoken - conversation, lecture, radio, television, and telephone. Once a type is selected, the user has to provide other details regarding the input sample text.

5.11.4 Sample Input

Once all the meta data related to the particular sample text is provided as above, the task is to input the particular sample. Then the particular sample with necessary meta data are stored in XML file.

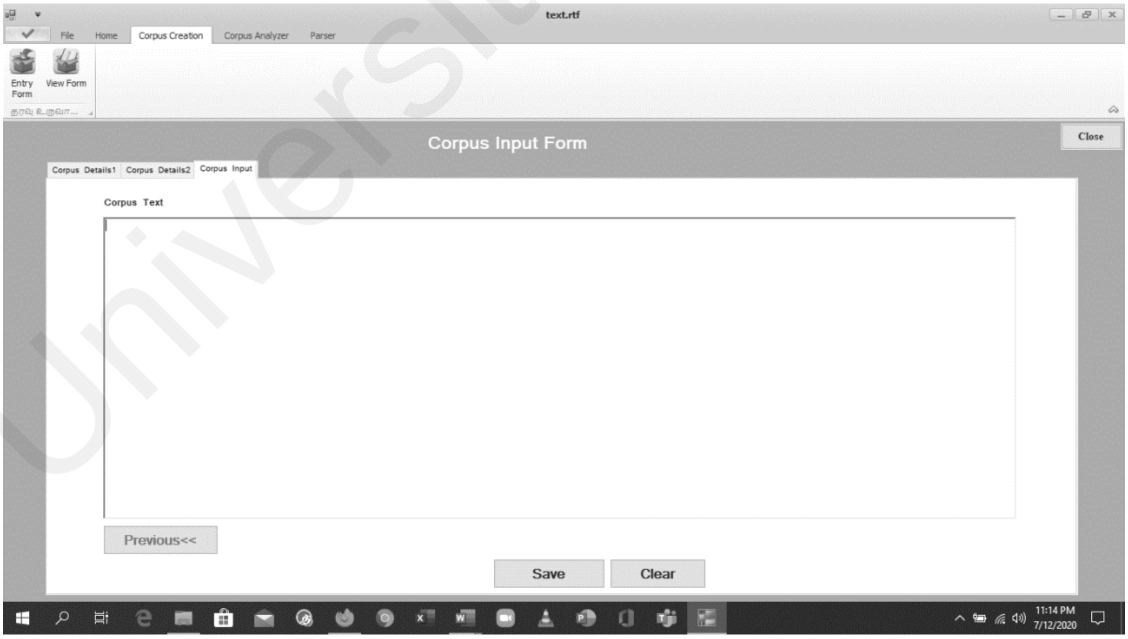
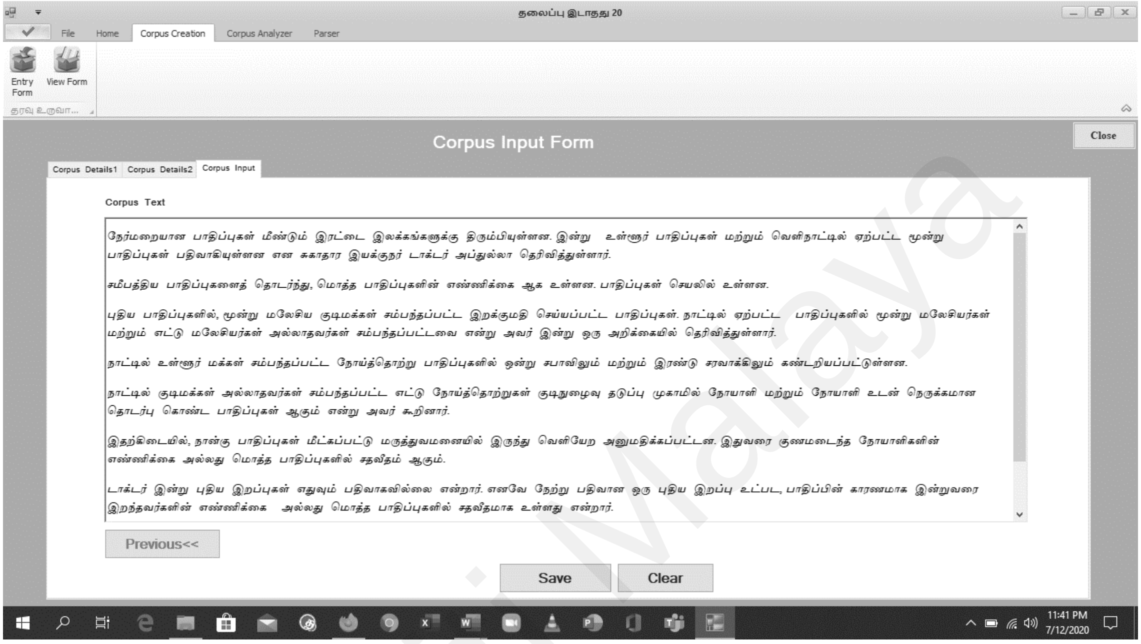


Figure 5.13: Sample Output

5.11.5 Sample Review

The user could retrieve and review the already input sample by using the meta data variables.



நேர்மறையான பாதிப்புகள் மீண்டும் இரட்டை இலக்கங்களுக்கு திரும்பியுள்ளன. இன்று உள்ளூர் பாதிப்புகள் மற்றும் வெளிநாட்டில் ஏற்பட்ட மூன்று பாதிப்புகள் பதிவாகியுள்ளன என சுகாதார இயக்குநர் டாக்டர் அப்துல்லா தெரிவித்துள்ளார்.

சமீபத்திய பாதிப்புகளைத் தொடர்ந்து, மொத்த பாதிப்புகளின் எண்ணிக்கை ஆக உள்ளன. பாதிப்புகள் செயலில் உள்ளன.

புதிய பாதிப்புகளில், மூன்று மலேசிய குடிமக்கள் சம்பந்தப்பட்ட இறக்குமதி செய்யப்பட்ட பாதிப்புகள். நாட்டில் ஏற்பட்ட பாதிப்புகளில் மூன்று மலேசியர்கள் மற்றும் எட்டு மலேசியர்கள் அல்லாதவர்கள் சம்பந்தப்பட்டவை என்று அவர் இன்று ஒரு அறிக்கையில் தெரிவித்துள்ளார்.

நாட்டில் உள்ளூர் மக்கள் சம்பந்தப்பட்ட நோய்த்தொற்று பாதிப்புகளில் ஒன்று சபாவிலும் மற்றும் இரண்டு சரவாக்கிலும் கண்டறியப்பட்டுள்ளன.

நாட்டில் குடிமக்கள் அல்லாதவர்கள் சம்பந்தப்பட்ட எட்டு நோய்த்தொற்றுகள் குடிநுழைவு தடுப்பு முகாமில் நோயாளி மற்றும் நோயாளி உடன் நெருக்கமான தொடர்பு கொண்ட பாதிப்புகள் ஆகும் என்று அவர் கூறினார்.

இதற்கிடையில், நான்கு பாதிப்புகள் மீட்கப்பட்டு மருத்துவமனையில் இருந்து வெளியேற அனுமதிக்கப்பட்டன. இதுவரை குணமடைந்த நோயாளிகளின் எண்ணிக்கை அல்லது மொத்த பாதிப்புகளில் சதவீதம் ஆகும்.

டாக்டர் இன்று புதிய இறப்புகள் எதுவும் பதிவாகவில்லை என்றார். எனவே நேற்று பதிவான ஒரு புதிய இறப்பு உட்பட, பாதிப்பின் காரணமாக இன்றுவரை இறந்தவர்களின் எண்ணிக்கை அல்லது மொத்த பாதிப்புகளில் சதவீதமாக உள்ளது என்றார்.

Figure 5.14: Sample Review

The input data could be viewed as above. All the stored sample files with necessary details are shown as above. The user can choose the needed file from the list.

5.11.6 Corpus Analysis

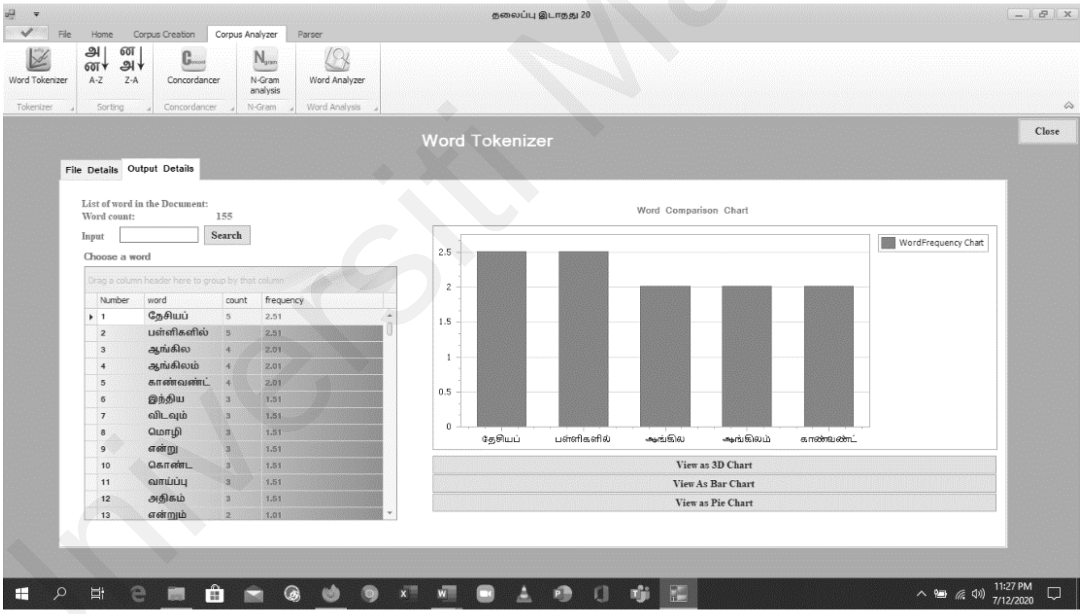
Now the constructed corpus could be analysed for various linguistic tasks. The following tasks could be done with the help of the tools provided here.

1. Tokenization and type retrieval
2. Concordancing

5.11.7 Word Processing (Spell Checking and Grammar Checking)

Before proceeding with the task of tokenization, first of all, all the sample texts in the corpus should undergo spell checking and grammar checking. In the present research project, these tasks are carried out by using Tamil Word Processor - MenTamizh. The mistakes might have been caused by typos or scanning and character recognition processing. Once this word processing task is done, the sample texts should be checked for normalization as per the grammar of written Malaysian Tamil (as discussed in detail elsewhere).

5.11.8 Tokenizer and Type Retrieval



List of word in the Document:
Word count: 155
Input
Choose a word
Drag a column header here to group by that column

Number	word	count	frequency
1	தேசியப்	5	2.51
2	பள்ளிகளில்	5	2.51
3	ஆங்கில	4	2.01
4	ஆங்கிலம்	4	2.01
5	காண்வண்ட்	4	2.01
6	இந்திய	3	1.51
7	விடவும்	3	1.51
8	மொழி	3	1.51
9	என்று	3	1.51
10	கொண்ட	3	1.51
11	வாய்ப்பு	3	1.51
12	அதிகம்	3	1.51
13	என்றும்	2	1.01

Figure 5.15: Tokenizer & Type

Now, the tokenizer tool could be applied. Based on the space criteria, this tool will identify all the tokens found in the sample text. Then the next task done by this tool is retrieval of types from the tokens. The identified types are listed in the table with their number of occurrences, and their frequencies as well are listed in the respective boxes. The frequency in ascending and descending order could be displayed. Also, the graphical representation of the same is displayed through three types of charts. The total number of tokens found in the sample texts is mentioned at the top. In this Window, a text box for searching any type among the listed ones is provided.

5.11.9 Concordancer

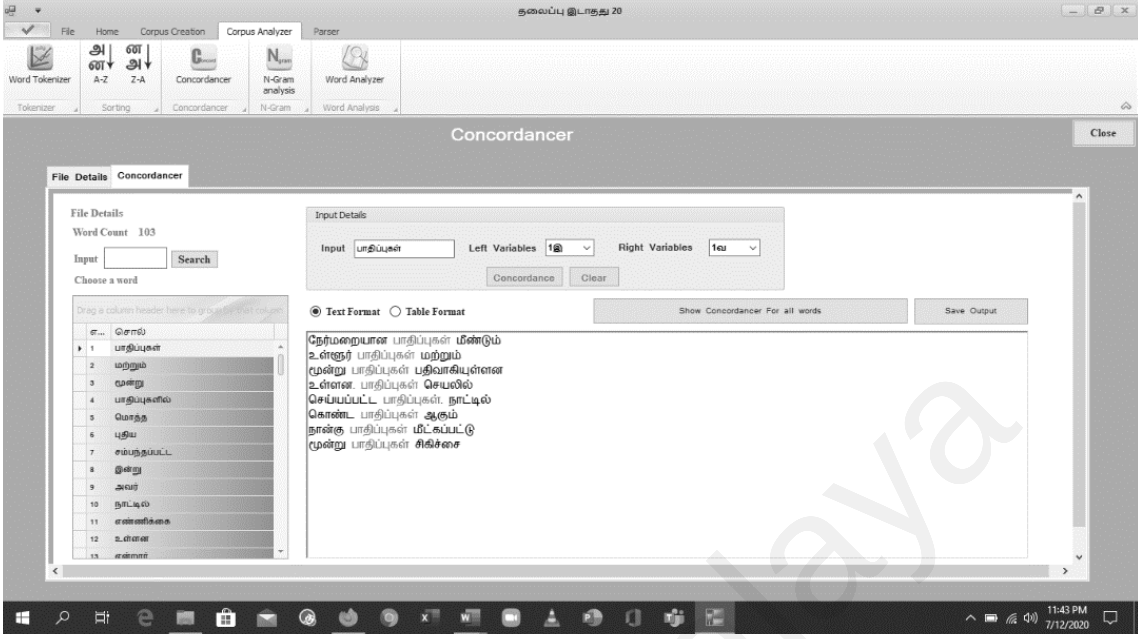
This concordancer tool works in two ways:

(1) Individual words could be given in the search box to get their concordances in the sample text.

(2) All the words found in the sample text could be concordanced.

Before starting the concordance for a word, the number of preceding and following words as variables could be selected. The concordance could be displayed in two ways - (1) in text format and (2) in table format.

5.11.10 Text Format



File Details

Word Count 103

Input

Choose a word

எ...	சொல்
1	பாதிப்புகள்
2	மற்றும்
3	மூன்று
4	பாதிப்புகளில்
5	மொத்த
6	புதிய
7	சம்பந்தப்பட்ட
8	இன்று
9	அவர்
10	நாட்டில்
11	எண்ணிக்கை
12	உள்ளன

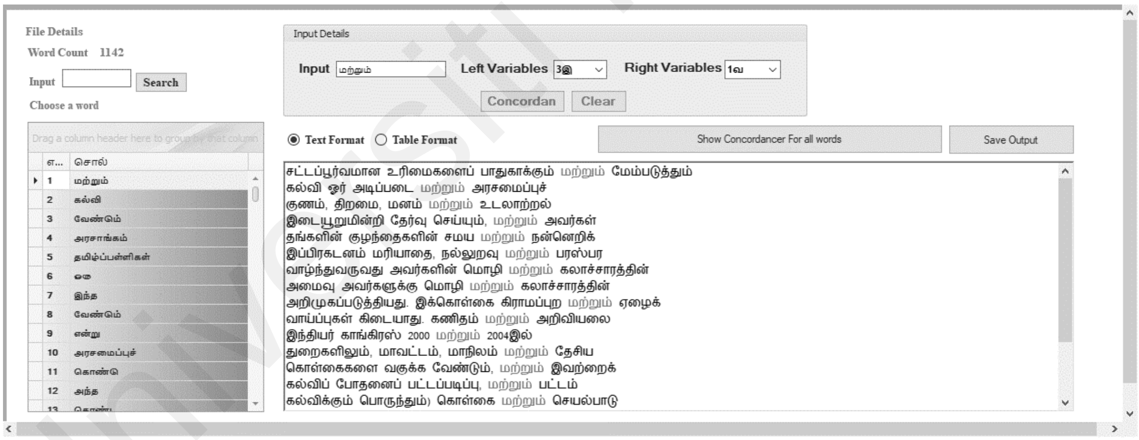
Input Details

Input பாதிப்புகள் Left Variables 1இ Right Variables 1வ

◉ Text Format ○ Table Format

```
நேர்மறையான பாதிப்புகள் மீண்டும்
உள்ளூர் பாதிப்புகள் மற்றும்
மூன்று பாதிப்புகள் பதிவாகியுள்ளன
உள்ளன. பாதிப்புகள் செயலில்
செய்யப்பட்ட பாதிப்புகள். நாட்டில்
கொண்ட பாதிப்புகள் ஆகும்
நான்கு பாதிப்புகள் மீட்கப்பட்டு
மூன்று பாதிப்புகள் சிகிச்சை
```

Figure 5.16: Concordance Text Format (-N1+1)



File Details

Word Count 1142

Input

Choose a word

எ...	சொல்
1	மற்றும்
2	கல்வி
3	வேண்டும்
4	அரசாங்கம்
5	தமிழ்ப்பள்ளிகள்
6	ஒரு
7	இந்த
8	வேண்டும்
9	என்று
10	அரசமைப்புச்
11	கொண்டு
12	அந்த

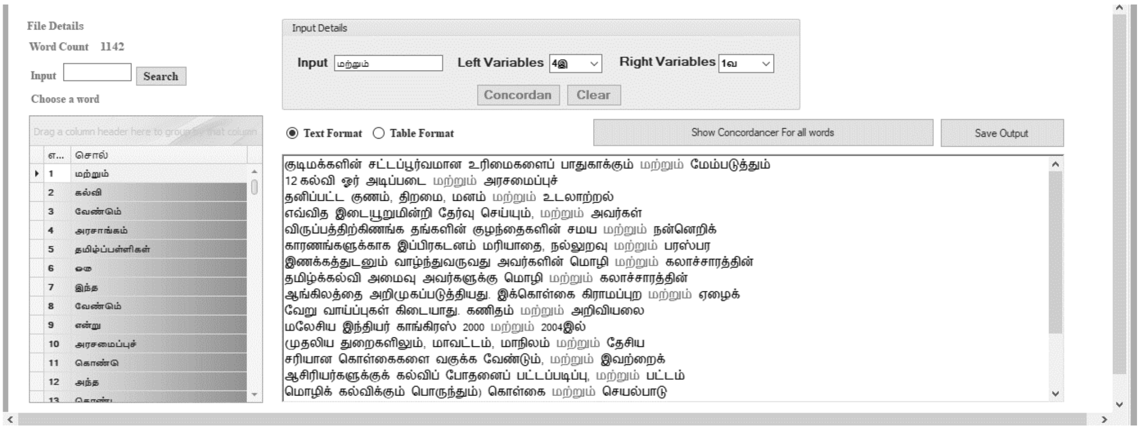
Input Details

Input மற்றும் Left Variables 3இ Right Variables 1வ

◉ Text Format ○ Table Format

```
சட்டப்பூர்வமான உரிமைகளைப் பாதுகாக்கும் மற்றும் மேம்படுத்தும்
கல்வி ஓர் அடிப்படை மற்றும் அரசமைப்புச்
குணம், திறமை, மனம் மற்றும் உடலாற்றல்
இடையூறுமின்றி தேர்வு செய்யும், மற்றும் அவர்கள்
தங்களின் குழந்தைகளின் சமய மற்றும் நன்னெறிக்
இப்பிரகடனம் மரியாதை, நல்லுறவு மற்றும் பரஸ்பர
வாழ்ந்துவருவது அவர்களின் மொழி மற்றும் கலாச்சாரத்தின்
அமைவு அவர்களுக்கு மொழி மற்றும் கலாச்சாரத்தின்
அறிமுகப்படுத்தியது. இக்கொள்கை கிராமப்புற மற்றும் ஏழைக்
வாய்ப்புகள் கிடையாது. கணிதம் மற்றும் அறிவியலை
இந்தியர் காங்கிரஸ் 2000 மற்றும் 2004இல்
துறைகளிலும், மாவட்டம், மாநிலம் மற்றும் தேசிய
கொள்கைகளை வகுக்க வேண்டும், மற்றும் இவற்றைக்
கல்விப் போதனைப் பட்டப்படிப்பு, மற்றும் பட்டம்
கல்விக்கும் பொருந்தும்) கொள்கை மற்றும் செயல்பாடு
```

Figure 5.17: Concordance Text Format (-N3+N1)



File Details

Word Count 1142

Input Search

Choose a word

எ...	சொல்
1	மற்றும்
2	கல்வி
3	வேண்டும்
4	அரசாங்கம்
5	தமிழ்ப்பள்ளிகள்
6	ஒரு
7	இந்த
8	வேண்டும்
9	என்று
10	அரசமைப்புச்
11	கொண்டு
12	அந்த
13	கொண்டு

Input Details

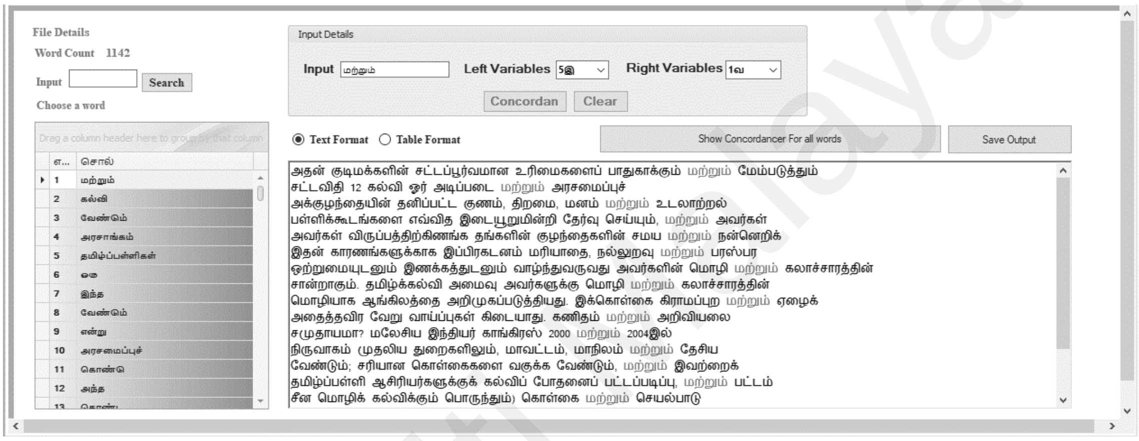
Input மற்றும் Left Variables 4இ Right Variables 1வ

● Text Format ○ Table Format

```
குடிமக்களின் சட்டப்பூர்வமான உரிமைகளைப் பாதுகாக்கும் மற்றும் மேம்படுத்தும்
12 கல்வி ஓர் அடிப்படை மற்றும் அரசமைப்புச்
தனிப்பட்ட குணம், திறமை, மனம் மற்றும் உடலாற்றல்
எவ்வித இடையூறுமின்றி தேர்வு செய்யும், மற்றும் அவர்கள்
விருப்பத்திற்கிணங்க தங்களின் குழந்தைகளின் சமய மற்றும் நன்னெறிக்
காரணங்களுக்காக இப்பிரகடனம் மரியாதை, நல்லுறவு மற்றும் பரஸ்பர
இணக்கத்துடனும் வாழ்ந்துவருவது அவர்களின் மொழி மற்றும் கலாச்சாரத்தின்
தமிழ்க்கல்வி அமைவு அவர்களுக்கு மொழி மற்றும் கலாச்சாரத்தின்
ஆங்கிலத்தை அறிமுகப்படுத்தியது. இக்கொள்கை கிராமப்புற மற்றும் ஏழைக்
வேறு வாய்ப்புகள் கிடையாது. கணிதம் மற்றும் அறிவியலை
மலேசிய இந்தியர் காங்கிரஸ் 2000 மற்றும் 2004இல்
முதலிய துறைகளிலும், மாவட்டம், மாநிலம் மற்றும் தேசிய
சரியான கொள்கைகளை வகுக்க வேண்டும், மற்றும் இவற்றைக்
ஆசிரியர்களுக்குக் கல்விப் போதனைப் பட்டப்படிப்பு, மற்றும் பட்டம்
மொழிக் கல்விக்கும் பொருந்தும்) கொள்கை மற்றும் செயல்பாடு
```

Drag a column header here to group by that column
Concordan
Clear
Show Concordancer For all words
Save Output

Figure 5.18: Concordance Text Format (-N4+N1)



File Details

Word Count 1142

Input Search

Choose a word

எ...	சொல்
1	மற்றும்
2	கல்வி
3	வேண்டும்
4	அரசாங்கம்
5	தமிழ்ப்பள்ளிகள்
6	ஒரு
7	இந்த
8	வேண்டும்
9	என்று
10	அரசமைப்புச்
11	கொண்டு
12	அந்த
13	கொண்டு

Input Details

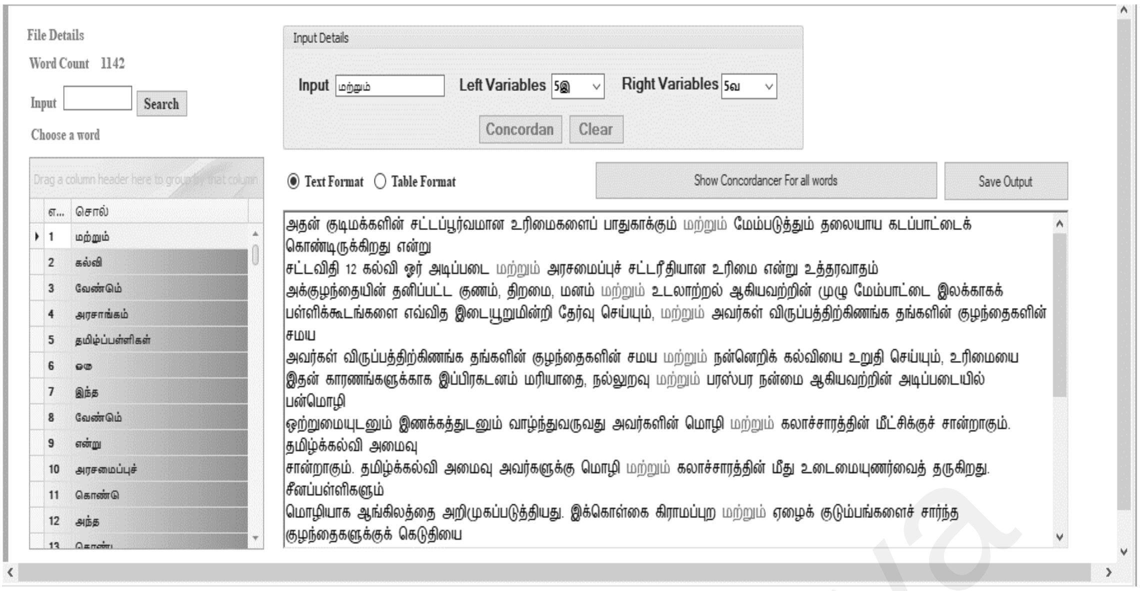
Input மற்றும் Left Variables 5இ Right Variables 1வ

● Text Format ○ Table Format

```
அதன் குடிமக்களின் சட்டப்பூர்வமான உரிமைகளைப் பாதுகாக்கும் மற்றும் மேம்படுத்தும்
சட்டவிதி 12 கல்வி ஓர் அடிப்படை மற்றும் அரசமைப்புச்
அக்குழந்தையின் தனிப்பட்ட குணம், திறமை, மனம் மற்றும் உடலாற்றல்
பள்ளிக்கூடங்களை எவ்வித இடையூறுமின்றி தேர்வு செய்யும், மற்றும் அவர்கள்
அவர்கள் விருப்பத்திற்கிணங்க தங்களின் குழந்தைகளின் சமய மற்றும் நன்னெறிக்
இதன் காரணங்களுக்காக இப்பிரகடனம் மரியாதை, நல்லுறவு மற்றும் பரஸ்பர
ஒற்றுமையுடனும் இணக்கத்துடனும் வாழ்ந்துவருவது அவர்களின் மொழி மற்றும் கலாச்சாரத்தின்
சான்றாகும். தமிழ்க்கல்வி அமைவு அவர்களுக்கு மொழி மற்றும் கலாச்சாரத்தின்
மொழியாக ஆங்கிலத்தை அறிமுகப்படுத்தியது. இக்கொள்கை கிராமப்புற மற்றும் ஏழைக்
அதைத்தவிர வேறு வாய்ப்புகள் கிடையாது. கணிதம் மற்றும் அறிவியலை
சமுதாயமா? மலேசிய இந்தியர் காங்கிரஸ் 2000 மற்றும் 2004இல்
நிருவாகம் முதலிய துறைகளிலும், மாவட்டம், மாநிலம் மற்றும் தேசிய
வேண்டும்; சரியான கொள்கைகளை வகுக்க வேண்டும், மற்றும் இவற்றைக்
தமிழ்ப்பள்ளி ஆசிரியர்களுக்குக் கல்விப் போதனைப் பட்டப்படிப்பு, மற்றும் பட்டம்
சீன மொழிக் கல்விக்கும் பொருந்தும்) கொள்கை மற்றும் செயல்பாடு
```

Drag a column header here to group by that column
Concordan
Clear
Show Concordancer For all words
Save Output

Figure 5.19: Concordance Text Format (-N5+N1)



File Details

Word Count 1142

Input

Choose a word

எ...	சொல்
1	மற்றும்
2	கல்வி
3	வேண்டும்
4	அரசாங்கம்
5	தமிழ்ப்பள்ளிகள்
6	ஒரு
7	இந்த
8	வேண்டும்
9	என்று
10	அரசமைப்புச்
11	கொண்டு
12	அந்த

Input Details

Input: மற்றும் Left Variables: 5இ Right Variables: 5வ

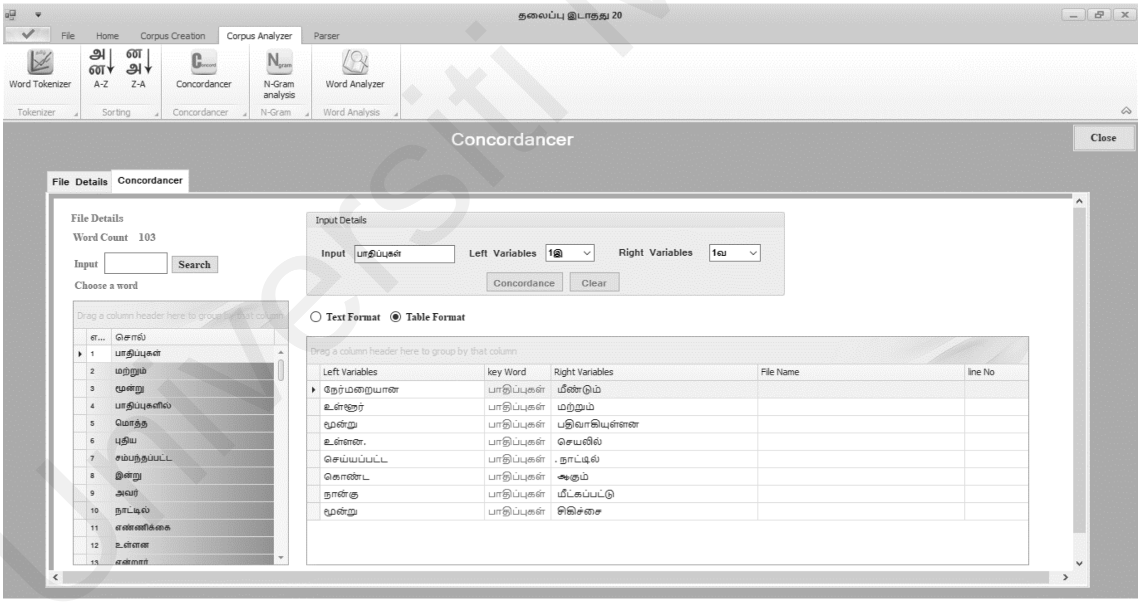
◉ Text Format ○ Table Format

அதன் குடிமக்களின் சட்டப்பூர்வமான உரிமைகளைப் பாதுகாக்கும் மற்றும் மேம்படுத்தும் தலையாய கடப்பாட்டைக் கொண்டிருக்கிறது என்று
சட்டவிதி 12 கல்வி ஓர் அடிப்படை மற்றும் அரசமைப்புச் சட்டரீதியான உரிமை என்று உத்தரவாதம்
அக்குழந்தையின் தனிப்பட்ட குணம், திறமை, மனம் மற்றும் உடலாற்றல் ஆகியவற்றின் முழு மேம்பாட்டை இலக்காகக்
பள்ளிக்கூடங்களை எவ்வித இடையூறுமின்றி தேர்வு செய்யும், மற்றும் அவர்கள் விருப்பத்திற்கிணங்க தங்களின் குழந்தைகளின் சமய
அவர்கள் விருப்பத்திற்கிணங்க தங்களின் குழந்தைகளின் சமய மற்றும் நன்னெறிக் கல்வியை உறுதி செய்யும், உரிமையை
இதன் காரணங்களுக்காக இப்பிரகடனம் மரியாதை, நல்லுறவு மற்றும் பரஸ்பர நன்மை ஆகியவற்றின் அடிப்படையில் பன்மொழி
ஒற்றுமையுடனும் இணக்கத்துடனும் வாழ்ந்துவருவது அவர்களின் மொழி மற்றும் கலாச்சாரத்தின் மீட்சிக்குச் சான்றாகும். தமிழ்க்கல்வி அமைவு
சான்றாகும். தமிழ்க்கல்வி அமைவு அவர்களுக்கு மொழி மற்றும் கலாச்சாரத்தின் மீது உடைமையுணர்வைத் தருகிறது. சீனப்பள்ளிகளும்
மொழியாக ஆங்கிலத்தை அறிமுகப்படுத்தியது. இக்கொள்கை கிராமப்புற மற்றும் ஏழைக் குடும்பங்களைச் சார்ந்த குழந்தைகளுக்குக் கெடுதியை

Search
Drag a column header here to group by that column
Concordan
Clear
Show Concordancer For all words
Save Output

Figure 5.20: Concordance Text Format (-N5+N5)

5.11.11 Table Format



தலைப்பு இடாதது 20

File Details

Word Count 103

Input

Choose a word

எ...	சொல்
1	பாதிப்புகள்
2	மற்றும்
3	மூன்று
4	பாதிப்புகளில்
5	மொத்த
6	புதிய
7	சம்பந்தப்பட்ட
8	இன்று
9	அவர்
10	நாட்டில்
11	எண்ணிக்கை
12	உள்ளன

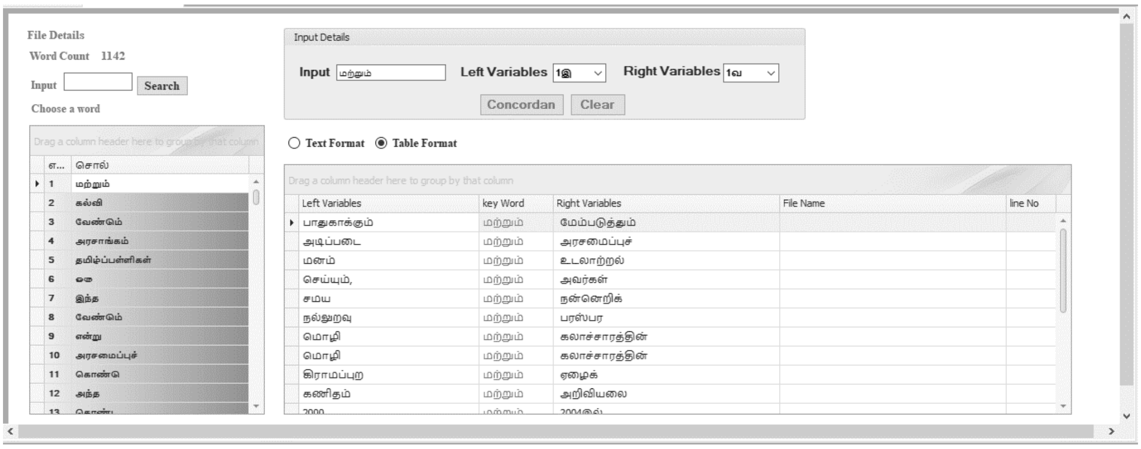
Input Details

Input: பாதிப்புகள் Left Variables: 1இ Right Variables: 1வ

○ Text Format ◉ Table Format

Left Variables	key Word	Right Variables	File Name	line No
நேர்மறையான	பாதிப்புகள்	மீண்டும்		
உள்ளூர்	பாதிப்புகள்	மற்றும்		
மூன்று	பாதிப்புகள்	பதிவாகியுள்ளன		
உள்ளன.	பாதிப்புகள்	செயலில்		
செய்யப்பட்ட	பாதிப்புகள்	. நாட்டில்		
கொண்ட	பாதிப்புகள்	ஆகும்		
நான்கு	பாதிப்புகள்	மீட்கப்பட்டு		
மூன்று	பாதிப்புகள்	சிகிச்சை		

Figure 5.21: Concordance Table Format 1



Input: மற்றும் Left Variables: 1இ Right Variables: 1வ

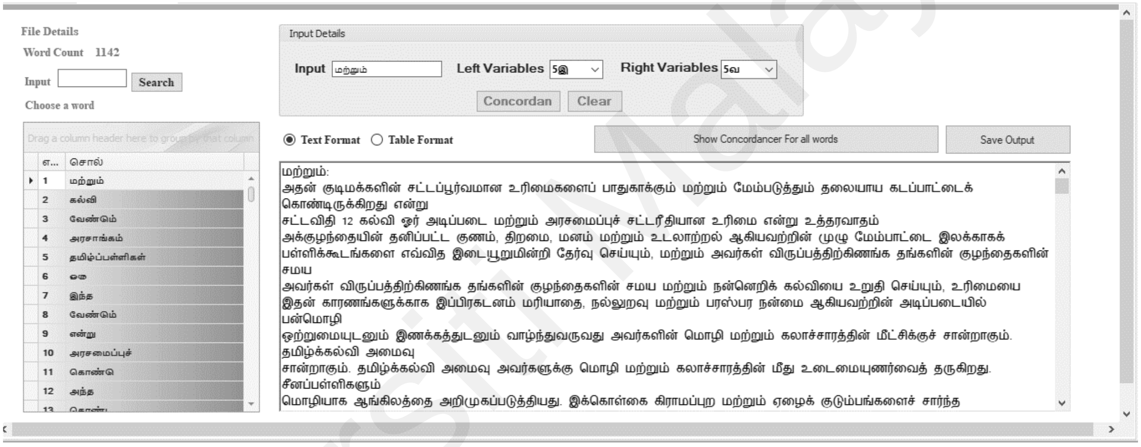
○ Text Format ● Table Format

எ...	சொல்
1	மற்றும்
2	கல்வி
3	வேண்டும்
4	அரசாங்கம்
5	தமிழ்ப்பள்ளிகள்
6	ஒை
7	இந்த
8	வேண்டும்
9	என்று
10	அரசமைப்புச்
11	கொண்டு
12	அந்த
13	கொண்

Left Variables	key Word	Right Variables	File Name	line No
பாதுகாக்கும்	மற்றும்	மேம்படுத்தும்		
அடிப்படை	மற்றும்	அரசமைப்புச்		
மனம்	மற்றும்	உடலாற்றல்		
செய்யும்,	மற்றும்	அவர்கள்		
சமய	மற்றும்	நன்னெறிக்		
நல்லுறவு	மற்றும்	பரஸ்பர		
மொழி	மற்றும்	கலாச்சாரத்தின்		
மொழி	மற்றும்	கலாச்சாரத்தின்		
கிராமப்புற	மற்றும்	ஏழைக்		
கணிதம்	மற்றும்	அறிவியலை		
2000	மற்றும்	2004இல்		

Figure 5.22: Concordance Table Format 2

5.11.12 Concordance for all the Tokens in the Sample Text



Input: மற்றும் Left Variables: 5இ Right Variables: 5வ

● Text Format ○ Table Format

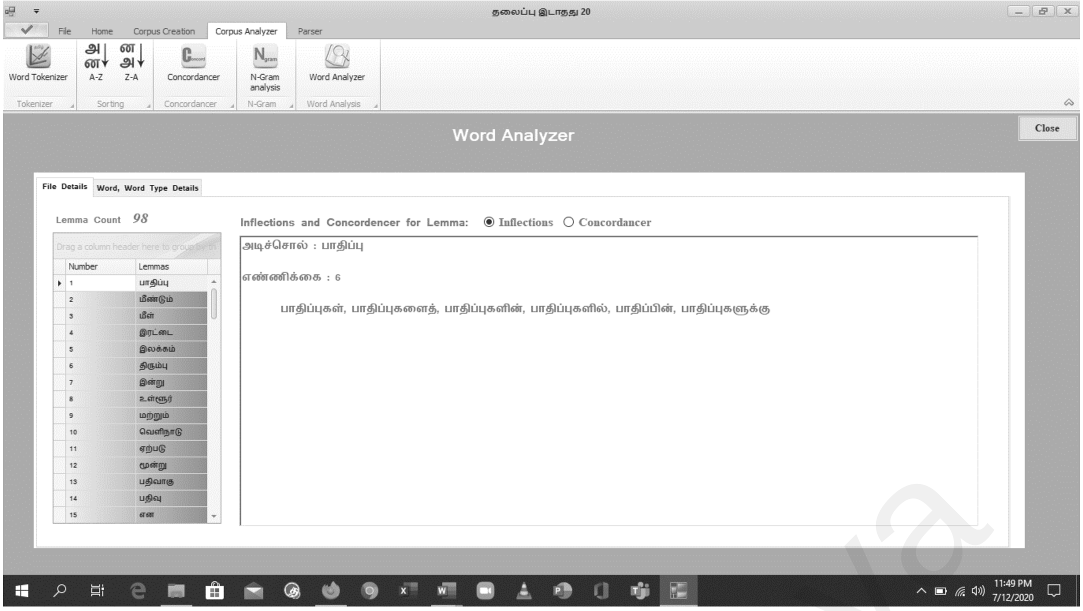
எ...	சொல்
1	மற்றும்
2	கல்வி
3	வேண்டும்
4	அரசாங்கம்
5	தமிழ்ப்பள்ளிகள்
6	ஒை
7	இந்த
8	வேண்டும்
9	என்று
10	அரசமைப்புச்
11	கொண்டு
12	அந்த
13	கொண்

```
மற்றும்:
அதன் குடிமக்களின் சட்டப்பூர்வமான உரிமைகளைப் பாதுகாக்கும் மற்றும் மேம்படுத்தும் தலையாய கடப்பாட்டைக்
கொண்டிருக்கிறது என்று
சட்டவிதி 12 கல்வி ஓர் அடிப்படை மற்றும் அரசமைப்புச் சட்டரீதியான உரிமை என்று உத்தரவாதம்
அக்குழந்தையின் தனிப்பட்ட குணம், திறமை, மனம் மற்றும் உடலாற்றல் ஆகியவற்றின் முழு மேம்பாட்டை இலக்காகக்
பள்ளிக்கூடங்களை எவ்வித இடையூறுமின்றி தேர்வு செய்யும், மற்றும் அவர்கள் விருப்பத்திற்கிணங்க தங்களின் குழந்தைகளின்
சமய
அவர்கள் விருப்பத்திற்கிணங்க தங்களின் குழந்தைகளின் சமய மற்றும் நன்னெறிக் கல்வியை உறுதி செய்யும், உரிமையை
இதன் காரணங்களுக்காக இப்பிரகடனம் மரியாதை, நல்லுறவு மற்றும் பரஸ்பர நன்மை ஆகியவற்றின் அடிப்படையில்
பன்மொழி
ஒற்றுமையுடனும் இணக்கத்துடனும் வாழ்ந்துவருவது அவர்களின் மொழி மற்றும் கலாச்சாரத்தின் மீட்சிக்குச் சான்றாகும்.
தமிழ்க்கல்வி அமைவு
சான்றாகும். தமிழ்க்கல்வி அமைவு அவர்களுக்கு மொழி மற்றும் கலாச்சாரத்தின் மீது உடைமையுணர்வைத் தருகிறது.
சீனப்பள்ளிகளும்
மொழியாக ஆங்கிலத்தை அறிமுகப்படுத்தியது. இக்கொள்கை கிராமப்புற மற்றும் ஏழைக் குடும்பங்களைச் சார்ந்த
```

Figure 5.23: Concordance for all the Tokens

5.11.13 Word Analysis

With the help of the morphological parser, all the types are being morphologically parsed into root and other affixes. In Tamil, all affixes are suffixes only except a few prefixes. The morphological parser in this software extracts all the lemmas from the types.



Word Analyzer

File Details | Word, Word Type Details

Lemma Count *98*

Number	Lemmas
1	பாதிப்பு
2	மீண்டும்
3	மீள்
4	இரட்டை
5	இலக்கம்
6	திரும்பு
7	இன்று
8	உள்ளூர்
9	மற்றும்
10	வெளிநாடு
11	ஏற்படு
12	மூன்று
13	பதிவாகு
14	பதிவு
15	என

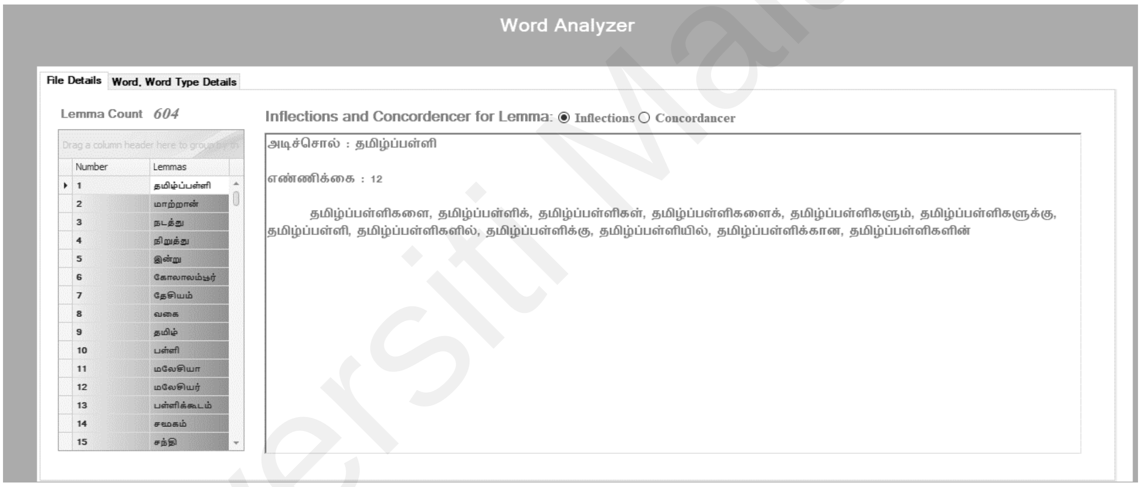
Inflections and Concordencer for Lemma: ● Inflections ○ Concordancer

அடிச்சொல் : பாதிப்பு

எண்ணிக்கை : 6

பாதிப்புகள், பாதிப்புகளைத், பாதிப்புகளின், பாதிப்புகளில், பாதிப்பின், பாதிப்புகளுக்கு

Figure 5.24: Word Analyzer-Lemma



Word Analyzer

File Details | Word, Word Type Details

Lemma Count *604*

Number	Lemmas
1	தமிழ்ப்பள்ளி
2	மாற்றான்
3	நடத்து
4	நிறுத்து
5	இன்று
6	கோலாலம்பூர்
7	தேசியம்
8	வகை
9	தமிழ்
10	பள்ளி
11	மலேசியா
12	மலேசியர்
13	பள்ளிக்கூடம்
14	சமகம்
15	சந்தி

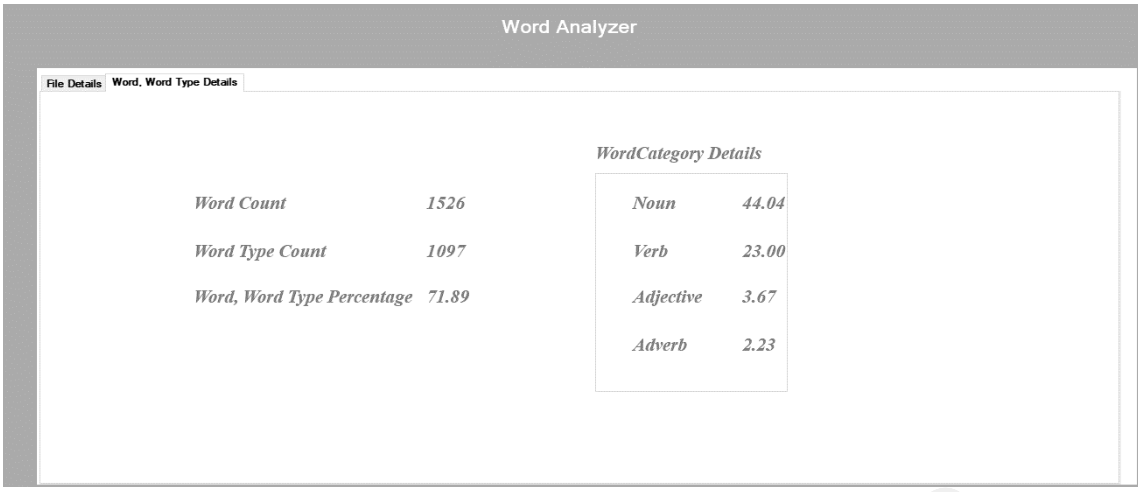
Inflections and Concordencer for Lemma: ● Inflections ○ Concordancer

அடிச்சொல் : தமிழ்ப்பள்ளி

எண்ணிக்கை : 12

தமிழ்ப்பள்ளிகளை, தமிழ்ப்பள்ளிக், தமிழ்ப்பள்ளிகள், தமிழ்ப்பள்ளிகளைக், தமிழ்ப்பள்ளிகளும், தமிழ்ப்பள்ளிகளுக்கு, தமிழ்ப்பள்ளி, தமிழ்ப்பள்ளிகளில், தமிழ்ப்பள்ளிக்கு, தமிழ்ப்பள்ளியில், தமிழ்ப்பள்ளிக்கான, தமிழ்ப்பள்ளிகளின்

Figure 5.25: Word Analyzer - Inflectional Lemma



Word Analyzer

File Details | Word, Word Type Details

Word Count	1526
Word Type Count	1097
Word, Word Type Percentage	71.89

WordCategory Details

Noun	44.04
Verb	23.00
Adjective	3.67
Adverb	2.23

Figure 5.26: Word Analyzer - Word category and Details

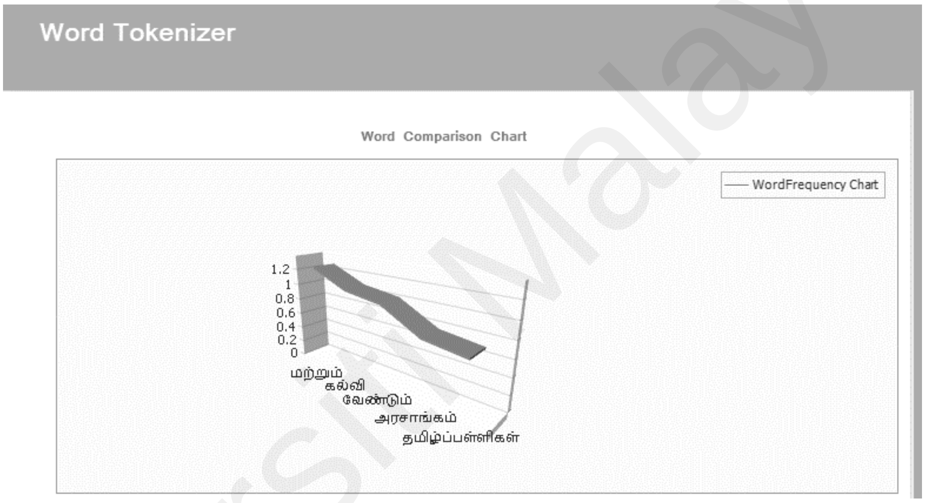
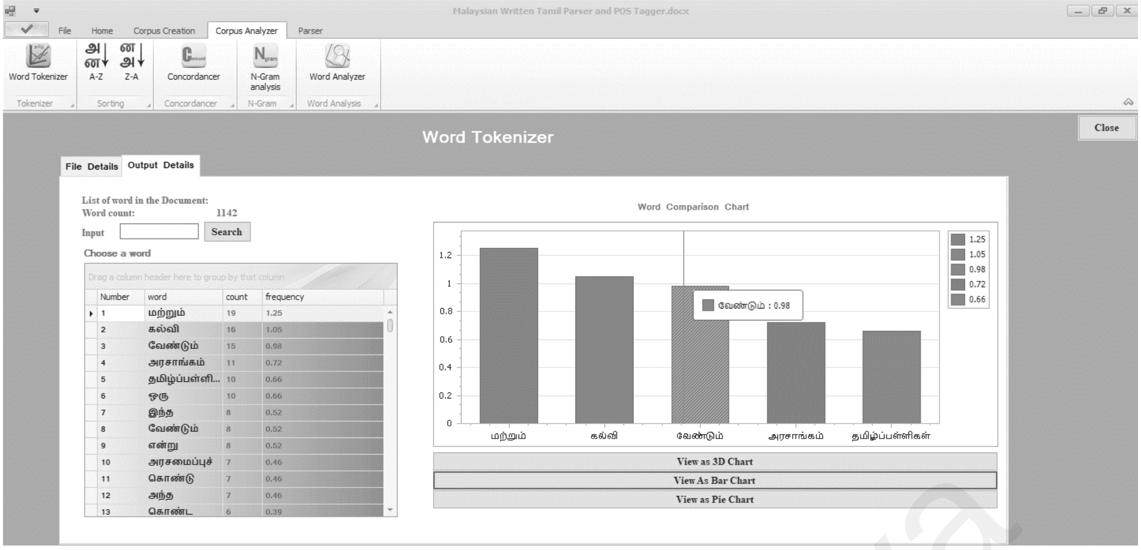


Figure 5.27: Word Tokenizer -Word comparison chart (a)



List of word in the Document:
Word count: 1142
Input Search
Choose a word

Drag a column header here to group by that column

Number	word	count	frequency
1	மற்றும்	19	1.25
2	கல்வி	16	1.05
3	வேண்டும்	15	0.98
4	அரசாங்கம்	11	0.72
5	தமிழ்ப்பள்ளி...	10	0.66
6	ஒரு	10	0.66
7	இந்த	8	0.52
8	வேண்டும்	8	0.52
9	என்று	8	0.52
10	அரசமைப்புச்	7	0.46
11	கொண்டு	7	0.46
12	அந்த	7	0.46
13	கொண்ட	6	0.39

Figure 5.28: Word Tokenizer -Word comparison chart (b)

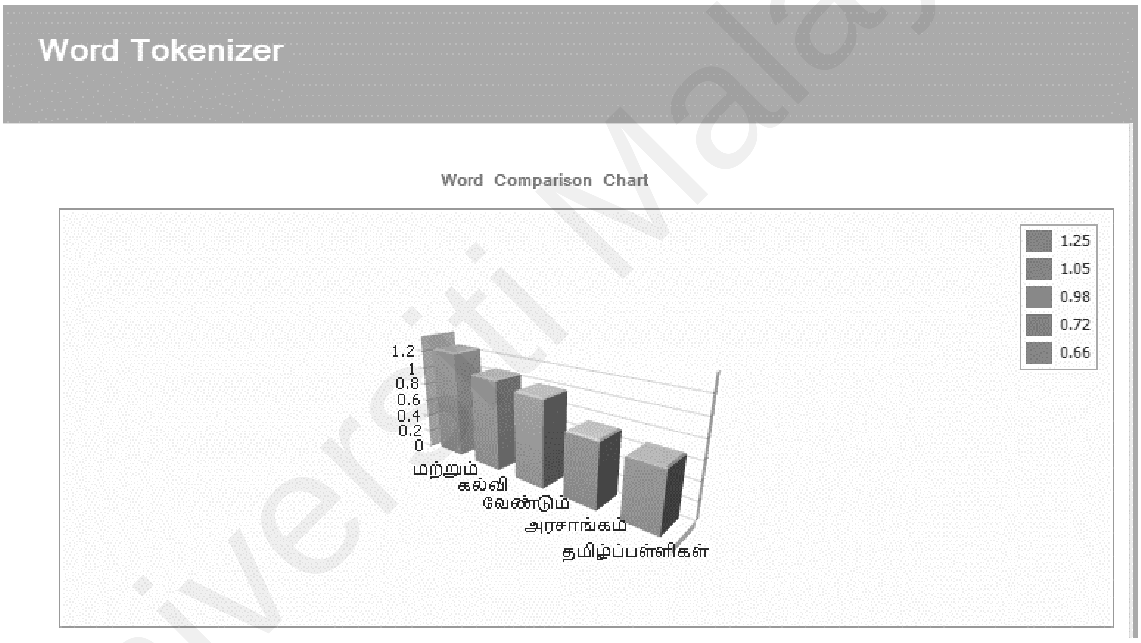
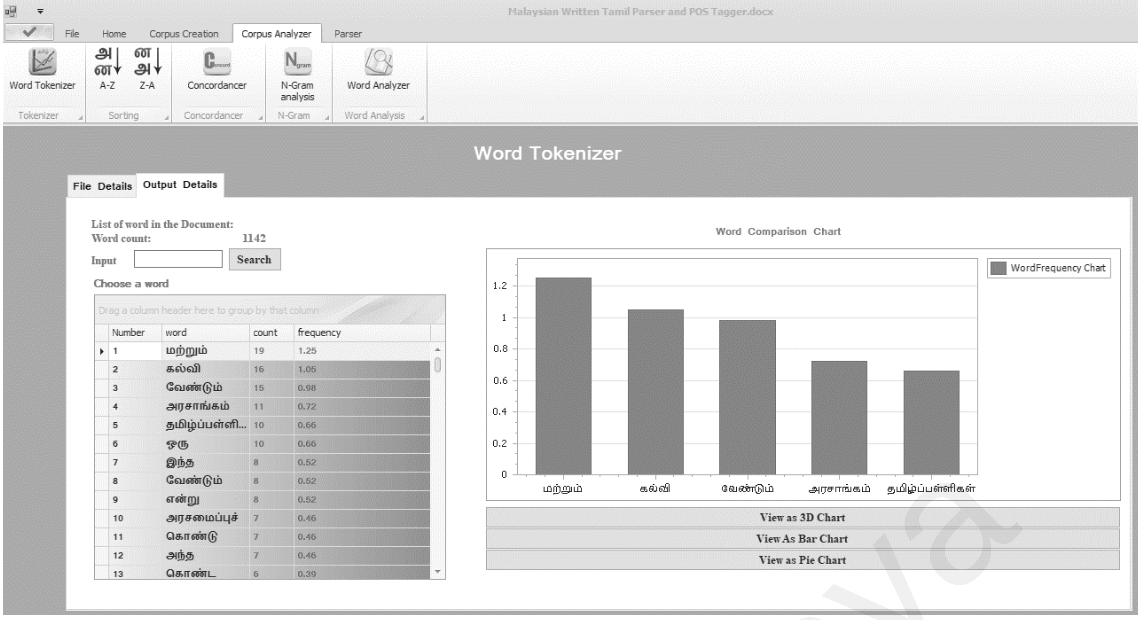


Figure 5.29: Word Tokenizer -Word comparison chart (c)



List of word in the Document:
Word count: 1142
Input
Choose a word
Drag a column header here to group by that column

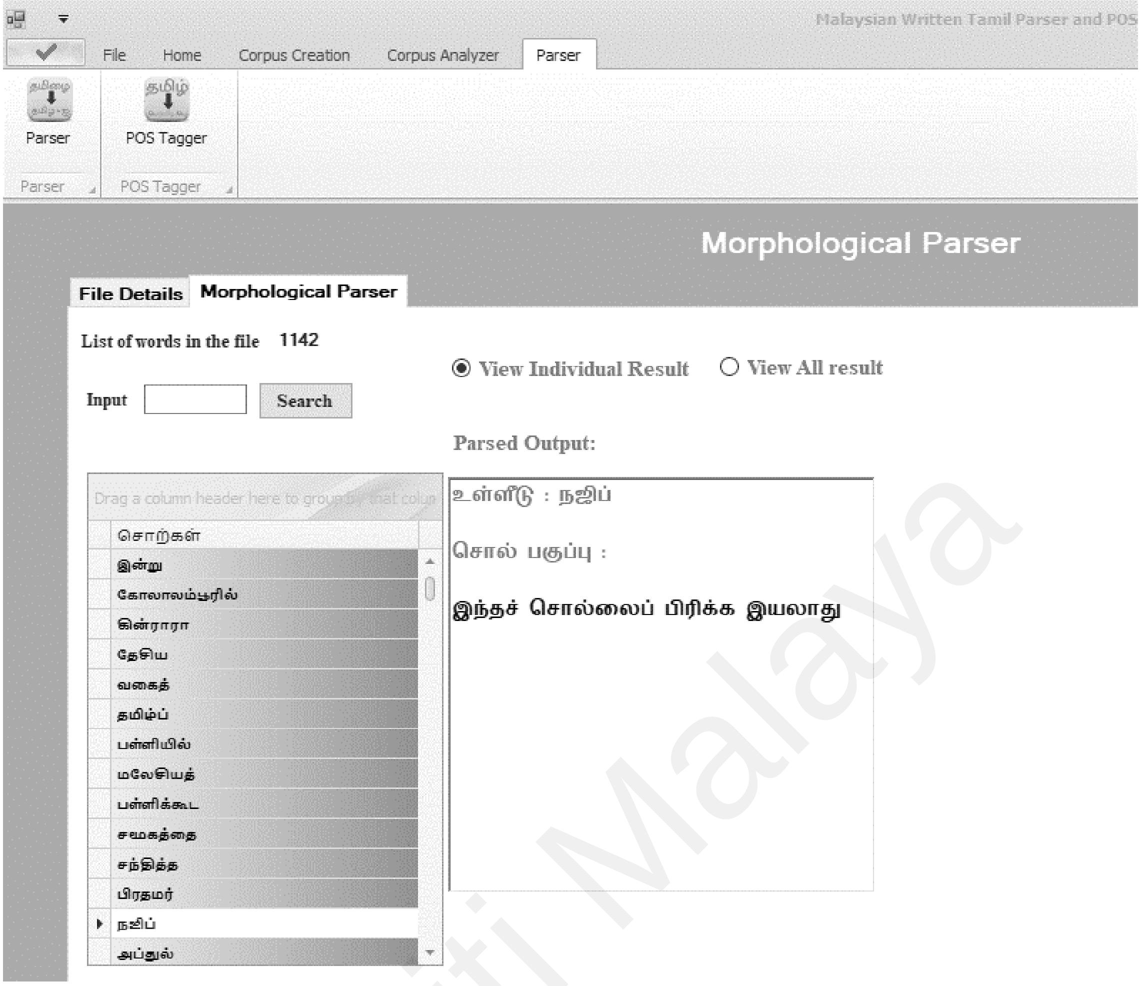
Number	word	count	frequency
1	மற்றும்	19	1.25
2	கல்வி	16	1.05
3	வேண்டும்	15	0.98
4	அரசாங்கம்	11	0.72
5	தமிழ்ப்பள்ளி...	10	0.66
6	ஒரு	10	0.66
7	இந்த	8	0.52
8	வேண்டும்	8	0.52
9	என்று	8	0.52
10	அரசமைப்புச்	7	0.46
11	கொண்டு	7	0.46
12	அந்த	7	0.46
13	கொண்ட	6	0.39

Figure 5.30: Word Tokenizer -Word comparison chart (d)

In the left Table of the Window, all the lemmas extracted from the types are being displayed and on the right side of the Window, there are two tools: one is the listing of inflected words or types of a particular Lemma, and the second one provides the concordances for the particular Types.

5.11.14 Morphological Analysis

One of the objectives of this research is to develop a morphological parser for written Malaysian Tamil. Based on a review of the existing Tamil morphological studies and the morphological features of WMTC, an automatic morphological parser has been developed here. The unparsed word Najib is illustrated below before adding this into the lexicon of the parser.



List of words in the file 1142

● View Individual Result ○ View All result

Input

Parsed Output:

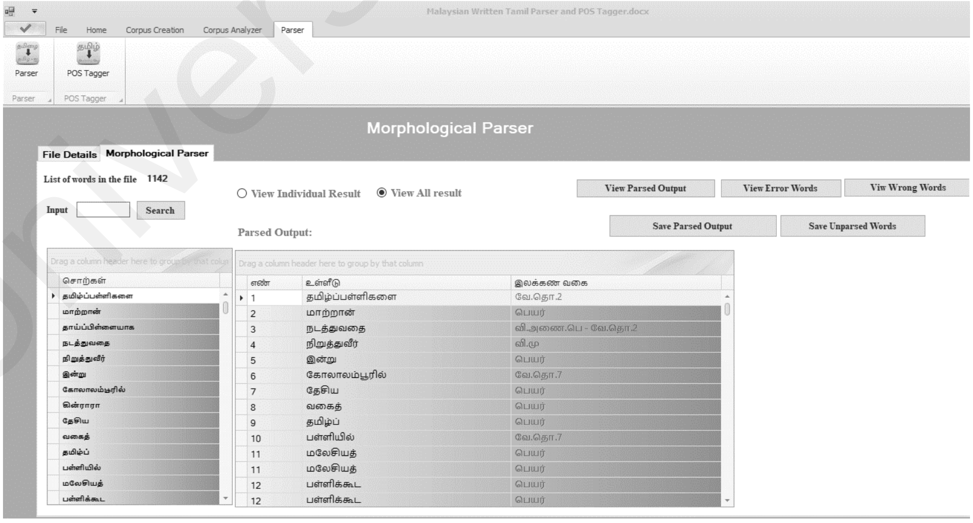
சொற்கள்
இன்று
கோலாலம்பூரில்
கின்ராரா
தேசிய
வகைத்
தமிழ்ப்
பள்ளியில்
மலேசியத்
பள்ளிக்கூட
சமூகத்தை
சந்தித்த
பிரதமர்
நஜிப்
அப்துல்

உள்ளீடு : நஜிப்

சொல் பகுப்பு :

இந்தச் சொல்லைப் பிரிக்க இயலாது

Figure 5.31: Parsed and Unparsed Word



List of words in the file 1142

Input

○ View Individual Result ● View All result

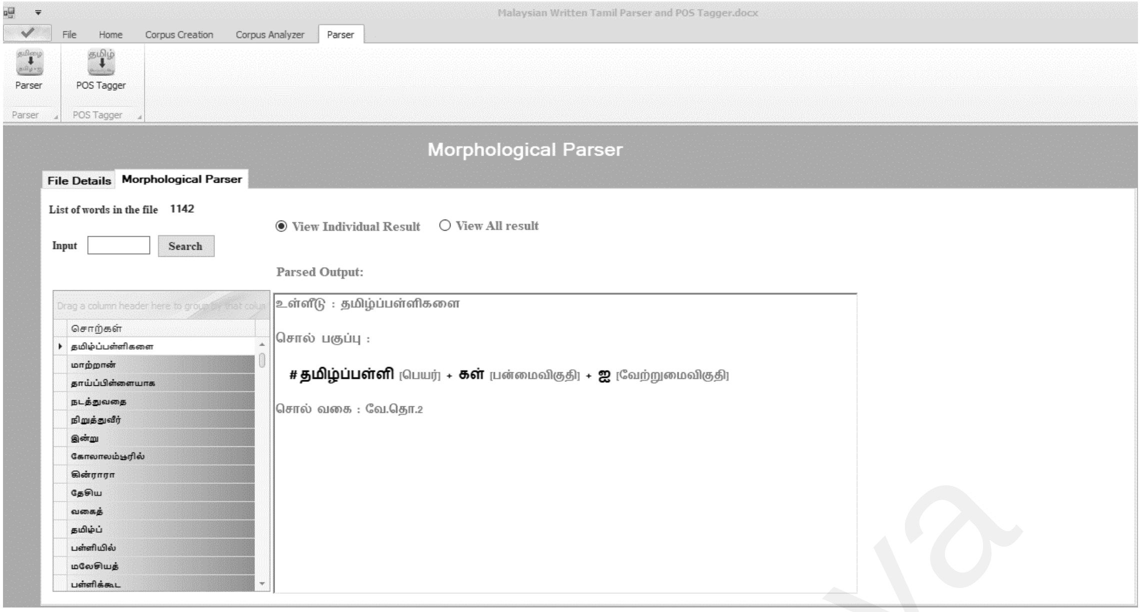
Parsed Output:

சொற்கள்
தமிழ்ப்பள்ளிகளை
மாற்றான்
தாய்ப்பிள்ளையாக
நடத்துவதை
நிறுத்துவீர்
இன்று
கோலாலம்பூரில்
கின்ராரா
தேசிய
வகைத்
தமிழ்ப்
பள்ளியில்
மலேசியத்
பள்ளிக்கூட

எண்	உள்ளீடு	இலக்கண வகை
1	தமிழ்ப்பள்ளிகளை	வே.தொ.2
2	மாற்றான்	பெயர்
3	நடத்துவதை	வி.அணை.பெ - வே.தொ.2
4	நிறுத்துவீர்	வி.மு
5	இன்று	பெயர்
6	கோலாலம்பூரில்	வே.தொ.7
7	தேசிய	பெயர்
8	வகைத்	பெயர்
9	தமிழ்ப்	பெயர்
10	பள்ளியில்	வே.தொ.7
11	மலேசியத்	பெயர்
11	மலேசியத்	பெயர்
12	பள்ளிக்கூட	பெயர்
12	பள்ளிக்கூட	பெயர்

Figure 5.32: Parsed Words

All the parsed words are tagged individually as illustrated below:


Malaysian Written Tamil Parser and POS Tagger.docx
File
Home
Corpus Creation
Corpus Analyzer
Parser
Parser
POS Tagger
Parser
POS Tagger

Morphological Parser

File Details | Morphological Parser

List of words in the file 1142

Input [ ] Search

● View Individual Result ○ View All result

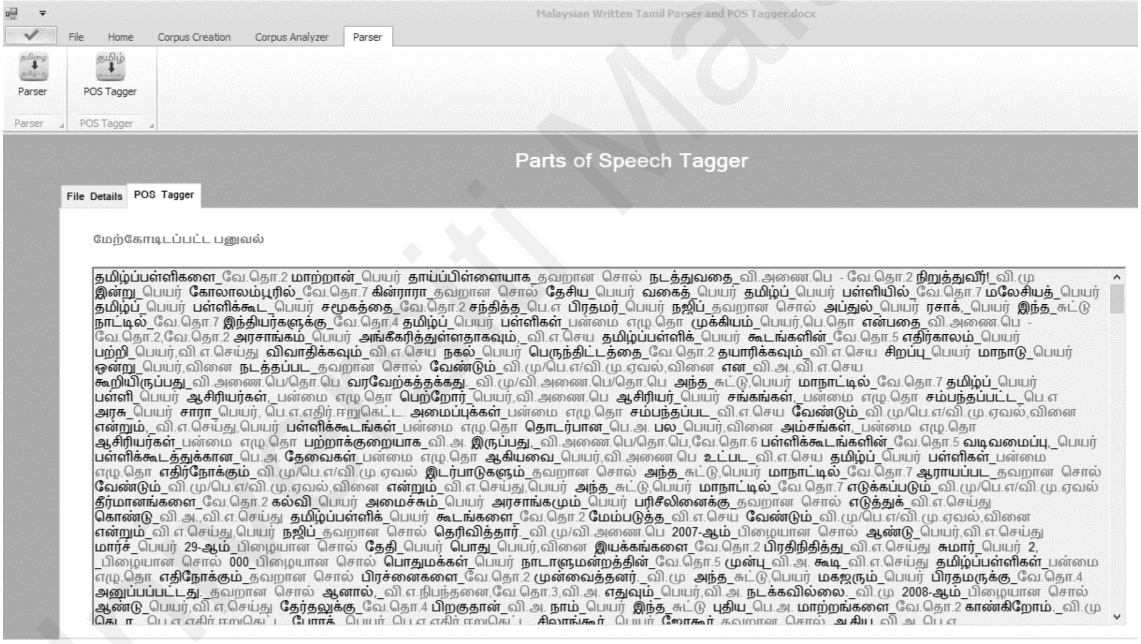
Parsed Output:

```
உள்ளீடு : தமிழ்ப்பள்ளிகளை

சொல் பகுப்பு :

 # தமிழ்ப்பள்ளி [பெயர்] + கள் [பன்மைவிகுதி] + ஐ [வேற்றுமைவிகுதி]

சொல் வகை : வே.தொ.2
```

Drag a column header here to group by that colun

சொற்கள்
தமிழ்ப்பள்ளிகளை
மாற்றான்
தாய்ப்பிள்ளையாக
நடத்துவதை
நிறுத்துவீர்
இன்று
கோலாலம்பூரில்
கின்ராரா
தேசிய
வகைத்
தமிழ்ப்
பள்ளியில்
மலேசியத்
பள்ளிக்கூட

Figure 5.33: Invidual POS Tagged Words



Parts of Speech Tagger

File Details | POS Tagger

மேற்கோடிடப்பட்ட பனுவல்

```
தமிழ்ப்பள்ளிகளை_வே.தொ.2 மாற்றான்_பெயர் தாய்ப்பிள்ளையாக_தவறான சொல் நடத்துவதை_வி.அணை.பெ - வே.தொ.2 நிறுத்துவீர்!_வி.மு
இன்று_பெயர் கோலாலம்பூரில்_வே.தொ.7 கின்ராரா_தவறான சொல் தேசிய_பெயர் வகைத்_பெயர் தமிழ்ப்_பெயர் பள்ளியில்_வே.தொ.7 மலேசியத்_பெயர்
தமிழ்ப்_பெயர் பள்ளிக்கூட_பெயர் சமூகத்தை_வே.தொ.2 சந்தித்த_பெ.எ பிரதமர்_பெயர் நஜிப்_தவறான சொல் அப்துல்_பெயர் ரசாக்,_பெயர் இந்த_சுட்டு
நாட்டில்_வே.தொ.7 இந்தியர்களுக்கு_வே.தொ.4 தமிழ்ப்_பெயர் பள்ளிகள்_பன்மை எழு.தொ முக்கியம்_பெயர்,பெ.தொ என்பதை_வி.அணை.பெ -
வே.தொ.2,வே.தொ.2 அரசாங்கம்_பெயர் அங்கீகரித்துள்ளதாகவும்,_வி.எ.செய தமிழ்ப்பள்ளிக்_பெயர் கூடங்களின்_வே.தொ.5 எதிர்காலம்_பெயர்
பற்றி_பெயர்,வி.எ.செய்து விவாதிக்கவும்_வி.எ.செய நகல்_பெயர் பெருந்திட்டத்தை_வே.தொ.2 தயாரிக்கவும்_வி.எ.செய சிறப்பு_பெயர் மாநாடு_பெயர்
ஒன்று_பெயர்,வினை நடத்தப்பட_தவறான சொல் வேண்டும்_வி.மு/பெ.எ/வி.மு.ஏவல்,வினை என_வி.அ.,வி.எ.செய
கூறியிருப்பது_வி.அணை.பெ/தொ.பெ வரவேற்கத்தக்கது._வி.மு/வி.அணை.பெ/தொ.பெ அந்த_சுட்டு,பெயர் மாநாட்டில்_வே.தொ.7 தமிழ்ப்_பெயர்
பள்ளி_பெயர் ஆசிரியர்கள்,_பன்மை எழு.தொ பெற்றோர்_பெயர்,வி.அணை.பெ ஆசிரியர்_பெயர் சங்கங்கள்,_பன்மை எழு.தொ சம்பந்தப்பட்ட_பெ.எ
அரசு_பெயர் சாரா_பெயர், பெ.எ.எதிர்.ஈறுகெட்ட அமைப்புக்கள்_பன்மை எழு.தொ சம்பந்தப்பட_வி.எ.செய வேண்டும்_வி.மு/பெ.எ/வி.மு.ஏவல்,வினை
என்றும்,_வி.எ.செய்து,பெயர் பள்ளிக்கூடங்கள்_பன்மை எழு.தொ தொடர்பான_பெ.அ. பல_பெயர்,வினை அம்சங்கள்,_பன்மை எழு.தொ
ஆசிரியர்கள்_பன்மை எழு.தொ பற்றாக்குறையாக_வி.அ. இருப்பது,_வி.அணை.பெ/தொ.பெ,வே.தொ.6 பள்ளிக்கூடங்களின்_வே.தொ.5 வடிவமைப்பு,_பெயர்
பள்ளிக்கூடத்துக்கான_பெ.அ. தேவைகள்_பன்மை எழு.தொ ஆகியவை_பெயர்,வி.அணை.பெ உட்பட_வி.எ.செய தமிழ்ப்_பெயர் பள்ளிகள்_பன்மை
எழு.தொ எதிர்நோக்கும்_வி.மு/பெ.எ/வி.மு.ஏவல் இடர்பாடுகளும்_தவறான சொல் அந்த_சுட்டு,பெயர் மாநாட்டில்_வே.தொ.7 ஆராயப்பட_தவறான சொல்
வேண்டும்_வி.மு/பெ.எ/வி.மு.ஏவல்,வினை என்றும்_வி.எ.செய்து,பெயர் அந்த_சுட்டு,பெயர் மாநாட்டில்_வே.தொ.7 எடுக்கப்படும்_வி.மு/பெ.எ/வி.மு.ஏவல்
தீர்மானங்களை_வே.தொ.2 கல்வி_பெயர் அமைச்சும்_பெயர் அரசாங்கமும்_பெயர் பரிசீலினைக்கு_தவறான சொல் எடுத்துக்_வி.எ.செய்து
கொண்டு_வி.அ.,வி.எ.செய்து தமிழ்ப்பள்ளிக்_பெயர் கூடங்களை_வே.தொ.2 மேம்படுத்த_வி.எ.செய வேண்டும்_வி.மு/பெ.எ/வி.மு.ஏவல்,வினை
என்றும்_வி.எ.செய்து,பெயர் நஜிப்_தவறான சொல் தெரிவித்தார்._வி.மு/வி.அணை.பெ 2007-ஆம்_பிழையான சொல் ஆண்டு_பெயர்,வி.எ.செய்து
மார்ச்_பெயர் 29-ஆம்_பிழையான சொல் தேதி_பெயர் பொது_பெயர்,வினை இயக்கங்களை_வே.தொ.2 பிரதிநிதித்து_வி.எ.செய்து சுமார்_பெயர் 2,
_பிழையான சொல் 000_பிழையான சொல் பொதுமக்கள்_பெயர் நாடாளுமன்றத்தின்_வே.தொ.5 முன்பு_வி.அ. கூடி_வி.எ.செய்து தமிழ்ப்பள்ளிகள்_பன்மை
எழு.தொ எதிநோக்கும்_தவறான சொல் பிரச்னைகளை_வே.தொ.2 முன்வைத்தனர்._வி.மு அந்த_சுட்டு,பெயர் மகஜரும்_பெயர் பிரதமருக்கு_வே.தொ.4
அனுப்பப்பட்டது._தவறான சொல் ஆனால்,_வி.எ.நிபந்தனை,வே.தொ.3,வி.அ. எதுவும்_பெயர்,வி.அ. நடக்கவில்லை._வி.மு 2008-ஆம்_பிழையான சொல்
ஆண்டு_பெயர்,வி.எ.செய்து தேர்தலுக்கு_வே.தொ.4 பிறகுதான்_வி.அ. நாம்_பெயர் இந்த_சுட்டு புதிய_பெ.அ. மாற்றங்களை_வே.தொ.2 காண்கிறோம்._வி.மு
```

Figure 5.34: POS Tagged Words

The Tools involved in the above-mentioned internal analysis are:

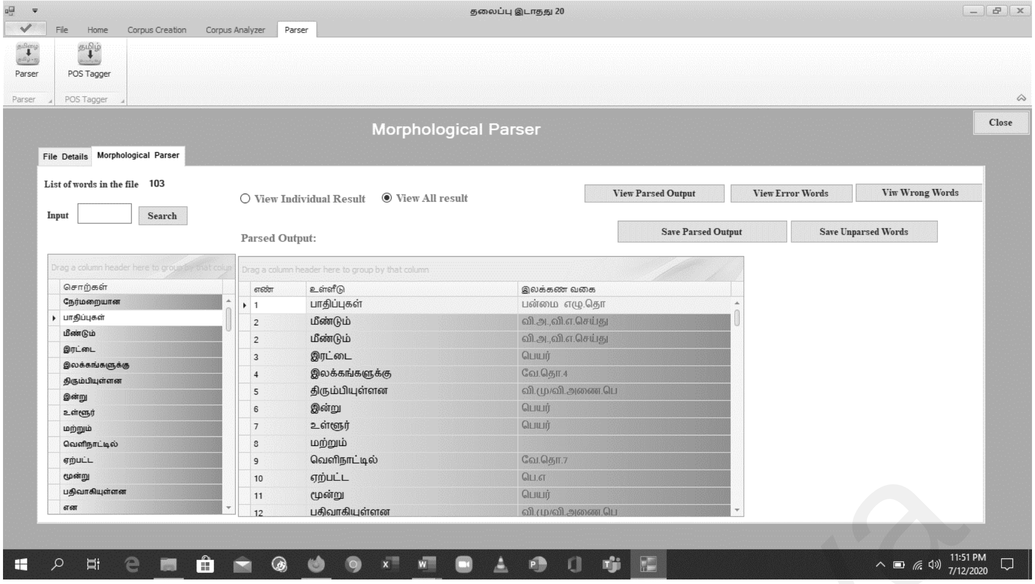
1. Tamil Root Words (Tamil Lexicon)
2. Tamil Grammatical Affixes
3. Tamil Morphotactic Rules
4. Tamil Morphophonemic Rules
5. Suffix Disambiguation Tool
6. Tamil Word Class (POS) Category Tagger
7. Word Category Disambiguation Tool

5.11.15 Tamil Grammatical Affixes

The following grammatical affixes are placed in a separate component:

(1) suffixes,
(2) postpositions,
(3) verbal particles,
(4) clitics and
(5) fillers.

The grammatical affixes are provided in a separate column as found below:



Morphological Parser

File Details | Morphological Parser

List of words in the file 103

Input [ ] Search

○ View Individual Result ● View All result

View Parsed Output | View Error Words | Viw Wrong Words

Save Parsed Output | Save Unparsed Words

Parsed Output:

Drag a column header here to group by that column

சொற்கள்
நேர்மறையான
பாதிப்புகள்
மீண்டும்
இரட்டை
இலக்கங்களுக்கு
திரும்பியுள்ளன
இன்று
உள்ளூர்
மற்றும்
வெளிநாட்டில்
ஏற்பட்ட
மூன்று
பதிவாகியுள்ளன
என

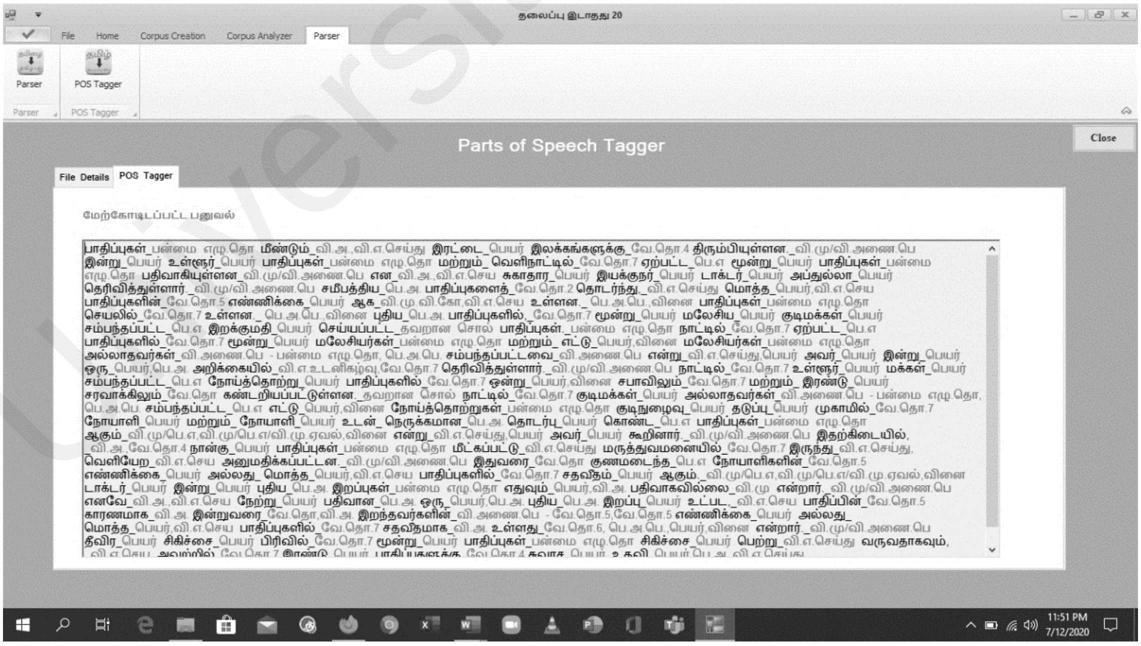
Drag a column header here to group by that column

எண்	உள்ளீடு	இலக்கண வகை
1	பாதிப்புகள்	பன்மை எழு.தொ
2	மீண்டும்	வி.அ.,வி.எ.செய்து
2	மீண்டும்	வி.அ.,வி.எ.செய்து
3	இரட்டை	பெயர்
4	இலக்கங்களுக்கு	வே.தொ.4
5	திரும்பியுள்ளன	வி.மு/வி.அணை.பெ
6	இன்று	பெயர்
7	உள்ளூர்	பெயர்
8	மற்றும்	
9	வெளிநாட்டில்	வே.தொ.7
10	ஏற்பட்ட	பெ.எ
11	மூன்று	பெயர்
12	பதிவாகியுள்ளன	வி.மு/வி.அணை.பெ

Figure 5.35: Tamil Grammatical Affixes

5.11.16 Tamil Word Class (POS) Category Tagger

All the types found in the corpus have been tagged for their word class category. Tamil word class (POS) category tagger:



Parts of Speech Tagger

File Details | POS Tagger

மேற்கோடிடப்பட்ட பனுவல்

பாதிப்புகள்_பன்மை எழு.தொ மீண்டும்_வி.அ.,வி.எ.செய்து இரட்டை_பெயர் இலக்கங்களுக்கு_வே.தொ.4 திரும்பியுள்ளன._வி.மு/வி.அணை.பெ
இன்று_பெயர் உள்ளூர்_பெயர் பாதிப்புகள்_பன்மை எழு.தொ மற்றும்_ வெளிநாட்டில்_வே.தொ.7 ஏற்பட்ட_பெ.எ மூன்று_பெயர் பாதிப்புகள்_பன்மை
எழு.தொ பதிவாகியுள்ளன_வி.மு/வி.அணை.பெ என_வி.அ.,வி.எ.செய சுகாதார_பெயர் இயக்குநர்_பெயர் டாக்டர்_பெயர் அப்துல்லா_பெயர்
தெரிவித்துள்ளார்._வி.மு/வி.அணை.பெ சமீபத்திய_பெ.அ. பாதிப்புகளைத்_வே.தொ.2 தொடர்ந்து,_வி.எ.செய்து மொத்த_பெயர்,வி.எ.செய
பாதிப்புகளின்_வே.தொ.5 எண்ணிக்கை_பெயர் ஆக_வி.மு.வி.கோ,வி.எ.செய உள்ளன._ பெ.அ.பெ.,வினை பாதிப்புகள்_பன்மை எழு.தொ
செயலில்_வே.தொ.7 உள்ளன._ பெ.அ.பெ.,வினை புதிய_பெ.அ. பாதிப்புகளில்,_வே.தொ.7 மூன்று_பெயர் மலேசிய_பெயர் குடிமக்கள்_பெயர்
சம்பந்தப்பட்ட_பெ.எ இறக்குமதி_பெயர் செய்யப்பட்ட_தவறான சொல் பாதிப்புகள்._பன்மை எழு.தொ நாட்டில்_வே.தொ.7 ஏற்பட்ட_பெ.எ
பாதிப்புகளில்_வே.தொ.7 மூன்று_பெயர் மலேசியர்கள்_பன்மை எழு.தொ மற்றும்_ எட்டு_பெயர்,வினை மலேசியர்கள்_பன்மை எழு.தொ
அல்லாதவர்கள்_வி.அணை.பெ - பன்மை எழு.தொ, பெ.அ.பெ. சம்பந்தப்பட்டவை_வி.அணை.பெ என்று_வி.எ.செய்து,பெயர் அவர்_பெயர் இன்று_பெயர்
ஒரு_பெயர்,பெ.அ. அறிக்கையில்_வி.எ.உடனிகழ்வு,வே.தொ.7 தெரிவித்துள்ளார்._வி.மு/வி.அணை.பெ நாட்டில்_வே.தொ.7 உள்ளூர்_பெயர் மக்கள்_பெயர்
சம்பந்தப்பட்ட_பெ.எ நோய்த்தொற்று_பெயர் பாதிப்புகளில்_வே.தொ.7 ஒன்று_பெயர்,வினை சபாவிலும்_வே.தொ.7 மற்றும்_ இரண்டு_பெயர்
சரவாக்கிலும்_வே.தொ கண்டறியப்பட்டுள்ளன._தவறான சொல் நாட்டில்_வே.தொ.7 குடிமக்கள்_பெயர் அல்லாதவர்கள்_வி.அணை.பெ - பன்மை எழு.தொ,
பெ.அ.பெ. சம்பந்தப்பட்ட_பெ.எ எட்டு_பெயர்,வினை நோய்த்தொற்றுகள்_பன்மை எழு.தொ குடிநுழைவு_பெயர் தடுப்பு_பெயர் முகாமில்_வே.தொ.7
நோயாளி_பெயர் மற்றும்_ நோயாளி_பெயர் உடன்_ நெருக்கமான_பெ.அ. தொடர்பு_பெயர் கொண்ட_பெ.எ பாதிப்புகள்_பன்மை எழு.தொ
ஆகும்_வி.மு/பெ.எ,வி.மு/பெ.எ/வி.மு.ஏவல்,வினை என்று_வி.எ.செய்து,பெயர் அவர்_பெயர் கூறினார்._வி.மு/வி.அணை.பெ இதற்கிடையில்,
_வி.அ.,வே.தொ.4 நான்கு_பெயர் பாதிப்புகள்_பன்மை எழு.தொ மீட்கப்பட்டு_வி.எ.செய்து மருத்துவமனையில்_வே.தொ.7 இருந்து_வி.எ.செய்து,
வெளியேற_வி.எ.செய அனுமதிக்கப்பட்டன._வி.மு/வி.அணை.பெ இதுவரை_வே.தொ குணமடைந்த_பெ.எ நோயாளிகளின்_வே.தொ.5
எண்ணிக்கை_பெயர் அல்லது_ மொத்த_பெயர்,வி.எ.செய பாதிப்புகளில்_வே.தொ.7 சதவீதம்_பெயர் ஆகும்._வி.மு/பெ.எ,வி.மு/பெ.எ/வி.மு.ஏவல்,வினை
டாக்டர்_பெயர் இன்று_பெயர் புதிய_பெ.அ. இறப்புகள்_பன்மை எழு.தொ எதுவும்_பெயர்,வி.அ. பதிவாகவில்லை_வி.மு என்றார்._வி.மு/வி.அணை.பெ
எனவே_வி.அ.,வி.எ.செய நேற்று_பெயர் பதிவான_பெ.அ. ஒரு_பெயர்,பெ.அ. புதிய_பெ.அ. இறப்பு_பெயர் உட்பட,_வி.எ.செய பாதிப்பின்_வே.தொ.5
காரணமாக_வி.அ. இன்றுவரை_வே.தொ,வி.அ. இறந்தவர்களின்_வி.அணை.பெ - வே.தொ.5,வே.தொ.5 எண்ணிக்கை_பெயர் அல்லது_
மொத்த_பெயர்,வி.எ.செய பாதிப்புகளில்_வே.தொ.7 சதவீதமாக_வி.அ. உள்ளது_வே.தொ.6, பெ.அ.பெ.,பெயர்,வினை என்றார்._வி.மு/வி.அணை.பெ
தீவிர_பெயர் சிகிச்சை_பெயர் பிரிவில்_வே.தொ.7 மூன்று_பெயர் பாதிப்புகள்_பன்மை எழு.தொ சிகிச்சை_பெயர் பெற்று_வி.எ.செய்து வருவதாகவும்,
_வி.எ.செய அவற்றில்_வே.தொ.7 இரண்டு_பெயர் பாதிப்புகளுக்கு_வே.தொ.4 சுவாச_பெயர் உதவி_பெயர்,பெ.அ.,வி.எ.செய்து

Figure 5.36: Parts of Speech Tagger

5.12 Lemma and other related Statistical studies

From the corpus, with the help of the tokenization, morphological parsing and disambiguation lemmas/root words with their grammatical categories could be arrived at, paving the way for further frequency studies of the following:

1. Type/Token ratio
2. Lemma/Type ratio

The above findings have been discussed with reference to the development of a comprehensive grammar of Tamil Word Forms/Types as well as to the development of morphological parser.

5.13 C# Coding, used for morphological parsing

The C# coding used for morphological parsing.

5.14 Conclusion

In this chapter, software design and architecture for the following processes have been explained in detail, together with the relevant charts and illustrations:

1) Corpus Database construction (lexicon and grammar parts)
2) Tokenisation and type selection
3) Morphological parsing (including momorphotacticsmor and morphophonemics)
4) POS tagging
5) N-gram analysis (for disambiguation)
6) Concordance analysis (for disambiguation)

7) Statistical analysis (word count, frequency, etc.)

All the above tools have been integrated into one unified software for this project. Here it is to be mentioned that initial morphological parser was based on the inflectional morphology of Tamil as described in various modern Tamil grammars; and subsequently this was modified and enriched based on the result of the initial parser applied over the present constructed corpus. In this way, the final Tamil inflectional computational morphology is a corpus based one. This modified and enriched Tamil computational morphological algorithm will be discussed in the next chapter.

CHAPTER 6

ALGORITHMS FOR THE MORPHOLOGICAL PARSER AND POS TAGGER

6.1 Introduction

In the previous chapter, the software development at each stage was explained step by step. The development of the morphological parser and POS tagger requires a suitable algorithm which forms the basis for the third research question of this thesis: How might a suitable algorithm be designed for developing the morphological parser and the POS tagger? This is answered in the present chapter.

It must be noted that the present research project involves two knowledge domains: (1) computer science and (2) computational linguistic algorithm of Tamil morphology. The first domain, that is, computer science comprising system architecture and software design, was already covered in the previous chapter, and the second domain will be addressed in this chapter. Some linguistic considerations in Tamil to develop a computational linguistic algorithm behind the software development are discussed in this chapter.

6.2 Computational Algorithm of Tamil Morphology

From the corpus and computational linguistic point of view (Jurafsky, 2002), the development of a morphological parser for Tamil would, as noted in the previous chapter, require the following:

1. A lexicon/dictionary containing the root words of Tamil.
2. A list of grammatical suffixes which are involved in the inflection process of Tamil lexemes.
3. A list of morphotactic rules of Tamil
4. A list of morphophonemic rules involved in Tamil.

6.3 Inflectional and agglutinative nature of Tamil

6.3.1 Inflectional morphology

As noted in earlier chapters, Tamil is an inflectional and agglutinative language (Devi, 2011; Rajendran et al., 2002). Regarding inflection, Huddleston and Pullum (2002) note that it deals with the inflectional forms of variable lexemes. Whereas the syntax tells us when a lexeme may or must carry a certain inflectional property, the inflectional morphology tells us what form it takes when it carries that inflectional property (Aronoff, 1993).

For example, a rule of syntax stipulates that a verb in construction with the perfect auxiliary *have* must carry the past participle inflection (as in *They have killed it, She had rung the bell*), while inflectional morphology describes how the past participles of verbs are formed from the base: *killed* is formed from the base *kill* by adding the suffix *-ed*, *rung* from *ring* by changing the vowel, and so on.

Anderson also describes (1982) that inflection is the morphology that is relevant to the syntax. It realizes all the morphosyntactic features of a word (Plural, Indicative, Active, etc., each specifying a morphosyntactic category such as Number, Mood, and Voice) depending on the syntactic context in which the word is inserted. Inflection adjusts the words provided by the lexicon to the morphosyntactic requirements of the syntax.

6.3.2 Agglutinativeness

Regarding the “agglutinativeness” nature of languages, Aikhenvald (2007, p. 4) describes that:

> word may consist of several morphemes but the boundaries between them are clear cut. There is typically a one-to-one correspondence between a morpheme and its meaning, and a morpheme has an invariant shape which makes it easy to identify.

6.3.3 Tamil as an Inflectional and agglutinative language

In Tamil, every word or lexeme inherits the grammatical features provided by its syntactic context in a sentence through either affixes or grammatical words. These grammatical affixes, postpositions, particles, and clitics are added to the stem - the lexeme - like beads on a string. Hence, the Tamil language is said to be one of the inflectional and agglutinative languages (Agesthialingom & Varma, 1980; Devi, 2011; Sheshasaayee & VR, 2015).

6.3.4 Grammaticalization

In Tamil, no grammatical feature is expressed by individual words - that is, as free morphemes. All the grammatical features are expressed by bound morphemes only - affixes, postpositions, and particles. Here it is to be noted that the postpositions and particles were originally lexemes. Later, these lexemes are grammaticalized and function as grammatical words in the history of Tamil language. That is why, even after grammaticalization, they maintain their “word like” shapes. The original lexemes behind these continue as the lexemes - free morphemes - also. But their grammaticalized forms cannot occur as free morphemes; they have to be added with the lexemes as bound forms. This is the demarcation between a word or lexeme form and its grammaticalized form.

6.3.5 Postpositions in Tamil

Kotandaraman (1997, p. 27) observes that: “A form which is historically traceable either to a noun or a verb and which does the function of a case suffix is a postposition.” That is, those nouns or verbs were grammaticalized to undertake some grammatical functions. But even after this grammaticalization process, these forms continue to exist as individual lexemes also. For example, the postposition “paRRi” is the verbal participle form of the the Verb Lexeme “paRRu” ‘to hold’. But now this form is grammaticalized as a postposition which occurs with the accusative case suffix “ai” giving the meaning ‘about’.

1. “naaN avar kaiyaip paRRi izuththeeN” ‘Holding his hand, I pulled him’
2. “naaN avaraippaRRip peeciNeeN” ‘I talked about him’

In (1), “paRRi” is one of the variants (verbal participle) of the Verb lexeme “paRRu” ‘to hold”; hence it occurs as a free morpheme. In (2), this is a grammatical item - a postposition which occurs along with the accusative case suffix as a bound morpheme, to express the casal inflection of the lexeme “avar” ‘he’. In this example (2), it is not the verbal participle of “paRRu”’to hold’, but it is a grammatical word “paRRi” ‘about’ - a postposition.

6.3.6 Particles in Tamil

“A form which is historically traceable to a full word, and which is used for conjugating the verbs is a verbal particle, e.g. -piRaku, utaN, etc.” (Kothandaraman, 1997). For example, the word form “piRaku” is an adverb. Later it is grammaticalized as a verbal particle which could be added with a Relative Participle to function as a Verbal Participle. However, it continues to retain its original lexical status also in Tamil.

6.3.7 Auxiliary Verbs in Tamil

Like postpositions which are originally lexemes, but grammaticalised later, many auxiliary verbs - aspectual, modal, and voice - exist in Tamil, which are grammaticalised forms from original lexemes (Agesthialingom, 2004). However, along with these grammaticalised forms, they maintain their lexical status also. For example, the verb root “paar” ‘to see’ now has been grammaticalised as an aspectual auxiliary verb also.

1. “naaN avaNaip paarththeeN” ‘I saw him’
2. “naan athai ezuthippaarththeeN” ‘I tried writing it’

In the first example (1), “paar” in the finite verb “paarththeeN” ‘saw-I’ is a lexeme; in the example (2), “paar” in “ezuthippaarththeeN” is an aspectual auxiliary giving the meaning ‘attempted to do something’. But since it represents a grammatical feature - aspectual auxiliary - it cannot exist as an independent unit (as a free morpheme), but to be added with the verbal participle “ezuthi” as a bound form (the root is “ezuthu” ‘to write’).

6.3.8 Clitics in Tamil

In addition to the above-mentioned grammatical suffixes and particles, there are some clitics such as “thaaN” (only) and “kuuta” (too), which are originally separate lexemes, are grammaticalised and added to the lexical words - noun, verb, adverb - as bound forms. Here, they add some grammatical or semantic features to the lexemes.

Examples:

1. “avar maaNnavarthaaN” > “avar maaNnavar-thaaN” ‘yes, he is a student’
2. “avarkuuta coNNaar” > “avar-kuuta coNNaar” ‘he too said’

Under oun inflection, plural, cases and postpositions are analysed; under verb inflection, tense, person, number, gender, auxiliary verbs, verbal participles, adjectival participles, participial nouns, adjectival nouns and verbal nouns are analysed; and the inflections of adjectives and adverbs are discussed in the following sections. In the addition to above, expect the adjective, with other categories various clitics are added. In fact, clitics are not inflectional category; they add some semantic content to the words with which they occur.

6.4 Inflection (Noun)

6.4.1 Noun and Pronoun inflection

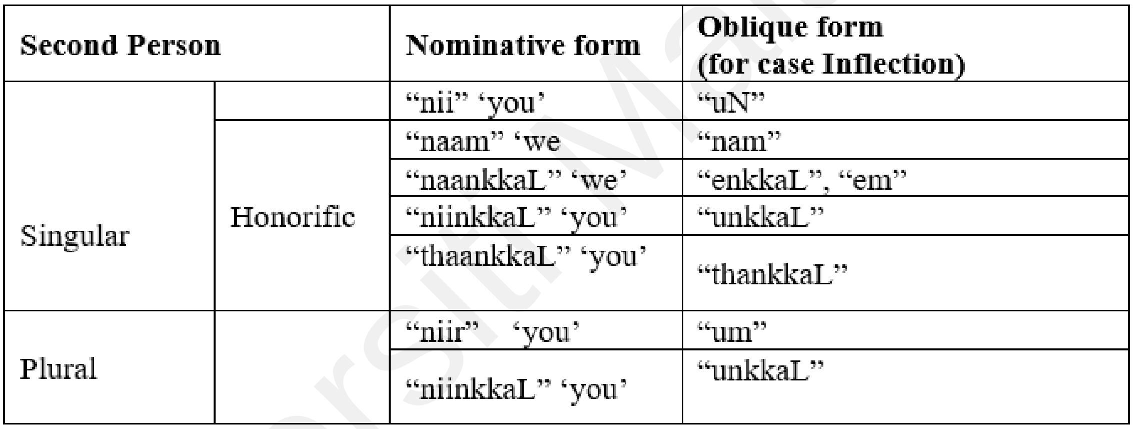
Tamil nouns and pronouns are inflected for number (singular and plural), case, and clitics (Mohanlal et al.). Some noun lexemes, not all, through some suffixes, show the gender also (“maaNnavaN” ‘student - male’; “maaNnavi” ’student -female’). The lexeme “muthalaaLi” ‘owner’ stands for both genders. Generally, the gender of a Subject noun in a particular sentence could be understood by the person - number - gender (PNG) suffix

- as an ‘agreement’ or ‘concord’ feature - occurring in the finite verb of the sentence. Nouns belong to the open class, whereas pronouns are closed ones which are tabled below:

Table 6.1: Pronoun (First Person)

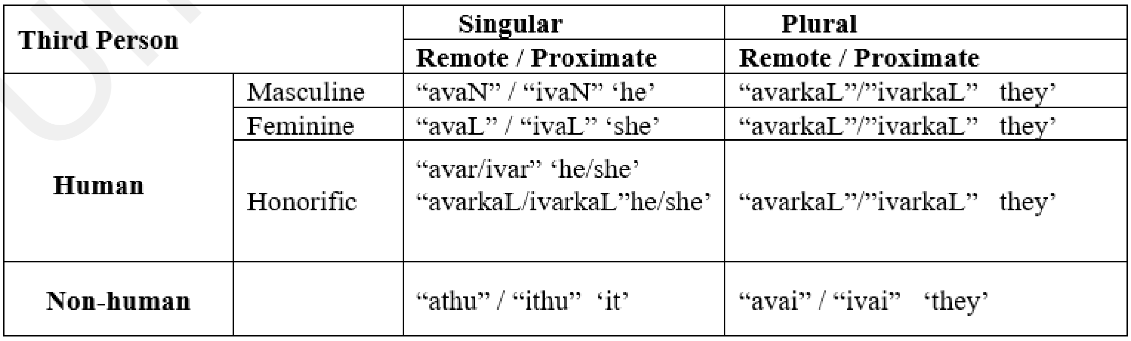
First Person	**Nominative form**	**Oblique form (for case Inflection)**
Singular	“naaN” ‘I’	“eN”
Plural (inclusive)	“naam” ‘we	“nam”
Plural (exclusive)	“naankkaL” ‘we’	“enkkaL”, “em”

Table 6.2: Pronoun (Second Person)



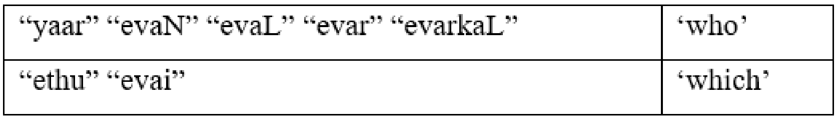
Second Person		**Nominative form**	**Oblique form (for case Inflection)**
Singular		“nii” ‘you’	“uN”
	Honorific	“naam” ‘we	“nam”
		“naankkaL” ‘we’	“enkkaL”, “em”
		“niinkkaL” ‘you’	“unkkaL”
		“thaankkaL” ‘you’	“thankkaL”
Plural		“niir” ‘you’	“um”
		“niinkkaL” ‘you’	“unkkaL”

Table 6.3: Pronoun (Third Person)



Third Person		**Singular**	**Plural**
		Remote / Proximate	**Remote / Proximate**
Human	Masculine	“avaN” / “ivaN” ‘he’	“avarkaL”/”ivarkaL” they’
	Feminine	“avaL” / “ivaL” ‘she’	“avarkaL”/”ivarkaL” they’
	Honorific	“avar/ivar” ‘he/she’ “avarkaL/ivarkaL”he/she’	“avarkaL”/”ivarkaL” they’
Non-human		“athu” / “ithu” ‘it’	“avai” / “ivai” ‘they’

Table 6.4: Pronoun (Interrogative Pronoun)



“yaar” “evaN” “evaL” “evar” “evarkaL”	‘who’
“ethu” “evai”	‘which’

6.4.2 Caariyai

Like the above oblique forms for pronouns which are used in case inflection, other nouns ending in certain phonemes will take oblique forms by adding some increments (“caariyai”) or doubling of the final phonemes in case inflection.

1) Nouns ending in “- am “: “maram” ‘tree’ + “thth” (“caariyai”) + “ai” ‘acc.case’

2) Nouns having the syllabic structure “(C) V: Cu”:

“kaatu” ‘forest’ + “t” (doubling of C) + “ai” ‘acc.case’

“aaRu” ‘river’ + “R” (doubling of C) + “ai” ‘acc.case’

Here, **C** stands for Plosive or Stop consonants and **V**: for long vowels.

6.4.3 Other Increments

There are other increments in Tamil that occur between some morphemes. Some increments are obligatory, and some are optional.

1) “athu” ‘it’ + “aN” (increment) + “ai” (acc.case) = “athaNai” ‘it (acc.case)’

2) “kaathu” ‘ear’ + “iN” (increment) + “ai” (acc.case) = “kaathiNai” ‘ear (acc.case)’

3) “avai” ‘they’ + “aRRu” (increment) + “ai” (acc.case) = “avaRRai” ‘they (acc.case)’

4) “puLi” ‘tamarind tree’ + “am” (increment) + “pazam” ‘fruit’ = “puLiyampazam” ‘tamarind fruit’

6.4.4 Case

6.4.4.1 Case suffixes

There are eight main cases in Tamil: nominative, accusative, instrumental, dative, locative, ablative, genitive, and vocative

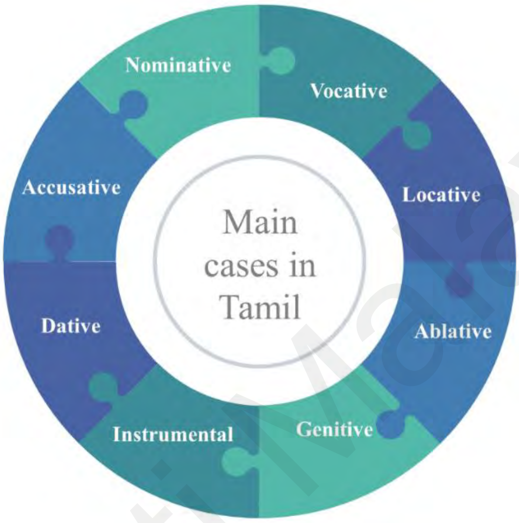


Figure 6.1: Case Suffixes

For nominative there is no separate suffix, but in some sentence constructions, it could be observed that there are some words such as “eNpathu”, “eNpavar”, “aaNathu”, “aaNavai” implying their precedent words are subject of the sentences (Agesthialingom, 1976) (Kothandaraman, 2006).

“kooyil eNpathu vazipatum itam”	‘Temple is a place of worshiping.’
“aaciriyar eNpavar nam vaazviN vazikaatti”	‘A teacher is a guide to our life’
“kutumpam aaNathu oor palkalaikkazakam”	‘A family is a University’

"utalpayiRchikaL aanavai aNaivarukkum mikavum theevaiyaaNavai"	'Exercises are very much needed ones to everybody'

Regarding the vocative case, there are two suffixes, "-ee", "-aa"; also, some prosodic features added with the respective nouns express the vocative case.

"Murukaa, inkkee vaa" 'Murugan, come here'

"naNparee, inkkee vaarunkkaL" 'Friend, come here'

For other seven cases, there are explicit case suffixes.

Accusative:	"ai"
Associative:	"ootu"
Instrumental:	"aal"
Casual:	"aal"
Dative:	"ku"
Genetive:	"athu"
Locative:	"il"

There is one more case - "ablative" - which is represented by a postposition "irunthu". This postposition occurs after the locative case "il" as well as some other nouns denoting place such as "ankkee" 'there' and "meelee" 'above'.

With some cases, without having the explicit presence of suffixes, by the context, the nouns could be inflected; that is, the cases are expressed by zero suffixes or zero allomorphs; the presence of case suffixes is optional.

Though the nouns are inflected for gender and "thinai" ('human - non-human'), they are mostly inherent ones. These inherent features of a nominative noun (subject of a sentence) are explicitly expressed by the Person-Number-Gender suffixes (PNG) of finite verbs of that particular sentence. With the imperative verb of an utterance, even the nominative noun could be absent or dropped.

For example, “nii poo” ‘you go’, the nominative noun “nii” could be dropped. Simply, “poo” ‘go’ can be used. Here, the subject noun could be dropped as in English. Among other cases, the accusative and genitive cases are also may not be expressed by explicit case markers. However, by the context, they could be understood. “avar patam paarththaar” ‘he saw the picture’. Here “patam” is the noun inflected for accusative case. However, the marker of this case could be dropped optionally. “ithu avar viitu” ‘This is his house’: Here, the word “avar” is inflected for Genitive case. But the case marker “athu”or “utaiya” is dropped.

6.4.4.2 Postpositions

In addition to the above-mentioned cases, there are more than fifty postpositions occur to express more case relations. There are two types of postpositions: one is, the postpositions which have to follow some case suffixes. The other one is the postpositions which could occur without any preceding case suffixes.

“avar maanavarkaLaippaRRip peecinaar” ‘he talked about the students’.

Here, the word “maanavarkaLaippaRRip” (“maanavar + kaL + ai + (p) + paRRi(p) has a postposition “paRRi” after the accusative case “-ai”. They jointly express the meaning “about’. But, with the following sentence, the postposition “itam” occurs without any preceding case suffix: “avaritam panam irukkiRathu” ‘With him, there is money’ > ‘he has money’.

6.4.4.3 Clitics

Regarding clitics, a maximum of four clitics could occur with a noun. However, if more than two clitics occur with a noun, there are some orders of occurrence to be followed. This should not be violated. With a word, a maximum of four clitics may be added with the nouns, verbs, and adverbs. However, there are some orders of occurrences among the clitics.

6.5 Verb inflection

There are two types of verbs in Tamil: one is, Finite verb; the other one is non-finite, that is, Participles.

6.5.1 Tense inflection

Except for a few, all the finite verbs are conjugated for tense, in addition to the PNG suffixes. There are three tenses in Tamil: present, past, and future. However, there are more allomorphs for these tenses. So, the verbs are generally grouped under thirteen types. This classification should not be violated by the verbs; otherwise, the resultant form would be ungrammatical.

6.5.2 PNG inflection

The finite verbs that occur in the 'predicate' of a sentence are mostly inflected for Person-Number- Gender (PNG). This inflection depends upon the 'PNG' of the 'Subject' of the sentence. That is, the finite verbs should be in agreement with the "Subject" of the sentences. In other words, there is a "concord" between the 'Subject' and the 'finite verb.' The PNG suffixes are listed in the table below:

Table 6.5: PNG First Person

First Person		
Singular	“-eeN”	“paarththeeN” ‘saw -I’
Plural	“-oom”	“paarththoom” ‘saw -We’

Table 6.6: PNG Second Person

Second Person			
Singular		“-aay”	“paarththaay” ‘saw -You’
	Honorific -1	“-iirkaL”	“paarththiirkaL” ‘saw - You’
	Honorific -2	“-iir”	“paarththiir” ‘saw -You’

Table 6.7: PNG Third Person

Third Person				
Human	Singular	Masculine	“-aaN”	“paarththaaN” ‘saw - he’
		Feminine	“-aaL”	“paarththaaL” ‘saw - she’
		Honorific I	“-aar”	“paarththaar” ‘saw - he/she’
		Honorific II	“-aarkaL”	“paarththaarkaL” ‘saw - he/she’
	Plural		“-aarkaL”	“paarththaarkaL” ‘saw - they’
Non-Human	Singular		“-athu”	“paarththathu” ‘saw -it’
	Plural		“-aNa”	“paarththaNa” ‘saw -they’

6.5.3 Auxiliary verbs

The verbs are divided into main verbs and auxiliaries. These auxiliary verbs denote the aspectual, modal and voice features. Therefore, the verbs in the predicate, in addition to the inflection of tenses and PNG, could take these features (Agesthialingom, 1964).

6.5.3.1 Aspectual

Aspectual stands to express the way in which the action of the main verbs takes place. Here, it is to be mentioned that the auxiliary aspectual verb could follow only the past participle form of the main verb. In Tamil, it is called as “ceythu” participle. There are many aspectual auxiliaries.

Table 6.8: Aspectual

Present Perfect Aspectual	
avar vanthirukkiRaar	‘he has come’
vanthu + irukkiRaar	‘(having) come + has’

6.5.3.2 Modal

Modal stands to denote the attitude of the speaker towards the action of the main verb. The modal auxiliaries in Tamil could follow only the infinitive forms of the main verb. In Tamil, it is called “ceya” participle. Some modal auxiliaries may occur without tense and PNG suffixes. There are many modal auxiliaries.

Table 6.9: Possibility modal

Possibility modal	
avar varalaam	‘he may come’
vara + laam	‘come + may’

6.5.3.3 Voice

Tamil has three voices: (1) active, (2) passive and (3) causative.

Passive verb construction occurs when the “subject” of the main verb undergoes the action. Passive voice auxiliary verbs could follow only the infinitive forms of the main verb. These morphotactic rules should follow the above order. There are two passive auxiliaries: “patu” and “peRu”. The suffix “vai” is the causative one.

Table 6.10: Passive voice

Passive Voice	
athu vaankappattathu	‘that was bought’
vaanka + pattathu	‘bought + was’

Table 6.11: Causative voice

Causative Voice	
avan patikkavaiththaaN	‘he made (someone) study’
patikka + vai + thth + aaN	‘to study + made’

Some more forms of Causative voice

“naaN avaraip patikkac ceytheeN’ ‘I made him study’ - Causative voice.

“naaN avaNaip patippiththeeN” ‘I made him study’. - Causative voice

“naaN avanukkuk kaaNnpiththeeN” ‘I showed (caused him to see) him’ - Causative voice

“naan avaraic ceyviththeeN” ‘I caused him to do”

6.5.4 Verbal Participles

In Tamil, verbs are inflected for various participles. In modern Tamil, there are six forms of verbal participles: Infinitive, Conjunctive, Conditional, Consecutive, Simultaneous, and Negative.

6.5.4.1 Infinitive verbal participle

“-a” is the infinitive marker in Tamil. Its alternants are “-ka” and “-kka”.

It can occur independently as well as when a verb is inflected for Voice or Modal Auxiliary, the verb will take this form.

Table 6.12: Infinitive markers

(“-a”)	“ceya” > “cey - a “	‘to do’
(“-ka”)	“pooka” > “poo - ka”	‘to go’
(“-kka”)	“patikka” > “pati - kka”	‘to study’

6.5.4.2 Conjunctive verbal participle:

There are six markers for this verbal participle. It can occur independently as well as when a verb is inflected for Aspectual Auxiliary, the verb will take this form.

Table 6.13: verbal participle markers

(“-thu”)	“ceythu” > “cey - thu”	‘having done’
(“-ththu”)	“patiththu” > “pati - ththu”	‘having studied’
(“-nthu”)	“natanthu” > nata - nthu”	‘having walked’
(“-i”)	“ooti” > “oot(u) - i”	‘having run’
(“-y”)	“pooy” > “poo - y”	‘having gone’
(“-tu”)	“thottu” > “thotu - tu”	‘having touched’

6.5.4.3 Conditional participle

There are two alternants for this participle:

Table 6.14: Conditional participle markers

(“-aal”)	“ceythaal” > “ceythu - aal”	‘if done’
(“-Naal”)	“aatiNaal” > “aati - Naal”	‘If played’

6.5.4.4 Consecutive participle

There two alternants for this participle:

Table 6.15: Consecutive participle

(“-athum”)	“ceythathum” > “ceythu -athum”	‘after having done’
(“-Nathum”)	“pooNathum” > “poo - Nathum”	‘after having gone’

6.5.4.5 Simultaneous participle

There are three alternants for this participle:

Table 6.16: Simultaneous participle

(“-kaiyil”)	“ceykaiyil” > “cey - kaiyil”	‘while doing’
(“-kkaiyil”)	“patikkaiyil” > “pati - kkaiyil”	‘while studying’
(“-um”)	“ceyyavum” > “cey(y)a - (v)um”	‘while doing’

6.5.4.6 Negative participle

There are six markers for this participle:

Table 6.17: Negative participle

(“-aamal”)	“ceyyaamal” > “cey(y) - aamal”	‘without doing’
(“-kaamal”)	“pookaamal” > “poo - kaamal”	‘without going’
(“-kkaamal”)	“patikkaamal” > “pati - kkaamal”	‘without studying’
(-aathu”)	“ceyyaathu” > “cey(y) - aathu”	‘without doing’
(“-kaathu”)	“pookaathu” > “poo - kaathu”	‘without going’
(“-kkaathu”)	“patikkaathu” > “pati - kkaathu”	‘without studying’

6.5.5 Adjectival or Relative participle

In Tamil, there are four types of Relative participles, which are derived from verbs:

Table 6.18: Past Relative participle marker

(“-past tense suffix + -a”)	“ceytha” > “cey - th - a”	‘that which had been done’

Table 6.19: Present Relative participle

(“-Present tense suffix + a”)	“ceykiRa” > “cey -kiR - a”	‘that which has been done’

Table 6.20: Future Relative Participle

(“-um”)	“ceyyum” > “cey(y) - um”	‘that which will be done’

Table 6.21: Negative Relative participle

(-aa (tha)”)	“ceyyaa(tha) > “cey(y) - aa(tha)”	‘that which is not done’

6.5.5.1 Complex Adjectival participles

Some complex Adjectival participles are derived from Infinitive verbal participles by adding the relative participle form of modals such as “veeNntiya” ‘be needed’, “kuutiya” ‘be capable,’ “thakka” ‘be suitable,’ “mutintha” ‘be able’. Here, it is to be noted that these adjectival participles would precede nouns.

Table 6.22: Complex Adjectival participles

“patikka-veeNntiya” > “patikkaveeNntiya”	‘be needed to study’
“patikka-kuutiya” > “patikka(k)kuutiya”	‘be capable to study’
“patikka-thakka” > “patikka(th)thakka”	‘be suitable to study’
“patikka-mutintha” > “patikkamutintha”	‘be able to study’

6.5.5.2 Complex Adverbial participles

Likewise, some complex Adverbial participles are derived from verbal participles by adding some other verbal participles. Here, it is to be noted that these adverbial participles would precede finite verb or any other verbal participles.

Table 6.23: Complex Adverbial participles 1

“ceythu - irukka” > “ceythirukka”	‘should have been done’
“ceyyaamal - irukka” > “ceyyaamalirukka”	‘should not have been done’
“ceythu - koNntu” > “ceythukoNntu”	‘have been doing’

6.5.5.3 Complex Adverbial participles 2

Some more complex adverbial participles could be formed by adding some verbal particles to the Relative participles.

Table 6.24: Complex Adverbial participles 2

“vantha - piN” > “vanthapin”	‘after having come’
“coNNa - pati” > “coNNapati”	‘as told’
“patiththa - varai” > “patiththavarai”	‘be enough to read’

6.5.6 Participial Nouns

Verb stem - tense / negative suffix - Gender Number suffix

Gender Number suffixes: “avaN”, “avaL”, “avar”, “avarkaL”, “athu”, “avai”

Table 6.25: Participial Nouns

“patiththavaN” > “pati - thth - avaN”	‘he who studied’
“patikkiRavaN” > “pati - kkir - avaN”	‘he who studies’
“patippavaN” > “pati - pp - avaN”	‘he who will study’
“patikkaathavaN” > “pati - kk -aath - avaN”	‘he who does /did not study’

After a verb takes the form of a “Participial Noun”, it behaves like a noun for further inflection.

6.5.7 Adjectival Nouns

Adjectives can also be added with some pronominal suffixes (“avaN” “avaL” “avar” “avarkaL” “athu” “avai”) to generate nouns. All these suffixes express the gender, number along with “thinai” (Human - non-human distinction). Though these forms seem to be third person, they stand for all persons - first, second and third persons- in this inflection. Example:

Table 6.26: Adjectival Nouns

“nalla - avaN” ‘good - person’	“nallavaN” ‘a good person’
“naaN nallavaN”	‘I am a good person’
“naankkaL nallavarkaL”	‘We (exclusive) are good persons’
“naam nallavarkaL”	We (inclusive) are good persons’
“nii nallavaN”	‘you are a good person (masculine)’
“nii nallavaL”	‘you are a good person (feminine)’
“niinkkaL nallavar”	“You (singular) are a good person’
“niinkkaL nallavarkaL”	‘you (plural) all are good persons’
“avaN nallavar”	“He is a good person’
“avaL nallavaL”	‘She is a good person’
“avarkaL nallavarkaL”	“They are good persons”
“athu nallathu”	‘That is a good one’
“avai nallavai”	‘They are good ones’

6.5.8 Verbal Nouns:

Table 6.27: Verbal Noun

1.“ceythal” ‘doing’	
“cey - thal”	‘do - VN suffix’ > ‘doing’
2. “ceykiRathu” ‘the act of doing’ (present)	
“cey - kiR - athu”	‘do - present tense - VN suffix’ > ‘doing’
3. “ceyyaathathu” ‘the act of not doing’	
“cey - y - aath - athu”	‘do - sandhi - NEG suffix - VN suffix’ > ‘not doing’

Here, the first one has no tense; that is, it is not inflected for any tense. However, the second and third types could be inflected for all the three tenses and negative, respectively.

After a verb takes the form of a "Verbal Noun," it behaves like a noun for further inflection. Here, it is to be mentioned that though both participial nouns and verbal nouns could be inflected for cases as like pure nouns, they cannot be preceded by adjectives but be preceded by adverbs. In this aspect, both these forms behave like verbs. In other words, both the participial nouns and verbal nouns, in one sense - inflected for cases - behave like nouns, but in taking modifiers, they behave like verbs. Between participial nouns and verbal nouns, there is a difference in taking plural suffixes. Where the participial nouns can take plural suffix "kaL", the verbal nouns will not take this plural suffix.

6.5.9 Adjectives and Adverbs

In Tamil, adjectives won't be inflected for any grammatical feature. But adverbs could take clitics such as emphasis.

Table 6.28: Adjectives and Adverbs

"avar veekamaaka varukiRaar"	'he comes fast'
"avar veekamaakaththaan varukiRaar"	'he comes fast only'
"avar veekamaakavee varukiRaar"	'he comes faster'

6.5.9.1 Relative Participles - Adverbs

However, Relative participles which function as adjectives one could combine with some verbal particles to become adverbs. But it is not inflection; but it is derivation. But Pure adjectives (“nalla” ‘good’; “ketta” ‘bad’), including derived adjectives (“azhakaaNa” > “azhaku + aaNa” ‘beautiful’) will not behave like this.

“avar vanthapiRaku naaN vantheeN” ‘I came after he did’

(“vantha + piRaku” ‘after he came’)

“avar varummuN naaN vantheen” ‘I came before he did’

(“varum + mun” ‘before he came’)

6.5.9.2 Adjectives and Adverbs

Adjectives and Adverbs from Nouns: In Tamil, nouns can be changed into adjectives and adverbs by adding the adjectival suffix and adverbial suffix “-aaNa” and “-aaka” respectively.

“azhaku” + “aaNa” > “azhakaaNa” ‘beautiful’

“azhaku” + “aaka” > “azhakaaka” ‘beautifully’

All the above discussions have clearly explained the inflection of the major lexical categories in Tamil - Noun, Verb, Adjective, and Adverb. All these inflection processes and their allomorphs are accommodated in the present morphological parser. The occurrence of these lexical categories obeys certain order which are treated as under.

6.6 Morphotactic rules

As explained before, in the inflection of words, every lexical word is inflected wherever necessary. Here, it is to be mentioned that since Tamil belongs to Head-last language, all the inflectional affixes occur as suffixes only. In Sanskrit borrowed words, some prefixes occur. Example: “niithi” ‘justice’ - “aniithi” ‘injustice’; “cuththam” ‘clean’ - “acuththam” ‘unclean’. Here “a-” in both instances is the prefix denoting the negation.

In modern Tamil, some lexicons are becoming as prefixes, if we consider their productivity. “muRpooku” ‘progressive’, “muRkaalam” ‘ancient period’, “muNNooti” ‘pioneer’.

It may be necessary to add more than one suffix to a particular word. In that case, there are rules to say in which order those suffixes are added with the lexical word. That is, there are some specific rules for the arrangement of suffixes in a lexical word. These are known as morphotactic rules.

6.6.1 Noun Morphotactics

For example, with a lexical noun, four kinds of suffixes can occur. Plural suffix, case suffix, postposition and clitics. Here, it is to be mentioned that the order of occurrence of suffixes in noun word forms, that is, the morphotactic rules of Tamil word inflection plays a crucial role in getting the correct grammatical word. Here, it is to be mentioned that the clitic component mainly occurs after the inflection, though there are some exemptions.

Between a noun lexeme and plural suffix, no clitic could occur. However, in between a case and a postposition, some clitics may occur.

1. “naaN avaraipaRRiyum peeciNeeN” ‘I talked about him too’
2. “naaN avaraiyumpaRRip peeciNeeN” ‘I talked about him too’

Here, in (1), the clitic “-um” occurs after the postposition “paRRi” whereas in (2), it occurs in between the case suffix “-ai” and the postposition “paRRi.”

Even among the inflectional suffixes, there is a clear-cut order for the occurrences of suffixes. For example, if a noun takes both a plural suffix and a case suffix, the plural suffix should precede the case suffix. Example: “paiyankaLai” ‘boys (acc.)’ > “paiyan” ‘boy’ - “kaL” ‘plural suffix’ - “ai” ‘acc. case suffix’.

The following representation clearly shows the noun morphotactics discussed above:

Noun + (Number) + (Filler) + (Case suffix) + (Clitic) + (Postposition) + (Clitic)

Noun: “maram” ‘tree’

Noun + Number: “maram - kaL” > “marankkaL” ‘trees’

Noun + Clitic: “maram - ee” > “maramee” ‘It is tree only’

Noun + Filler: “maram - ththu” > “marathththu” ‘tree’

Noun + Filler + Case: “maram - ththu - ai” > “maraththai” ‘tree (acc.)’

Noun + Filler + Case + Clitics: “maram - ththu - ai - (y)um” > “maraththaiyum” ‘tree (acc.) too’

Noun + Filler + Case + Postposition: “maram - ththu - ai - (p)paRRi”

➔ “maraththaippaRRi” ‘about the tree’

Noun + Filler + Case + PP + Clitics: “maram - ththu - ai - (p)paRRi - (y) aa”

➔ “maraththaippaRRiyaa” ‘Is it about the tree?

The outcome of the above-mentioned algorithm is represented as the noun inflectional chart (noun morphotactic rules) as follows:

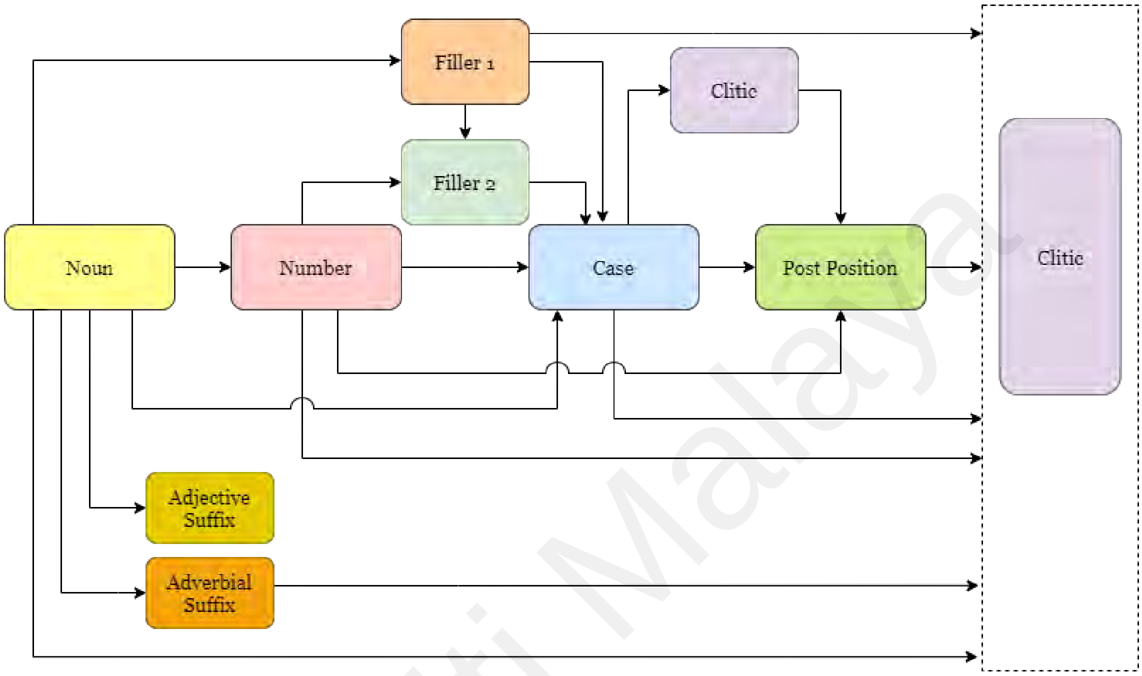


Figure 6.2: Tamil Noun inflectional chart

6.6.2 Verb Morphotactics

When a Tamil verb undergoes inflection, the morphotactic rules as shown in the following representation should be followed:

Verb + (Asp. Aux) + (Voice Aux) + (Modal Aux) + (Tense) + PNG + Clitics

1. Verb + Tense + PNG: “pati -thth -aaN” > “patiththaaN” ‘studied - he’
2. Verb (“ceythu” form) + Asp. Aux + PNG: “patiththu - iru - kiR - aaN” ‘has studied - he’
3. Verb (“ceya” form) + Voice Aux + Tense + PNG: “patikka(p) - patu - t - athu” ‘ is studied - it’

4. Verb (“ceya” form) + Modal Aux: “patikka - laam” > “patikkallam” ‘may be studied’
5. Verb (“ceya” form) + Modal Aux + Tense + PNG: “patikka - paar - thth -aaN” > “patikkappaarththaaN” ‘attempted to study - he’
6. Verb (“ceythu” form) + Asp Aux (“ceya” form) + Voice Aux + Tense + PNG: “eluthi(k) - kaatta (“ceya” form) - vai - thth - aaN” > “eluthikkaattavaiththaaN” ‘caused (someone) to show writing’
7. Verb (“ceya” form) + Voice Aux + Modal Aux: “patikka(p) - pata (“ceya” form) - laam” > “patikkappatalaam” ‘might be studied’
8. Verb (“ceya” form) + Voice Aux + Asp Aux + Modal Aux: “patikka(p) - pattu (“ceythu” form) - irukka (“ceya” form) - laam” > “patikkappattirukkalaam” ‘might have been studied’

In Verb inflection also, in some contexts, the clitics could occur between the main verb and the following auxiliaries (Aspectuals and Modals) as follows:

1. “naaN avaraip paarththuyirukkiReeN**aa**?” ‘Have I seen him?’
2. “naaN avaraip paarthth**aa**yirukkireeN?” ‘Have I seen him?’

In (1), to emphasize the whole verb phrase “paarththuyirukkiReeN”, the clitic occurs finally whereas in (2), since the emphasis is shifted to the main verb - “paarththu”, the clitic is added to this main verb before the auxiliary verb “irukkiReeN”. The outcome of the above-mentioned algorithm is represented as the verb inflectional chart (verb morphotactic rules) as follows:

The verb inflectional chart (verb morphotactic rules)

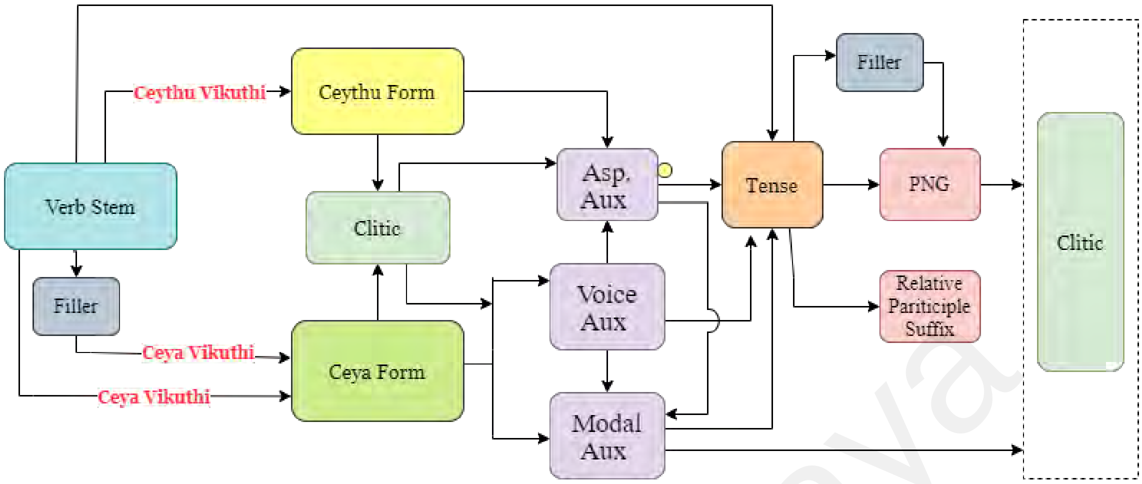


Figure 6.3: Tamil Verb inflectional chart

6.7 Adjective

In derived adjectives (adjectives derived from nouns, by adding the suffix “-aaNa”), between the noun and adjective suffix, clitic may occur.

“azaku - aaNa” > “azakaaNa ‘beautiful’

“azaku -um - aaNa” > “azakumaaNa” - ‘beautiful too’

Here, it is to be mentioned that from the adjectives, by adding respective Number - Gender suffixes, nouns could be derived.

“nalla - (v)aN” > “nallavaN” ‘handsome person (male)’

“azakaaNa - (v)aN” > “azakaaNavaN” ‘handsome person (male)’

“azakaaNa - (v)aL” > “azakaaNavaL” ‘beautiful person (female)’

“azakaana - (v)arKaL” > “azakaaNavarkaL” ‘handsome/beautiful persons (male & female)’.

“azakaaNa - athu” > “azakaaNathu” ‘beautiful thing(non-human)’

“azakaaNa -(v)ai” > “azakaaNavai” ‘beautiful things (non-human)’

Adjective:

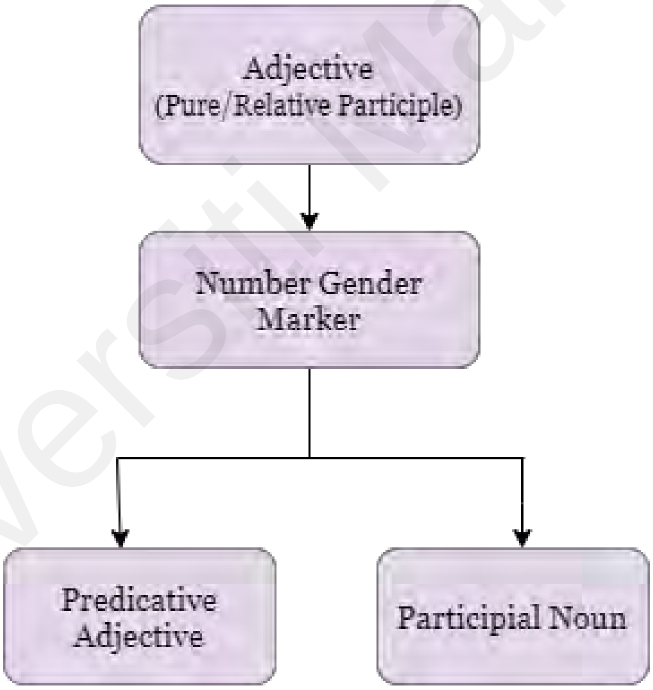


Figure 6.4: Adjective Chart

6.8 Adverb

Like adjectives, in the derived adverbs also, some clitics may occur between the noun and adverbial suffix.

“collum ceyalum - aaka” > “collum ceyalumaaka” ‘with word and action’

Adverb:

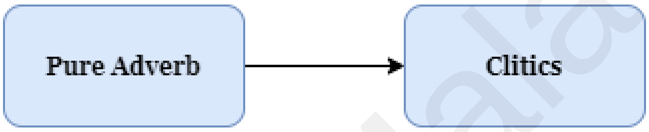


Figure 6.5: Adverb Chart

The above flow chart of Tamil inflection is the backbone for the development of morphological parsing and POS tagging. This kind of robust flow charts would pave the way for further research in corpus linguistics, computational linguistics, language analysis, lexicography, pedagogical methods, and Language technology.

6.9 Cohesive devices

Various words used in discourse as cohesive devices such as “aakavee” ‘that is why,’ “eeNeNRaal” ‘because,’ “athaNaal” ‘therefore’ are placed in the lexicon under the category ‘adverb’. In traditional grammars, they are described as “itaiccol” ‘word in between’. That is, they are neither nouns nor verbs. At the same time, they are not grammatical suffixes also.

6.10 Morphophonemic rules "Sandhi"

There are some phonologically changing rules that operate between lexical word and following suffix and between suffix and suffix (Kothandaraman, 1997). These rules are called here as sandhi rules.

In Tamil, both in inflection as well as in derivation, there are some sandhi rules that operate. They should not be avoided. In some contexts, the occurrence of sandhi helps to disambiguate the words having similar phonological or orthographic words. "avar veelai paarkkiRaar" 'he works' (here "veelai" means "work" or "job" - a single word.); "avar veelai**p** paarkkiRaar" 'he sees the spear' (here "veelai "is "veel + ai"; "veel" 'spear', "ai" is accusative case suffix and "**p**" is the sandhi.

So, the phonological word "veelai" is an ambiguous one. In this context, the sandhi "p" helps us to disambiguate the meaning of these two forms. And in Tamil, two vowel phonemes cannot occur in adjacent position within a word - a single word or a compound word. That is, a vowel should not be followed by another vowel. In such a situation, the consonant "y" or" v" occurs. These two consonants are called as "utampatu mey" 'glide' in Tamil grammar.

6.10.1 Three kinds of Sandhi rules

1. Deletion of phoneme
2. Addition of phoneme
3. Change in phoneme.

Deletion: “maram” ‘tree’ + “veer” ‘root’ > “maraveer” ‘the root of the tree’; here, the final phoneme “m” of “maram” is deleted.

Addition: “poor” ‘war’ + “kappal” ‘ship’ > “poorkkappal” ‘war ship’: here , in between the two words , one phoneme ‘k’ is added.

Change: “maram” ‘tree’ + “kaL” ‘plural suffix’ > “marangaL” ‘trees’; here, the final phoneme of “m” in “maram” changes into “ng” phoneme

Deletion and addition: “maram” ‘tree’ + “palakai” ‘plank’ > “mara+palakai” > “mara + p + palakai”; here, first the ending of “maram” is deleted and then the addition of “p” takes place between “mara” and “palakai”.

Doubling of phoneme: “aaRu” ‘river’ + Genitive case “0” + “thaNnNniir” ‘water’ > “aaRRu” + “thaNnNniir” > aaRRu + th + thaNnNniir” > “aaRRuththaNnNniir” ‘river water’; here, first, the final ending of “aaRu” is doubled; then in between “aaRRu” and “ thaNnNniir” , one addition of “th” takes place.

6.10.2 Occurrence of inflectional increment “Caariyai”

In addition, with some nouns, some empty morphemes are called “caariyai” in Tamil. They may be called inflectional increments in some grammars. When all these suffixes are added with a lexical noun word, there are some clear-cut orders in the arrangement of their occurrence. If this order of arrangement is not followed, the resultant inflected word will become ungrammatical.

“maraththaippaRRiyaa” > mara(m) + aththu + ai + (p) + paRRi + (y) + aa >
‘tree’ + ‘caariyai’ + ‘acc. case’ + ‘(sandhi)’ + ‘postposition’ + ‘(sandhi)’ + ‘clitic’ > ‘is it about the tree?’

In the above example, the root word “maram” before being inflected for accusative case, its “m” ending is dropped and then the resultant form takes the Caariyaai “aththu”.

6.10.3 Occurrence of Glide

To avoid vowel clusters in Tamil, some addition of consonant (or called as “semi-vowel”) takes place. They are: “y” or “v.”

Not only inside a word and compounds (word + word), but also between words, there is sandhi. That is, there are internal as well as external sandhi processes.

External Sandhi: “avarukku paNnam theevai” > “avarukku-p paNnam theevai” ‘he (dative case)’ + ‘money’ + ‘needs’ > ‘he needs money’.

So, it is important to know the morphotactic and morphophonemic rules during morphological parsing. That is, the parser program should be able to identify the morpho tactic rules and morphophonemic rules during the parsing process of a word.

6.11 Part-of-Speech Tagging

Based on the main grammatical categories and then their sub-categories existing in Tamil, in the present thesis, the words (Word forms/Types) are tagged into 52 sets. The norms followed in giving the abbreviations for the word categories are explained here.

Nouns are sub-categorized for the properties - number, case, and postpositions. Verbs are sub-categorized for six properties - tense, negative, PNG, aspectual, modal, and voice. The clitics are added with nouns, verbs, adjectives and adverbs.

Tamil is a Head-last language not only in syntax but also in morphology (Deivasundaram, 2021).That is, the final suffix in a word form/type decides the category of the respective word form. For example, the word form "patiththaaN" 'studied - he' is morphologically parsed into three parts: "pati - thth - aaN". Here, there are three components in this word form. The first one is the verb stem; the second one is tense; the final component is PNG.

Table 6.29: Component of the world

Verb Stem	**Tense**	**PNG**
"pati"	"-thth-"	"-aaN"
'study'	'past tense suffix'	'PNG suffix'
'studied - he'		

This final suffix determines the word category of the whole word as a finite verb. Here, it is to be noted that the addition of clitics will not change the word level category. For example, both "patiththaaN" 'studied - he' and "patiththaaNaa?" 'Did he study?' are finite verbs, though the second one contains an interrogative clitic "aa." With the nouns also, it is the same. However, the clitics add some semantic features to the words with which they are added.

The difference between transitive and intransitive verbs is lexical based. That is, this difference is not marked by any grammatical suffix.

With nouns, the difference between human and non-human is not explicitly marked by any suffix. It is also a lexical based one. That is, this difference should be noted in the lexical database. For example, the word "selvam" generally stands for the meaning 'wealth'. But it is also used to name a person as a proper noun. The difference could be understood only by the PNG of the finite verb, or some pronouns in the particular sentence.

"selvam vanthaar" 'Selvam came'
"selvam vanthathu" '(someone) got wealth'

Here, the PNGs occurring in the finite verbs in the above sentences help to disambiguate the meaning in the word "selvam". In the first sentence, the finite verb "vanthaar" has "aar" as PNG which is in agreement with the subject "selvam" and in the second one, the PNG "athu" occurring in the finite verb helps us to know that the subject of this sentence is a non-human one.

6.12 POS Tagger

As mentioned in the previous 5th chapter on ‘The development of morphological parser and POS tagger’, once the word forms are segmented into lexical roots or stems, the next task was to tag every word form with its POS. This was done by the POS tagger program which was based on the output segments with their respective categories of every word form.

To decide the POS of a Tamil word form or type, the following knowledge of Tamil grammar was provided to the POS tagger program:

1. The POS of every root word in Tamil is given in the lexical database.
2. The grammatical category of all the suffixes, postpositions, and particles.
3. The deciding principle to guide the tagger in assigning the correct POS category to the parsed word form - the inflected word.

Based on the analysis of Tamil Inflection, for the present research project, 51 (fifty-one) major POS categories are identified and included to tag the inflected words. Among them, for noun word forms, there are 29 (twenty-nine) categories, for verb 17 (seventeen) categories, and for adjective and adverbs 5 (five).

Table 6.30: WMTC Noun Tag List

WMTC Noun Tag List			
no	**Noun**	**Suffix**	**Tag**
1	katalkaL	kaL	N- NH-PL
2	aaciriyarkaLa	kaL	N-H-PL
3	man	0	N-NH-C1/2/4/6
4	aaciriyar	0	N-H-C1/6/8
5	aaththiyai	Ai	N-C2
6	viittaiccuRRi	aiccuRRi	N-C2-P
7	kaththiyaal	Aal	N-C3
8	thaRiyil, viittil	Il	N-C3/7
9	viittukku	Kku	N-C4
10	viittukkaaka	Kkaaka	N-C4-P
11	viittilirunthu	Ilirunthu	N-C5
12	avaNutaiya	Utaiya	N-C6
13	viittiNmeel	iNmeelee	N-C6-P
14	avaNutaN	utaN	N-P
15	kaththikaLai	kaLai	N-PL-C2
16	viitukaLaiccuRRi	kaLaiccuRRi	N-PL-C2-P
17	kaththikaLaal	kaLaal	N-PL-C3
18	thaRikaLil, viitukaLil	kaLil	N-PL -C3/7
19	viitukaLukku	kaLukku	N-PL-C4
20	viitukaLukkaaka	kaLukkaaka	N-PL-C4-P
21	viitukaLilirunthu	kaLilirunthu	N-PL-C5
22	avarkaLutaiya	kaLutaiya	N-PL-C6
23	viitukaLiNmeel	kaLiNmeelee	N-PL-C6-P
24	avarkaLutaN	kaLutaN	N-SG/PL-P
25	naaLkaLaaka	kaLaaka	N-PL-P
26	naaN, avaN	kaL	PN
27	athu, ithu	kaL	DPN
28	yaar, eNNa	0	IPN
29	eN, em, enkkaL, nam, uN, unkkaL, um	0	OPN

Table 6.31: WMTC Verb Tag List

Verb Tag List		
1	cari, thaNi	V-F-D
2	illai, uNntu	V-F-A
3	ceythaaN	V F
4	Ceyya	V-IN
5	Ceythu	V-VP
6	Ceytha	RP
7	Ceyyaatha	RP-NEG
8	Ceyyaa	RP-DN
9	Ceythaal	VP-COND
10	Ceythathum	VP-CONS
11	ceykaiyil-	VP-SIM
12	Ceyyaamal	VP-NEG
13	Ceythaalum	VP-CONC
14	Patiththal	VN
15	Patippathu	VN
16	patiththavaN	PAR-N
17	nallavaN	AN

Table 6.32: Adjectival and Adverbial Tag List

Adjectival and Adverbial Tag List	
mika, nirampa	INT
nalla, azakaaNa	ADJ
mella,veekamaaka, veekamaay	ADV
allathu, aakavee, eNavee, eeNeNRaal	SADV
ayyoo, ammaa	IJ

For noun inflection, there are 5 suffix types and for verb inflection, there are 50 suffix types.

Table 6.33: Noun Suffix List

Noun Suffix List		
1	kaL	PL
2	veeRRumai	C
3	peyarppiNNottu	N-P
4	muNNottu-ema	PF-NEG
5	muNNottu	PF

Table 6.34: Verb Suffix List

Verb Suffix List		
1	kaalam – nikaz	TNS-PR
2	kaalam – iRappu	TNS-PT
3	kaalam – ethir	TNS-FU
4	ethirmaRai	NEG
5	thie-thao	PN(F -SG)
6	thie-thapa	PN(F -PL)
7	thie-muo	PN(S-SG)
8	thie-muo/mupa	PN(S-SG/PL)
9	thiepaa(aa-o)	PNG(T-SG-M)
10	thiepaa(pe-o)	PNG(T-SG-F)
11	thiepaa(pothu-o)	PNG(T-SG-Co)
12	thiepaa(aahRi-o)	PNG(T-SG-NH)
13	thiepaa(palar)	PNG(T-PL)
14	thiepaa(aahRi-pa)	PNG(T-PL-NH)
15	Peyareccavikuthi	RPS
16	peyareccavikuthi -ema	RPS-NEG
17	peyareccavikuthi-iike	RPS-DNEG
18	viNaiyeccavikuthi-ceya	VP-INF
19	viNaiyeccavikuthi-ceythu	VP-CONJ
20	viNaiyeccavikuthi-nipa	VP-COND
21	viNaiyeccavikuthi-utaNnikazvu	VP-CONS
22	viNaiyeccavikuthi-thotarcci	VP-SIMU
23	viNaiyeccavikuthi-ethirmaRai	VP-NEG
24	Vikuu	VA

25	Vinoo	VM
26	Vipaa	VV
27	Mio	CL
28	viNaiyati	VS
29	viNaippiNNottu	V-PAR
30	iNnaippuvikuthi	CONJ
31	Viyappu	INTJ
32	viyankkooL	OPT
33	thotarviNaiyatai	SADV
34	Iyakkuvikuthi	CS
35	viNaiyaalaNnaiyumpeyar vikuthi	PAR-N(SG-M)
36	viNaiyaalaNnaiyumpeyar vikuthi	PAR-N(SG-F)
37	viNaiyaalaNnaiyumpeyar vikuthi	PAR-N(SG-C)
38	viNaiyaalaNnaiyumpeyar vikuthi	PAR-N(SG-NH)
39	viNaiyaalaNnaiyumpeyar vikuthi	PAR-N(PL-C)
40	viNaiyaalaNnaiyumpeyar vikuthi	PAR-N(PL-NH)
41	peyarataippeyar vikuthi	ADJ-N(SG-M)
42	peyarataippeyar vikuthi	ADJ-N(SG-F)
43	peyarataippeyar vikuthi	ADJ-N(SG-C)
44	peyarataippeyar vikuthi	ADJ-N(SG-NH)
45	peyarataippeyar vikuthi	ADJ-N(PL-C)
46	peyarataippeyar vikuthi	ADJ-N(PL-NH)
47	Thozilpeyarvikuthi	VN
48	Thozilpeyarvikuthi	VN-PR
49	Thozilpeyarvikuthi	VN-PT
50	Thozilpeyarvikuthi	VN-FU

Adjectives could be inflected for two major POS: one is, predicative adjective and verbal participle. Example: Predicative Adjective: “nalla-avaN” ‘good-he’. These predicative adjective suffixes show the Number (singular, plural) and Gender (Masculine, Feminine, Epicene and Neuter). The Person feature (1st, 2nd and 3rd) is not marked here.

The relative participle which is formed from verbs could also take the above suffixes to change into participial nouns, because they resemble adjectives in their function. Adverbs could take only clitics which is mostly semantic in nature.

Mostly the POS of Tamil word forms are determined by their last suffix (except clitics). This is due to the nature of Tamil, which is a Head-last one. For example, a verb root could take tense suffixes and PNG marker. According to the verb morphotactics, the verb is first inflected for tense, then followed by PNG to become a finite verb.

Example: “pati-thth-aaN” ‘study - past tense - PNG’.

Here, the finiteness characteristics are provided to the verb by the PNG suffix, which occurs finally next to the past tense suffix. Then only it is tagged as a finite verb. If there is no PNG suffix, then the verb could only become a non-finite verb which has many sub-types. There is one exemption here to be noticed. Some negative finite verb forms in Tamil could occur without the presence of PNG suffix. So, it is called as a negative defective finite verb.

Example: “patikk -aa” ‘study -won’t’. Here it is be mentioned that the tense suffix and negative suffix are complementary in nature. That is, with finite verbs, either tense suffixes may come or negative suffix. The negative suffix occupies the tense suffix slot here. The above negative finite verb resembles the negative defective relative participle. This leads to ambiguity problems that can be solved only with the concordance program.

Examples:

(1) “athu vaaraa” ‘It won’t come” - Finite verb

(2) “vaaraa paiyaN” the boy who does not come” - Relative participle

All the above knowledge of Tamil inflectional suffixes was provided to the POS tagger program. So, once a word form is segmented into their segments - morphs, the POS tagger could tag it for its legitimate POS.

6.13 Conclusion

The previous chapter reported on the technological process of developing a morphological parser and POS tagger. This chapter has discussed the algorithm of the computational morphological features found in the WMTC, that is, the computational linguistic rules that underlie the parser and the POS tagger, especially the inflectional morphology of Tamil. The discussion resulted in a noun and verb flow chart representing the inflectional processes of nouns and verbs taking place in the generation of word forms or types. Now that it is reported on the constructed corpus, the corpus tool developed, and the linguistic rules underlying the corpus tool, the third research question: How might a suitable algorithm be designed for developing the morphological parser and the POS tagger? has been addressed in this chapter.

CHAPTER 7

DISCUSSION

7.1 Introduction

In this chapter, responses and solutions to the issues involved in the construction of MWTC, development of the morphological parser, the POS tagger and the design of the algorithm required for the development of the parser and tagger are discussed. The discussion suggests how in the process of the creation and development of the corpus, the morphological parser and the POS tagger, some new linguistic features which were not documented in past research were uncovered.

7.2 Corpus construction - Issues and solutions for WMTC

With respect to the present project for the development of WMTC, the thesis has followed some corpus linguistics features, as follows:

7.2.1 Representative Corpus

Since the present corpus is only for the written Malaysian Tamil, published Tamil materials available in Malaysia both in printed media as well as the electronic media were searched. Care was taken in this respect. Some of the journal publishers, in printed media domain, had discontinued temporarily or stopped their publications permanently, for various reasons. Attempts were made to access the relevant materials by meeting the publishers and obtaining copies of the old journals.

Regarding the works of creative literature and literary criticism, many libraries, especially Universiti Malaya Tamil Library were visited and with the help of the staff there the scanned copies of them were obtained. Also, many Tamil writers, literary critics and publishers were personally approached to get their publications. For school Tamil textbooks, many teachers and students helped much. All the above efforts helped to contribute to the successful construction of WMTC.

7.2.2 Balanced Corpus

After the collection of all the needed materials, some considerations in relation to proportion of materials from various domains were made for their inclusion in the present corpus. Various aspects - domains, number of readers, importance of the publications in the society - were considered to decide their proportion in the corpus. As far as possible, neither the personnel preference nor the subjective opinion was allowed in deciding the selection of the materials or their proportion in the corpus.

7.2.3 Selection of samples in context

To be fed as an input to the automatic parser and POS tagger, the constructed corpus data were pre-processed to get an error-free and normalized one according to the Tamil grammar. It is to be mentioned here that though the object of analysis is Tamil words, they were not collected as individual words, but words as they occur in actual sentences or utterances as part of coherent texts. That is, the aspect of "cohesiveness" among the collected sentences in a sample was given due importance. Here, 'text' means not only individual sentence, but sentences in combination - a coherent text (Widdowson, 2007). This language piece is called here as a "sample". For example, to get all the properties of Tamil cohesive words or devices such as "aakavee" / "athaaNaalthaaN" 'so / therefore / that is why', "eeNeNRaal" 'because', "eNRaalum" 'even though', not only the properties of these words and not only the sentence in which they occur, but also the previous or following sentences are also important.

Examples:

1) “neeRRu mazai peythathu. aakavee naaN paLLikkuc cellavillai”
   ‘Yesterday it rained. So, I didn’t go to school’
2) “neeRRu naaN paLLikkuc cellavillai. eeNeNRaal mazai peythathu”
   ‘I didn’t go to school yesterday. Because it rained’
3) “neeRRu mazai peythathu. eNRaalum naaN paLLikkuc ceNReeN”
   ‘Eventhough, it rained yesterday, I went to school’.

That is, every word or token in their linguistic context - syntactic and textual contexts - was analyzed. This really helped the researcher to understand the lexical and inflectional properties of every word, decided by both the lexical and the morpho-syntactic properties. These morpho-syntactic properties in fact attribute the necessary inflectional properties to every word or word-forms.

In some coherent texts, there may not be explicit or surface level cohesive devices to select the coherent texts. For example, in some texts, there may not be surface level antecedent for the anaphora (“avaN” ‘he’; “avaL” ‘she’ etc.,) resolution to be carried out in getting the cohesive sample passage. By reading and comprehending the passage only, the samples could be selected as a coherent one. Since at present there is no such solution-given devices readly available to sort out these problems, the entire materials had to be manually gone through to get the necessary coherent sample passages.

7.3 Specific Issues and solutions for WMTC development

From data collection to getting final sample texts for the present project, there were many general problems and some specific problems. These issues were mentioned in chapter 4 and their solutions are discussed here.

7.3.1 Scanning and OCR problem

The data collection has several hurdles like cost, copyright, non-availability of old documents in digitized form, keying large texts and issues in scanning. Among these, in scanning the printed materials and getting the final electronic text samples, many problems were faced. There was no fool-proof OCR (Optical Character Recognizer) software for Tamil, it was much difficult to convert the image files scanned with the scanner into readable Tamil texts.

Among the OCR software available for Tamil, the OCR software provided by the Google was better for this task. Hence this OCR software was used for the present project. However, after this conversion into texts with the help of Google software, the manual verification of the texts had to be done, of course along with the Tamil spellchecker tools such as Mentamizh Word processor.

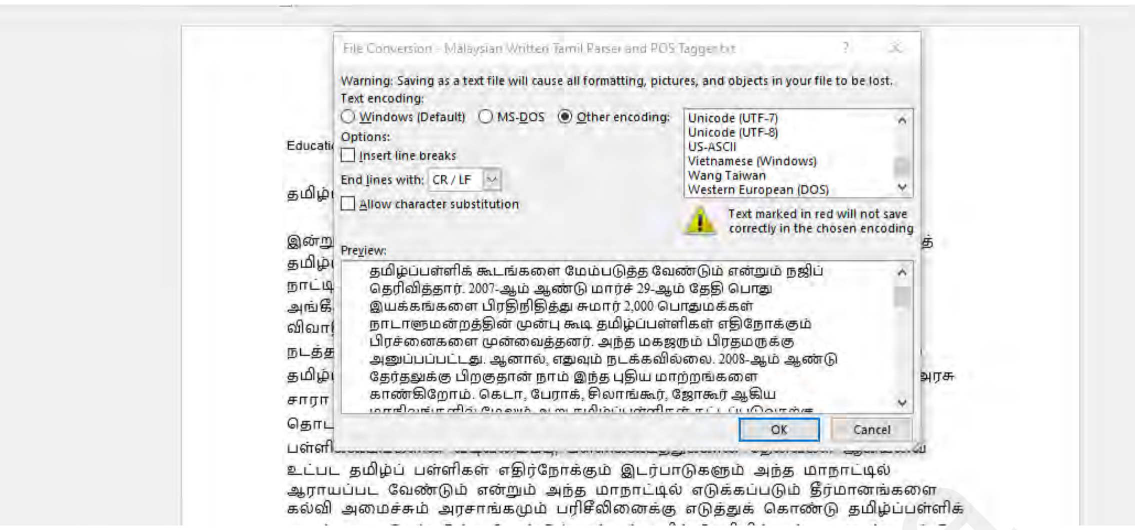
7.3.2 Overlapping of Texts over graphic images

Another important issue in the above process was the placement of some texts over the graphic images. In many instances the graphics were intermingled with the texts. Due to this, the contents of the graphics contained in the texts are not scannable. But, here, the Google OCR could not help. It required re-work on the text by manually keying the text. The texts had to be separated from these graphic images. It is to be mentioned here that especially in the textbooks a quite number of graphic images were intermixed with the texts. Also the high resolution smart phone was very much helpful.

7.3.3 Encoding problem

There were some encoding problems faced in this sample selection task. Before the year 2000, the encodings followed for Tamil texts were not uniform. Different propriety encoding formats were used by different software companies. Some of them were: TISCII, TAB, TAM. These all were ASCII based ones. Around 2000, the new Unicode encoding was introduced. Now it is the only encoding almost used for Tamil electronic texts. But even after this, some people used to adopt the old non-unicode encoding. The corpus developed for the present project is Unicode. All these non-unicode texts were converted into Unicode texts.

Even with the Unicode, there was some inconsistency with some characters such as ‘ko’(கொ) and ‘koo’(கோ). These characters could be represented either as two characters combination (‘k + o’ “க்+ஒ” or ‘k + oo’ “க்+ஓ”) and as three characters (‘k + e +aa’ “க்+எ+ஆ” or ‘k+ ee + aa’ “க்+ஏ+ஆ”) combination. The present morphological parser initially adapted the two-character encoding for the above Tamil graphemes. So, whenever it faced the texts which adapted three character encoding, it failed to recognize them. To come out from this problem, the parser was modified to accommodate the three character encoding also for the above Tamil characters.



File Conversion - Malaysian Written Tamil Parser and POS Tagger.txt

Warning: Saving as a text file will cause all formatting, pictures, and objects in your file to be lost.

Text encoding:

○ Windows (Default) ○ MS-DOS ● Other encoding: Unicode (UTF-7) / Unicode (UTF-8) / US-ASCII / Vietnamese (Windows) / Wang Taiwan / Western European (DOS)

Options:

□ Insert line breaks

End lines with: CR / LF

□ Allow character substitution

Text marked in red will not save correctly in the chosen encoding

Preview:

தமிழ்ப்பள்ளிக் கூடங்களை மேம்படுத்த வேண்டும் என்றும் நஜிப் தெரிவித்தார். 2007-ஆம் ஆண்டு மார்ச் 29-ஆம் தேதி பொது இயக்கங்களை பிரதிநிதித்து சுமார் 2,000 பொதுமக்கள் நாடாளுமன்றத்தின் முன்பு கூடி தமிழ்ப்பள்ளிகள் எதிநோக்கும் பிரச்னைகளை முன்வைத்தனர். அந்த மகஜரும் பிரதமருக்கு அனுப்பப்பட்டது. ஆனால், எதுவும் நடக்கவில்லை. 2008-ஆம் ஆண்டு தேர்தலுக்கு பிறகுதான் நாம் இந்த புதிய மாற்றங்களை காண்கிறோம். கெடா, பேராக், சிலாங்கூர், ஜோகூர் ஆகிய

உட்பட தமிழ்ப் பள்ளிகள் எதிர்நோக்கும் இடர்பாடுகளும் அந்த மாநாட்டில் ஆராயப்பட வேண்டும் என்றும் அந்த மாநாட்டில் எடுக்கப்படும் தீர்மானங்களை கல்வி அமைச்சும் அரசாங்கமும் பரிசீலினைக்கு எடுத்துக் கொண்டு தமிழ்ப்பள்ளிக்

Figure 7.1: Unicode (UTF-8)

7.3.4 File format problem

Regarding file format, the present project used plain text file format to store the sample texts for further analysis. But, many of the original text materials were in different file formats such as .doc,.docs. These propriety format commands in these file formats had to be removed to get the Plain Text File format uniformly.

7.3.5 Spoken variety mixing problem

Sometimes spoken variety words that intervened had to be removed. This resulted in non-coherence of the texts or reduction of sample size, which had to be managed.

7.3.6 Storage media problem

Sometimes CDs may be available that had to be transferred which required much time and care. While copying the CDs, since some of them were old and damaged, many problems were encountered. The e-mails and blogs were so personal and due permission from the concerned persons were obtained with much difficulty.

7.3.7 Issues from Web sources

The web sources are not that much straight forward. Sometimes the needed text materials had to be searched in a circular way and they were in different formats. However, web documents were removed from their original contexts. For example, multimedia texts are stripped of their multiple contents and reduced to plain word documents (txt.doc).

7.3.8 Issues in tokenisation and type selection tasks

7.3.8.1 Normalisation problem

After getting the cleaned samples to be included in the corpus, the next problem faced was a linguistic issue related to tokenisation. In Tamil, the grammatical affixes are in two forms: one is purely bound forms such as plural suffix, case suffix, tense suffix etc. The second one is the words grammaticalised from individual lexicon.

The problem here is, even after their grammaticalisation, they function as lexicons also. But wherever they do grammatical function they should be fixed - that is, as bound morphemes. However, it is found that in their grammatical function also, they are mistakenly written as free forms. Hence there is a need for normalisation of the texts. Otherwise, it would create some problem in the process of tokenisation. This problem was discussed in detail in the chapter 5. This problem was solved with the help of morphological parser and concordance program, accompanied by manual work.

7.3.8.2 Homonymy problem

The existence of homonyms in Tamil also gave problem in type selection from the tokens. For example, as explained in detail in earlier chapter, the token “malar” ‘flower’/ ‘to blossom’ represents for two different words: one is a noun, the other one is a verb. That is, they are two different “types”. Just by taking the outward forms, type selection should not be done. Along with their grammatical categories the type selection should be proceeded. The type selection was carefully carried out by taking into account not only the outward forms but also the grammatical categorization. Only after POS tagging, duplication if any, can be removed.

7.3.8.3 Ambiguity problem

The homophonous forms of some grammatical suffixes also must be disambiguated in the process of grammatical - POS - tagging. For example, the orthographic form of the suffix -அது “-athu” may stand for either PNG (அது வருகிறது “athu varukiRathu” ‘it comes’), or verbal noun suffix (அது வருகிறது எனக்குத் தெரிகிறது “athu varukiRathu eNakkuth therikiRathu” ‘I could see its coming’) or participial noun (அங்கே வருகிறது எது? “ankkee varukiRathu ethu?” ‘Which is coming there?’) in Tamil. Only the linguistic contexts could help to solve this type problems, with the help from concordance program.

7.4 Issues in the development of morphological parser and POS tagger and solutions

In the morphological parsing and POS tagging, the following issues encountered, and the solutions provided in the present project:

7.4.1 Emergence of new linguistic features

New linguistic features emerged in written Malaysian Tamil, which were not available in the earlier stage of written Tamil nor discussed in the grammar works available. They are two kinds: one is, new lexicons; the other one is, new grammatical features.

The significant outcome of the implementation of the present corpus analysis is that some important aspects are observed which missed the attention of other grammarians due to the non-availability of a well constructed corpus - some regarding the postpositions, auxiliaries and Sandhi (morphophonomics rules). Some of the instances are given below:

The Tamil linguistics scholars have recognized the postpositions emerged in the modern Tamil. However, there are some problems with these postpositions which were not answered by them. For example, though the postposition '-paRRi' was recognized but some of which modified forms such as '-paRRiya', '-paRRiyathu' were not discussed.

"avaNaippaRRi peeciNeeN"	'I spoke about him'
"avaNaippaRRiya peeccu"	'The talk about him'
"peeccu avanaippaRRiyathu"	'The talk is about him'

The various inflectional properties are added to the root lexicons such as noun and verb; that is, only the root lexicons are inflected. No inflectional suffix is inflected for some other inflectional properties. In the above case, though '-paRRi' is the grammatical suffix - postposition - it is inflected for adjectives or participial nouns. The reason may be this postposition is a grammaticalised one from the verbal participle of 'paRRu' (hold). Because of this relation with the lexicon 'paRRu', the adjectival form 'paRRiya' and the participial noun 'paRRiyathu' have emerged.

That is, even after the grammaticalisation, this 'paRRiya' is inflected for adjectival and participial properties. Here the problem is, in these contexts, whether the forms 'paRRiya' and 'paRRiyathu' are the inflected forms of the postposition 'paRRi' or the adjectival and participial noun forms derived from the root lexicon 'paRRu'. In the present thesis, they are considered as separate and independent postpositions.

Likewise, there is another problem found in the auxiliary verbs. Between the main verbs and the auxiliary verbs, usually no other words could be inserted. For example, in the verb phrase "pookap paarththaaN", "pooka" is the infinitive form of the root verb "poo" 'go' and "paar" 'attempt' is a modal auxiliary verb. Here, no other word can be inserted between them. This is a test to identify this "paar" is an auxiliary verb and to identify the lexical word "paar" in the phrase "eNNaip paarththaaN" 'He saw me'. In the latter one, some words may occur between "eNNai" and "paarththaaN" in "eNNai iNRu viittil paarththaaN" 'He saw me in the house today'. However, with some modal auxiliaries such as "veeNntum" in the finite verb "vara veeNntum" 'should come', some clitics may be inserted between the infinitive form of the main verb "vaa" 'come' and the modal auxiliary "veeNntum" 'should' as shown below:

“avaN varath**thaaN**veeNntum” ‘He should come’. Here “-thaaN-” is the clitic.

“avaN varav**ee**veeNntum” ‘He should come’. Here “-ee-” is the clitic.

“avaN vara**mattumee**veNntum” ‘He should come (only)’. Here “-mattum-” and “-ee-” are the clitics.

The above noted behaviour could be seen in some aspectual auxiliary verbs also.

“avaN vanthu**koNntiru**kkaveeNntum” ‘He should have been coming’. Here “-koNntiru-” is the aspectual auxiliary.

“avaN vanthukoNntu**thaaN**irukkaveeNntum” ‘He should have been coming’. Here “-thaaN-” is a clitic that is inserted inside the single aspectual “-koNntiru”.

The above new features regarding aspectual and modal auxiliaries are observed in the constructed corpus. Another problem is with the ‘caariyai’ (inflectional increment).

The grammatical rule is that when a noun ends with “-am” and it is followed by a case suffix, a ‘caariyai’ “aththu” should occur between them.

“avaN maraththaip paarththaaN” ‘He saw the tree’

In the above example, the phrase “maraththai” has three parts as follows:

“maram + aththu + ai” ‘tree + inflectional increment + accusative case marker’.

“maram” is a noun ends with “-am”; “-ai” is an accusative case suffix. Hence the inflectional increment “-aththu-” occurs between them. This takes place as per the standard grammatical rule. But the following example shows some new features; that is, before some non-case suffixes also the inflectional increment ‘aththu’ occurs.

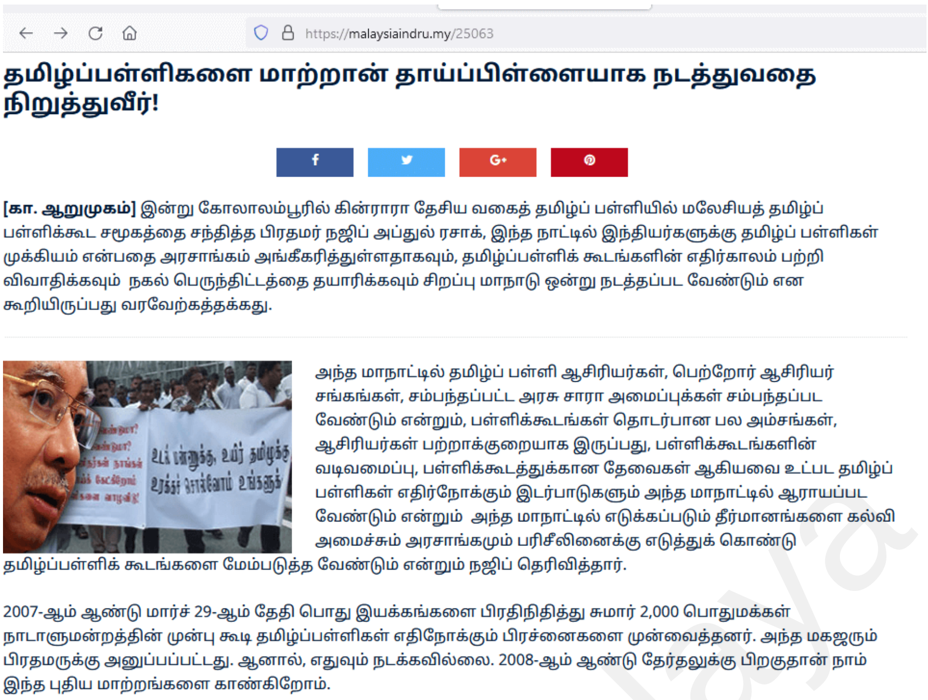
“maNRaththaar”	‘association members’
“cuRRaththaar”	‘relatives’
“kutumpaththaar”	‘family members’

In the above example, after the main lexical nouns end with “-am”, the derivational suffix “-aar” ‘people’ occurs which is not a case suffix; even then the inflectional increment occurs.

Thus, the present morphological parser developed for this project included all the above grammatical features which are not explained or given in the available Tamil grammars.

7.4.2 Issue of Named Entity Recognition

Another important issue is with the proper nouns and some common nouns. Identification of these words as proper and common nouns by the morphological parser is not possible. Hence all the proper and common nouns used in WMTC were added in the lexicon with their inherent grammatical properties. Especially Malay, Chinese and English proper and common names were added in the lexicon. To mention a few ‘Najib, Kinrara, Johore’ which are shown below:



https://malaysiaindru.my/25063

தமிழ்ப்பள்ளிகளை மாற்றான் தாய்ப்பிள்ளையாக நடத்துவதை நிறுத்துவீர்!

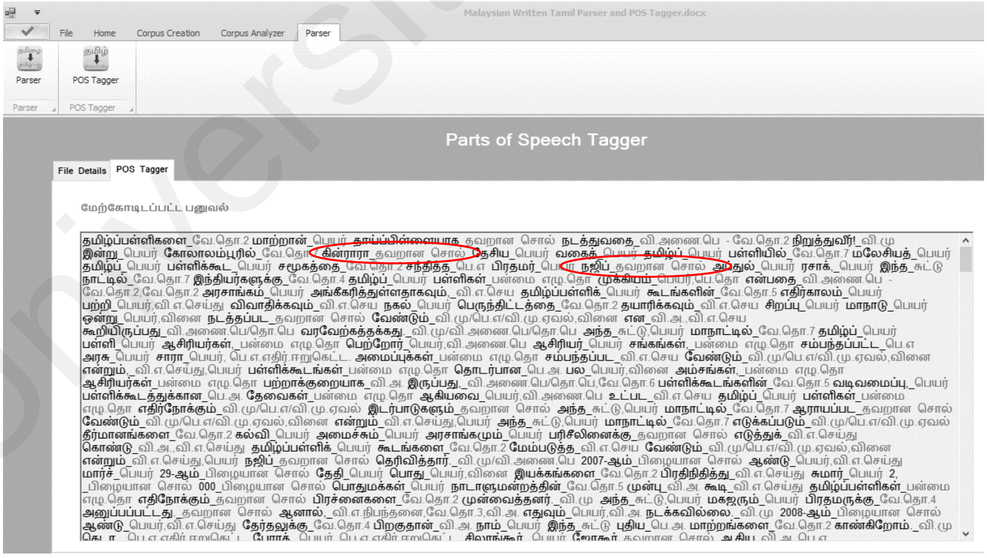
[கா. ஆறுமுகம்] இன்று கோலாலம்பூரில் கின்ராரா தேசிய வகைத் தமிழ்ப் பள்ளியில் மலேசியத் தமிழ்ப் பள்ளிக்கூட சமூகத்தை சந்தித்த பிரதமர் நஜிப் அப்துல் ரசாக், இந்த நாட்டில் இந்தியர்களுக்கு தமிழ்ப் பள்ளிகள் முக்கியம் என்பதை அரசாங்கம் அங்கீகரித்துள்ளதாகவும், தமிழ்ப்பள்ளிக் கூடங்களின் எதிர்காலம் பற்றி விவாதிக்கவும் நகல் பெருந்திட்டத்தை தயாரிக்கவும் சிறப்பு மாநாடு ஒன்று நடத்தப்பட வேண்டும் என கூறியிருப்பது வரவேற்கத்தக்கது.

அந்த மாநாட்டில் தமிழ்ப் பள்ளி ஆசிரியர்கள், பெற்றோர் ஆசிரியர் சங்கங்கள், சம்பந்தப்பட்ட அரசு சாரா அமைப்புக்கள் சம்பந்தப்பட வேண்டும் என்றும், பள்ளிக்கூடங்கள் தொடர்பான பல அம்சங்கள், ஆசிரியர்கள் பற்றாக்குறையாக இருப்பது, பள்ளிக்கூடங்களின் வடிவமைப்பு, பள்ளிக்கூடத்துக்கான தேவைகள் ஆகியவை உட்பட தமிழ்ப் பள்ளிகள் எதிர்நோக்கும் இடர்பாடுகளும் அந்த மாநாட்டில் ஆராயப்பட வேண்டும் என்றும் அந்த மாநாட்டில் எடுக்கப்படும் தீர்மானங்களை கல்வி அமைச்சும் அரசாங்கமும் பரிசீலினைக்கு எடுத்துக் கொண்டு தமிழ்ப்பள்ளிக் கூடங்களை மேம்படுத்த வேண்டும் என்றும் நஜிப் தெரிவித்தார்.

2007-ஆம் ஆண்டு மார்ச் 29-ஆம் தேதி பொது இயக்கங்களை பிரதிநிதித்து சுமார் 2,000 பொதுமக்கள் நாடாளுமன்றத்தின் முன்பு கூடி தமிழ்ப்பள்ளிகள் எதிநோக்கும் பிரச்னைகளை முன்வைத்தனர். அந்த மகஜரும் பிரதமருக்கு அனுப்பப்பட்டது. ஆனால், எதுவும் நடக்கவில்லை. 2008-ஆம் ஆண்டு தேர்தலுக்கு பிறகுதான் நாம் இந்த புதிய மாற்றங்களை காண்கிறோம்.

Figure 7.2: Electronic sample text from the Malaysia indru web news portal

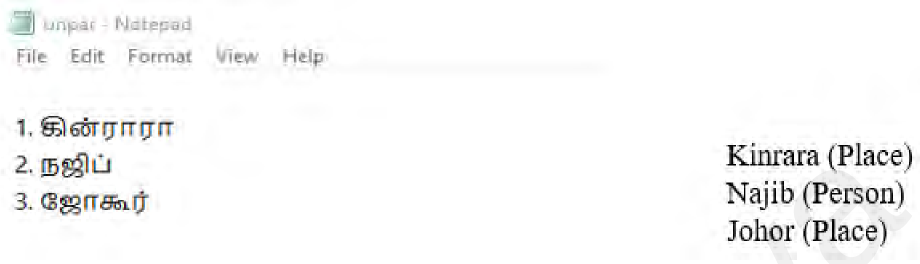
The above sample text is fed into moerphological parser and tagger and the output is shown below.



```
தமிழ்ப்பள்ளிகளை_வே.தொ.2 மாற்றான்_பெயர் தாய்ப்பிள்ளையாக_தவறான சொல் நடத்துவதை_வி.அணை.பெ - வே.தொ.2 நிறுத்துவீர்!_வி.மு
இன்று_பெயர் கோலாலம்பூரில்_வே.தொ.7 கின்ராரா_தவறான சொல் தேசிய_பெயர் வகைத்_பெயர் தமிழ்ப்_பெயர் பள்ளியில்_வே.தொ.7 மலேசியத்_பெயர்
தமிழ்ப்_பெயர் பள்ளிக்கூட_பெயர் சமூகத்தை_வே.தொ.2 சந்தித்த_பெ.எ பிரதமர்_பெயர் நஜிப்_தவறான சொல் அப்துல்_பெயர் ரசாக்,_பெயர் இந்த_சுட்டு
நாட்டில்_வே.தொ.7 இந்தியர்களுக்கு_வே.தொ.4 தமிழ்ப்_பெயர் பள்ளிகள்_பன்மை எழு.தொ முக்கியம்_பெயர்,பெ.தொ என்பதை_வி.அணை.பெ -
வே.தொ.2,வே.தொ.2 அரசாங்கம்_பெயர் அங்கீகரித்துள்ளதாகவும்,_வி.எ.செய தமிழ்ப்பள்ளிக்_பெயர் கூடங்களின்_வே.தொ.5 எதிர்காலம்_பெயர்
பற்றி_பெயர்,வி.எ.செய்து விவாதிக்கவும்_வி.எ.செய நகல்_பெயர் பெருந்திட்டத்தை_வே.தொ.2 தயாரிக்கவும்_வி.எ.செய சிறப்பு_பெயர் மாநாடு_பெயர்
ஒன்று_பெயர்,வினை நடத்தப்பட_தவறான சொல் வேண்டும்_வி.மு/பெ.எ/வி.மு.ஏவல்,வினை என_வி.அ.,வி.எ.செய
கூறியிருப்பது_வி.அணை.பெ/தொ.பெ வரவேற்கத்தக்கது._வி.மு/வி.அணை.பெ/தொ.பெ அந்த_சுட்டு,பெயர் மாநாட்டில்_வே.தொ.7 தமிழ்ப்_பெயர்
பள்ளி_பெயர் ஆசிரியர்கள்,_பன்மை எழு.தொ பெற்றோர்_பெயர்,வி.அணை.பெ ஆசிரியர்_பெயர் சங்கங்கள்,_பன்மை எழு.தொ சம்பந்தப்பட்ட_பெ.எ
அரசு_பெயர் சாரா_பெயர், பெ.எ.எதிர்.ஈறுகெட்ட. அமைப்புக்கள்_பன்மை எழு.தொ சம்பந்தப்பட_வி.எ.செய வேண்டும்_வி.மு/பெ.எ/வி.மு.ஏவல்,வினை
என்றும்,_வி.எ.செய்து,பெயர் பள்ளிக்கூடங்கள்_பன்மை எழு.தொ தொடர்பான_பெ.அ. பல_பெயர்,வினை அம்சங்கள்,_பன்மை எழு.தொ
ஆசிரியர்கள்_பன்மை எழு.தொ பற்றாக்குறையாக_வி.அ. இருப்பது,_வி.அணை.பெ/தொ.பெ,வே.தொ.6 பள்ளிக்கூடங்களின்_வே.தொ.5 வடிவமைப்பு,_பெயர்
பள்ளிக்கூடத்துக்கான_பெ.அ. தேவைகள்_பன்மை எழு.தொ ஆகியவை_பெயர்,வி.அணை.பெ உட்பட_வி.எ.செய தமிழ்ப்_பெயர் பள்ளிகள்_பன்மை
எழு.தொ எதிர்நோக்கும்_வி.மு/பெ.எ/வி.மு.ஏவல் இடர்பாடுகளும்_தவறான சொல் அந்த_சுட்டு,பெயர் மாநாட்டில்_வே.தொ.7 ஆராயப்பட_தவறான சொல்
வேண்டும்_வி.மு/பெ.எ/வி.மு.ஏவல்,வினை என்றும்_வி.எ.செய்து,பெயர் அந்த_சுட்டு,பெயர் மாநாட்டில்_வே.தொ.7 எடுக்கப்படும்_வி.மு/பெ.எ/வி.மு.ஏவல்
தீர்மானங்களை_வே.தொ.2 கல்வி_பெயர் அமைச்சும்_பெயர் அரசாங்கமும்_பெயர் பரிசீலினைக்கு_தவறான சொல் எடுத்துக்_வி.எ.செய்து
கொண்டு_வி.அ.,வி.எ.செய்து தமிழ்ப்பள்ளிக்_பெயர் கூடங்களை_வே.தொ.2 மேம்படுத்த_வி.எ.செய வேண்டும்_வி.மு/பெ.எ/வி.மு.ஏவல்,வினை
என்றும்_வி.எ.செய்து,பெயர் நஜிப்_தவறான சொல் தெரிவித்தார்._வி.மு/வி.அணை.பெ 2007-ஆம்_பிழையான சொல் ஆண்டு_பெயர்,வி.எ.செய்து
மார்ச்_பெயர் 29-ஆம்_பிழையான சொல் தேதி_பெயர் பொது_பெயர்,வினை இயக்கங்களை_வே.தொ.2 பிரதிநிதித்து_வி.எ.செய்து சுமார்_பெயர் 2,
_பிழையான சொல் 000_பிழையான சொல் பொதுமக்கள்_பெயர் நாடாளுமன்றத்தின்_வே.தொ.5 முன்பு_வி.அ. கூடி_வி.எ.செய்து தமிழ்ப்பள்ளிகள்_பன்மை
எழு.தொ எதிநோக்கும்_தவறான சொல் பிரச்னைகளை_வே.தொ.2 முன்வைத்தனர்._வி.மு அந்த_சுட்டு,பெயர் மகஜரும்_பெயர் பிரதமருக்கு_வே.தொ.4
அனுப்பப்பட்டது._தவறான சொல் ஆனால்,_வி.எ.நிபந்தனை,வே.தொ.3,வி.அ. எதுவும்_பெயர்,வி.அ. நடக்கவில்லை._வி.மு 2008-ஆம்_பிழையான சொல்
ஆண்டு_பெயர்,வி.எ.செய்து தேர்தலுக்கு_வே.தொ.4 பிறகுதான்_வி.அ. நாம்_பெயர் இந்த_சுட்டு புதிய_பெ.அ. மாற்றங்களை_வே.தொ.2 காண்கிறோம்._வி.மு
```

Figure 7.3: Initial output of the tagged sample from WMTC Parser

Here, some of the words (நஜிப், கின்ராரா, ஜோகூர்) denoting persons and places present in the sample text above could not be handled because of their absence in the initial Lexicon.



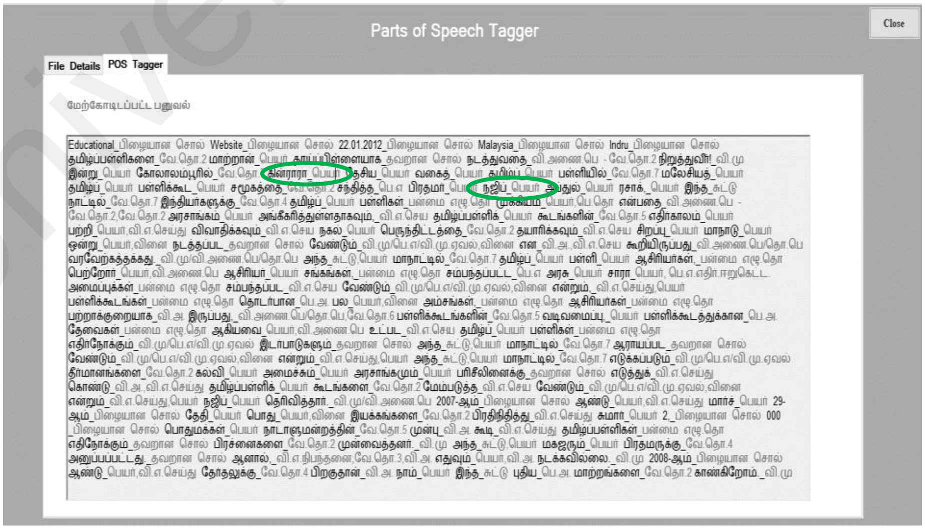
1. கின்ராரா
2. நஜிப்
3. ஜோகூர்

Kinrara (Place)
Najib (Person)
Johor (Place)

Figure 7.4 Unrecognized Words

Here they (Kinrara, Najib, Johor) were added in the final lexicon with their grammatical properties - human vs nonhuman, male vs female, singular vs plural. This issue with these words could be handled in the future when named Entity Recognizer software for Tamil will be available.

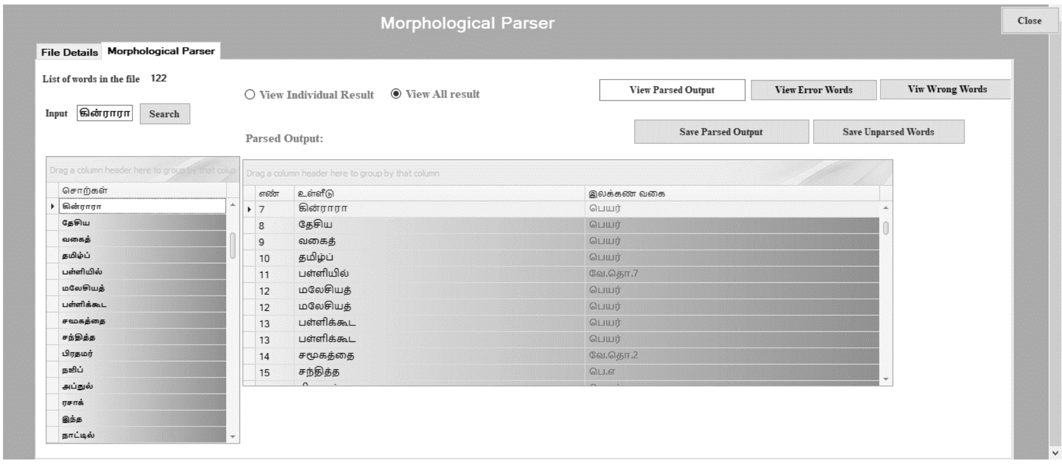
The output of the above sample from the enhanced WMTC morphological parser and tagger:



Educational_பிழையான சொல் Website_பிழையான சொல் 22.01.2012_பிழையான சொல் Malaysia_பிழையான சொல் Indru_பிழையான சொல்
தமிழ்ப்பள்ளிகளை_வே.தொ.2 மாற்றான்_பெயர் தாய்ப்பிள்ளையாக_தவறான சொல் நடத்துவதை_வி.அணை.பெ - வே.தொ.2 நிறுத்துவீர்!_வி.மு
இன்று_பெயர் கோலாலம்பூரில்_வே.தொ.7 கின்ராரா_பெயர் தேசிய_பெயர் வகைத்_பெயர் தமிழ்ப்_பெயர் பள்ளியில்_வே.தொ.7 மலேசியத்_பெயர்
தமிழ்ப்_பெயர் பள்ளிக்கூட_பெயர் சமூகத்தை_வே.தொ.2 சந்தித்த_பெ.எ பிரதமர்_பெயர் நஜிப்_பெயர் அப்துல்_பெயர் ரசாக்_பெயர் இந்த_சுட்டு
நாட்டில்_வே.தொ.7 இந்தியர்களுக்கு_வே.தொ.4 தமிழ்ப்_பெயர் பள்ளிகள்_பன்மை எழு.தொ முக்கியம்_பெயர்,பெ.தொ என்பதை_வி.அணை.பெ -
வே.தொ.2,வே.தொ.2 அரசாங்கம்_பெயர் அங்கீகரித்துள்ளதாகவும்,_வி.எ.செய தமிழ்ப்பள்ளிக்_பெயர் கூடங்களின்_வே.தொ.5 எதிர்காலம்_பெயர்
பற்றி_பெயர்,வி.எ.செய்து விவாதிக்கவும்_வி.எ.செய நகல்_பெயர் பெருந்திட்டத்தை_வே.தொ.2 தயாரிக்கவும்_வி.எ.செய சிறப்பு_பெயர் மாநாடு_பெயர்
ஒன்று_பெயர்,வினை நடத்தப்பட_தவறான சொல் வேண்டும்_வி.மு/பெ.எ/வி.மு.ஏவல்,வினை என_வி.அ.,வி.எ.செய கூறியிருப்பது_வி.அணை.பெ/தொ.பெ
வரவேற்கத்தக்கது._வி.மு/வி.அணை.பெ/தொ.பெ அந்த_சுட்டு,பெயர் மாநாட்டில்_வே.தொ.7 தமிழ்ப்_பெயர் பள்ளி_பெயர் ஆசிரியர்கள்,_பன்மை எழு.தொ
பெற்றோர்_பெயர்,வி.அணை.பெ ஆசிரியர்_பெயர் சங்கங்கள்,_பன்மை எழு.தொ சம்பந்தப்பட்ட_பெ.எ அரசு_பெயர் சாரா_பெயர், பெ.எ.எதிர்.ஈறுகெட்ட
அமைப்புக்கள்_பன்மை எழு.தொ சம்பந்தப்பட_வி.எ.செய வேண்டும்_வி.மு/பெ.எ/வி.மு.ஏவல்,வினை என்றும்,_வி.எ.செய்து,பெயர்
பள்ளிக்கூடங்கள்_பன்மை எழு.தொ தொடர்பான_பெ.அ. பல_பெயர்,வினை அம்சங்கள்,_பன்மை எழு.தொ ஆசிரியர்கள்_பன்மை எழு.தொ
பற்றாக்குறையாக_வி.அ. இருப்பது,_வி.அணை.பெ/தொ.பெ,வே.தொ.6 பள்ளிக்கூடங்களின்_வே.தொ.5 வடிவமைப்பு,_பெயர் பள்ளிக்கூடத்துக்கான_பெ.அ.
தேவைகள்_பன்மை எழு.தொ ஆகியவை_பெயர்,வி.அணை.பெ உட்பட_வி.எ.செய தமிழ்ப்_பெயர் பள்ளிகள்_பன்மை எழு.தொ
எதிர்நோக்கும்_வி.மு/பெ.எ/வி.மு.ஏவல் இடர்பாடுகளும்_தவறான சொல் அந்த_சுட்டு,பெயர் மாநாட்டில்_வே.தொ.7 ஆராயப்பட_தவறான சொல்
வேண்டும்_வி.மு/பெ.எ/வி.மு.ஏவல்,வினை என்றும்_வி.எ.செய்து,பெயர் அந்த_சுட்டு,பெயர் மாநாட்டில்_வே.தொ.7 எடுக்கப்படும்_வி.மு/பெ.எ/வி.மு.ஏவல்
தீர்மானங்களை_வே.தொ.2 கல்வி_பெயர் அமைச்சும்_பெயர் அரசாங்கமும்_பெயர் பரிசீலினைக்கு_தவறான சொல் எடுத்துக்_வி.எ.செய்து
கொண்டு_வி.அ.,வி.எ.செய்து தமிழ்ப்பள்ளிக்_பெயர் கூடங்களை_வே.தொ.2 மேம்படுத்த_வி.எ.செய வேண்டும்_வி.மு/பெ.எ/வி.மு.ஏவல்,வினை
என்றும்_வி.எ.செய்து,பெயர் நஜிப்_பெயர் தெரிவித்தார்._வி.மு/வி.அணை.பெ 2007-ஆம்_பிழையான சொல் ஆண்டு_பெயர்,வி.எ.செய்து மார்ச்_பெயர் 29-
ஆம்_பிழையான சொல் தேதி_பெயர் பொது_பெயர்,வினை இயக்கங்களை_வே.தொ.2 பிரதிநிதித்து_வி.எ.செய்து சுமார்_பெயர் 2,_பிழையான சொல் 000
_பிழையான சொல் பொதுமக்கள்_பெயர் நாடாளுமன்றத்தின்_வே.தொ.5 முன்பு_வி.அ. கூடி_வி.எ.செய்து தமிழ்ப்பள்ளிகள்_பன்மை எழு.தொ
எதிநோக்கும்_தவறான சொல் பிரச்னைகளை_வே.தொ.2 முன்வைத்தனர்._வி.மு அந்த_சுட்டு,பெயர் மகஜரும்_பெயர் பிரதமருக்கு_வே.தொ.4
அனுப்பப்பட்டது._தவறான சொல் ஆனால்,_வி.எ.நிபந்தனை,வே.தொ.3,வி.அ. எதுவும்_பெயர்,வி.அ. நடக்கவில்லை._வி.மு 2008-ஆம்_பிழையான சொல்
ஆண்டு_பெயர்,வி.எ.செய்து தேர்தலுக்கு_வே.தொ.4 பிறகுதான்_வி.அ. நாம்_பெயர் இந்த_சுட்டு புதிய_பெ.அ. மாற்றங்களை_வே.தொ.2 காண்கிறோம்._வி.மு

Figure 7.5 Added Lexical Words

These unparsed names were added to the lexicon of the enhanced parser as shown below:



List of words in the file 122

Input கின்ராரா Search

○ View Individual Result ● View All result

Parsed Output:

Drag a column header here to group by that colun

சொற்கள்
கின்ராரா
தேசிய
வகைத்
தமிழ்ப்
பள்ளியில்
மலேசியத்
பள்ளிக்கூட
சமூகத்தை
சந்தித்த
பிரதமர்
நஜிப்
அப்துல்
ரசாக்
இந்த
நாட்டில்

Drag a column header here to group by that column

எண்	உள்ளீடு	இலக்கண வகை
7	கின்ராரா	பெயர்
8	தேசிய	பெயர்
9	வகைத்	பெயர்
10	தமிழ்ப்	பெயர்
11	பள்ளியில்	வே.தொ.7
12	மலேசியத்	பெயர்
12	மலேசியத்	பெயர்
13	பள்ளிக்கூட	பெயர்
13	பள்ளிக்கூட	பெயர்
14	சமூகத்தை	வே.தொ.2
15	சந்தித்த	பெ.எ

Figure 7.6: Output of the enhanced WMTC morphological parser and tagger

7.5 Enhancement and finalization of the morphological parser and Tagger

The morphological parser initially developed was enhanced through various stages to arrive at the final satisfactory morphological parser. Again and again the corpus samples were put into the parser to identify the insufficiency in the following:

1. Lexicon
2. Grammatical suffixes
3. Morphotactics
4. Morphophonemics

7.5.1 Enhanced Parser

As explained in the earlier chapter, the morphological parser is a corpus based one. That is, the lexicon and inflection morphological aspects of WMTC were developed initially with the help of initial lexicon and grammar books available in market and then they were modified based on the constructed corpus.

The samples with finalized types were sent to the morphological parser. After this process, the word forms were parsed as either lexical roots or stems with the help of the enhanced morphological parser. That is, the samples were fed at regular intervals and the resultant residual issues were remedied by modifying the lexicon and suffixes lists, morphotactics and morphophonemic rules accordingly.

7.5.2 Issues in designing the algorithm for Tamil morphological structure

7.5.2.1 Noun Inflection

Lexically inherent features: Nouns are inflected for number and case. However, some nouns and pronouns show the 'ThiNnai' (Human - nonhuman) inherently through some orthographic forms.

Examples:

1)	"veelaikkaaraN"	'male servant'
	"veelaikkaari"	'female servant'
	"veelaikkaarar"	'servant (male or female)'
2)	"kuRavaN"	'male gypsy'
	"kuRaththi"	'female gypsy'
3)	"aaciriyaN"	'male teacher'
	"aaciriyai"	'female teacher'
	"aaciriyar"	'teacher (male or female)'
4)	"maaNnavaN"	'male student'
	"maaNnavi"	'female student'
	"maaNnavar"	'student (male or female)'

5)	"avaN"	'he'
	"avaL"	'she'
	"avar"	'he/she'

6)	"oruvaN"	'one (male)'
	"oruththi"	'one(female)'

In the above examples, the gender is expressed not by any separate segmentable suffix, but by the phonemic endings of the words; that is, the gender feature is inherent as part of the lexeme. In the Lexicon, this difference in gender should be accommodated. When these lexemes are tagged for POS, based on the lexicon only, the gender could be provided.

b) Quantifiers: There are some quantifier words in Tamil which express the plurality. This is inherent one.

Examples:

"elloorum"	'all (human)'
"ovvoruvarum"	'each one of (human)'
"aNaivarum"	'all (human)'
"ellaam"	'all (non-human)'
"ovvoNRum"	'each one of (non-human)'
"aNaiththum"	'all (non-human)'

When the above kinds of nouns are faced by the POS tagger, the lexicon comes to help. These grammatical features are inherent in these words. There is no segmentable suffix to be handled by the morphological parser.

c) Inherent Plurality: Some nouns such as “makkaL” ‘people’, “pala” ‘many’, “ellaarum” ‘all persons’, “ellaam” ‘all (non-human)’ are inherently plural. So, in tagging them for POS, the lexicon has to be referred.

d) Case: Regarding case inflection, there are some problems. Some case suffixes are optional and may be absent in the surface. That is, zero allomorphs have to be set for these kinds of cases.

Examples:

1) “ithu avar(utaiya) viitu” ‘This - his - house’ > ‘This is his house’
   Here, the possessive case suffix “-utaiya” is absent.
2) “naaN puththakam (thth- ai) patiththeeN” ‘I - book(acc.) - read-I’ > ‘I read the book’. Here, accusative case suffix “-ai” is absent.
3) “naaN paLLikkuutam(thth-ukku) ceNReeN” ‘I -school (dative) - went-I’ > ‘I went to the school’. Here, the dative case suffix “-ukku” is absent.

In (2) and (3) the “-thth-” is an increment (empty morph) which was already explained in the earlier chapter. In all the above examples, the case suffixes are absent. But by the linguistic context, they are understood. For this problem of suffix absence, the concordance program helped.

e) Postposition: Regarding some postpositions, there are some issues to categorise them. For example, the accusative postposition “paRRi” ‘about’ in “avaNaippaRRi” ‘about him’ is originally a verbal participle form of the verb lexicon “paRRu” ‘to hold’. But it is grammaticalised into a postposition in the above phrase to give the meaning ‘about’. In general, once a lexicon is grammaticalised, it won’t be inflected for any other inflectional property. But this postposition “paRRi” takes a relative participle form in some sentences.

1) “naaN avaraippaRRip peeciNeeN” ‘I - about him - talked’ > ‘I talked about him’

2) “avaraippaRRiya eNathu peeccu” ‘about him - my - talk’ > ‘My talk about him’

The postposition “paRRi” in (1) is changed into “paRRiya” which is a relative participle form. The problem here is, how to tag it for POS? Whether as a postposition or as a relative participle? In the present research project, it is also considered as a postposition on par with “paRRi” since they both express same meaning. The point to be noticed here is, since the “paRRi” is in the form of a verbal participle of the verb “paRRu”, it is followed by a verb - in this example by “peeciNeeN” ‘talked-I’. But “paRRiya” is in a relative participle form, it is followed by a noun “peeccu” ‘talk’. The test is whether they occur as bound morphemes or as free morphemes. In these cases, they occur as bound forms only - that is they follow the accusative suffix within the words. Like this, there are problems with other postpositions also. For example, the dative postposition “uriya” ‘belongs to’ could take the participial noun marker and change into “uriyathu”. Here, the same problem explained for the accusative postposition “paRRiya” exists.

This problem exists for the postpositions which are grammaticalised forms of the lexicons. The main reason is, after their grammaticalisation, those words exist as lexicons also along with the grammaticalised forms. By their property whether they are ‘free form’ or ‘bound form’, here, they are tagged as postpositions. Here also, the manual intervention using the concordance program helped to solve this type of issues.

f) Grammatical category and function: In some contexts, there are problems in deciding the correct POS category; whether it should be based on the grammatical category or on grammatical function.

1) “avar koopaththil peeciNaar”

2) “avar koopaththootu peeciNaar”

3) “avar koopaththutaN peeciNaar”

4) “avar koopamaaka peeciNaar”

5) “avar koopamaay peeciNaar”

All the above have same meaning “He spoke angrily. But there are five wordforms for ‘angrily’ in Tamil. The root word is “koopam”; but they differ in the suffix for adverb - (1) “il” (2) “ootu” (3) utaN” (4) “aaka” (5) “aay”. From (1) to (3), there are increments “thth” since the noun ends with the phoneme “m”.

The suffixes from (1) to (3) are in fact case suffixes whereas in (4) and (5) they are adverbial suffixes. So, if we tag them based on the suffixes, the wordforms for the meaning “angrily’ from (1) to (3) should be tagged as casal phrases and (4) and (5) should be tagged as adverbs. On the other hand, if we tag them based on function, they all should be tagged as simply adverbs. The present research project has adopted the latter option.

g) Homophonous forms and inflection: With some nouns, though their phonological forms are similar, the case inflection behaves differently. The word “aaRu” which has the meaning ‘river’ won’t take any case suffix without doubling the final consonant “-R-”. But the same phonological form “aaRu” which has the meaning “six”, the numeral, during the case inflection the final “-R-” would not be doubled.

1) “avar aaRRaik katanthaar” ‘He crossed the river’

2) “paththilirunthu aaRaik kazi” ‘Deduct six from ten’

The phonological form “aavathu”, as a suffix which could be added with nouns gives different meanings and does different functions.

1) “avaN muthalaavathu maaNnavaN” ‘He is the first student’

2) “avaNaavathu ceyyattum” ‘Let him atleast do it’

3) “avaNaavathu ceyvathaavathu” ‘(I don’t think) he will do it!’

Here, these different “aavathu” forms are three different kinds of suffixes. That is, they belong to different grammatical categories. There is one more “aavathu” which is not a suffix, but an inflected verb wordform.

4) “uNNaal aavathu oNRum illai” ‘There is nothing which you could do.’

The above discussion on “aavathu” emphasizes the importance of concordance program to identify the correct grammatical categorisation of this form which depends upon the discourse meaning of the sentences. This needs the human intervention.

7.5.2.2 Verb Inflection

As explained in the earlier chapter, verbs in Tamil are inflected for tense, PNG, participles, participial nouns and verbal nouns. With these, some problems are being faced.

a) Tense: In Tamil, there are three tenses: past, present and future. However, in some contexts, the present tense itself is used for future.

1) “avar iNRu varukiRaar” ‘He comes today”

2) (a) “avar naaLai varuvaar” ‘He will come tomorrow’

(b) “avar naaLai varukiRaar” ‘He will come tomorrow’

In (1) the verb “vaa” ‘come’ is inflected for present tense taking the due suffix “kiR”. In (2) (a), the verb is inflected for future tense, taking the future tense suffix “v”. But, in (2)(b), though this sentence expresses future activity, the present tense suffix “kiR” is used instead of future tense suffix “v”. So, only by the linguistic contexts such as some adverbs “iNRu” ‘today’, “naaLai” ‘tomorrow’, the verb should be interpreted whether they denote present or future activity.

“avar naaLai varukiRaar” ‘He comes tomorrow’

Likewise, both the present activity and the habitual activity of could be expressed by present tense marker:

1) “avar ippoothu patikkiRaar” (present activity)

2) “avar thiNanthoorum patikkiRaar” (habitual activity)

So, here also, the usage of adverbs and the concordance program help to disambiguate whether the present tense suffix is used for ‘present or ‘habitual’ activity.

Another problem is with the future tense suffix for non-human and relative participle forms.

1) “athu naaLaikku varum” ‘It will come tomorrow’

2) “avai naaLaikku varum” ‘They will come tomorrow’

3) “ippoothu varum paiyan” ‘the boy who comes now’

4) “naaLai varum paiyan” ‘the boy who will come tomorrow’

With (1) and (2), the finite verb form “varum” is the same. Subject - Verb concord is same for both singular and plural. It could be decided by the Subject noun or pronoun. With (3) and (4), the relative participle form (which has the same future finite verb form) is same for both present and future. Here, deciding whether this form is future finite verb or relative participle form is also a problem. These ambiguities could be solved only with the help from concordance program.

b) Simultaneous verbal participle and verbal noun:

1) “avar varukaiyil naaN paarththeeN” ‘I saw him when he was coming”

2) “avar varukaiyil enakku makizcci” ‘I am happy about his coming’.

In (1) “varukaiyil (varu-kaiyil)”, “kaiyil” is the verbal participle marker whereas in (2) “varukaiyil (varukai-(y)il)”, “kai” is part of the noun “varukai” ‘coming’ and “il” is the case suffix. That is, the different ways followed in segmentation or parsing of these words decide the exact POS tagging. This also could be done with the help of the concordance program.

c) Participial noun and relative participal

Already one problem involved in participial noun identification was explained in the earlier chapter.

1) "patiththavarai" "patiththavar-ai" 'the one who is educated (acc.)'
2) "patiththavarai" "patiththa-varai" 'upto one's reading'

The (1) is a participial noun and the (2) is relative participle ("patiththa" + verbal particle. Here, two ways segmentation. Only the concordance program could help in getting correct segmentation and POS tagging.

d) Temporal Vs Spatial difference in the suffix

With verbal participle categorization, there is one more problem because of the difference between temporal and spatial lexicons.

1) "naaN avar viittukkumuNpu vantheeN"	'I came in front of his house'
2) "naaN avar pecuvathaRkumuNpu vantheeN"	'I came before he spoke'

In (1), the case suffix + postposition "ukku + muNpu" occurs with a spatial noun - "viitu" 'house' whereas in (2), the same form "ukku + muNpu" occurs as a temporal adverbial marker before the action noun ('verbal noun'). This ambiguity also could be solved only with the concordance program.

7.5.2.3 Participial noun and adjectival noun

In general school grammar books, the word “patiththavan” ‘one who studied’ is defined as a participial noun -வினையாலணையும் பெயர் “viNaiyaalaNnaiyum peyar”. This wordform consists of the following segments: “pati -thth -avaN” ‘study - past tense suffix - pronominal suffix’. The basic root word is a verb “pati” ‘study’. This verb takes a tense suffix followed by the pronominal suffixes. That is, it is derived from a verb inflection.

There is one more wordform which resembles the above participial noun because of the presence of the pronominal suffix. For example, “nallavaN” ‘a good person’ consists of the above suffix “avaN”; but the root word or lexicon is not a verb. It is an adjective - “nalla” ‘good’ and it has to be categorized in a different way by the term ‘adjectival noun’ -பெயரடைப் பெயர் “peyarataippeyar”. This grammatical categorization is generally absent in popular grammar books. Only some linguistic reseachers have dealt with it (Kothandaraman, 1997). In the present research project, this difference in the POS categorization is maintained.

7.5.2.4 Adjective and Adverbs in Tamil

In Tamil, the adverbs could take almost all the clitics whereas there are some restrictions with the adjective. Pure adverbs such as “mella” ‘slowly’ can take clitics. But pure adjectives such “nalla” ‘good’, “ketta” ‘bad’ won’t take clitics directly.

Examples:

1) “avar mella varukiRaar”	‘He comes slowly’
2) “avar urakka peecukiRaar”	‘He speaks loudly’

Here, the above two adverbs take most of the clitics as follows:

1.	“mellaththaaN”	‘slowly only’
2.	“mellavee”	‘very slowly’
3.	“mellakkuuta”	‘slowly too’
4.	“mellavum”	‘slowly too’

But the pure adjectives would not take any clitic directly. But if an adjectival noun suffix is added with this, then the resultant forms could take clitics.

“nalla” ‘good’ + “athu” ‘it’ > “nallathu” ‘it is good’ could take clitics.

1.	“nallathaa?”	‘is it good?’	Here “aa” is the clitic.
2.	“nallathee”	‘only good’	Here “ee” is the clitic.
3.	“nallathuthaaN”	‘it is good only’	Here “thaaN” is the clitic.
4.	“nallathum”	‘and good also’	Here “um” is the clitic.

Likewise, the derived adverbs such as “azakaaka” ‘beautifully’ “veekamaaka” ‘fast’ could take clitics as like pure adverbs.

1)	“azakaakaththaaN”	‘beautifully only’	Here “thaaN” is the clitic.
2)	“veekamaakakkuuta”	‘fast too’	Here “kuuta” is the clitic.

But derived adjectives also would not take any clitic. This is the one major difference between an adjective and an adverb. But inbetween a noun and an adjectival or an adverbial suffix, some clitics could occur.

1) “aaciriyarumaaNa” ‘be also as a teacher’ (“aaciriyar -um-aaNa”); here, “um” is the clitic and “aaNa” is the adjectival suffix.

2) “aaciriyarumaaka” ‘also as a teacher’ (“aaciriyar - um - aaka”); here, “um” is the clitic and “aaka” is the adverbial suffix.

7.6 Conclusion

All the challenges discussed in this chapter, together with the responses highlighted, are related to computational and corpus linguistics: corpus construction and corpus analysis, using various language software tools such as the morphological parser, POS tagger, concordancer, N-gram and the relevant algorithm.

Most of the issues considered shed lights on the emergence of new linguistic features in WMTC and the need to accommodate these features in the future research on Tamil grammar. All these will contribute to enhanced Tamil grammar teaching in schools, teacher training colleges and universities, which in turn benefit the Tamil-speaking community, a point which will be considered in the next chapter. There are also implications for language technological applications such as spell checker and grammar checker, which are also considered in the next chapter.

CHAPTER 8

CONCLUSION

8.0 Introduction

This research project has presented an account of corpus building for written Malaysian Tamil and discussed the development of a morphological parser and a POS tagger to analyze the designed corpus. The corpus construction work was informed by previous work in corpus linguistics (e.g., BNC, COCA, the Brown corpus) and it took into consideration the status quo of the Tamil language in Malaysia.

The corpus access tools developed here to process and study the Tamil corpus share some main functions in such corpus tools as WordSmith and AntConc. It has also been shown that additional functions for the processing of the Tamil language have been integrated in a larger software package comprising the parser and the POS tagger. Following the common practice of technology development, the present project has also evaluated and optimized the corpus tools by data feeding.

The present chapter gives a synopsis of the characteristics of the finalized built-up corpus and the developed corpus tools; it discusses their contributions and limitations and considers implications for future research and practice.

8.1 The contributions of WMTC and corpus tools for Tamil corpus linguistics

The present project has accomplished two things: (1) the creation of a corpus of written Tamil in Malaysia, and (2) corpus access tools for the Tamil corpus and other similar corpora.

The constructed corpus has 1 million orthographic words of written Tamil. These words were collected from 500 samples containing about 2000 words each. These samples were collected from a range of domains and media to make this corpus a representative and balanced one as possible, although admittedly, for practical reasons, there are some limitations, which are discussed later.

The corpus access software developed in this project contains multiple affordances that can serve two major purposes: (1) corpus annotation and (2) corpus analysis. For the first purpose, the software includes a morphological parser and a POS tagger. For the second purpose, the software features mainstream affordances such as a concordancer, a keyword extractor and a wordlist function. All these functions are integrated into one single software. More details about the corpus and corpus access tools are summarized as follows:

8.1.1 Corpus development of written Malaysian Tamil

This is the pioneer research project on written Malaysian Tamil. This developed corpus consists of texts containing about one million orthographic words from various media and domains. The corpus linguistics features “representativeness” and “balanced” have been adopted in constructing the corpus for further corpus analysis. Based on the relative frequency of the sources and domains, the corpus texts were selected.

8.1.2 Corpus access software development

The second major contribution of the research project is the development of a corpus access software that can serve purposes of corpus access and annotation. The software developed in this project is an integrated one, having all the corpus linguistic tools - corpus construction tool, corpus view tool, tokenization tool, type selection tool, morphological parsing tool, POS tagging tool, concordance tool, N-gram tool, lemma extraction tool and the necessary statistical tools.

In any corpus project, the tokenization and the type selection are very important tasks. For this, in addition to the morphological parser, two more important tools - Concordance programme and N-gram - are seriously needed. The reason is the meaning of a word/word form is dependent not only on the phonological form of the word but also on the POS category of the particular word form. For example, the meaning of the word "malar" in a particular sample depends upon the POS category of this word. This word may be a noun or a verb. Here, the POS ambiguity must be solved first. To solve this ambiguity, the linguistic context of a particular word should be studied. Also, some words may have a single POS tag, but their meaning may be different. For example, the noun word "pati" has different unrelated meanings. It may be 'stepping-stone' or 'measure' or 'copy of the material'. To solve this, the Word Sense Disambiguation (WSD) tool such as found in AntConc and Wordsmith tools are necessary. The present project has two such tools: one is, Concordance programme and another one is, N-gram programme. These two tools developed for this project are useful to disambiguate both POS ambiguity and meaning ambiguity of the ambiguous words present in the project.

The corpus access software also includes other useful functions. Also, for this project, some statistical tools such as frequency count are much useful for analyzing the data objectively. Another important tool developed for this project is to extract the root/lemma of the inflected word forms. This may be useful to construct new Tamil lexicons/dictionaries.

Here it is to be mentioned that the success of any morphological parser depends upon the morphotactics of words. The rules of morphotactics in any language are finite. Also, their occurrences are ordered. In this sense, any morphological parser for any language is based on finite number of states of morphemes in its words or wordforms. That is why, in computational morphology of any language, the word form or inflected form may be captured by definite morphological rules or mathematically and computationally either by a Regular Expression (RE) or by a Finite State Automata (FSA).

8.1.3 Flow charts for morphological inflection

The study of all the grammatical affixes and their order of occurrence in a language is essential to develop morphological parser for that language. This study is the basis for the study of morphological inflection or inflection morphology of any language. The present project has made an elaborate study of Tamil inflectional morphology and designed two flow charts - one is for noun inflection and another for verb inflection. Also, adverbial and adjectival inflections are also captured and the flow charts for the same are developed. These flow charts are helpful not only for this project but also for any linguistic projects which need any Tamil inflectional study.

8.1.4 Proofing and Normalization

For this, the input wordforms or types should be error-free grammatical ones; that is, the collected text samples should be proof-read and normalized. This normalization is necessary for Tamil. Because in the history of Tamil, many lexicons have undergone grammaticalization process and even after their grammaticalization the above lexicons have been existing as lexicon side by side with their grammaticalized forms. Since these grammaticalized words retain their original lexical status also, the Tamil language users in many contexts write these grammatical words not as bound morphemes but as free morphemes. But in Tamil, all the grammatical forms are bound morphemes. They have no independent existence from the words to which they add grammatical properties.

Since the development of a morphological parser is one of the objectives of the present project, the necessary tokenization of the corpus texts and the selection of types have been done. For these two tasks, the necessary text cleaning and normalization processes are done over the corpus texts. The normalization process could be done with the help of a preliminary morphological parser and a check list of grammatical words - post positions and auxiliaries.

8.1.5 Development of final morphological parser and POS tagger

Based on the result of proofing and normalization, the tasks of final selection of tokens and type selection were completed. The final list of word types was ready for final parsing and POS tagging. The final list of the types found in the corpus were sent for further morphological parsing. In this process, the morphological parser went through many intermediate stages to reach its final stage - that is, to handle the types or word forms of Tamil. However, when the tentative morphological parser was used, its insufficiency to handle WMTC word structure was found. The final morphological parser was then developed which to handle Tamil morphological features, including new suffixes and morphophonemic features.

Here, it is to be mentioned that this is the pioneer attempt for the development of a morphological parser for written Malaysian Tamil based on corpus - that is, authentic data, collected from the real use of written Malaysian Tamil.

Moreover, the present developed morphological parser is based on the modern computational and corpus linguistic features, in addition to the adaptation of the latest concepts of database design as well as information retrieval and extraction from computer science. This developed morphological parser would certainly help the future research in computational morphology for written Malaysian Tamil.

It is the first time in Malaysia that major inflectional features of WMTC words are computerized in the form of a morphological parser. Based on this, the necessary flow-chart for Tamil morphotactics - especially for noun and verb inflection would certainly be useful for further language technology application such as spell-checker and sandhi-checker.

Moreover, with the help of the morphological parser, all the types (word forms) in the corpus were linguistically annotated for their grammatical categories, including their sub-categorization. This would be much useful for the development of syntactic parser for written Malaysian Tamil in the future.

The list of all the lemmas (root words) extracted with the help of the developed morphological parser would certainly help the lexicographers in their development of dictionaries, including pedagogical ones, for written Malaysian Tamil.

In addition to the primary task of the development of a morphological parser for written Malaysian Tamil, some more corpus linguistic tools like concordancer, N-gram analysis tool and frequency study tool were developed and included in this research project. All these corpus linguistic tools would certainly help to develop more language technology tools in the areas such as E-Lexicography, Word-level automatic Machine Translation for written Malaysian Tamil.

Finally, the corpus texts consisting of about one million types (word forms) along with their linguistic annotations would help not only the Tamil corpus and computational linguists and language technologists but would also be useful for researchers and teachers in Tamil pedagogy in Malaysia.

The contribution of annotated and unannotated corpus materials and tag sets for exploring the Tamil language.

The project has created both an annotated and an unannotated corpus. The annotated corpus, with a 51 POS tag set based on Tamil grammar developed especially for this purpose, provides researchers with rich information on the Tamil morphology. The POS tag set, which helps to distinguish the various grammatical categories in contemporary written Tamil, forms one of the significant contributions to the field.

The unannotated corpus, on the other hand, allows researchers to access individual texts in the corpus and to analyze the corpus from their own aims and approaches and to develop their own POS tags based on their own interest. The texts in the Tamil corpus developed for this research project form a meaningful dataset for Tamil-related research.

The selected texts, based on the design of the corpus (e.g., its authenticity, metadata, and annotation), allow for the exploration of a wide range of research topics on the Tamil language.

Further, the sample texts gathered for this project are not individual words, but they are cohesive passages, consisting of about 2000 words each. The metadata added with each sample may help researchers to explore various socio-linguistic and communicative features of interest that characterize those texts.

8.2 The contribution of the Tamil morphological structure chart

For an inflectional and agglutinative language such as Tamil, the backbone for the development of a morphological parser is the morphotactics and morphophonomics. The Tamil morphotactics is exhaustively studied in the present corpus and is represented in a flow chart given below. In other words, the computational aspects of Tamil morphotactics have been fully studied in the present thesis which is one of the important contributions to the Tamil computational morphology. Certainly, it will help the researchers involved in Tamil computational linguistics and language technology to move forward in the right direction. In future, with the help of the Tamil computational morphotactics many more language technology tools ranging from spell checker to machine translation for WMT can be developed.

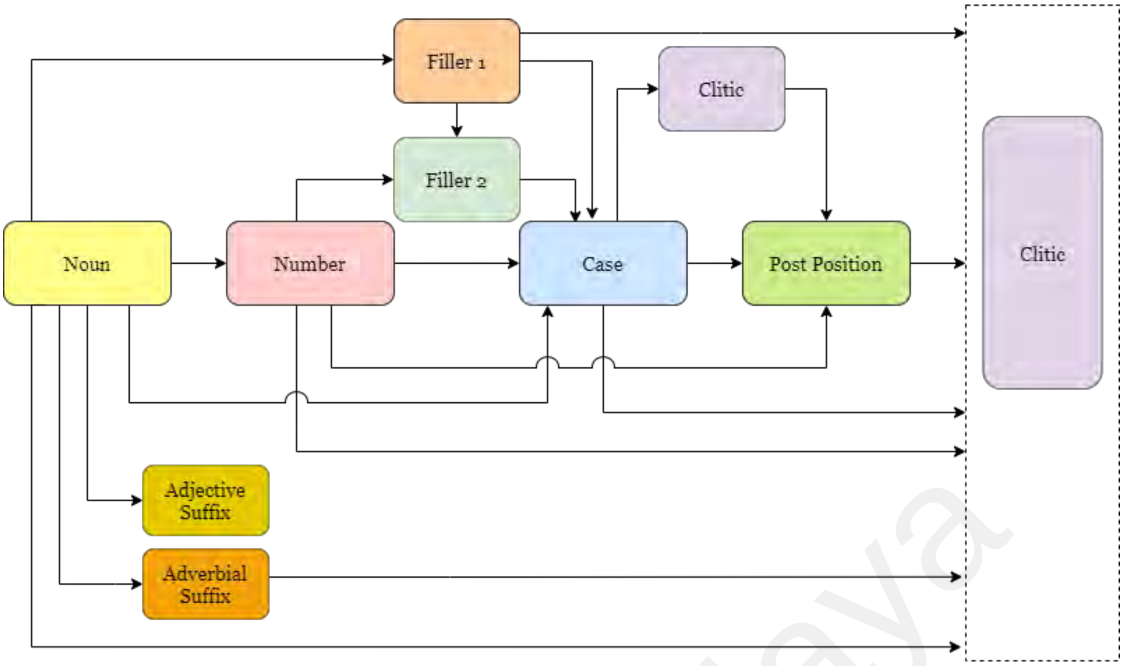


Figure 8.1: Diagrammatic representation of

Tamil Computational Morphotactics (Noun)

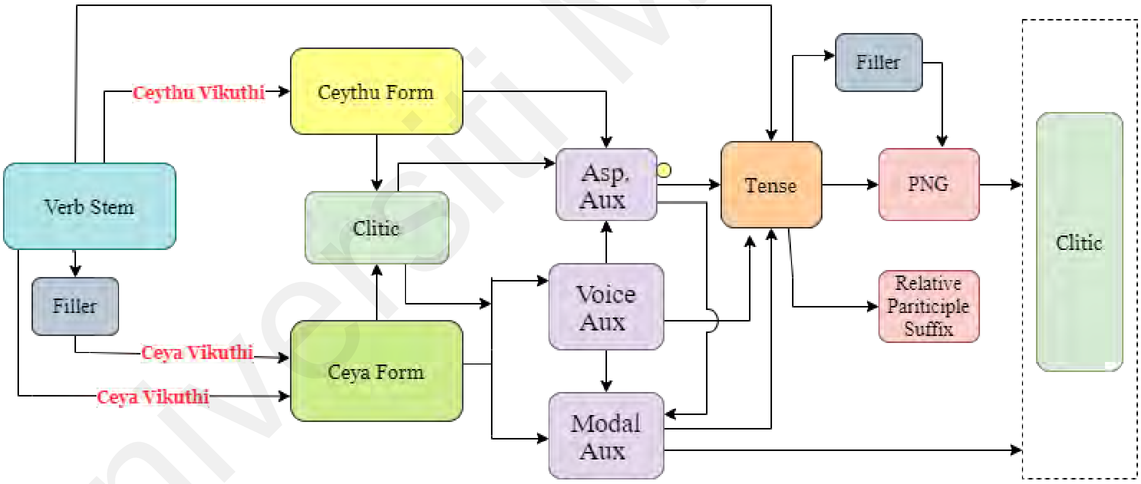


Figure 8.2: Diagrammatic representation of

Tamil Computational Morphotactics (Verb)

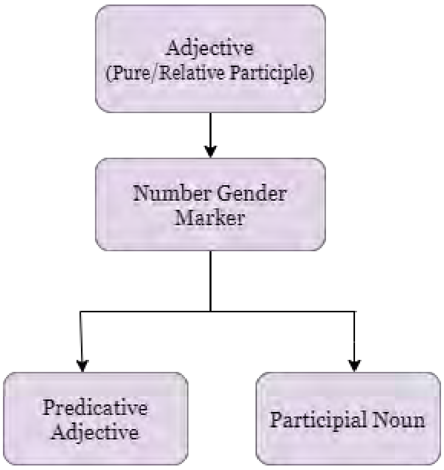


Figure 8.3: Diagrammatic representation of

Tamil computational Morphotactics (Adjective)

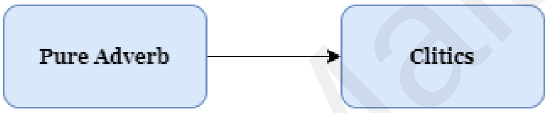


Figure 8.4: Diagrammatic representation of

Tamil computational Morphotactics (Adverb)

8.3 New features in written Malaysian Tamil

The research project has applied the corpus access tool to the constructed corpus. The corpus analyses have revealed a variety of changes in contemporary Tamil as compared with the ancient Tamil. These changes are reflected in grammar, morphotactics, morphophonemics, and lexicon.

Tamil, one of the ancient and classical languages, has been continuously growing and enriching itself to feed the new needs of the Tamil speech community. Though it preserves its originality, whenever necessary, accommodates new changes in lexicon and grammar. Here it is to be noticed that these changes do not affect in any way the basic structure of Tamil. That is why, it could be claimed that modern written Tamil in Malaysia is not a new one which fundamentally differs from the old one. On the other hand, it adds any new feature only, based on the new communicative needs, which does not change its basic structure.

Many new grammatical features have emerged in written Malaysian Tamil. Though some grammar works have paid attention to written Malaysian Tamil, they are not based on corpus. The main reason is, for them, no systematically collected corpus was available. Because of this, some new features are not accounted for by them. To accommodate these new features and to enrich the above grammar works, the present research project was ventured which is based on authenticated Tamil corpus.

The present systematically constructed corpus for contemporary Tamil paved the way to identify the unnoticed allomorphs for various grammatical categories in contemporary Tamil. In this respect, this corpus-based analysis is a pioneer one in the description of Tamil morphology in written Malaysian Tamil.

These newly emerged morphotactic changes have to be accommodated in the morphological parser to segment the individual components of the contemporary Tamil word forms/types. If the segmentation is not done in a right way, the respective word could not be tagged for its correct POS and further syntactic analysis could not be proceeded. In this aspect, the present research project has laid the important foundation for any higher-level linguistic analysis. This type of changes in the morphotactics has been discussed in a detailed way in this research project.

In Tamil, the internal sandhi / morphophonemics plays an important role in the agglutinative process involved in building the compound words (root + root) and word forms (root + inflectional suffixes) as well. In segmenting the inflected word form into proper segments / morphemes, the rules of morphophonemics should be given to the parser for correct morphological parsing task.

In written Malaysian Tamil, some old morphophonemic rules have disappeared. Likewise, there are many new changes in written Malaysian Tamil. These changes have been identified in the present corpus. With the help of the corpus, new lexicons emerged in written Malaysian Tamil have been identified. This emergence of new lexicons is to satisfy the new communicative needs of Tamil speaking people. The corpus could also reveal the influence of Malay, English and Chinese language over Tamil. All these changes and developments in contemporary written Tamil have been discussed in detail in the previous chapters.

8.4 The computational nature of Tamil Morphology

For a feature or a process to be computable, they should have mathematical nature. Likewise, if the structure of any language is to be computable, the rules behind the structure should have mathematical nature. In this present project, all the rules - morphotactic and morphophonemic rules - behind the Tamil morphological parser are computable. All these rules are explicit and systematically structured. Nothing in the Tamil morphological structure - the morphotactics, the morphophonemics and grammatical categorization - is beyond the rules. It shows that the nature or properties of Tamil word structure or Tamil morphology is a mathematical one. However, to make Tamil morphology a computable one, we need a deeper and microlevel analysis of Tamil word forms. This research project did this expected analysis which made this project a significant one.

8.5 Implications

The corpus and corpus access tools developed in this project have multiple implications.

8.5.1 Implications for Language Technology

The present morphological parser could be used as a backbone in the development of Tamil spell-checker and sandhi checker. Also, to develop a syntactic parser for Tamil in computational linguistics, this present morphological parser and POS tagger would be useful. This corpus-based project could be much useful in the development of Tamil dictionaries (including electronic dictionaries) for the written Malaysian Tamil.

In Tamil language technology, the development of a spell checker, sandhi checker and grammar checker are important. To develop all these tools which are useful in proofing of any text typed by the user, a well-built lexicon and a morphological parser are necessary. Moreover, the morphological parser should have necessary speed and accuracy. For the development of Tamil morphological parser, the rules of phonotactics, morphotactics and morphophonemics behind every Tamil word should be properly analyzed and accommodated in the parser, which has been addressed in this research project. Based on this present project, the researchers or developers interested in this task could proceed to develop the word processor tools such as spell checker, sandhi checker and word grammar checker.

The research described and reported in this thesis has implications for the development of a syntactic analyzer. To analyze a sentence, first of all, the grouping of words into phrases and grouping phrases into that sentence are needed. To arrive at the phrases correctly, we need POS tagged words present in that sentence. Hence all the words in that sentence, first of all, should be parsed and based on this parsing, they should be POS tagged. The present project helps to get the POS information for each and every word in any sentence.

This can be schematically represented as below:

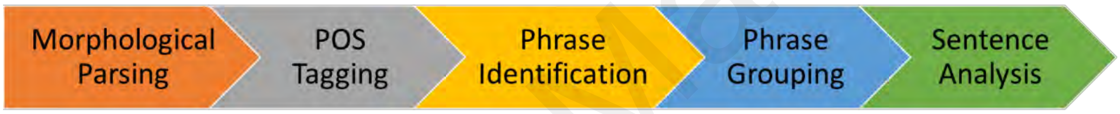


Figure 8.5 The process of syntactic analysis

In this way the present project lays the foundation or ways on which syntactic analysis could be done.

8.5.2 Implications for linguistic research

For Malaysian Tamil, there is no well-constructed lexicon for written Tamil. To have this, it is needed for a scientifically constructed Tamil corpus, on which the needed lexicon could be built. The word or word forms could be analysed by a morphological parser to get the root or basic lexicon. For this task, the present project advanced this line of research by constructing a well-designed corpus, a morphological parser, and a POS tagger to have all the roots or lexicon of Malaysian Tamil.

The study of discourse structure: In the discourse analysis of any text, the analysis of use and usage of sentences in that particular text is essential. In this discourse study, the inflectional morphology plays an important role.

8.5.3 Practical implications for Tamil teaching

The results from the thesis can be applied to Tamil language teaching. In Tamil language teaching (both in first language and second language teaching), textbook writers can use this corpus in writing lessons with the authenticated words and sentences available in present corpus. The present morphological parser and POS tagger would be useful for preparing and analyzing Tamil teaching materials.

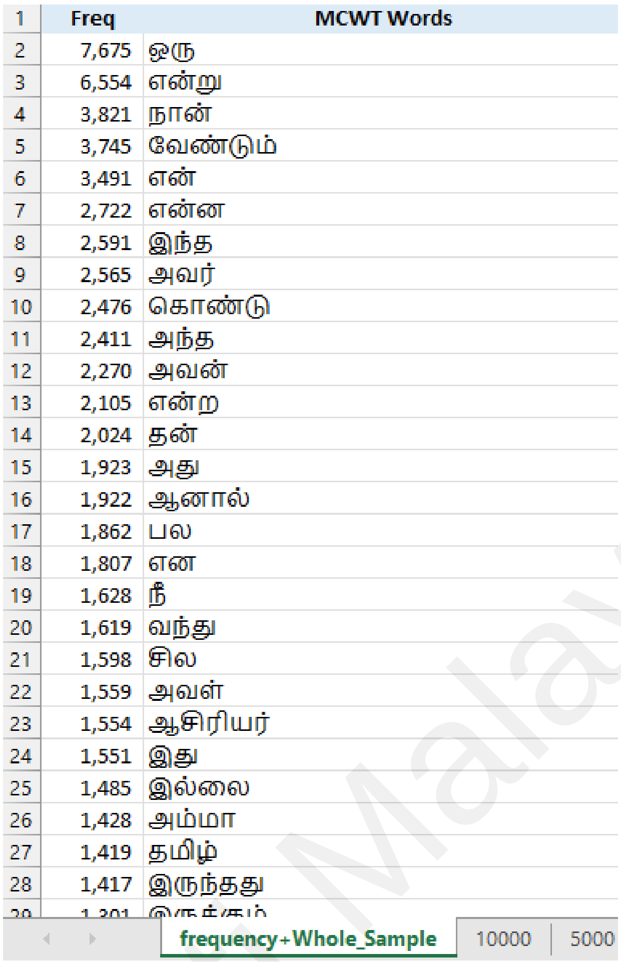
The corpus texts consisting of about one million orthographic words along with their linguistic annotations would help the researchers and teachers in Tamil pedagogy in Malaysia. This research project could also inform the development of Malaysian Tamil language curriculum.

In Tamil language pedagogy, the present situation is, the grammar being taught in schools is the old Tamil grammar. But the lessons are in contemporary Tamil. This gap hinders the students' language performance. The grammar taught in the classrooms should help the students to develop their language performance. Only if the above-mentioned new changes are accommodated in the old grammars, Tamil language teaching would be much useful for the students in the development of their Tamil linguistic competence and performance.

In classroom teachings, to teach various Tamil grammatical categories, word parsing and identification of grammatical category of any word, the output of the morphological parser and POS tagger could be more useful. The students could be taught Tamil words in real context using the concordance programme developed in the present research project.

8.5.4 Essential Vocabulary words for L1 and L2 Learners

It is well known that for the second language learners (L2) of English, a total of 5,000 root words/lexicons with elementary grammatical rules are sufficient to learn the English language and for the first language learners (L1) a total of 20,000 vocabulary words with appropriate grammatical rules are required to learn the English language. (Laufer & Nation, 1995; Nation & Waring, 1997). For Tamil language teaching - both as a first language and a second language - the same concept could be adopted. The results from the present corpus project can be used to inform the teaching of Tamil to Malay and Chinese L2 learners. The frequency of words in WMTC is represented in the following figure 8.6.

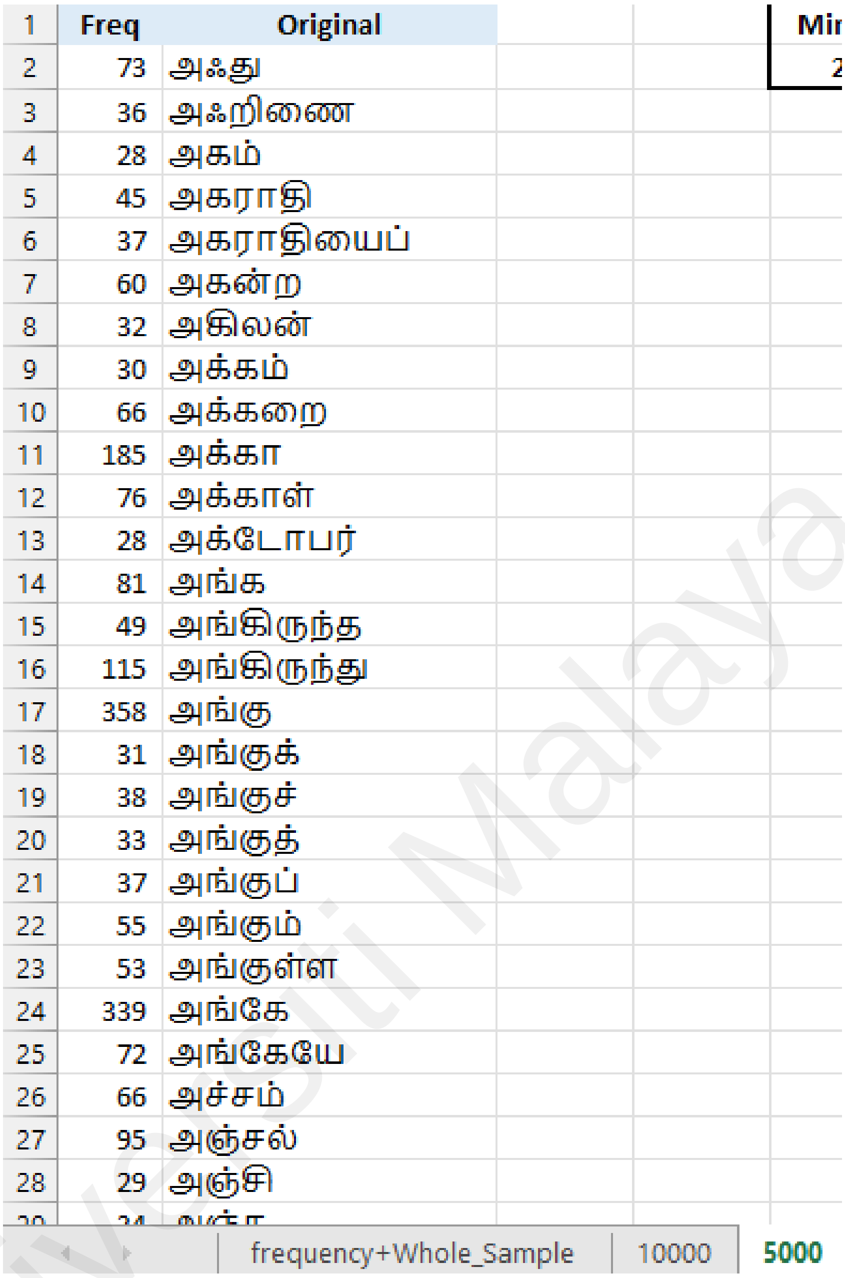


1	Freq	MCWT Words
2	7,675	ஒரு
3	6,554	என்று
4	3,821	நான்
5	3,745	வேண்டும்
6	3,491	என்
7	2,722	என்ன
8	2,591	இந்த
9	2,565	அவர்
10	2,476	கொண்டு
11	2,411	அந்த
12	2,270	அவன்
13	2,105	என்ற
14	2,024	தன்
15	1,923	அது
16	1,922	ஆனால்
17	1,862	பல
18	1,807	என
19	1,628	நீ
20	1,619	வந்து
21	1,598	சில
22	1,559	அவள்
23	1,554	ஆசிரியர்
24	1,551	இது
25	1,485	இல்லை
26	1,428	அம்மா
27	1,419	தமிழ்
28	1,417	இருந்தது
29	1,391	இருக்கும்

Figure 8.6: High frequency words

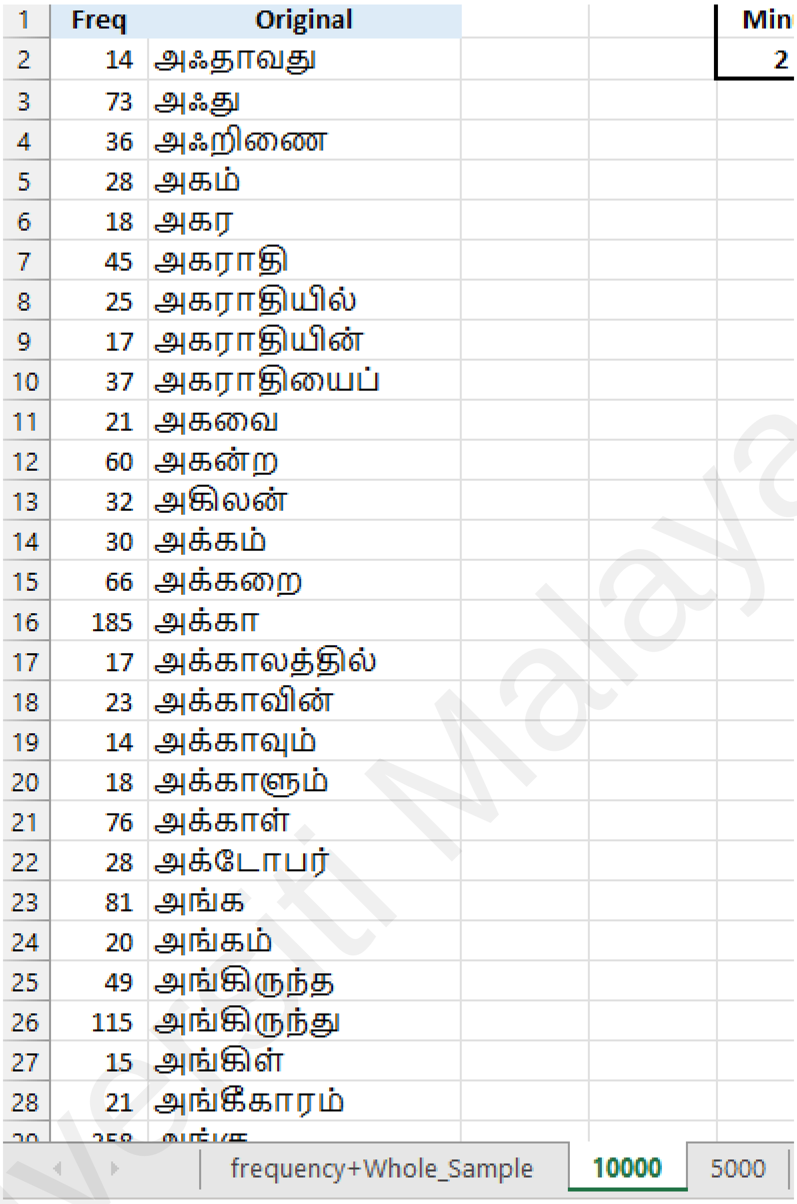
8.5.5 High frequency words for curriculum developers

The most frequent root words/lexicons identified in the present corpus project (Figure 8.7 and 8.8) can help textbook writers and planners in selecting the vocabularies to be included in textbooks for Tamil language learners based on their proficiency levels. Here, not only the basic lexicons, but also the most frequent inflected forms of the above Tamil lexicons may be considered to teach the inflectional grammar of Tamil. The first 1,000 high frequency words can be used in the production of pre-school text materials. The first 5,000 words high frequency words are recommended for the primary school children learning Tamil. A total number of 10,000 high frequency words can be included for the secondary school children learning Tamil Language.



	Freq	Original
2	73	அஃது
3	36	அஃறிணை
4	28	அகம்
5	45	அகராதி
6	37	அகராதியைப்
7	60	அகன்ற
8	32	அகிலன்
9	30	அக்கம்
10	66	அக்கறை
11	185	அக்கா
12	76	அக்காள்
13	28	அக்டோபர்
14	81	அங்க
15	49	அங்கிருந்த
16	115	அங்கிருந்து
17	358	அங்கு
18	31	அங்குக்
19	38	அங்குச்
20	33	அங்குத்
21	37	அங்குப்
22	55	அங்கும்
23	53	அங்குள்ள
24	339	அங்கே
25	72	அங்கேயே
26	66	அச்சம்
27	95	அஞ்சல்
28	29	அஞ்சி

Figure 8.7: High frequency words (5000)



1	Freq	Original	Min
2	14	அஃதாவது	2
3	73	அஃது	
4	36	அஃறிணை	
5	28	அகம்	
6	18	அகர	
7	45	அகராதி	
8	25	அகராதியில்	
9	17	அகராதியின்	
10	37	அகராதியைப்	
11	21	அகவை	
12	60	அகன்ற	
13	32	அகிலன்	
14	30	அக்கம்	
15	66	அக்கறை	
16	185	அக்கா	
17	17	அக்காலத்தில்	
18	23	அக்காவின்	
19	14	அக்காவும்	
20	18	அக்காளும்	
21	76	அக்காள்	
22	28	அக்டோபர்	
23	81	அங்க	
24	20	அங்கம்	
25	49	அங்கிருந்த	
26	115	அங்கிருந்து	
27	15	அங்கிள்	
28	21	அங்கீகாரம்	

Figure 8.8: High frequency Words (10,000)

The number of words required for assessing the standard of language learning is very well researched for English language. The number of high frequency words ranging from 500 the lowest to 16,000 the highest to grade the levels is fixed for English language (https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions).They are as follows:

Table 8.1: CEFR Language level

Language Level	Number of Base Words Needed
A1	500
A2	1000
B1	2000
B2	4000
C1	8000
C2	16000

The number of vocabularies for respective levels could be adopted for Tamil language learners, which is based on the frequency of words, as given above. The faculty of languages and linguistics of Universiti Malaya has requested to assess the competency level of Tamil language students belonging to first degree Linguistics on the lines of Common European Framework for Language Learning (CEFR) (https://www.coe.int/en/web/common-european-framework-reference-languages). The number of base words needed for each language level prescribed for CEFR may be adopted for Tamil language also by making use of the present corpus developed. This is represented in the following figure 8.9.

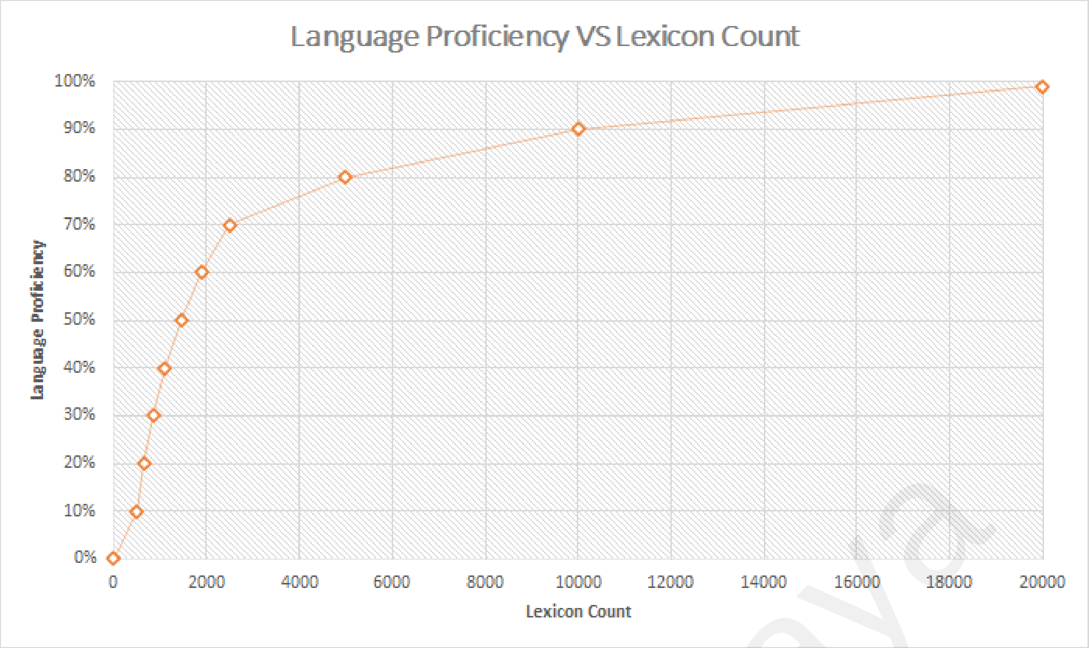


Figure 8.9: Language proficiency vs Lexicon Count

8.5.6 Lexile Measures for Reading

For English language teaching:

> The Lexile Framework for Reading makes test scores actionable by placing student reading ability on the same developmental scale as text complexity. Today, Lexile measures are recognized in the United States and around the world as the gold standard for helping students navigate the path to college and career readiness. (https://lexile.com/)

For Tamil teaching also, the above kind of Lexile framework for reading maybe adopted to improve the language competency. For this, the present corpus would be much useful. Also, here it is suggested to develop a special purpose corpus for Tamil language teaching.

8.6 Limitations of the present study

1) The present research project is concerned only with written Tamil but not spoken Tamil.
2) This project does not attempt either to analyze or to compare the written Tamil used in other countries such as Tamil Nadu, Sri Lanka.
3) The size of the present project is one million orthographic words only. It is not large enough (not in billions or trillions) to study all the linguistic aspects of written Malaysian Tamil such as morphophonology, morphosyntax, semantics and discourse structure. The present research project restricts itself only with the development of morphological parser and POS tagger.
4) There are some issues in morphological parsing which could be solved only with the further development of a syntactic parser.
5) The present research project restricts itself to Tamil computational morphology. Not all the grammatical features could be covered due to small size of the corpus. The efficiency of the parser is dependent on the size of the corpus. Though the features of the Brown corpus, BNC and COCA were studied, according to the local conditions of written Malaysian Tamil they were slightly modified and adopted. Since this research is first of its kind for written Malaysian Tamil, the ratio of samples in each domain could not match with BNC and COCA. Only two (0.4%) out of 500 samples comprising of different texts (each containing 100-300 words) in same domain were combined to make 2000 words accordingly.

8.7 Recommendations for Future Research

The thesis project is a pioneering effort to apply corpus linguistics to the language of Tamil in Malaysia. The study is a beginning for this enterprise and provides an avenue for future research. It is in this spirit that the thesis recommends further research in this direction. In the very first place, there is the need for a larger corpus of written Tamil up to billions of words or tokens. It is more likely for researchers to observe linguistic examples in a large-scale corpus. Future research may consider expanding the scope of data from written Tamil to spoken Tamil. Research based on spoken Tamil corpus has the potential to reveal real-time communication and other communicative functions. In addition, more computer tools are desired to process Tamil corpora. For example, computer software that can semantically analyse Tamil corpora and even translate Tamil to other languages such as English, Malay and Chinese would have practical implications.

Many new branches have been coming up in the field of corpus linguistics. Only a few have been dealt with using the corpus. Under the following topics in linguistics, various scholarly works have been done: discourse analysis (Baker, 2006; Cheng, 2013; Conrad, 2002; Flowerdew, 2013), pragmatics (Adolphs, 2008), lexis and grammar (Francis, 1993; Hunston, 1995, 2002b), syntax (Aarts, 1991; Mindt, 1995), (Oostdijk, 1991) historical linguistics (Davies, 2012; Janda et al., 2020; Jenset & McGillivray, 2017),language acquisition (Huat, 2012; Monaghan & Rowland, 2017), language teaching (Cheng, 2010; Gilquin, 2021; O'keeffe et al., 2007; Sinclair, 2014), language variation (Biber, 2010; Kachru, 2008), lexicography (Hanks, 2012; Hoey & O'Donnell, 2008; Hurskainen, 2003; Krishnamurthy, 2008; Sinclair, 2003), psycholinguistics (Ellis & Simpson-Vlach, 2009),semantics (Baroni & Lenci, 2010; Stubbs, 2001), social psychology (Balossi, 2014), sociolinguistics (Baker, 2010; Friginal & Hardy, 2013), and cultural studies (Shastri, 1988).

The present research project could pave the way for future researchers to explore more exciting projects in corpus and computational linguistics.

8.8 Conclusion

All the three research objectives: To construct a corpus of written Malaysian Tamil (i.e., WMTC); to develop a morphological parser and POS tagger to process WMTC; and to choose and design the relevant algorithms for the morphological parser and the POS tagger, have been fulfilled in this present thesis.

REFERENCES

Aarts, J. (1991). Intuition-based and observation-based grammars. In K. Aijmer & B. Altenberg (Eds.), *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. Longman.

Abdullah, I. H., Rahman, A. N. C. A., & Jaludin, A. (2021). The Development of the Malaysian Hansard Corpus: A Corpus of Parliamentary Debates 1959-2020. *Jurnal Linguistik*, *25*(1).

Adolphs, S. (2008). *Corpus and context: Investigating pragmatic functions in spoken discourse*. John Benjamins Publishing Company. http://www.loc.gov/catdir/toc/ecip084/2007045722.html.

Aduriz, I., Agirre, E., Aldezabal, I., Arregi, X., Arriola, J. M., Artola, X., Gojenola, K., Maritxalar, A., Sarasola, K., & Urkia, M. (2000). A Word-level Morphosyntactic Analyzer for Basque. LREC.

Agesthialingom, S. (1964). Auxiliary verbs in Tamil. *Tamil culture*, *11*(3).

Agesthialingom, S. (1976). *Dravidian case system* (Vol. 1). Annamalai University.

Agesthialingom, S. (2004). Numeral System in Tamil: Generation. *South-Indian Horizons: Felicitaion Volume for Francois Gros on the Occasion of His 70th Birthday*, *94*, 323.

Agesthialingom, S., & Varma, G. S. (1980). *Auxiliaries in Dravidian* (Vol. 70). Annamalai University.

Aikhenvald, A. Y. (2007). *Typological dimensions in word-formation*. In Shopen, Timothy, (ed.) Language Typology and Syntactic Description. Cambridge University Press.

Ajees, A., & Idicula, S. M. (2020). *Design and development of an integrated framework for pronominal anaphora resolution in malayalam* Cochin University of Science and Technology].

Allan, R. (2009). Can a graded reader corpus provide ‘authentic’input? *ELT Journal*, *63*(1), 23-32.

Anand Kumar, M., Dhanalakshmi, V., Rekha, R., Soman, K., & Rajendran, S. (2010). A novel data driven algorithm for Tamil morphological generator. *International Journal of Computer Applications*, *975*, 8887.

Anand Kumar, M., Dhanalakshmi, V., Soman, K., & Rajendran, S. (2010). A sequence labeling approach to morphological analyzer for tamil language. *International Journal on Computer Science and Engineering*, *2*(06), 1944-1951.

Anderson, S. R. (1982). Where's morphology? *Linguistic inquiry*, *13*(4), 571-612.

Ang, L. H., Rahim, H. A., Tan, K. H., & Salehuddin, K. (2011). Collocations in Malaysian English learners' writing: A corpus-based error analysis. *3L: The Southeast Asian Journal of English Language Studies*, *17*(Special Issues), 31-44.

Annamalai, E., & Steever, S. B. (2015). Modern Tamil. In Steever, S.B.(ED), *The dravidian languages* (pp. 118-175). Routledge.

Anthony, L. (2014). *AntConc*. In (Version 3.4.3) Waseda University. http://www.laurenceanthony.net/.

Antony, P., & Soman, K. (2012). Computational morphology and natural language parsing for Indian languages: a literature survey. *International Journal of Scientific and Engineering Research*, *3*.

Aronoff, M. (1993). *Morphology by itself: Stems and inflectional classes* (Vol. 22). MIT press.

Aston, G. (1998). The bnc handbook exploring the british national corpus with sara guy aston and lou burnard.

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and linguistic computing*, *7*(1), 1-16.

Awal, N. M., Zainuddin, I. S., & Ho-Abdullah, I. (2011). Use of comparable corpus in teaching translation. *Procedia-Social and Behavioral Sciences*, *18*, 638-642.

Baker, P. (2006). *Using corpora in discourse analysis*. A&C Black.

Baker, P. (2010). *Sociolinguistics and corpus linguistics*. Edinburgh University Press.

Balakrishnan, R. (2002). Morphology and Tamil Computing. Paper read in International Seminar on Tamil Computing, February.

Balasubramanian, T. (1980). Timing in Tamil. *Journal of Phonetics*, *8*(4), 449-467.

Balasundararaman, L. (2003). Context free grammar for natural language constructs-an implementation for venpa class of tamil poetry.

Balossi, G. (2014). A Corpus Linguistic Approach to Literary Language and Characterization. *Amsterdam and Philadelphia: John Benjamins*.

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational linguistics*, *36*(4), 673-721.

Bauer, M. W., & Aarts, B. (2000). Corpus construction: A principle for qualitative data collection. *Qualitative researching with text, image and sound: A practical handbook*, 19-37.

Beesley, K. R. (1998). Arabic morphology using only finite-state operations. Computational Approaches to Semitic Languages, Xerox Research Centre Europe.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243-257. https://doi.org/10.1093/llc/8.4.243

Biber, D. (2010). Corpus-based and corpus-driven analyses of language variation and use. In *The Oxford handbook of linguistic analysis*.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press. Publisher description.

Biber, D., & Reppen, R. (2015). *The Cambridge handbook of English corpus linguistics*. Cambridge University Press.

Botley, M. D. s. H., Faizal Metom, Puan Lilly Dillah, Puan Doreen (2007). A Corpus-Based Archive Of Learner English In Sabah/Sarawak (Cales Phase 2).

Bowker, L., & Pearson, J. (2002). *Working with specialized language: a practical guide to using corpora*. Routledge.

Burnard. (2000). Corpus resources and minority language engineering. LREC.

Campoy, M. C., Cubillo, M. C. C., Belles-Fortuno, B., & Gea-Valor, M. L. (2010). *Corpus-based approaches to English language teaching*. A&C Black.

Chaal, M. K. S. (2011). Syntax Ina Business Context: A Learner Corpus Analysis Manvender Kaur Sarjit Chaal. *ELT: Converging Approaches and Challenges*, 17.

Chau, M. H. (2015). *From language learners to dynamic meaning makers: A longitudinal investigation of Malaysian secondary school students' development of English from text and corpus* University of Birmingham]. Birmingham.

Chen, S.-H. (2018). *Big Data in Computational Social Science and Humanities* (1st ed.). Springer International Publishing : Imprint: Springer,. https://doi.org/10.1007/978-3-319-95465-3

Cheng, W. (2010). What can a corpus tell us about language teaching. *The Routledge handbook of corpus linguistics*, 319-332.

Cheng, W. (2011). *Exploring corpus linguistics: Language in action*. Routledge.

Cheng, W. (2013). Corpus-based linguistic approaches to critical discourse analysis. *The encyclopedia of applied linguistics*, 1353-1360.

Chomsky, N. (2004). Language and mind: current thoughts on ancient problems. In *Variation and universals in biolinguistics* (pp. 379-405). Brill.

Conrad, S. (2002). 4. Corpus linguistic approaches for discourse analysis. *Annual review of applied linguistics*, *22*, 75.

Cowie, A. P. (1998). *Phraseology: Theory, analysis, and applications*. OUP Oxford.

Crawford, W., & Csomay, E. (2015). *Doing corpus linguistics*. Routledge.

Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided error analysis. *System*, *26*(2), 163-174.

Dash, N. S. (2021). *Language Corpora Annotation and Processing*. Springer.

Davies, M. (2005). The advantage of using relational databases for large corpora: Speed, advanced queries, and unlimited annotation. *International Journal of Corpus Linguistics*, *10*(3), 307-334. https://doi.org/10.1075/ijcl.10.3.02dav.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English(1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*(2), 159-190. https://doi.org/10.1075/ijcl.14.2.02dav.

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, *25*(4), 447-464. https://doi.org/10.1093/llc/fqq018

Davies, M. (2012). Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, *7*(2), 121-157.

Davies, M., & Gardner, D. (2013). *A frequency dictionary of contemporary American English: Word sketches, collocates and thematic lists*. Routledge.

Deivasundaram, N. (2021). *Linguistics and Computational Linguistics (Tamil)* (Vol. 2.0). Amutha Nilayam.

Deivasundaram, N., & Gopal, A. (2003). Computational Morphology of Tamil. *Word Structure in Dravidian, Kuppam: Dravidian University*, 406-410.

Department of Statistics, M. (2020, 30.3.22). Department of Statistics, Malaysia. https://www.dosm.gov.my.

Devi, S. L. (2011). Text Extraction for an Agglutinative Language. *Language in India*, *11*(5).

Don, Z. M. (2010). Processing natural Malay texts: A data-driven approach. *Trames*, *14*(1), 90-103.

Duraipandi, R. (2006). The mophological generator and parsing engines of tamil verb forms. *Tamil Internet*.

Ellis, N. C., & Simpson-Vlach, R. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education.

Evangeline, M. M., & Shyamala, K. (2020). Verb Identification Using Morphophonemic Rules In Tamil Language. *ICTACT Journal on Soft Computing*, *11*(1), 2237-2243.

Feng, Z. (2006). Evolution and present situation of corpus research in China. *International Journal of Corpus Linguistics*, *11*(2), 173-207.

Flowerdew, L. (2013). Corpus-based discourse analysis. In *The Routledge handbook of discourse analysis* (pp. 200-214). Routledge.

Francis, G. (1993). A corpus-driven approach to grammar: Principles, methods and examples. *Text and technology: In honour of John Sinclair*, *1*, 137-156.

Francis, W. N., & Kucera, H. (1964). A standard corpus of present-day edited American English, for use with digital computers. *Brown University, Providence*.

Friginal, E., & Hardy, J. (2013). *Corpus-based sociolinguistics: A guide for students*. Routledge.

Garside, R. (1995). Using CLAWS to annotate the British National Corpus. *URL: http://info. ox. ac. uk/bnc/garside_allc. html.*

Gatto, M. (2014). *Web as corpus: Theory and practice*. A&C Black.

Gilquin, G. (2002). Automatic retrieval of syntactic structures. *International Journal of Corpus Linguistics*, *7*(2), 183-214. https://doi.org/10.1075/ijcl.7.2.03gil.

Gilquin, G. (2021). One norm to rule them all? Corpus-derived norms in learner corpus research and foreign language teaching. *Language Teaching*, 1-13.

Godfrey, J. J., & Zampolli, A. (1997). Language resources. *Survey of the state of the art in human language technology*, 101.

Granger, S. (2002). A bird's-eye view of learner corpus research. *Computer learner corpora, second language acquisition and foreign language teaching*, *6*, 3-33.

Grishina, E. (2006). Spoken Russian in the Russian national corpus (RNC). Proceedings of the fifth international conference on language resources and evaluation (LREC'06).

Hadley, G., & Charles, M. (2017). Enhancing extensive reading with data-driven learning. *Language Learning & Technology*, *21*(3), 131-152.

Hanks, P. (2012). The corpus revolution in lexicography. *International Journal of Lexicography*, *25*(4), 398-436.

Hart, G. L. (2015). *Poets of the Tamil Anthologies*. Princeton University Press.

Hart, G. L., & Heifetz, H. (2002). *The Four Hundred Songs of War and Wisdom: An Anthology of Poems from Classical Tamil, the Purananuru*. Columbia University Press.

Hoey, M., & O'Donnell, M. B. (2008). Lexicography, grammar, and textual position. *International Journal of Lexicography*, *21*(3), 293-309.

Hoffmann, S. (2008). *Corpus linguistics with BNCweb: A practical guide*. Peter Lang.

Hoffmann, S., Evert, S., Smith, N., Lee, D., & Berglund-Prytz, Y. (2008). *Corpus linguistics with BNCweb-a practical guide* (Vol. 6). Peter Lang.

Hoffmann, S., Smith, Nicholas,, & Rayson, P. (2008). Corpus tools and methods, today and tomorrow: Incorporating linguists' manual annotations. *Literary and linguistic computing*, *23*(2), 163-180.

Huang, C.-R., & Chen, K.-j. (2010). Academia sinica balanced corpus of modern Chinese 4.0. *Academia Sinica. RetrievedJanuary*, *13*, 2016.
Huat, C. M. (2012). Learner corpora and second language acquisition. *Corpus applications in applied linguistics*, 191-207.

Hudson, P. T., & Buijs, D. (1995). Left-to-right processing of derivational morphology. *Morphological aspects of language processing*, 383-396.

Hundt, M., Nesselhauf, N., & Biewer, C. (2007). *Corpus linguistics and the web*. Brill.

Hunston, S. (1995). Grammar in teacher education: The role of a corpus. *Language Awareness*, *4*(1), 15-31.

Hunston, S. (2002a). *Corpora in applied linguistics*. Ernst Klett Sprachen.

Hunston, S. (2002b). Pattern grammar, language teaching, and linguistic variation. *Using corpora to explore linguistic variation*, 167-183.

Hunston, S. (2008). Starting with the small words: Patterns, lexis and semantic sequences. *International Journal of Corpus Linguistics*, *13*(3), 271-295.

Hunston, S., & Gill, F. (1998). Verbs observed: A corpus-driven pedagogic grammar1. *Applied linguistics*, *19*(1), 45-72.

Hurskainen, A. (2003). New advances in corpus-based lexicography. *Lexikos*, *13*.

Ide, N. (2003). The American National Corpus: Everything you always wanted to know... and weren't afraid to ask. *Invited keynote, Corpus Linguistics*.

Ilson, R. (1982). The Survey of English Usage: past, present—and future. *ELT Journal*, *36*(4), 242-247.

Janda, R. D., Joseph, B. D., & Vance, B. S. (2020). *The Handbook of Historical Linguistics, Volume II*. John Wiley & Sons.

Jayan, J. P., Rajeev, R., & Rajendran, S. (2011). Morphological analyser and morphological generator for Malayalam-Tamil machine translation. *International Journal of Computer Applications*, *13*(8), 0975-8887.

Jenset, G. B., & McGillivray, B. (2017). *Quantitative historical linguistics: A corpus framework* (Vol. 26). Oxford University Press.

Joharry, S. A., & Rahim, H. A. (2014). Corpus research in Malaysia: a bibliographic analysis. *Kajian Malaysia*, *32*(1), 17.

Johns, T. (1991). *Should you be persuaded: Two samples of data-driven learning materials* (Vol. 4). na.

Jurafsky, D. (2009). MJ: Speech and Language Processing. Saddle River. In (pp. 45-46): NJ, USA: Pearson/Prentice Hall.

Jurafsky, G., Daniel. (2002). Automatic labeling of semantic roles. *Computational linguistics*, *28*(3), 245-288.

Jurafsky, M., Daniel, Taylor, P., Rachel, Ess-Dykema, C. V., Meteer, M., Stolcke, Andreas, Ries, K., Coccaro, N., Shriberg, E., & Bates. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, *26*(3), 339-373.

Kachru, Y. (2008). Language variation and corpus linguistics. *World Englishes*, *27*(1), 1-8.

Kaur, J. (2010). Achieving mutual understanding in world Englishes. *World Englishes*, *29*(2), 192-208.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. Addison Wesley Longman.

Knowles, G., Don, Z. M., Jan, J. M., Sargunan, R. A., Yong, J., Devi, S., Doshi, A., & ad Awab, S. (2006). The Malaysian Corpus of Learner English: a bridge from linguistics to ELT. In: Varieties of English in Southeast Asia and beyond. Kuala Lumpur University.

Koskenniemi, K. (1984). A general computational model for word-form recognition and production. Proceedings of the 4th Nordic Conference of Computational Linguistics (NODALIDA 1983).

Kothandaraman, P. (1997). *A grammar of contemporary literary Tamil*. Int. Inst. of Tamil Studies.

Kothandaraman, R. (2006). Beginnings of the Writing System in Tamil. *Negotiations with the Past: Classical Tamil in Contemporary Tamil*, *103*, 95.

Krishnamurthy, R. (2008). Corpus-driven lexicography. *International Journal of Lexicography*, *21*(3), 231-242.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics*, *16*(3), 307-322.

Lee, D. Y. (2001). Genres, registers, text types, domains and styles: Clarifying the concepts and nevigating a path through the BNC jungle.

Lee, L., & Low, H. (2011). Developing an online Malay language word corpus for primary schools. *International Journal of Education and Development Using ICT*, *7*(3).

Leech, G. (1999). The distribution and function of vocatives in American and British English conversation. *Language and Computers*, *26*, 106.

Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus Linguistics and the Web* (pp. 133-149). Rodopi.

Leech, G. N. (1992). 100 million words of English: the British National Corpus (BNC). *어학연구*.

Léon, J. (2005). Claimed and unclaimed sources of corpus linguistics. *Henry Sweet Society for the History of Linguistic Ideas Bulletin*, *44*(1), 36.

Liesenfeld, A. M. (2018). The Use of Janwai in the Management of Disagreement in Malaysian Cantonese Conversation: Evidence from MYCanCor Corpus. [19th Chinese Lexical Semantics Workshop](CLSW 2018), 26-28 May, National Chung Cheng University, Taiwan.

Lindquist, H. (2018). *Corpus linguistics and the description of English*. Edinburgh University Press.

Lu, X. (2014). *Computational methods for corpus annotation and analysis*. Springer.

Lushanthan, S., Weerasinghe, A., & Herath, D. (2014). Morphological analyzer and generator for Tamil language. 2014 14th International Conference on Advances in ICT for Emerging Regions (ICTer).

Mahadi, T. S. T., Vaezian, H., & Akbari, M. Universiti Sains Malaysia.

Mahlberg, M. (2005). English general nouns. *A corpus theoretical approach*, 370.

Malamatidou, S. (2017). *Corpus triangulation: Combining data and methods in corpus-based translation studies*. Routledge.

Mann, D., Weston, N., Frederic, K., Ogunshile, E., & Ramachandran, R. (2019). Tamil talk: What you speak is what you get! 2019 7th International Conference in Software Engineering Research and Innovation (CONISOFT).

Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

McCarthy, M., & Carter, R. (2001). Size isn't everything: Spoken English, corpus, and the classroom. *TESOL Quarterly*, *35*(2). https://doi.org/10.2307/3587654

McCarthy, M., & Carter, R. (2001). Size isn't everything: spoken English, corpus, and the classroom. *Tesol Quarterly*, *35*(2), 125.

McCarthy, M., & Carter, R. (2004). *TEANGA, the Journal of the Irish Association for Applied Linguistics*, *21*(null), 30.

McCarthy, M., & McCarten, J. (2012). Corpora and materials design. *Corpus applications in applied linguistics*, 225-241.

McEnery, A., & Xiao, R. (2003). The Lancaster Corpus of Mandarin Chinese.

McEnery, T., & Hardie, A. (2012a). *Corpus linguistic: Method, theory and practice*. Cambridge University Press.

McEnery, T., & Hardie, A. (2012b). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

McEnery, T., & Xiao, R. (2007). Parallel and comparable corpora: What is happening? In *Incorporating corpora* (pp. 18-31). Multilingual Matters.

McEnery, T., & Xiao, R. (2011). What corpora can offer in language teaching and learning. In *Handbook of research in second language teaching and learning* (pp. 382-398). Routledge.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Taylor & Francis.

Menaka, S., Ram, V. S., & Devi, S. L. (2010). Morphological generator for Tamil. *Proceedings of the Knowledge Sharing event on Morphological Analysers and Generators (March 22-23, 2010), LDC-IL, Mysore, India*, 82-96.

Mindt, D. (1995). *An empirical grammar of the English verb: Modal verbs*. Cornelsen Berlin.

Mitkov, R. (2004). *The Oxford handbook of computational linguistics*. Oxford University Press.

Mohanlal, S., Sharada, B., Fatihi, A., Gusain, L., Bayer, J. M., Ravichandran, S., Baskaran, G., Ramamoorthy, L., Subburaman, C., & Thirumalai, S. Parsing Noun Inflections: Tamil.

Monaghan, P., & Rowland, C. F. (2017). Combining language corpora with experimental and computational approaches for language acquisition research. *Language Learning*, *67*(S1), 14-39.

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. John Benjamins Publishing Company.

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. *Vocabulary: Description, acquisition and pedagogy*, *14*, 6-19.

Nelson, M. (2010). Building a written corpus: what are the basics? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics*. Routledge.

O'keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge University Press.

O’Keeffe, A. (2012). Corpora and media studies. *Corpus applications in applied linguistics*, 441-454, Continuum international Publishing Group.

Oostdijk, N. (1991). *Corpus linguistics and the automatic analysis of English*. Rodopi.

Parameshwari, K. (2011). An implementation of APERTIUM morphological analyzer and generator for Tamil. *Parsing in Indian Languages*, 41.

Pon, K. (1993). An Introduction to the Tamil Language and Literature. *学習院大学言語共同研究所紀要*, *16*, 71-86.

Pullum, G. K., & Huddleston, R. (2002). Adjectives and adverbs. *The Cambridge grammar of the English language*, 525-595.

Quirk, R. (1990). Language varieties and standard language. *English today*, *6*(1), 3-10.

Rahim, H. A. (2005). Impak konotasi budaya terhadap leksis: satu kajian semantik berasaskan korpus, ke atas perkataan perempuan dan wanita. *Jurnal Bahasa*, *5*(1), 83-111.

Rahim, H. A. (2014). Corpora in language research in Malaysia. *Kajian Malaysia*, *32*(1), 1.

Rajan, K., Ramalingam, V., & Ganesan, M. (2012). Machine Learning of Sandhi Rules for Tamil. Proceedings of the 11th International Conference INFITT,

Rajendran, S. (2006). Parsing in tamil: Present state of art. *Language in India*, *6*, 8.

Rajendran, S., Arulmozi, S., Shanmugam, B. K., Baskaran, S., & Thiagarajan, S. (2002). Tamil wordnet. Proceedings of the first international global WordNet conference. Mysore.

Rajendran, S., Viswanathan, S., & Kumar, R. (2003). Computational morphology of Tamil verbal complex. *Language in India*, *3*(4).

Ramasamy, L., & Žabokrtský, Z. (2011). Tamil dependency parsing: results using rule based and corpus based approaches. International Conference on Intelligent Text Processing and Computational Linguistics.

Ramaswamy, V. (2000). Morphological Generator for Tamil. *M. Phil Dissertation Submitted to University of Hyderabad*.

Ramaswamy, V. (2003). A morphological analyzer for Tamil. *PhD Dissertation Submitted to University of Hyderabad.*

Rayson, P. E. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison* Lancaster University].

Rayson, P. E. (2015). Computational tools and methods for corpus compilation and analysis. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 32-49). Cambridge University Press.

Renganathan, V. (2016). *Computational Approaches to Tamil Linguistics*. Cre-A.

Richardson, S. D. (1994). Bootstrapping statistical processing into a rule-based natural language parser. The Balancing Act: Combining Symbolic and Statistical Approaches to Language.

Rissanen, M., Kytö, M., Kahlas-Tarkka, L., Kilpiö, M., Nevanlinna, S., Taavitsainen, I., Nevalainen, T., & Raumolin-Brunberg, H. (1993). The helsinki corpus of english texts. *Kyttö et. al*, 73-81.

Samuel, J. G. (1994). Preservation of palm-leaf manuscripts in Tamil. *IFLA journal*, *20*(3), 294-305.

Saravanan, B. (1999). *Morphotactics: patterns in Tamil morphology*. State University of New York at Stony Brook.

Sardinha, A. B. (1996). WordSmith tools. *Computers & Texts 12 (1996)*.

Sarveswaran, K., Dias, G., & Butt, M. (2021). ThamizhiMorph: A morphological parser for the Tamil language. *Machine Translation*, *35*(1), 37-70.

Scott, M. (2010). What can corpus software do? In A. O'Keeffe & M. J. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 136-151). Routledge.

Scott, M. (2016). *WordSmith Tools*. In Lexical Analysis Software.

Selby, M. A. (2019). Tamil: A Biography. In: JSTOR.

Shapiro, M. C., & Schiffman, H. F. (2019). *Language and society in South Asia*. De Gruyter Mouton.

Shastri, S. (1988). Code mixing in the process of Indianization of English: a corpus based study. *Indian Linguistics*, *49*(1-4), 34-53.

Sheshasaayee, A., & Deepa, V. A. (2016). A Conceptual Model for Acquisition of Morphological Features of Highly Agglutinative Tamil Language Using Unsupervised Approach. In *Information Systems Design and Intelligent Applications* (pp. 499-507). Springer.

Sheshasaayee, A., & VR, A. D. (2015). Morpheme Segmentation for Highly Agglutinative Tamil Language by Means of Unsupervised Learning. *International Journal of Computer Applications*, *975*, 8887.

Sinclair, J. (2004). Intuition and annotation–the discussion continues. In *Advances in corpus linguistics* (pp. 39-59). Brill Rodopi.

Sinclair, J. (2005). Meaning in the framework of corpus linguistics. *Lexicographica (2004)*, *20*(2005), 20-32.

Sinclair, J., & Sinclair, L. (1991). *Corpus, concordance, collocation*. Oxford University Press, USA.

Sinclair, J. M. (1991). *Corpus, concordance and collocation*. Oxford University Press.

Sinclair, J. M. (1996). The empty lexicon. *International Journal of Corpus Linguistics*, *1*(1), 99-119.

Sinclair, J. M. (2003). 4.2 Corpus processing. In *A practical guide to lexicography* (pp. 179-193). John Benjamins.

Sinclair, J. M. (2014). Corpus evidence in language description. *Teaching and language corpora*, (pp. 27-39). Routledge.

Steever, S. B. (2018). Tamil and the Dravidian languages. In *The world's major languages* (pp. 653-671). Routledge.

Stubbs, M. (1993). British traditions in text analysis. *Text and technology: In honour of John Sinclair*, 23-24.

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Blackwell publishers Oxford.

Taylor, C. (2008). What is corpus linguistics? What the data says. *ICAME journal*, *32*, 179-200.

Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics*, *10*(1), 1-13.

Thirumalai, M. (2004). Tradition, modernity, and impact of globalization-Whither will Tamil go. *Language in India. Volume4*, *1*(1).

Thompson, G., & Hunston, S. (2006). System and corpus: Two traditions with a common ground. *System and corpus: Exploring connections*, 8.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. John Benjamins Publishing Company.

Tyler, A. (2010). Usage-based approaches to language and their applications to second language learning. *Annual Review of Applied Linguistics*, *30*, 270-291. https://doi.org/10.1017/s0267190510000140.

Veerappan, R., Antony, P., Saravanan, S., & Soman, K. (2011). A rule based kannada morphological analyzer and generator using finite state transducer. *International Journal of Computer Applications*, *27*(10), 45-52.

Weisser, M. (2016). *Practical corpus linguistics: An introduction to corpus-based language analysis* (Vol. 43). John Wiley & Sons.

Widdowson, H. G. (2007). *Discourse analysis* (Vol. 133). Oxford University Press Oxford.

Winskel, H. (2020). Learning to read in multilingual Malaysia: A focus on Bahasa Melayu, Tamil and Chinese. *GEMA Online Journal of Language studies*, *20*(1).

Xiao, R., McEnery, A., Baker, J., & Hardie, A. (2004). Developing Asian language corpora: standards and practice. The 4th Workshop on Asian Language Resources.

Zvelebil, K. (1960). Dialects of Tamil III. *Archiv Orientalni*, *28*(3), 414-456.

Zvelebil, K. (1973). *The Smile of Murugan: On Tamil Literature of South India*. Brill.

Zvelebil, K. (1974). *Tamil literature*. Otto Harrassowitz Verlag.

Zvelebil, K. (1992). *Companion studies to the history of Tamil literature* (Vol. 5). Brill.