Corpus Management and Multi-lingual Lexical Database Tools

CDAC, Thiruvananthapuram

Abstract

In this paper we will discuss about the tools for corpus management and Multi-lingual lexical database which were used for English to Malayalam translation system called AnglaMalayalam. AnglaMalayalam is an adaptation of AnglaBharti technology which was developed by IIT Kanpur.

1. Introduction

Corpus is a body of language. Corpus is an inevitable resource in Natural Language Computing. It is widely used to study the behaviour of Natural Languages, develop statistical language computing rules and building lexical resources. In the field of Machine Translation, a good quality corpus is very much essential for good quality translation. Thus corpus collection and its analysis form an essential task. In the following section we will discuss about the tools used for lexicon updation and domain specific lexicon entry for AnglaMalayalam system.

2. Corpus Collection and Analysis

The corpus is acquired from web resources, Journals, Magazines, text books, etc. The collected corpus had to be processed for analysis. The process includes cleaning, unique word Identification, sentence boundary marking, parsing, extracting word and POS, extracting bi-gram and tri-gram and extraction of NP and VP.

The corpus which is collected from web may contain certain junk characters, extra spaces after the words or sentences. Before analysing the corpus, such characters and spaces should be removed properly. The abbreviations and acronyms were extracted from the cleaned corpus based on certain heuristics. Accurate sentence splitting is an important building block of many NLP systems. Most Part of Speech taggers require input in the form of one sentence per line. We used a sentence tokenization tool developed in perl with a standard CPAN Module Lingua::EN::Sentence.

2.1 Unique Word Extraction

From the raw corpus, word and frequency count will be taken. Then the word is compared with the main lexicon. The compared words are subjected to the stemming process. Stemming is mainly done to remove the plural markers. After stemming process the words are again compared with the main lexicon. If the word is not found in the lexicon then it will be stored in to a separate file. Manual verification is required after the process to eliminate the nonsensical words.

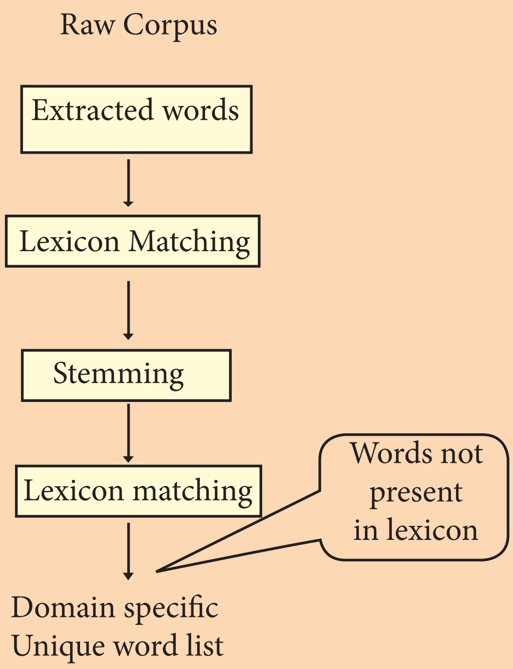
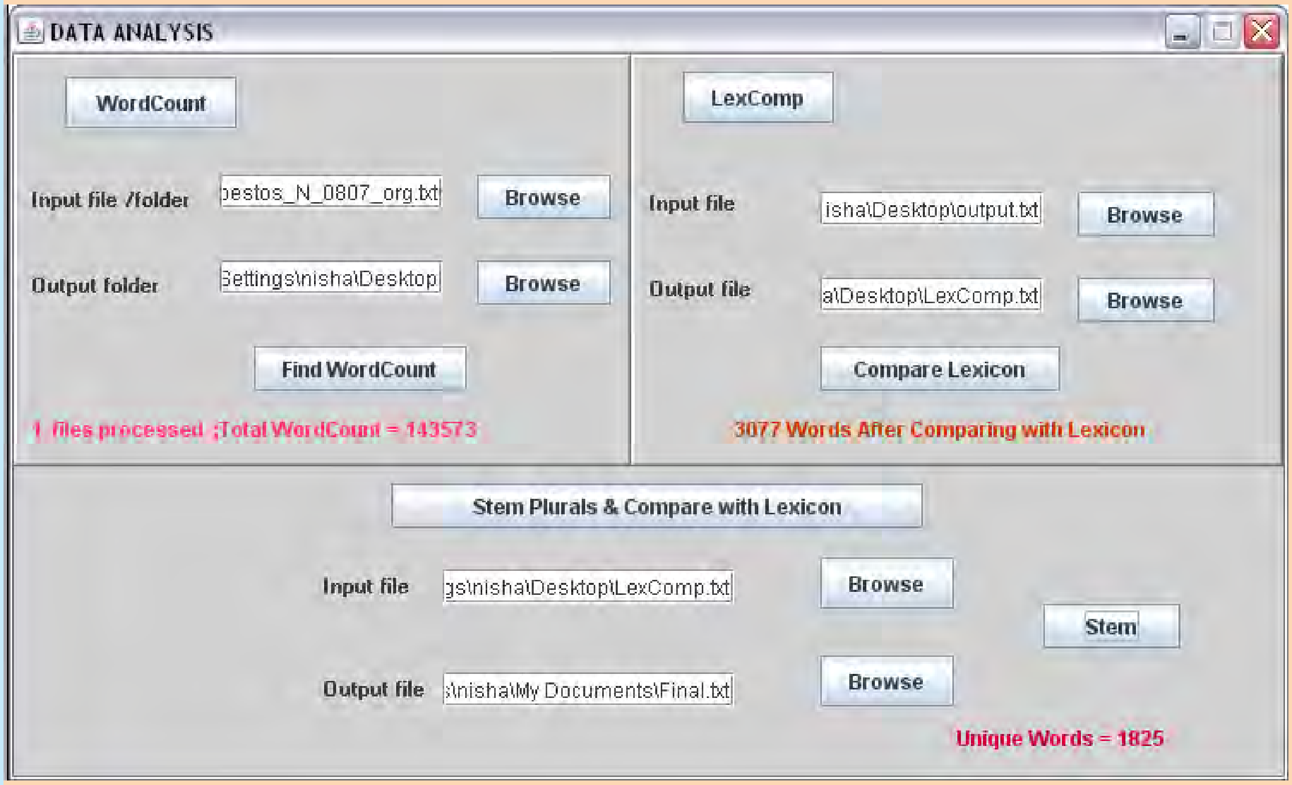


Figure 1: Flow Chart for Unique Word Extraction Tool



Input file /folder: oestos_N_0807_org.txt

Output folder: Settings\nisha\Desktop

1 files processed ;Total WordCount = 143573

Input file: isha\Desktop\output.txt

Output file: a\Desktop\LexComp.txt

3077 Words After Comparing with Lexicon

Input file: gs\nisha\Desktop\LexComp.txt

Output file: s\nisha\My Documents\Final.txt

Unique Words = 1825

Figure 2: Screenshot for Unique Word Extraction Tool

2.2 Multiword Extraction Tools

2.2.1 Sentence Boundary Detection

The Sentence Boundary Detection tool is developed in perl with a standard CPAN Module Lingua::EN::Sentence. The module is an implementation of standard sentence boundary detection algorithm.

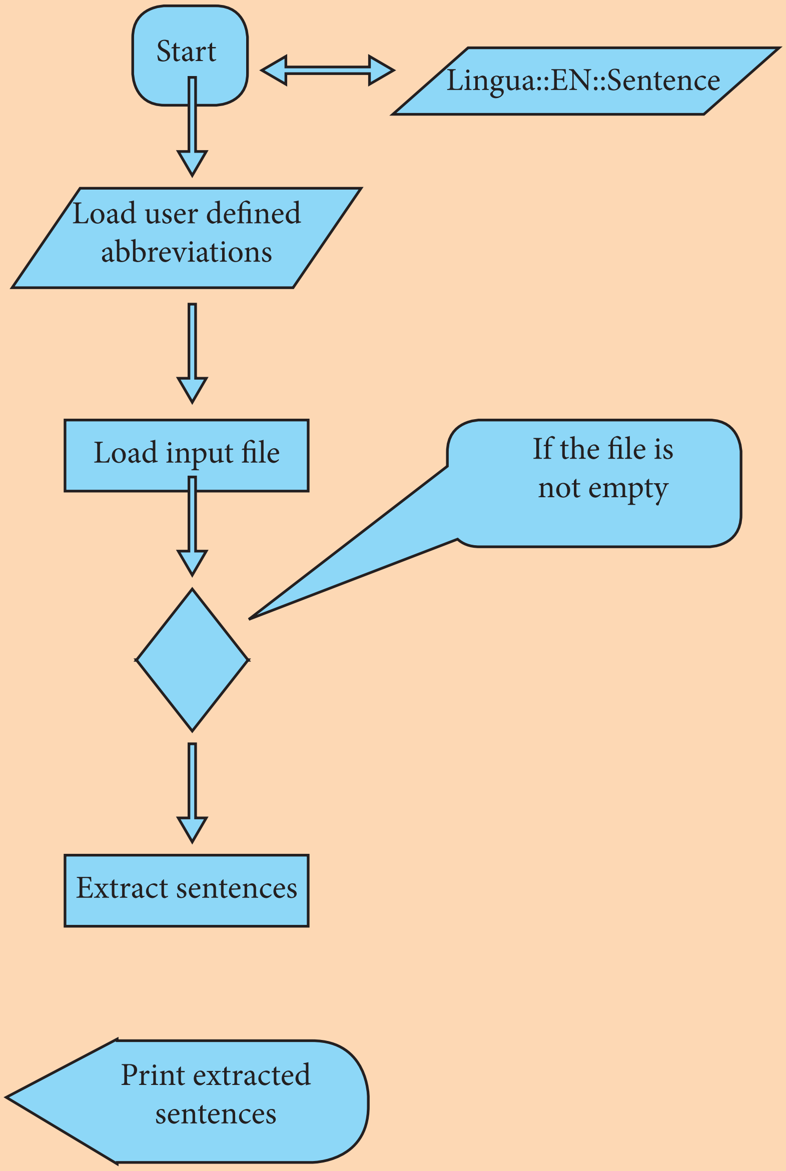


Figure 3: Flow Chart for Sentence Boundary Detection

2.2.2 Named Entity Extraction

The Named Entity extraction tool developed in Perl with a Standard CPAN (Comprehensive Perl Archive) Module called Lingua::EN::NamedEntity. It is a perl implementation of basic Named Entity Recognition Algorithm. The implementations can be installed in the system to use the algorithm in perl programs.

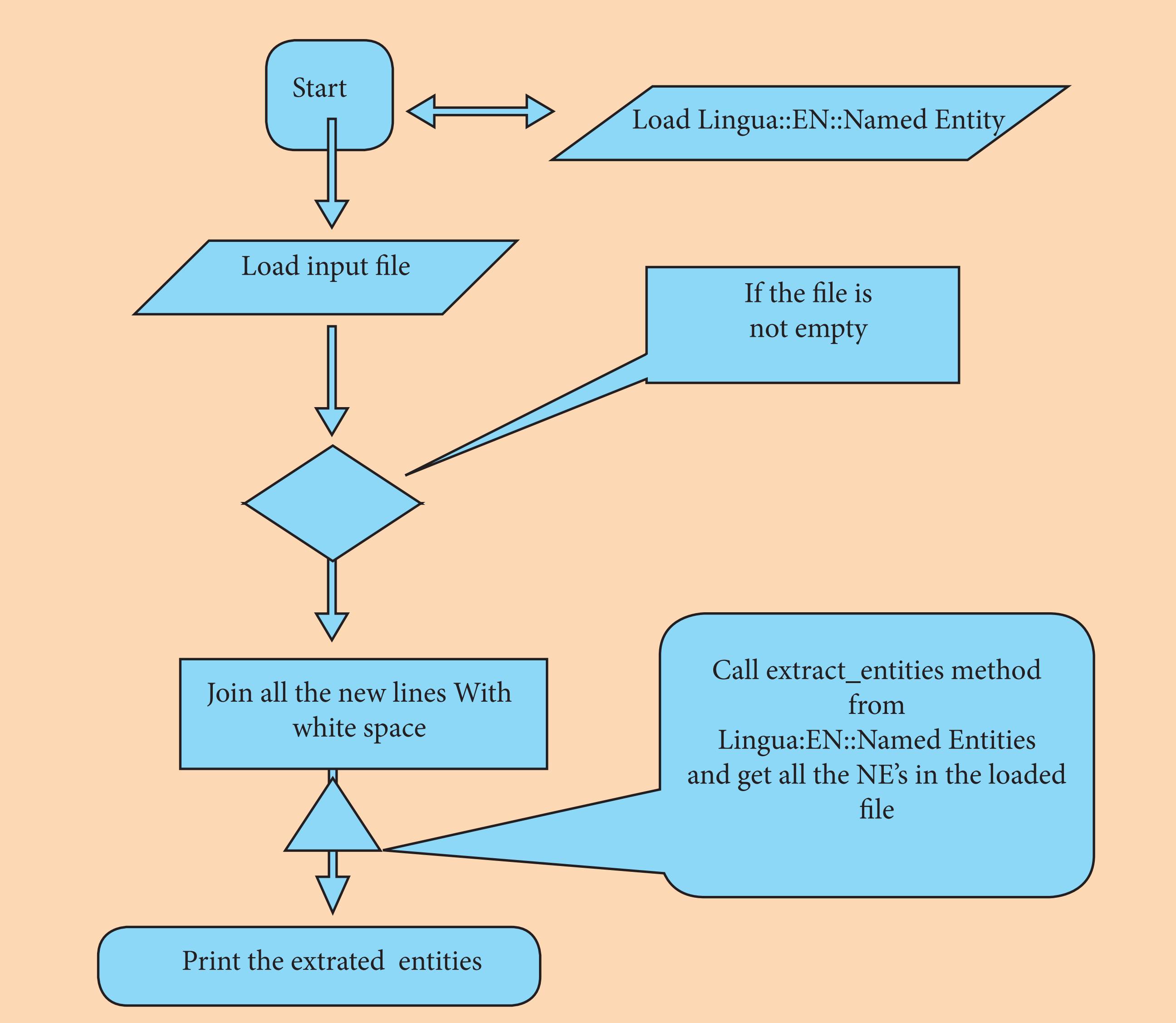

Start
Load Lingua::EN::Named Entity
Load input file
If the file is not empty
Join all the new lines With white space
Call extract_entities method from Lingua:EN::Named Entities and get all the NE's in the loaded file
Print the extrated entities

Figure 4: Flowchart for Named Entity Recognition

2.2.3 Parser

The tool is developed with Charniak parser and Perl programming language. Charniak parser is a statistical parser for English language. The parser can be downloaded from ftp://ftp.cs.brown.edu/pub/nlparser/ under the GNU GP license. The Perl program takes absolute path to sentence boundary marked corpus. It will take the files one by one. Processes and stores the output to the directory parsed.

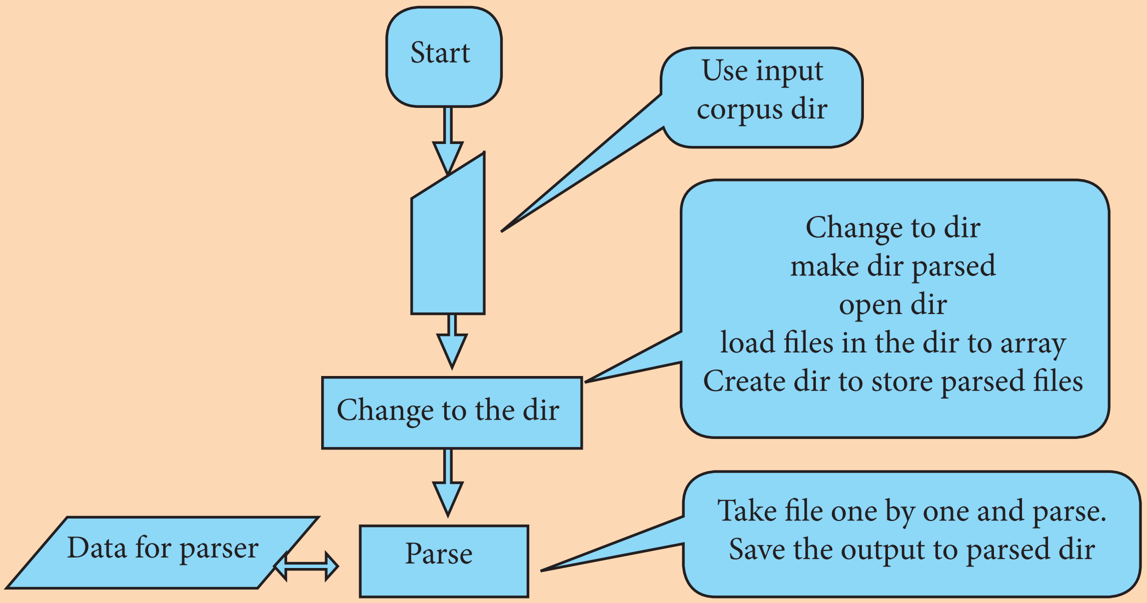

Start
Use input corpus dir
Change to dir
make dir parsed
open dir
load files in the dir to array
Create dir to store parsed files
Change to the dir
Data for parser
Parse
Take file one by one and parse.
Save the output to parsed dir

Figure 5: Flowchart for the Parser

2.2.4 MWE extraction

The program is developed in perl. It takes a java bigram and Trigram tool; which extracts bigrams and Trigrams along with POS from the parsed corpus. The tool is divided in to four different perl files. It recieves the absolute path to parsed corpus file and extracts NP word and POS, word and POS, bigram and trigaram and MWES.

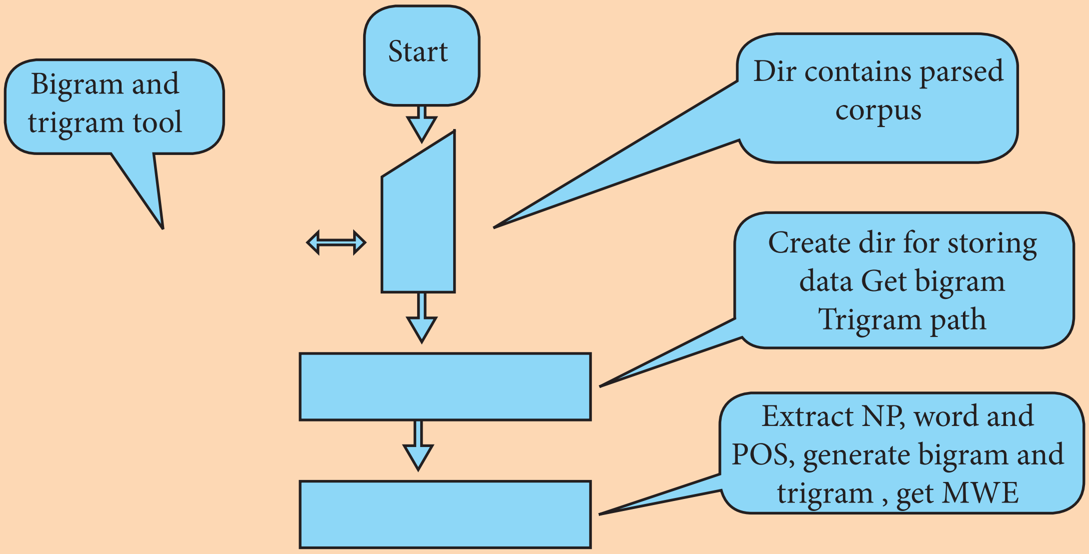
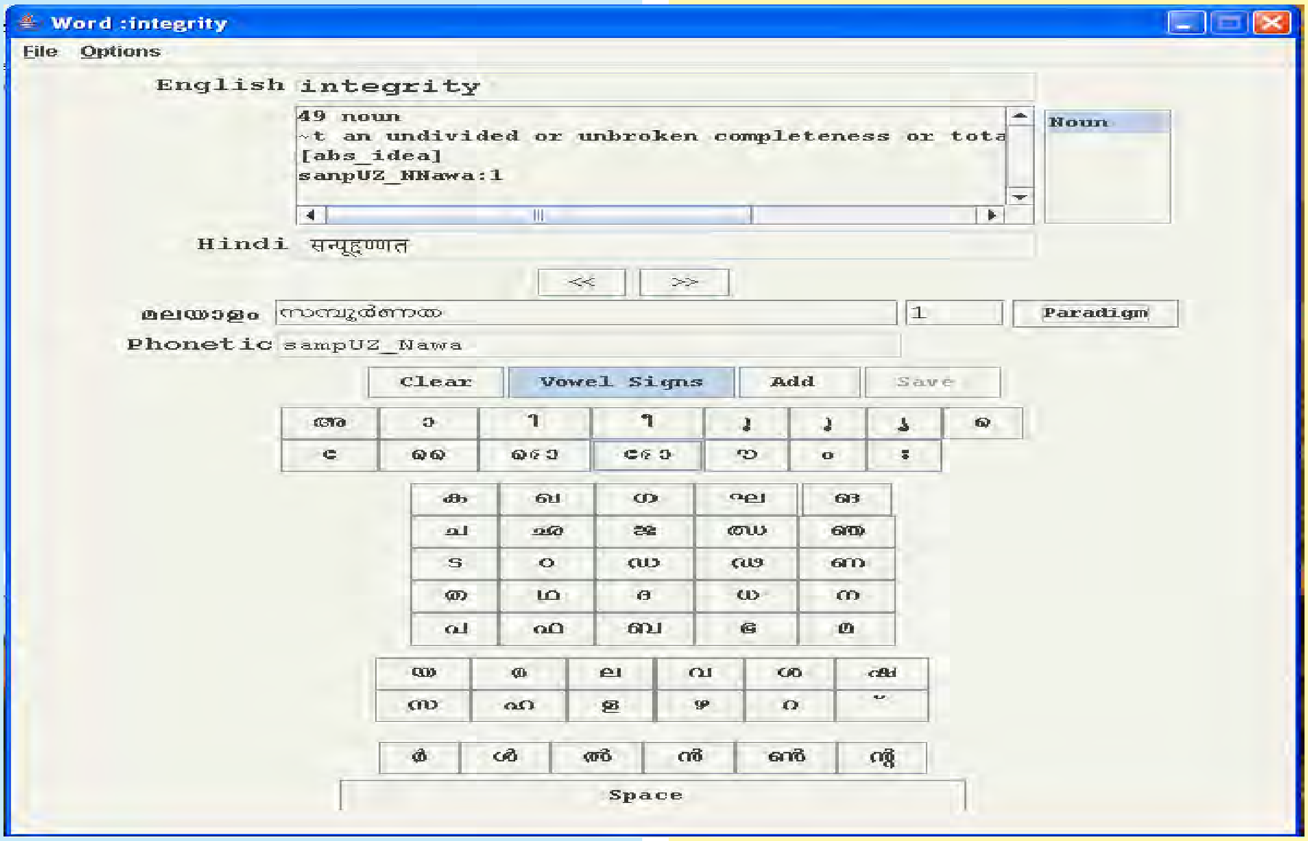

Bigram and trigram tool
Start
Dir contains parsed corpus
Create dir for storing data Get bigram Trigram path
Extract NP, word and POS, generate bigram and trigram , get MWE

Figure 6: Flowchart for Multiword Extraction

3. Lexicon Updation Tool

A lexicon is usually structured as a collection of English words along with their syntactic and semantic information. The syntactic information includes the POS, Tense, aspect, Modality, etc. Lexicon is one of the major parts in English to Malayalam translation process. We have used English – TL lexicon as a reference to develop English – Malayalam lexicon. Since the English – TL lexicon contains a round 45000 words. Manual entry will affect the dictionary structure. So a tool is developed to replace the

Here the user can enter the meaning by using the online keyboard provided along with the tool or he can use the inscript keyboard. The paradigm number can be assigned by using the different combinations of inflections for a particular category of word provided by the system. We can choose the apt paradigm number through manual selection. The pop up coming upon clicking the Paradigm button is given below.



English integrity

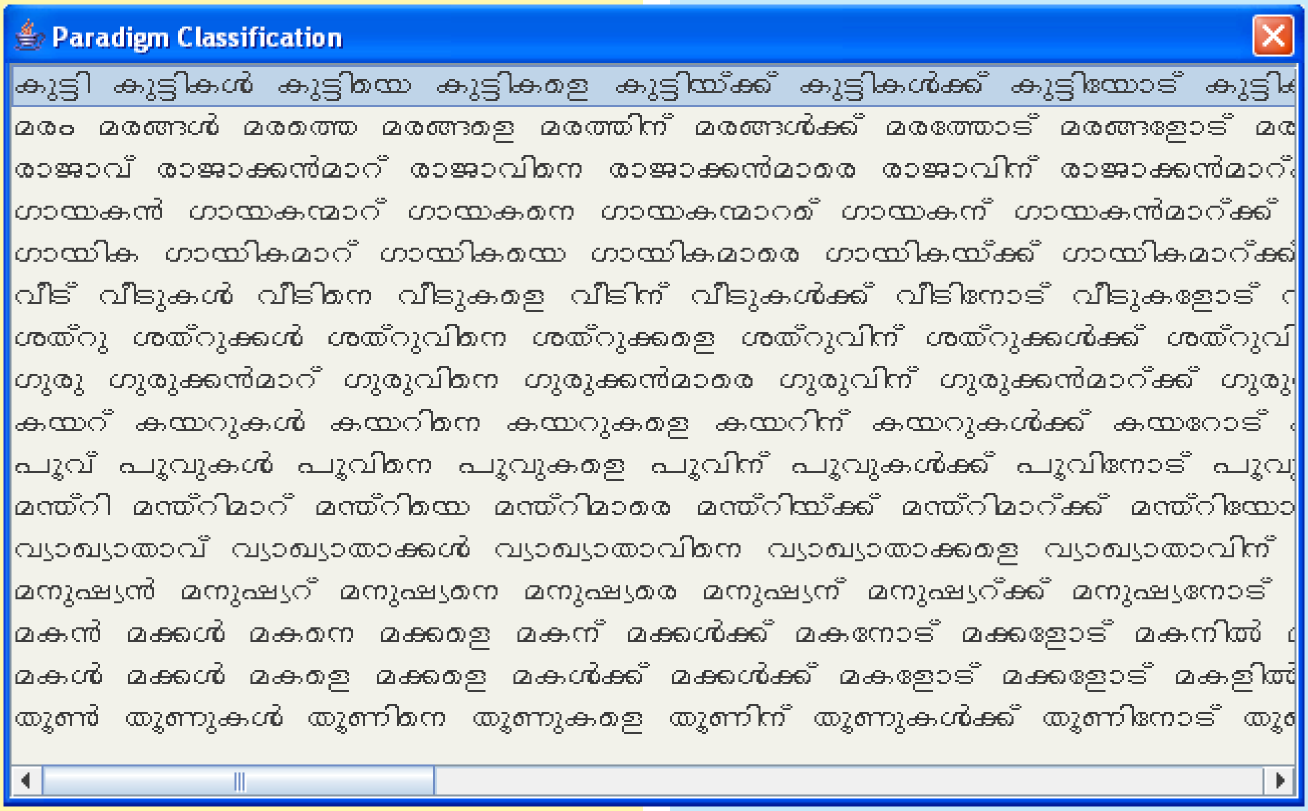
```
49 noun
~t an undivided or unbroken completeness or tota
[abs_idea]
sanpUZ_NNawa:1
```

Hindi सम्पूर्णता

മലയാളം സമ്പൂർണ്ണത 1

Phonetic sampUZ_Nawa

Figure 7: Lexicon Updation Tool



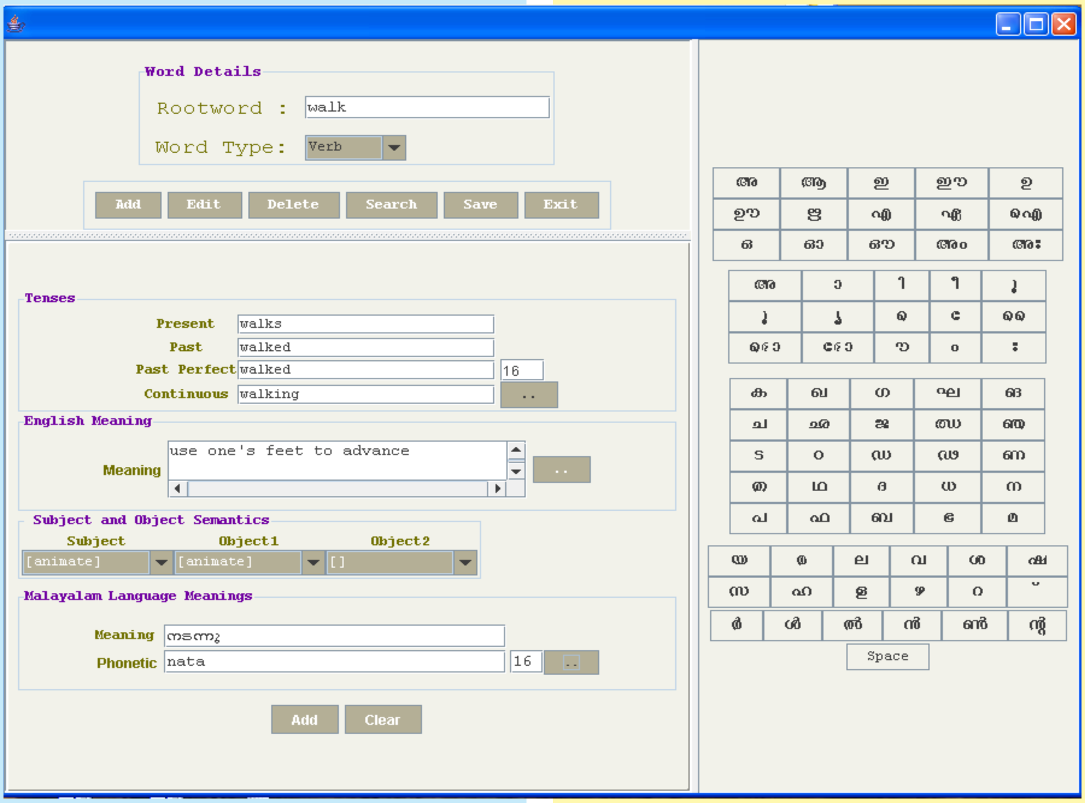
കുട്ടി കുട്ടികൾ കുട്ടിയെ കുട്ടികളെ കുട്ടിയ്ക്ക് കുട്ടികൾക്ക് കുട്ടിയോട് കുട്ടിക
മരം മരങ്ങൾ മരത്തെ മരങ്ങളെ മരത്തിന് മരങ്ങൾക്ക് മരത്തോട് മരങ്ങളോട് മര
രാജാവ് രാജാക്കൻമാറ് രാജാവിനെ രാജാക്കൻമാരെ രാജാവിന് രാജാക്കൻമാറ്
ഗായകൻ ഗായകന്മാറ് ഗായകനെ ഗായകന്മാറത് ഗായകന് ഗായകൻമാറ്ക്ക്
ഗായിക ഗായികമാറ് ഗായികയെ ഗായികമാരെ ഗായികയ്ക്ക് ഗായികമാറ്ക്ക്
വീട് വീടുകൾ വീടിനെ വീടുകളെ വീടിന് വീടുകൾക്ക് വീടിനോട് വീടുകളോട്
ശത്റു ശത്റുക്കൾ ശത്റുവിനെ ശത്റുക്കളെ ശത്റുവിന് ശത്റുക്കൾക്ക് ശത്റുവി
ഗുരു ഗുരുക്കൻമാറ് ഗുരുവിനെ ഗുരുക്കൻമാരെ ഗുരുവിന് ഗുരുക്കൻമാറ്ക്ക് ഗുരു
കയറ് കയറുകൾ കയറിനെ കയറുകളെ കയറിന് കയറുകൾക്ക് കയറോട്
പൂവ് പൂവുകൾ പൂവിനെ പൂവുകളെ പൂവിന് പൂവുകൾക്ക് പൂവിനോട് പൂവു
മന്ത്റി മന്ത്റിമാറ് മന്ത്റിയെ മന്ത്റിമാരെ മന്ത്റിയ്ക്ക് മന്ത്റിമാറ്ക്ക് മന്ത്റിയോ
വ്യാഖ്യാതാവ് വ്യാഖ്യാതാക്കൾ വ്യാഖ്യാതാവിനെ വ്യാഖ്യാതാക്കളെ വ്യാഖ്യാതാവിന്
മനുഷ്യൻ മനുഷ്യറ് മനുഷ്യനെ മനുഷ്യരെ മനുഷ്യന് മനുഷ്യറ്ക്ക് മനുഷ്യനോട്
മകൻ മക്കൾ മകനെ മക്കളെ മകന് മക്കൾക്ക് മകനോട് മക്കളോട് മകനിൽ
മകൾ മക്കൾ മകളെ മക്കളെ മകൾക്ക് മക്കൾക്ക് മകളോട് മക്കളോട് മകളിൽ
തൂൺ തൂണുകൾ തൂണിനെ തൂണുകളെ തൂണിന് തൂണുകൾക്ക് തൂണിനോട് തൂ

Figure 8: Inflection for nouns

4. Domain specific Lexical Entry

Lot of new words are extracted from the corpus and extracted domain specific words. Domain specific lexical entry tool was used to create domain specific lexicon. This tool can also be used to update the existing lexicon. The lexicon is having a specific format, which contains syntactic and semantic information. The tool is an XML based tool which has paradigm generators for Malayalam and English words. Since it is a domain specific lexical entry tool, there is a provision for selecting the domain like General, Health or Tourism. The other options are for entering the root word and POS tag of the word. Also, there were provisions for adding, editing, deleting and searching the desired root word and its information's.

In the Noun entry field, we have provided options for entering the details such as gender (masculine/feminine/neuter/don't care), number (singular/plural), person (first/second/third) and also the Target language details. In the field of Target language details, there were provisions for entering the meaning, semantic tag and paradigm number. For Adjective entries, there were provisions for entering the degrees of comparison (positive/comparative/superlative) and Target language details as mentioned above. For Verb entries, options were there for entering the tense forms (present/past/past participle/continuous), transitive, intransitive and bitransitive details. In the case of transitive verbs, enter the meaning, subject and object and paradigm number. In the case of intransitive verbs, enter the meaning, subject and paradigm number and for bitransitive verbs, enter the meaning, subject, first object, second object and paradigm number. From the below figure (9), we can see that the English paradigm "16" has been generated based on the tenses, which is strictly based on certain heuristics used in AnglaBharti technology. We have displayed available English meanings of the root word. The user can select the correct one which is similar in meaning to the root word. For this purpose, we have incorporated a dictionary of Word Net for getting the matching meanings corresponding to the words. The tool has the provisions to enter the Malayalam meaning using the online keyboard. The apt paradigm number for Malayalam can be assigned by using the different combinations of inflections for a particular category of word provided by the System.



Word Details

Rootword : walk

Word Type: Verb

Tenses

Present walks

Past walked

Past Perfect walked 16

Continuous walking

English Meaning

Meaning use one's feet to advance

Subject and Object Semantics

Subject	Object1	Object2
[animate]	[animate]	[]

Malayalam Language Meanings

Meaning നടന്നു

Phonetic nata 16

Figure 9: Lexicon Entry Tool

5. Summary

In this paper we have discussed about the tools for corpus management and multi-lingual lexicon database. We have seen the various tools used for corpus cleaning; identify unique words, sentence boundary marking, parsing, extracting word and POS, extracting bi-gram and tri-gram. We have also discussed about the tool for multi-lingual lexical database creation.

Abbreviations / Acronyms

NP: Noun Phrase, **VP:** Verb Phrase, **PP:** Preposition Phrase, **POS:** Part of Speech, **XML:** Extended Mark-up Language, **CPAN:** Comprehensive perl archive, **NN:** Noun singular, **NNS:** Noun plural, **NNP:** Proper Noun Singular