An Improvized Morphological Analyzer cum Generator for Tamil : A case of implementing the open source platform *APERTIUM*

Parameswari K,
Centre for Applied Linguistics and Translation Studies,
University of Hyderabad.
parameshkrishnaa@gmail.com

Overview

- Morphological Analyzer and Generator - two crucial tools in MT involving NLP.
- Deals with the improvized database implemented on Apertium for morphological analysis and generation.
- Discusses the evaluation of tools with large corpora to estimate the efficacy , coverage and speed.

• A Morphological analyzer is a computational tool to analyze word forms into their roots, categories along with their constituent functional elements and the generator is an inverse of it.

•The attempt involves a practical adoption of *lttoolbox* for the **Modern Standard Written Tamil** in order to develop an improvized open source Morphological Analyzer and generator.

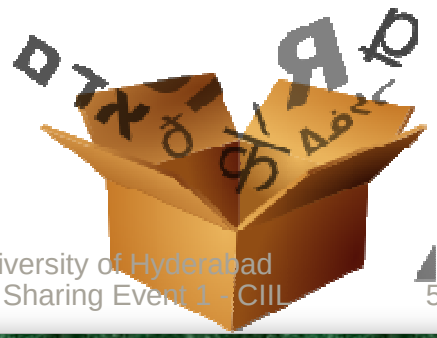
• The tool uses the computational algorithm **Finite State Transducers** for one-pass analysis and generation, and the database is developed in the morphological model called **Word and Paradigm .**

Need for improvized Morphological tools

- Open Source - Easily Accessible
- Handling Derivation Morphology
- Speed
- User friendly

APERTIUM -‘lttoolbox’

- Developed by the Tran`sducens research group at the **Universitat d’Alacant in Spain.**
- One of the open source machine translation systems has originated within the project **“Open-Source Machine Translation for the languages of Spain”.**
- A componant called ‘lttoolbox’ for performing lexical processing tasks of language like Morphological Analyzer, Generator and POS Tagger.



Downloading…

- The current versions of the Apertium toolbox as well as language data are available from the **Sourceforge page** sf.net/projects/apertium.

Data Organization In Apertium

Aperium 'lttoolbox' uses,

- **Word and Paradigm model** (Hockett,1958) for linguistic database
- **Finite-State Transducers** (Jurafsky, 2003 and Mikel L. Forcada and et.al, 2008) as a computational algorithm for processing the data.

Lexical Resources required,

- **Paradigms**
- **Root word Dictionary**

Data Organization in Apertium

- ✓ Primarily the data is adopted from CALTS Morph.
- ✓ The data is improved.
  1. Paradigm Improvization
  2. Dictionary Improvization

Tamil Morphology

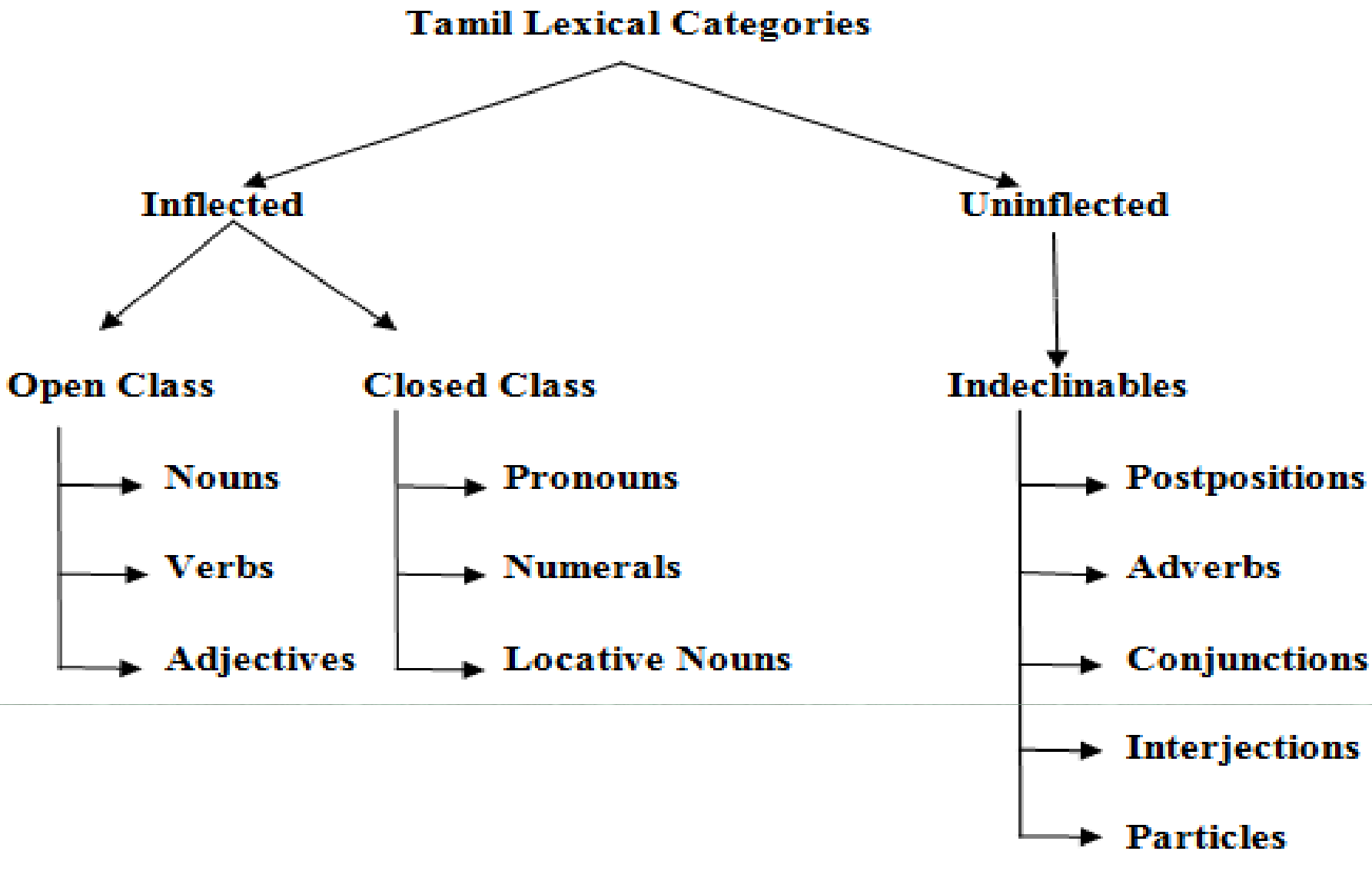
• Tamil, a Dravidian Language is known for its agglutinative morphology.

• Computing the analysis of Tamil Morphology demands a comprehensive but exhaustive analysis of its inflectional categories according to their functional properties.

• The present attempt classifies the morphological categories of Tamil based on their role in inflection. There are two classes,

Class A: The forms which anchor with suffixes or morphosyntactic elements.

Class B: The forms which are incapable of receiving such inflection.



Tamil Lexical categories

Distinct Category

- The present study considers pronouns as a distinct minor class because of its characteristic formation of oblique and idiosyncratic plural forms.

  ufkalYE- < nI (2p. pl)+ obl + Accusative>'us+ Accusative'

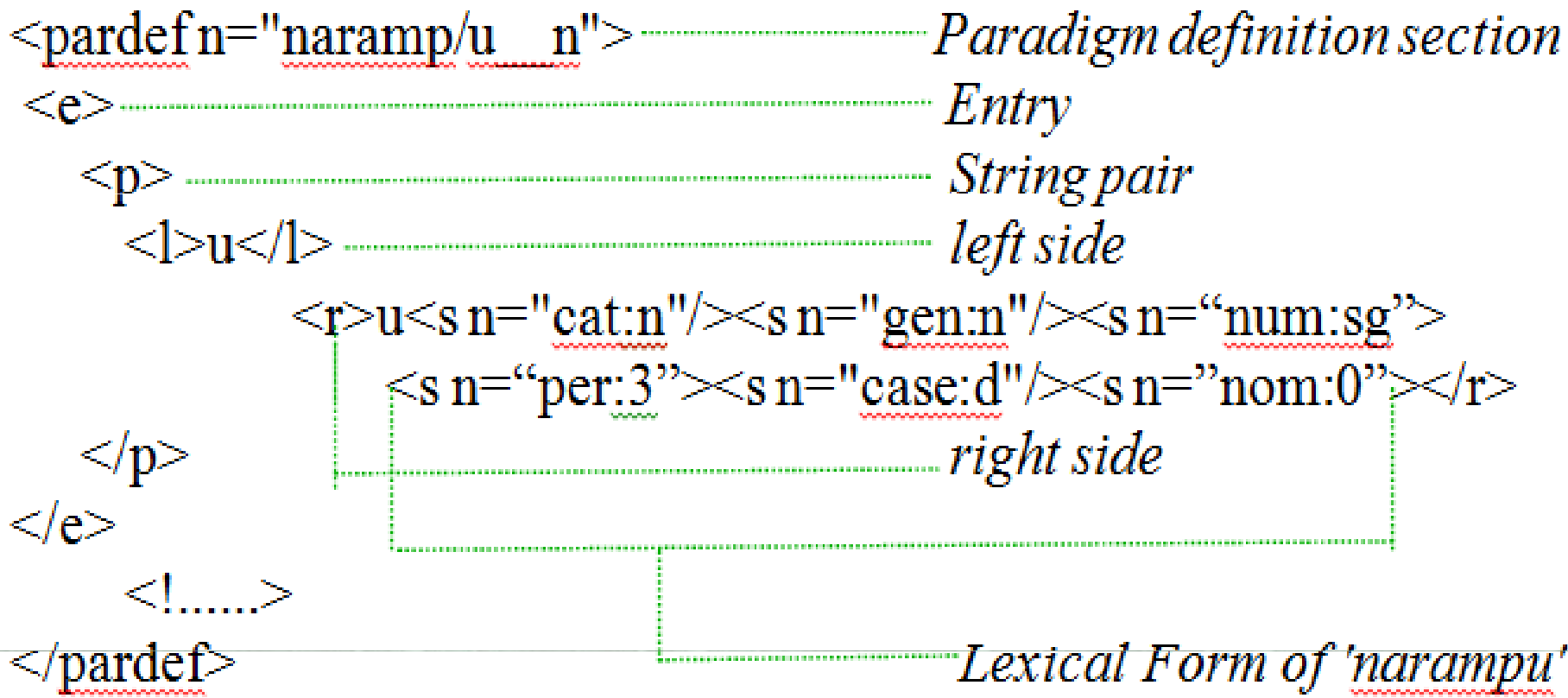
- The numerals have inflection with special particles such as **quantitative particles** (*kAl*, *arE*, *mukkAl etc)*, **attributive particles** (*per*) and **temporal particles** like time (*maNi*) which involve peculiar inflection when compared to nouns.

  walAvAyiram <walA-attributive particle + Ayiram ' a thousand per head‘ >

- The locative nouns are the indicators of time and space that are ascertained as a minor category because it exhibits an irregular inflection with morphosyntactic properties of noun. For instance,

  arukil aruku + Locative 'near + Locative'

HANDLING INFLECTION



The XML Format of inflectional paradigm for a noun .

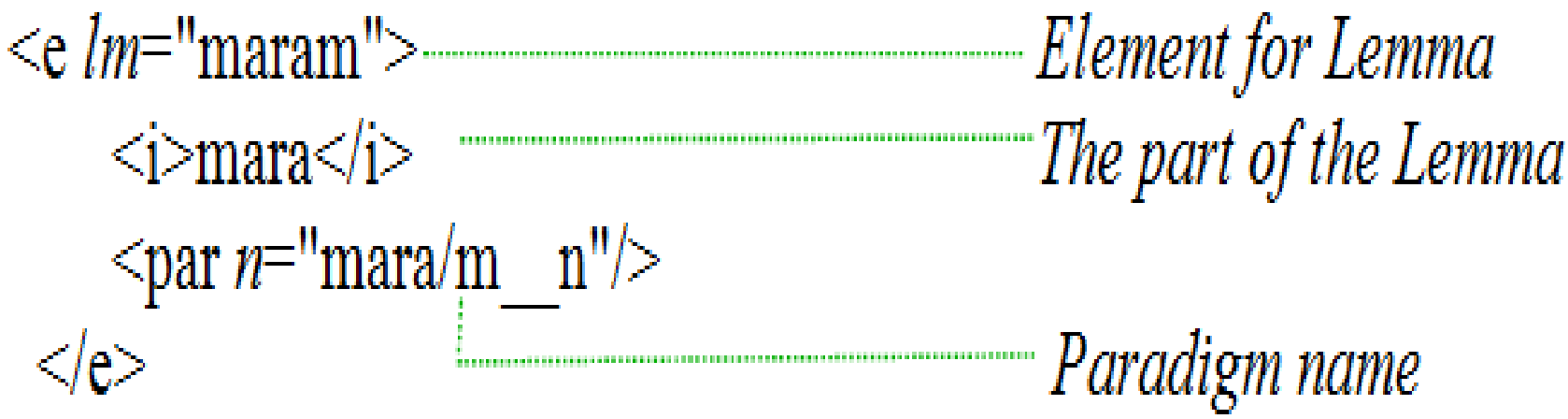
HANDLING DERIVATION

```
<pardef n="pati__v">
 <e>
   <p>
     <l>kirYavanY</l>
       <r><s n="v"/><s n="m"/><s n="sg"/><s n="3"/><s n="0"/>
          <s n="kirY_a"/></r>
     </p><i>kirYavanY</i><par n="avanY__P"/>
 </e>
     <!.....>
</pardef>
```

Linking paradigm for derivation

The XML Format of derivational paradigm for a deverbal pronoun.

Dictionary Entry



A DICTIONARY ENTRY OF THE LEXEME ' maram'

Database

Paradigmatic Database			Dictionary Size (lemma) Number of Words	
Category	Number of Inflectional Classes	Number of Inflections per class	Category wise	Total
Noun	20	743	57,322	68,060
Verb	29	934	10,114	
Adjective	2	372	209	
Pronoun	11	654	18	
Numeral	14	370	129	
NST	7	67	62	
Avy	-	-	206	

COMPILING AND PROCESSING

The data is compiled and processed by using the applications used in the lexical processing modules and tools (lttoolbox).

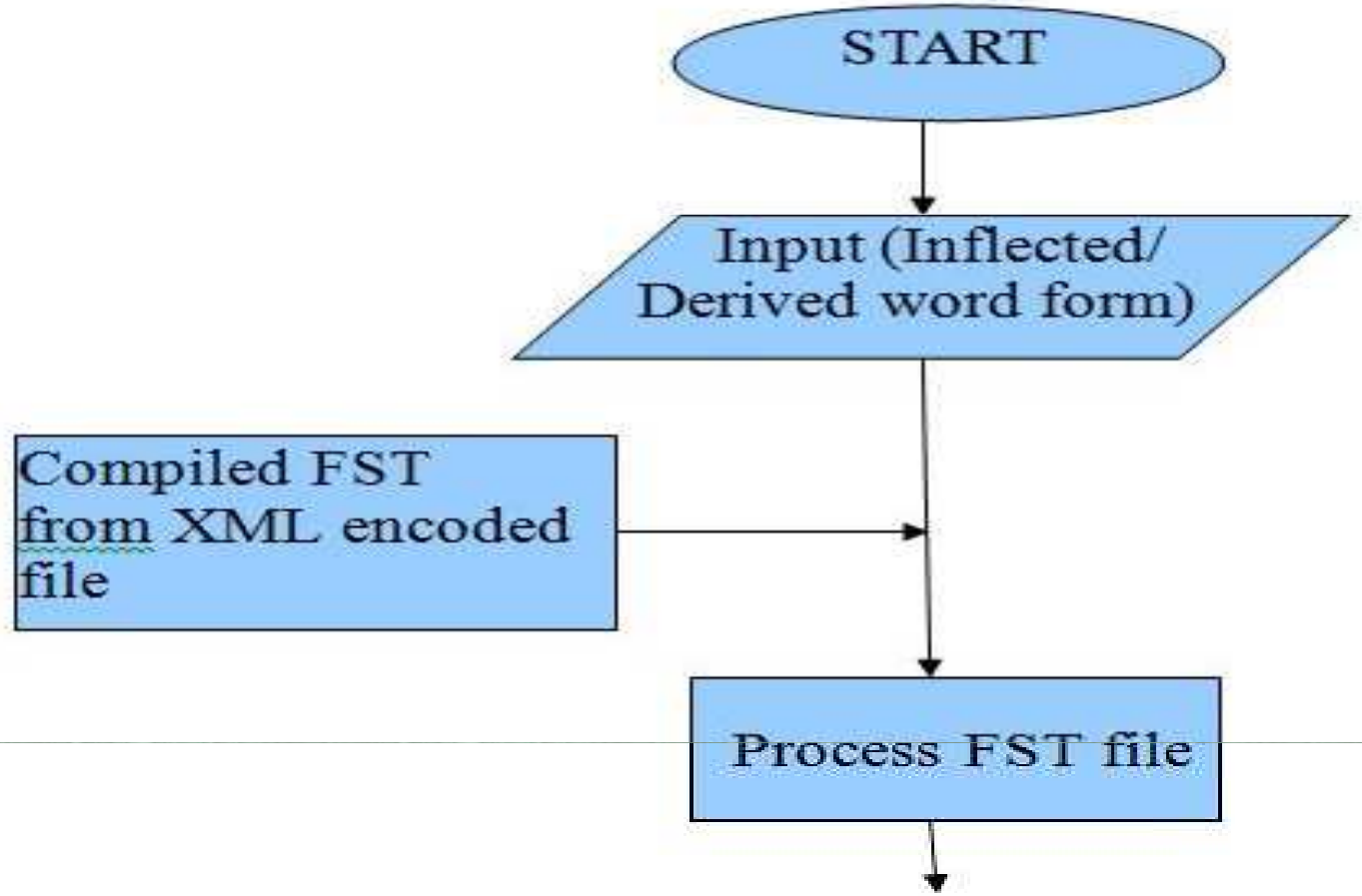
The ‘lt-comp’ is the application responsible for compiling dictionaries used by Apertium into a compact and efficient representation.

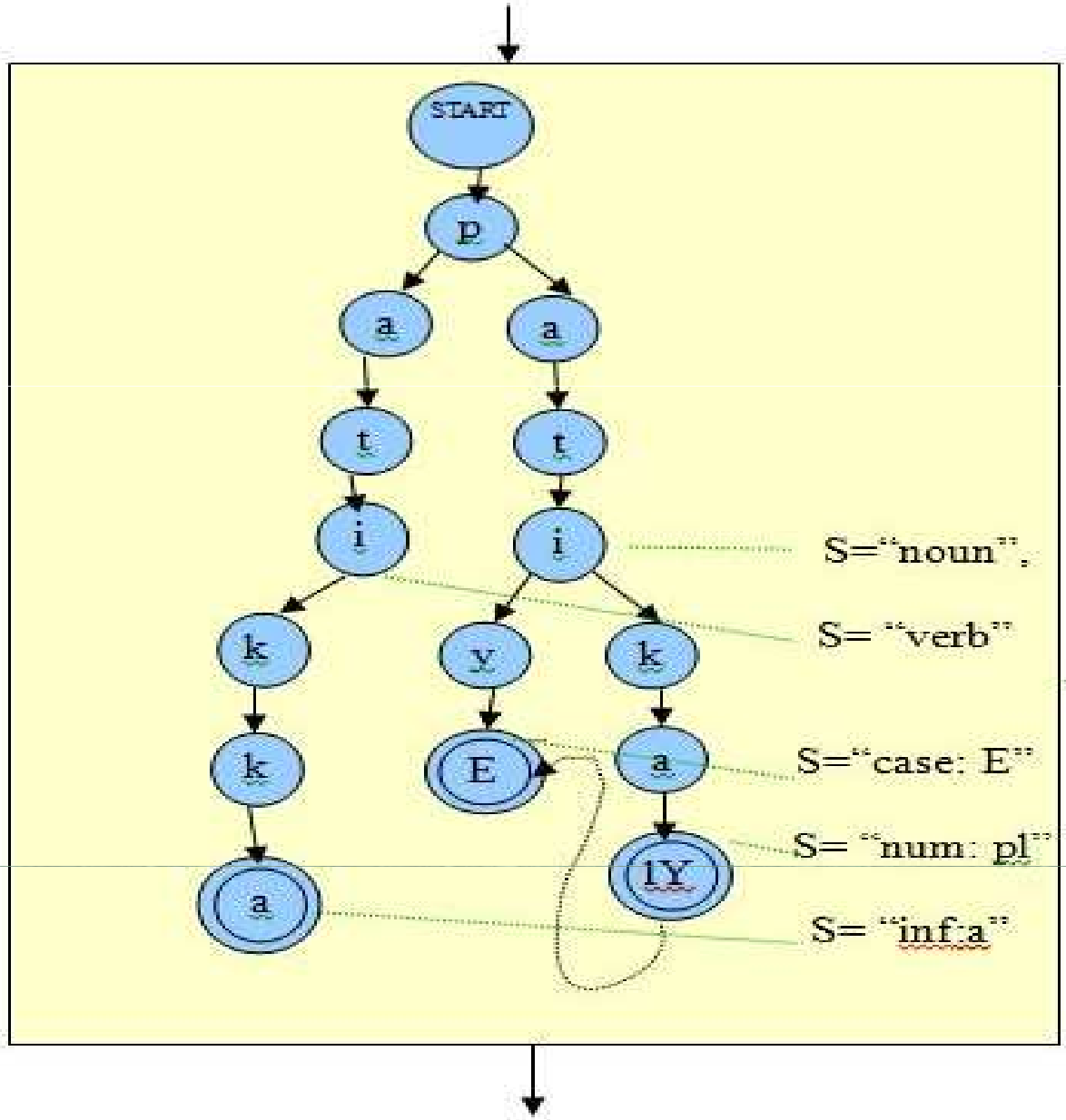
Synopsis : lt-comp [lr | rl] dictionary_file output_file

The dictionary which is compiled is processed by the application 'lt-proc' that is responsible for functioning the data.

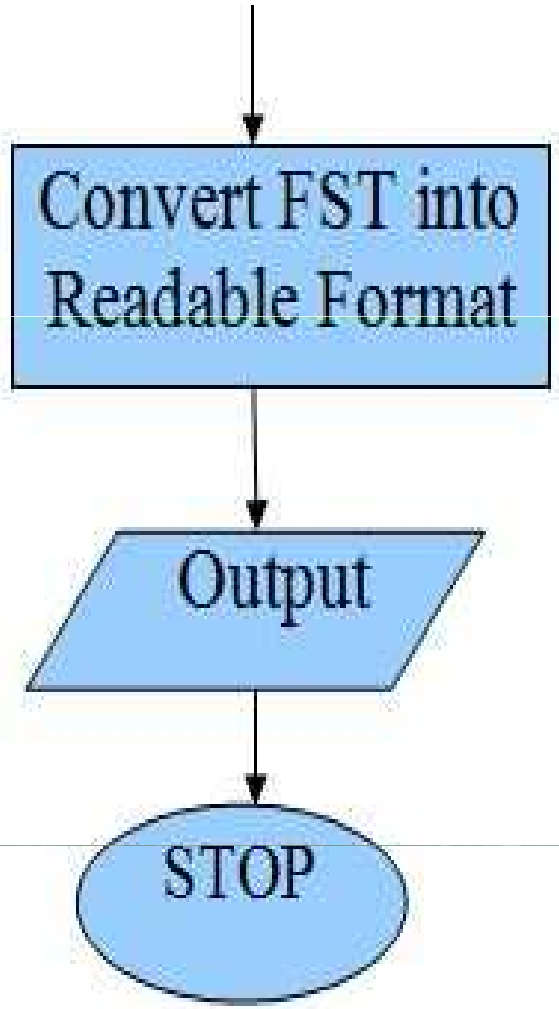
Synopsis : lt-proc [-c] [-a|-g] fst_file [input_file [output_file]]

DATA FLOW


START
Input (Inflected/ Derived word form)
Compiled FST from XML encoded file
Process FST file



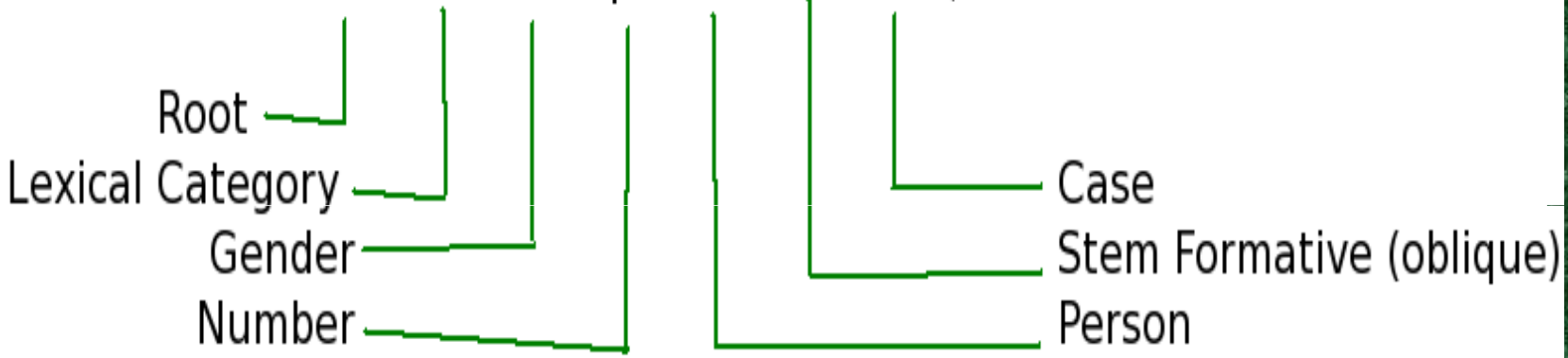
This is an Example FST for lexeme "pati" which can be a verb and noun


Convert FST into Readable Format
Output
STOP

Morphological Analyzer:

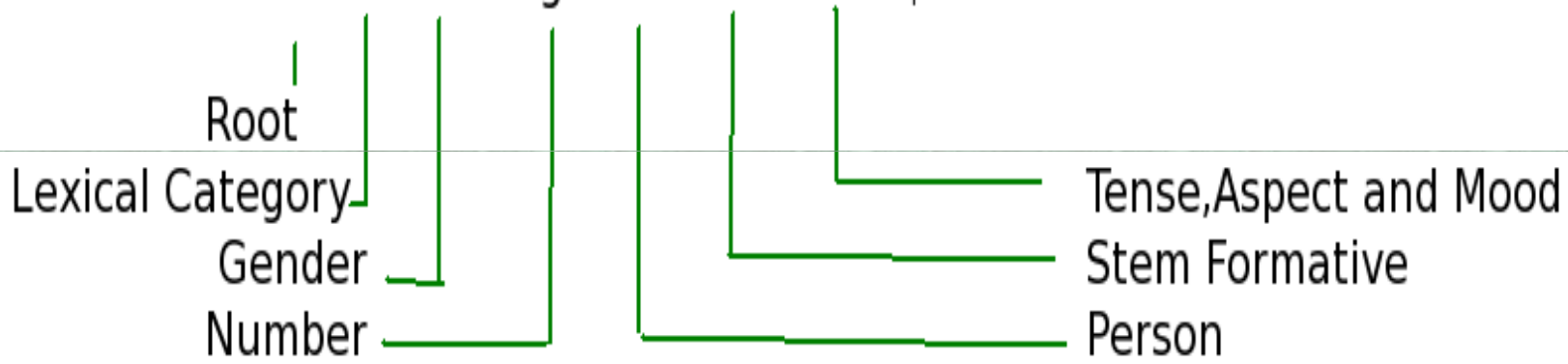
Input: marafkalYE

Output: marafkalYE/maram<n><n><pl><3><o><E>$



Morphological Generator:

Input: ^vA<v><f><sg><3><0><nw>$



Output: vanwAlY

Evaluation

- The Morphological analyzer tool was tested using the corpus (CALTS corpus of 4.4 million words and EMILLI CIIL corpus of 4.8 million words) in order to find out its the coverage of the corpus.

Corpus	Total words	Reg. words	coverage	Speed
CALTS Corpus	4,45,130	3,75,891	84.44%	0m0.289s
EMILLI CIIL Corpus	4,85,543	4,05,898	83.59%	0m0.297s

Speed

The morph analyzer is tested for its speed along with the other available Tamil Morphological analyzers which are developed in CALTS, University of Hyderabad and AU-KBC research Centre, Anna Unicersity. The speed of each modules for 1,00,000 words as follows.

Morph	Processing Speed
CALTS- MORPH	0m14.194s
AU-KBC MORPH	0m14.703s
CALTS-APERTIUM	0m0.198s

The above speed shows Calts-Apertium consumes less speed to analyze large number of data.

Error Analysis

Type	Word From	Frequency in the Copus
Orthographic Variation	*koyil* 'temple' *kovil* 'temple'	885 occurrences 204 occurrences
Inflectional Variation	*eVlYYuwwu-kkalY* 'letters' *eVlYYuwwu-kalY* 'letters'	57 occurrences 171 occurrences
Dialectal Variation	*vanwAy* 'You came' (standard) *vanweV 'You came'* (dialect)	*765* occurrences 6 occurrences
Naturalized English words	*polIS* 'police'	20070 occurrences
Proper nouns	*kaNNanY* 'male name' *wamilYYnAtu* 'Tamil Nadu'	211 occurances 364 occurrences

Salient features of Apertium

- Easily Accessible.
- Speed.
- Two in One tool (Both Generator and Analyzer).
- GNU License – allow us to modify.
- Uses XML format which is easy to modify.

Conclusion

- The Apertium tool for Tamil is efficient in terms of time for processing a large number of words.
- The combination of Finite State Transducers (letter transducer) and the paradigm approach works more efficient and speedy parsing.
- The other advantage of the Apertium is that the current morphological database can be used to create a parallel morphological generator for Tamil.
- In future, further work can be carried out in this area to meet the other morphological lapses for attaining maximum coverage and precision.

Reference

Arden, A.H. 1976. *A progressive Grammar of the Tamil language*. Madras : The Christian Literature Society.

Fercada, Mikel et.al. 2008. *Documentation of the Open-Source Shallow-Transfer Machine Translation platform Apertium.* Published in website: http://www.gnu.org/copyleft/fd1.html.

Ramaswamy, Vaishnavi. 2003. *A morphological Analyzer for Tamil.* Unpublished Ph.D. Thesis. Hyderabad: University of Hyderabad.

Uma Maheshwar Rao, G. 1999. *Morphological Analyzer for Telugu.* (electronic form). Hyderabad: University of Hyderabad.

Uma Maheshwar Rao, G. 2002. *A Computational Grammar of Telugu.* (Momeo) Hyderabad: University of Hyderabad.

Vaidhya, Ashwini and Dipti Misra Sharma. 2009. *Using Paradigms for Certain Morphological phenomena in Marathi.* Puplished in 7th International Conference on NLP-2009: India

Viswanathan. S et.al. 2003. *A Tamil Morphological Analyser* in Recent Advances in NLP. pp 31-39.

THANK YOU