# NLP TOOLS DEVELOPMENT FOR TAMIL LANGUAGE

Loganathan, UFAL

# Overview

- Introduction
- My work involving Tamil NLP
  - English – Tamil MT
  - Tamil Morphological Analyzer

# Introduction – Indian Languages

- **23** official languages
- **29** languages have more than **1 million** nat


**<u>Tamil</u>**

Approx 67 million speakers in India

# Introduction – Resources for Tamil

- Tamil Editing/Unicode Support - **Available**
- Dictionary – **Available**
  - Tamil lexicon, Winslow, Fabricius, McAlpin, Kathirvelu pillai – *Published online by Univ. Of Chicago*
- Morphological Analyzer/Tagger – **Partially Available**
- Phrase Structure/Dependency Parser – **NO**
- Parallel Corpora – **NO (publicly, readily)**
  - Active Bilingual websites: *www.wsws.org*, *www.cinesouth.com*
  - Tamilnadu government schoolbooks (with English translations)

# My Work involving Tamil NLP

- Enlgish – Tamil Translation System (Master's Thesis)
- Morphological Analyzer

# English – Tamil Translation System

# General Differences

- Morphological

| Noun Cases | Tamil suffixes | English words |
|---|---|---|
| Nominative | Ø | Ø |
| Accusative | ai | Ø |
| Dative | kku, ukku, ku | Ø, to, for, at |
| Benefactive | (u)kkuAka | for |
| Instrumental | Al | with, of, by |
| Sociative | Otu, utan | with |
| Locative | il, itam | in, on, among, to, with, from |
| Ablative | iliruwTu, itamiruwTu | from |
| Genitive | in, utaiya, aTu | ‘s, of |

# General Differences

- Syntactical difference

Kumar talked about linguistics

kumAr moziyiyalaip paRRip pEcinAn

குமார் மொழியியலைப் பற்றிப் பேசினான்

moziyiyalaip paRRip kumAr pEcinAn

# General Differences

- Syntactical (complex sentences)

Pollution Control Authority's regional officer said — MC
**that** his department is not agreeing with the central minister's opinion — SC
**that** Pollution Control Authority is not functioning — RC

MC SC RC-> RC SC MC

mAcuk kattuppAttu vAriyam ceyalpaTavillai **enRa**
maTTiya amaiccarin karuTTil TangkaLaTu TuRaikku utanpAtu illai **ena**
mAcuk kattuppAttu vAriya aTikAri TeriviTTAr

மாசுக்கட்டுப்பாட்டு வாரியம் செயல்படவில்லை என்ற
மத்திய அமைச்சரின் கருத்தில் தங்களது துறைக்கு உடன்பாடு இல்லை என
மாசுக்கட்டுப்பாட்டு வாரிய அதிகாரி தெரிவித்தார்

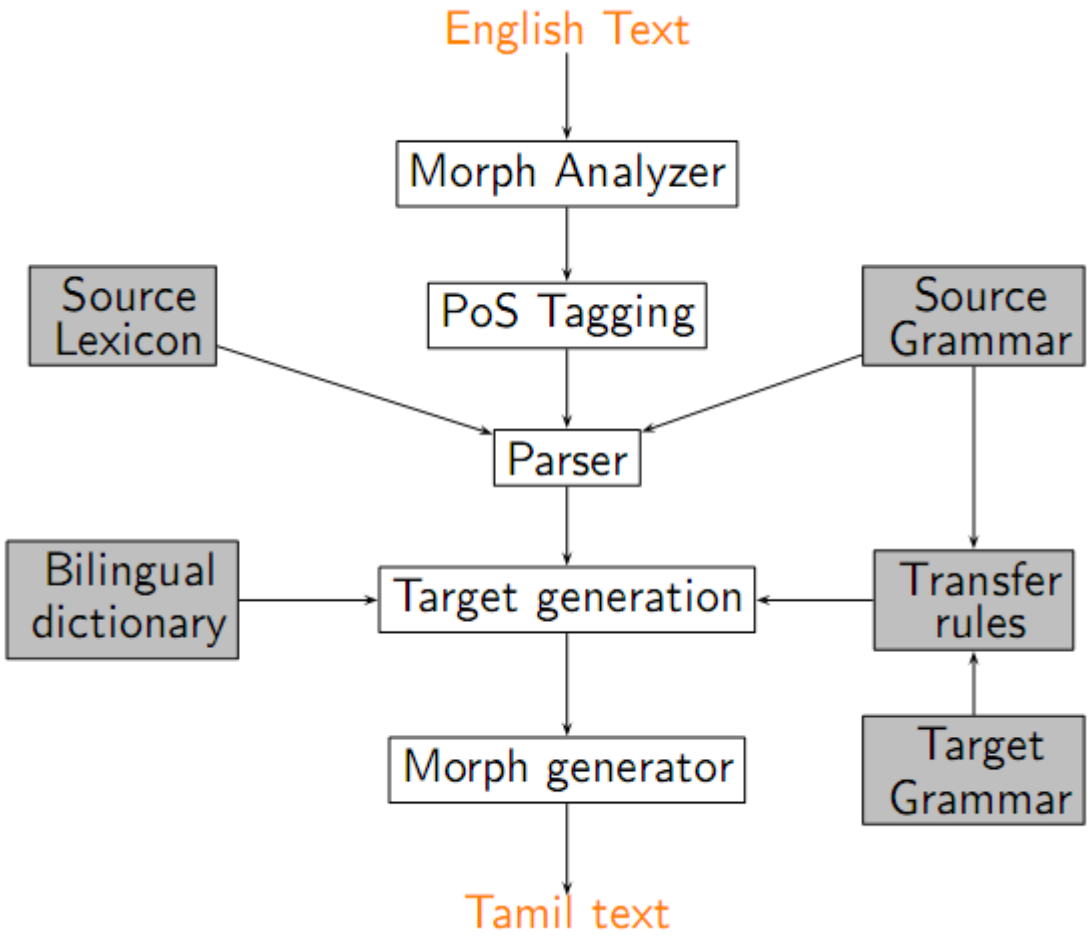
# How hard is English-Tamil MT

- The previous examples illustrate
  - Tamil -> SOV,  English -> SVO
  - Tamil is a restricted free word order language
  - Tamil is agglutinative
- Difference occurs in
  - Syntactical level i.e word ordering
  - Morphological level
- We need
  - An efficient syntax reordering module
  - Morphological generator

# Approaches

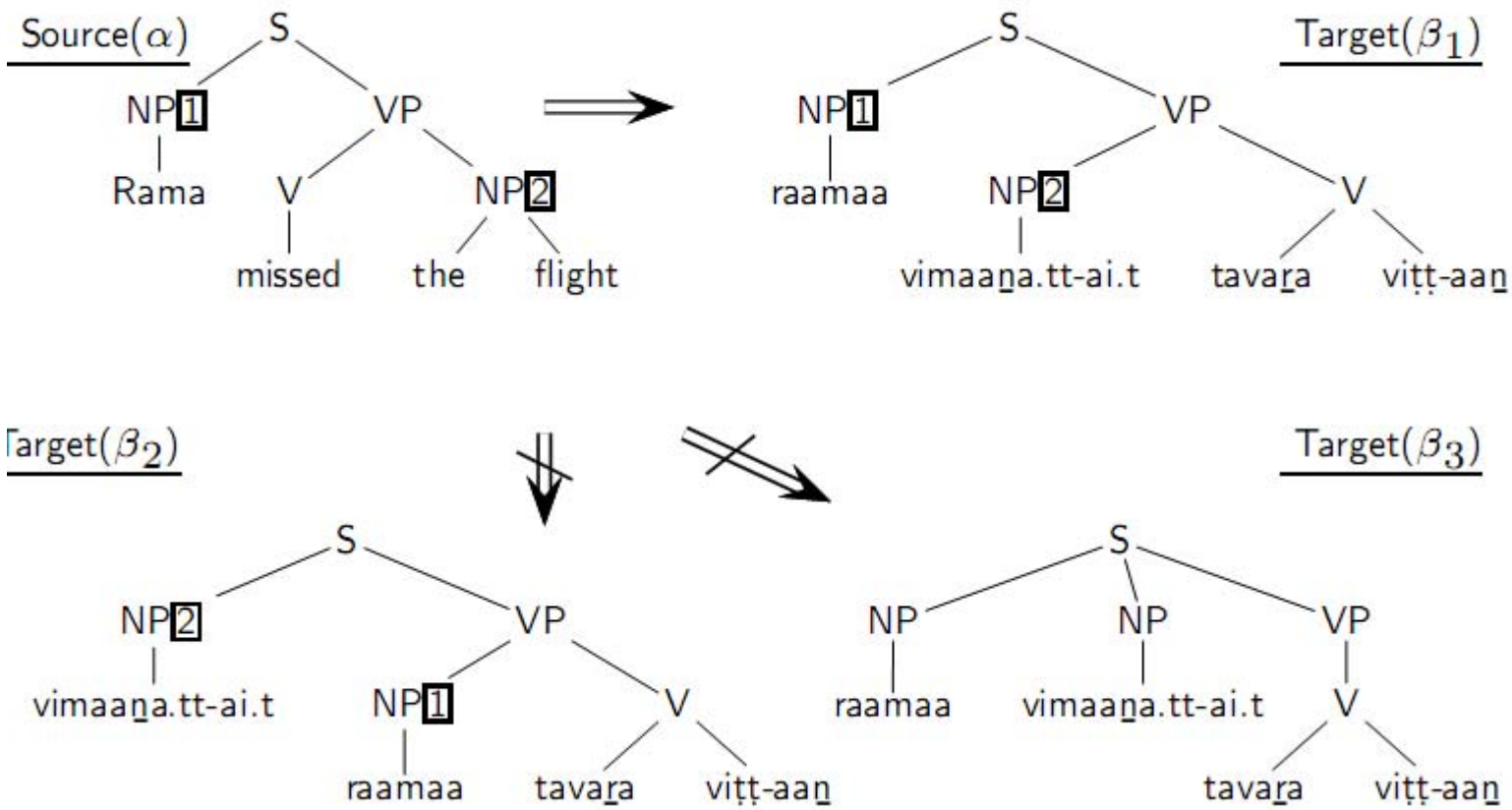
- Syntax Transfer Based MT
- Statistical Machine Translation (SMT)

# Syntax Transfer MT



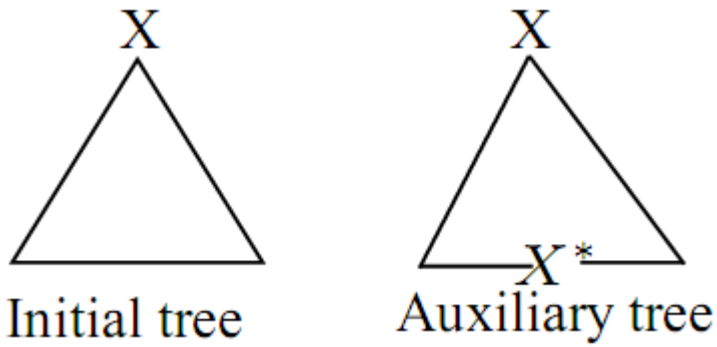

**Syntax Transfer MT: Architecture**

# MT using Synchronous CFG



# Source Grammar – Tree Adjoining Grammar

- Tree Generating System introduced by Aravind Joshi
- TAG – Multilevel tree rewriting system
- Basic units (Elementary trees)
  - Initial trees (Basic structures)
  - Auxiliary trees  (Recursive structures)

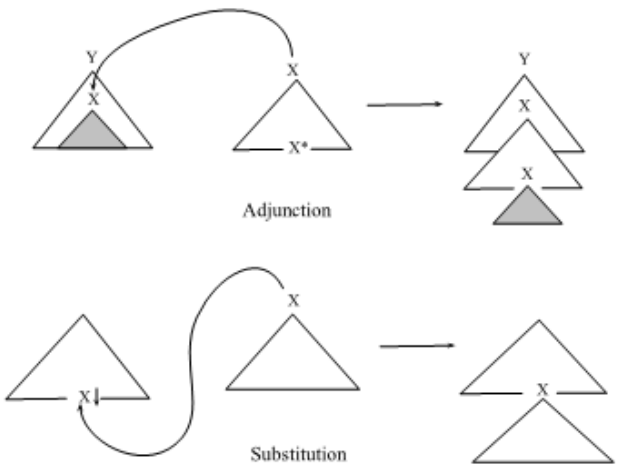


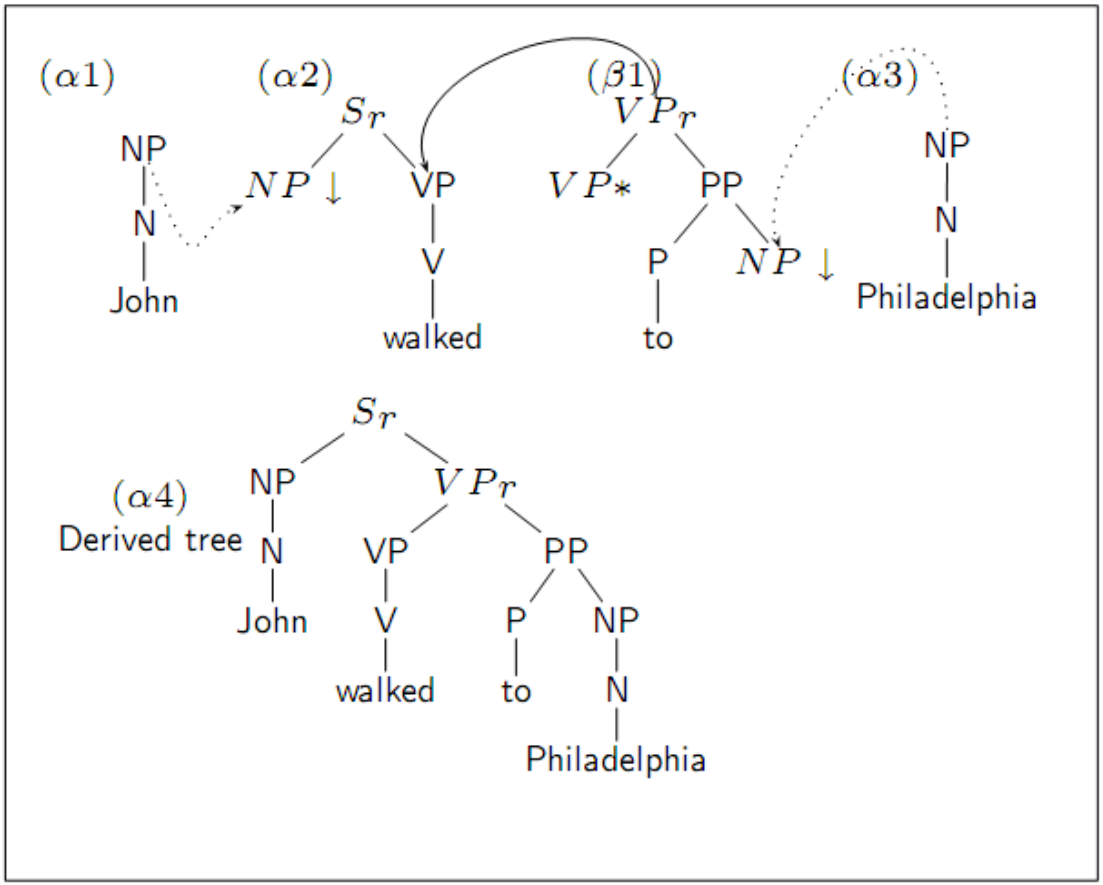
# TAG – Formal Definition

Definition: **(Tree Adjoining Grammar).** TAG consists of 5-tuples $(\Sigma, NT, I, A, S)$, where

1. $\Sigma$ is a finite set of terminal symbols;
2. $NT$ is a finite set of non-terminal symbols: $\Sigma \cap NT = \emptyset$;
3. $S$ is a distinguished non-terminal symbos: $S \in NT$;
4. $I$ is a finite set of trees, called *initial trees*, characterized as follows,
   - interior nodes are labeled by non-terminal symbols;
   - the nodes on the frontier of initial trees are labeled by terminal or non-terminals; non-terminal symbols on the frontier of trees in $I$ are marked for substitution; usually marked as $(\downarrow)$
5. $A$ is a finite set of trees, called *auxiliary trees*, characterized as follows,
   - interior nodes are marked by non-terminal symbols;
   - the nodes on the frontier of auxiliary trees are labeled by terminal symbols or non-terminal symbols. Non-terminal symbol on the frontier of the trees in $A$ are marked for substitution except for one node, called the *foot node*. The foot is node is marked with $(*)$; the label of the foot node must be identical to the label of the root node.

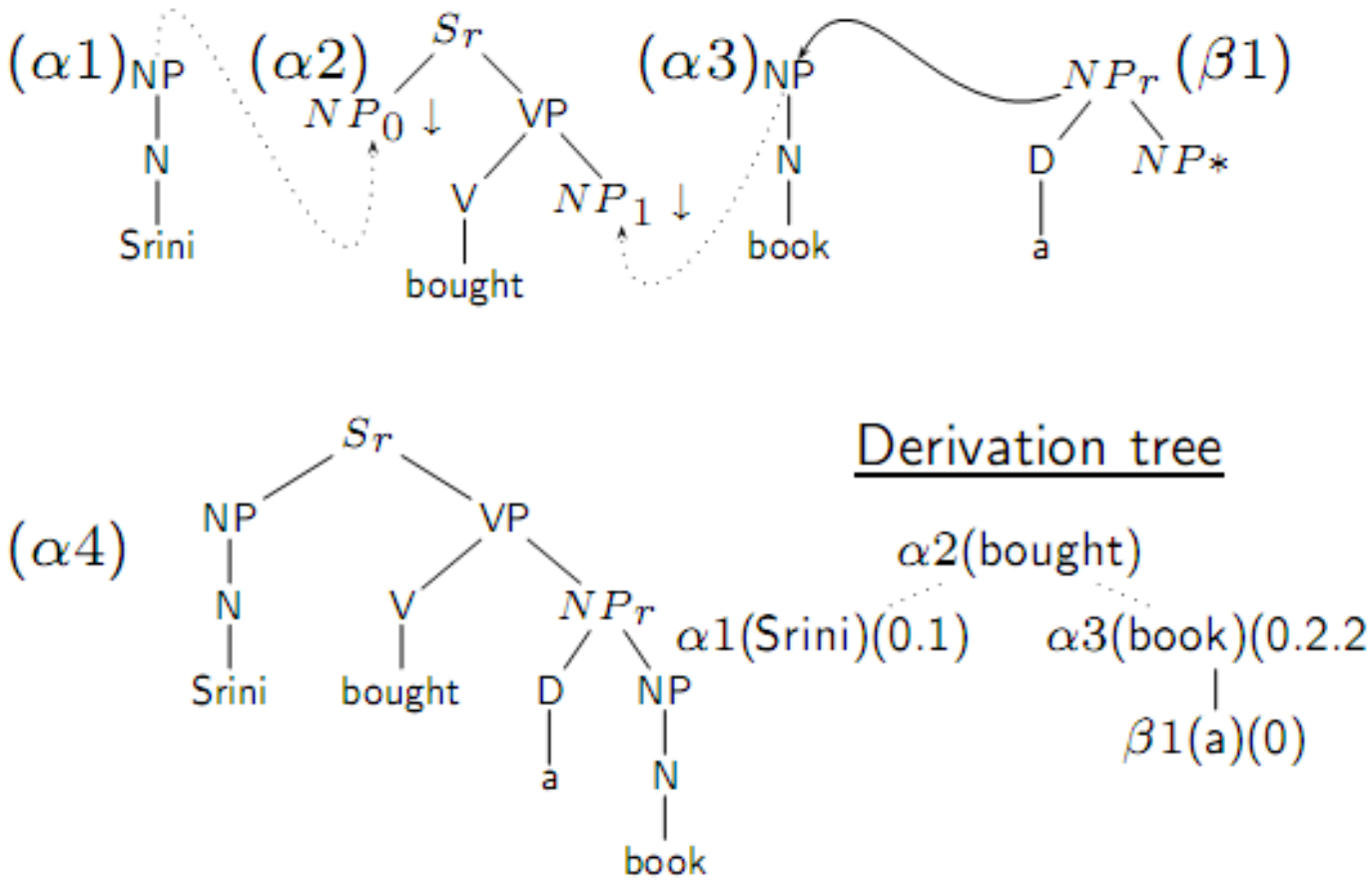
# TAG Operations




**Substitution and Adjunction/ Ex: from XTAG Manual**

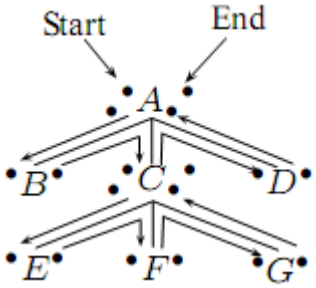
# TAG Derivation Structure



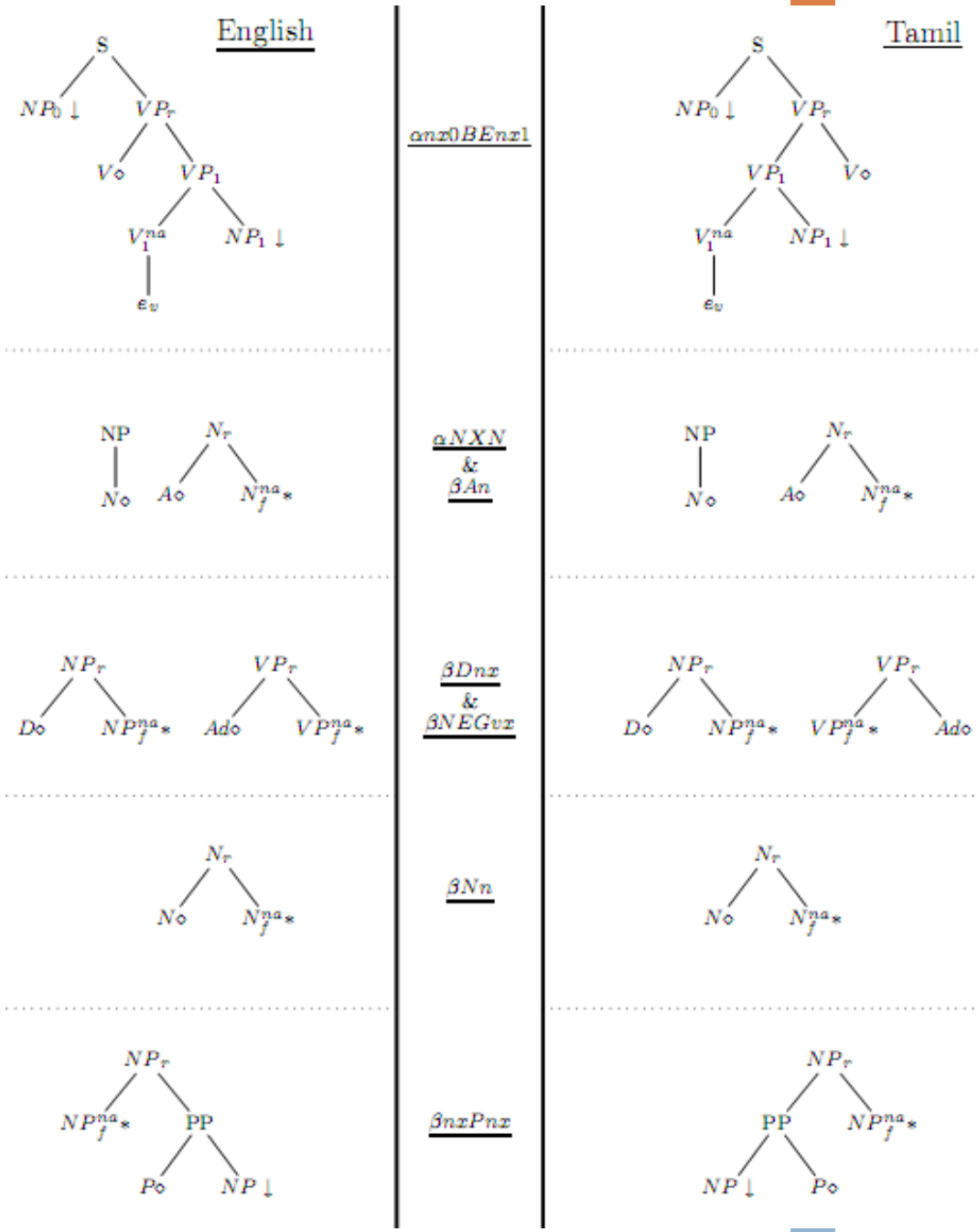

**Derivation Tree: "Srini bought a book"/ Ex: from XTAG Manual**

# Parsing Lexicalized TAGs

- Many parsing algorithms were suggested, including CYK parser for TAG, Head-Corner parsing algorithm, Bidirectional parsing algorithm and more recent work on Statistical LTAG parsing.
- For parsing source side, Yves Schabes algorithm was implemented in Java.



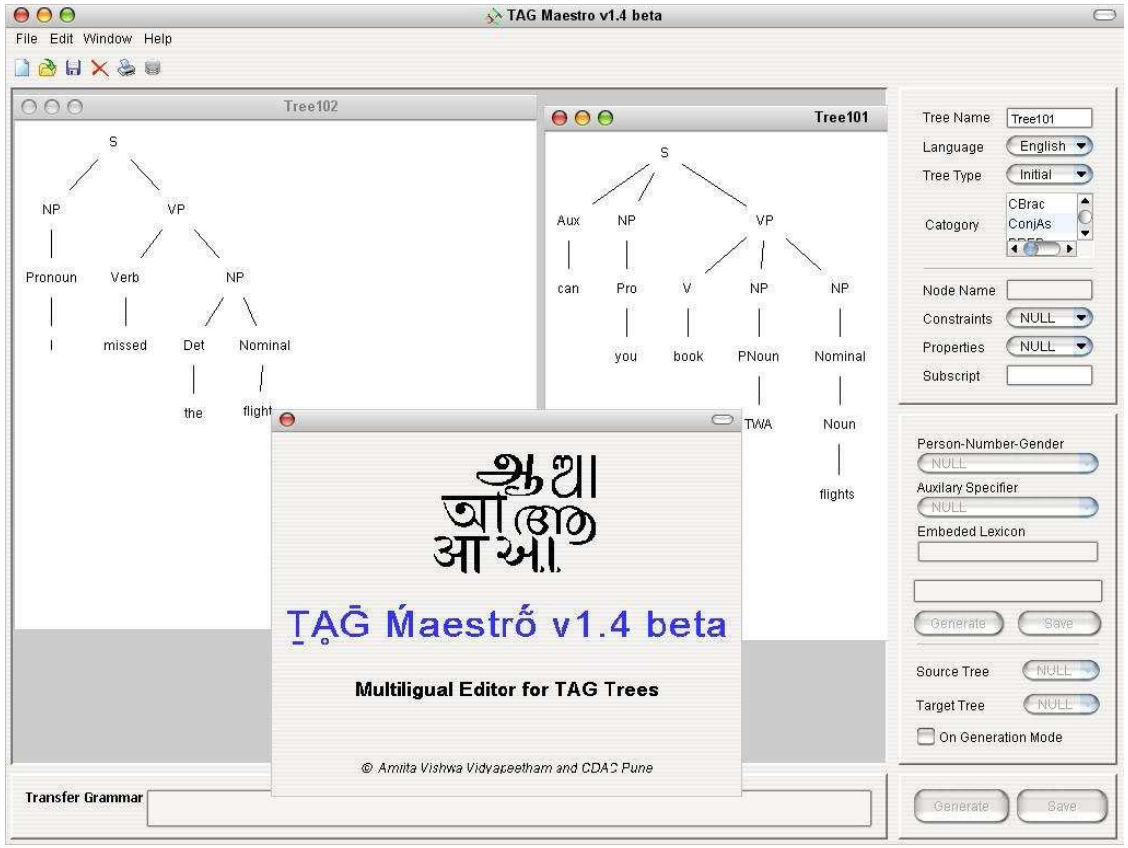
# Transfer Grammar



# Experiments and Results

- The entire translation system is written in Java.
- Implemented modules include LTAG parser for English, STAG system for syntax reordering of English into Tamil.
- Our system uses the same language resources developed for XTAG system for parsing the source side sentence. All XTAG related databases have been converted into Mysql format.

# LTAG Tree Editor for Visualization




Collaborative effort between Amrita and CDAC

# Experiments and Results

English: i met John
Tamil: நான் ஜான் ai சந்தித்த்

English: he met John yesterday
Tamil: அவன்/அவர் நேற்று ஜான் ai சந்தித்த்

English: John is a good boy
Tamil: ஜான் ஒரு நல்ல பையன்

English: John is not a good boy
Tamil: ஜான் ஒரு நல்ல பையன் இல்லை

English: Mary said that John said that Ram came yesterday
Tamil: Ram நேற்று வந்த் என்று ஜான் சொன்ன் என்று Mary சொன்ன்

**Sample Output**

# Statistical Machine Translation

- EILMT English-Tamil Parallel Corpus

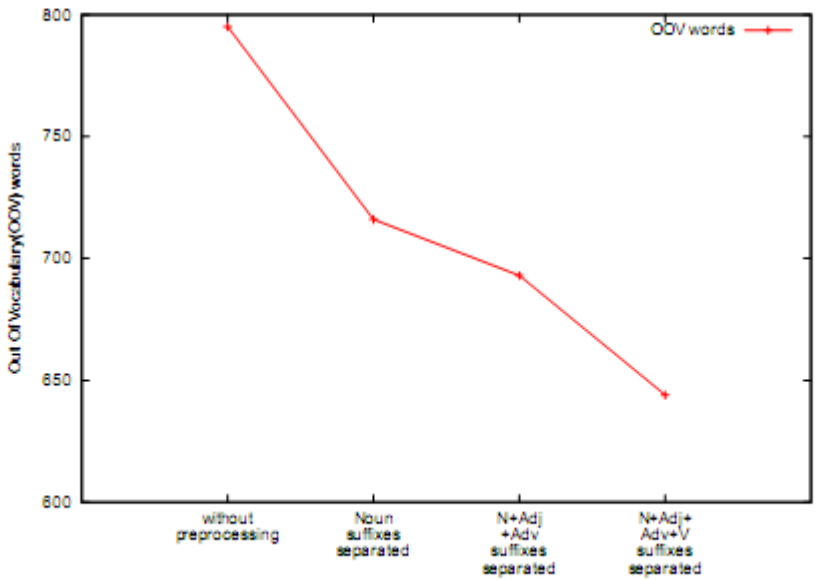
| Health | Tourism |
|---|---|
| 6000 | 15000 |

- Monolingual Tamil Data

| | #Sentences | #Words |
|---|---|---|
| Training data | 95464 | >1.2 million |
| Test data | 1000 | 12K |

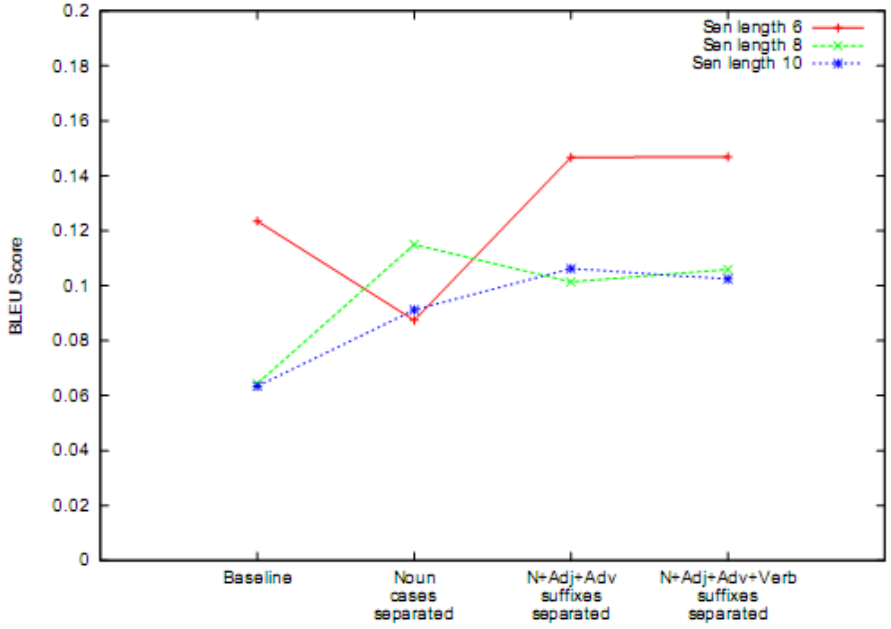
# Statistical Machine Translation

| POS | Suffix list |
|---|---|
| Noun | ai, (u)kku(aaka), aal, ooṭu, (il\|iṭam)(iruntu), uṭaṉ, iṉ, uṭaiya, atu |
| Adj+Adv | aaṉa, aaka |
| Verb | (kkiṟ\|kiṟ\|kiṉṟ\|tt\|t\|nt\|iṉ\|pp\|p\|v)(avar\|oor\|iir\|aar\|avaṉ\|avaḷ\|atu\|avai\|eeṉ\|aay\|aaṉ \|aaḷ\|um\|oom), aat(avaṉ\|avaḷ\|atu\|avai\|avar\|oor\|iir\|iirkaḷ\|avarkaḷ) |

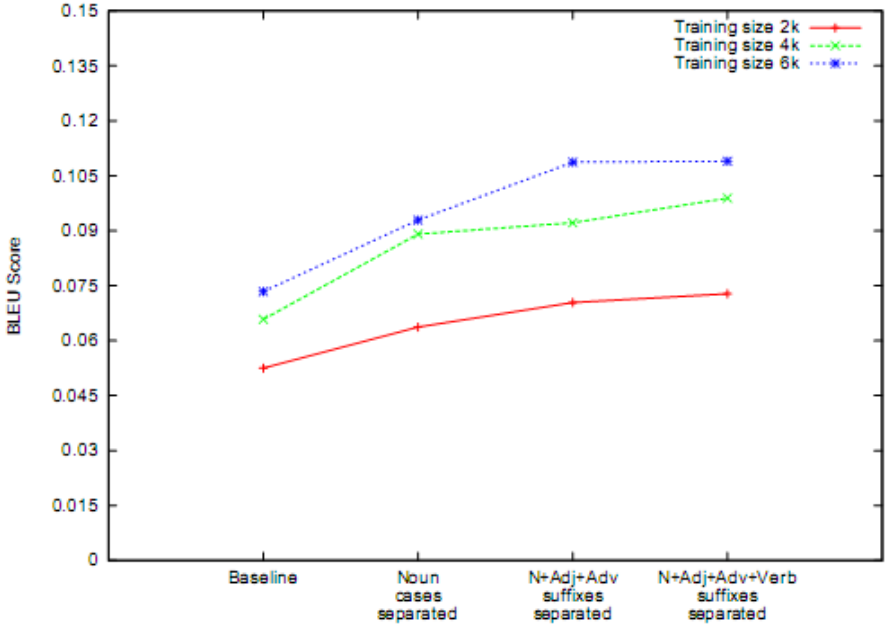



Monolingual data: Sensitivity to Morphology

# SMT – Results for Health Corpus

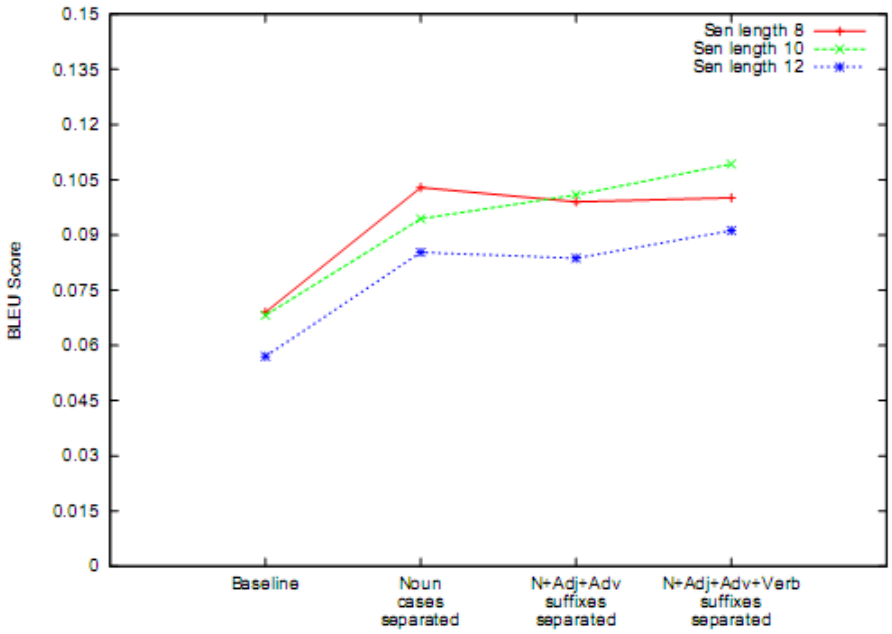



a) Sensitivity to sentence length & Morphology

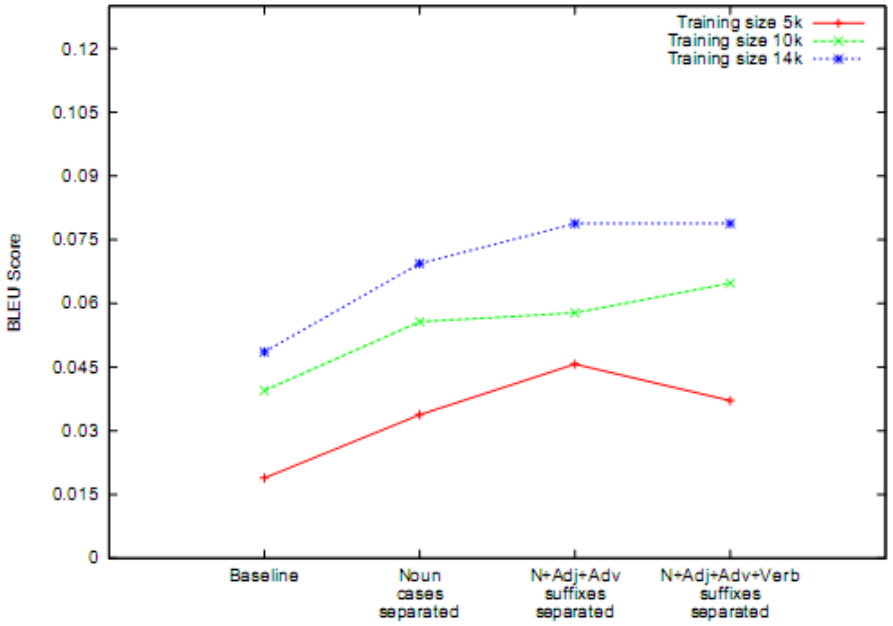



b) Sensitivity to training corpus size & Morphology

# SMT – Results for Tourism Corpus




a) Sensitivity to sentence length & Morphology




b) Sensitivity to training corpus size & Morphology

# SMT – Sample Output

English Text
==========

in karnataka there are the chamundi hills .
mumbai is a cluster of seven islands .
the building is illuminated on sunday evenings .
the 12th century temple is dedicated to lord shiva .

Baseline Tamil output
==========

பின்னர் கர்நாடக உள்ள சாமுண்டிண்டி மலைகள் உள்ளன .
மும்பை ஏழு தீவுகளின் தொகுப்பு . ப்பு . .
இந்த கட்டிடம் ஆகும் இல்லுமினஎஎட் ஞாயிற்று கிழமை காலை மாலை
12ம் நூற்றாண்டில் பகவான் சிவன்ர கோயில் அர்ப்பணிக்கப்பட்டுள்ளது .

Nouns+Adj+Adv+Verb suffixes separated
==========

கர்நாடகா இல் உள்ள சாமுண்டி மலைகள் உள்ளன .
மும்பை இல் உள்ள ஒரு ஏழு தீவுகவ்ுகள் இன் தொகுப்பு . .
இந்த கட்டிடம் விளக்குகள் ஆல் மநாலை இல் இருக்க் உம் . .
12ம் நூற்றாண்ட் இல் கடவுள் சிவவன் உக்கு அர்ப்பணிக்கப்பட்டுள்ளது .

# Tamil Morphological Analyzer

# NLP @ Amrita – Morphological Analyzer for Tamil

- Tamil is **agglutinative**
- The major inflectional categories in Tamil are nouns and verbs.
- Noun morphology of Tamil is simple compared to verb morphology.
- Extremely simple paradigms were used to categorize the root words.
- The lexicon includes **50000 nouns** and few hundred verbs.
- FSTs were used to build Morphological Analyzer/generator

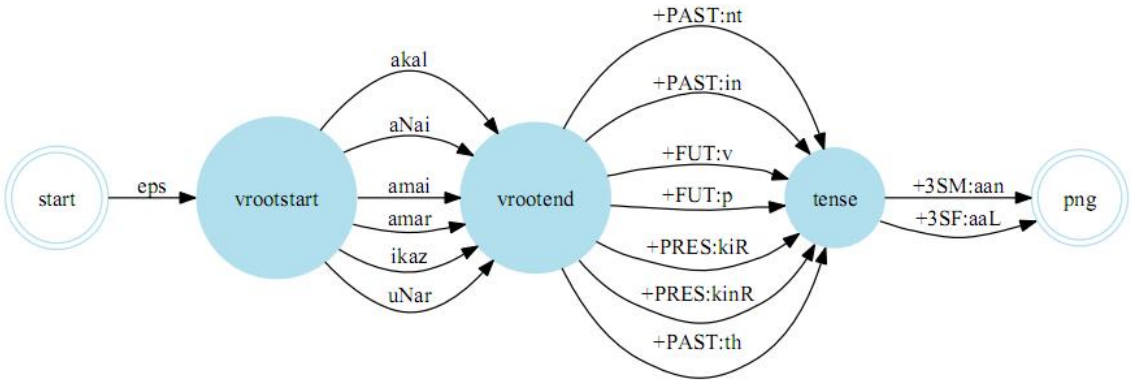



**Figure:** Morph Generator FST

# NLP @ Amrita – Morphological Analyzer for Tamil

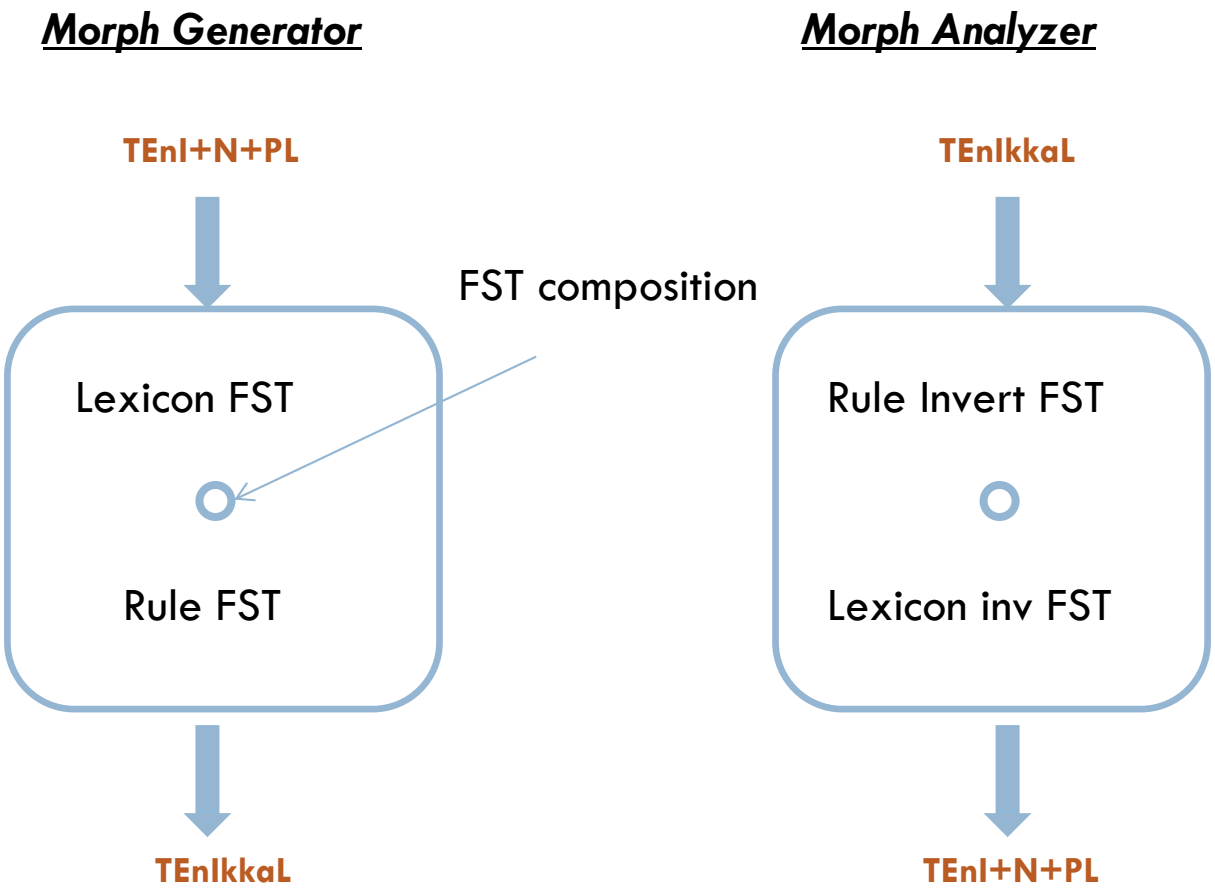



**Figure:** Finite State Transducer for Morphological Processing

# NLP @ Amrita – Morphological Analyzer for Tamil

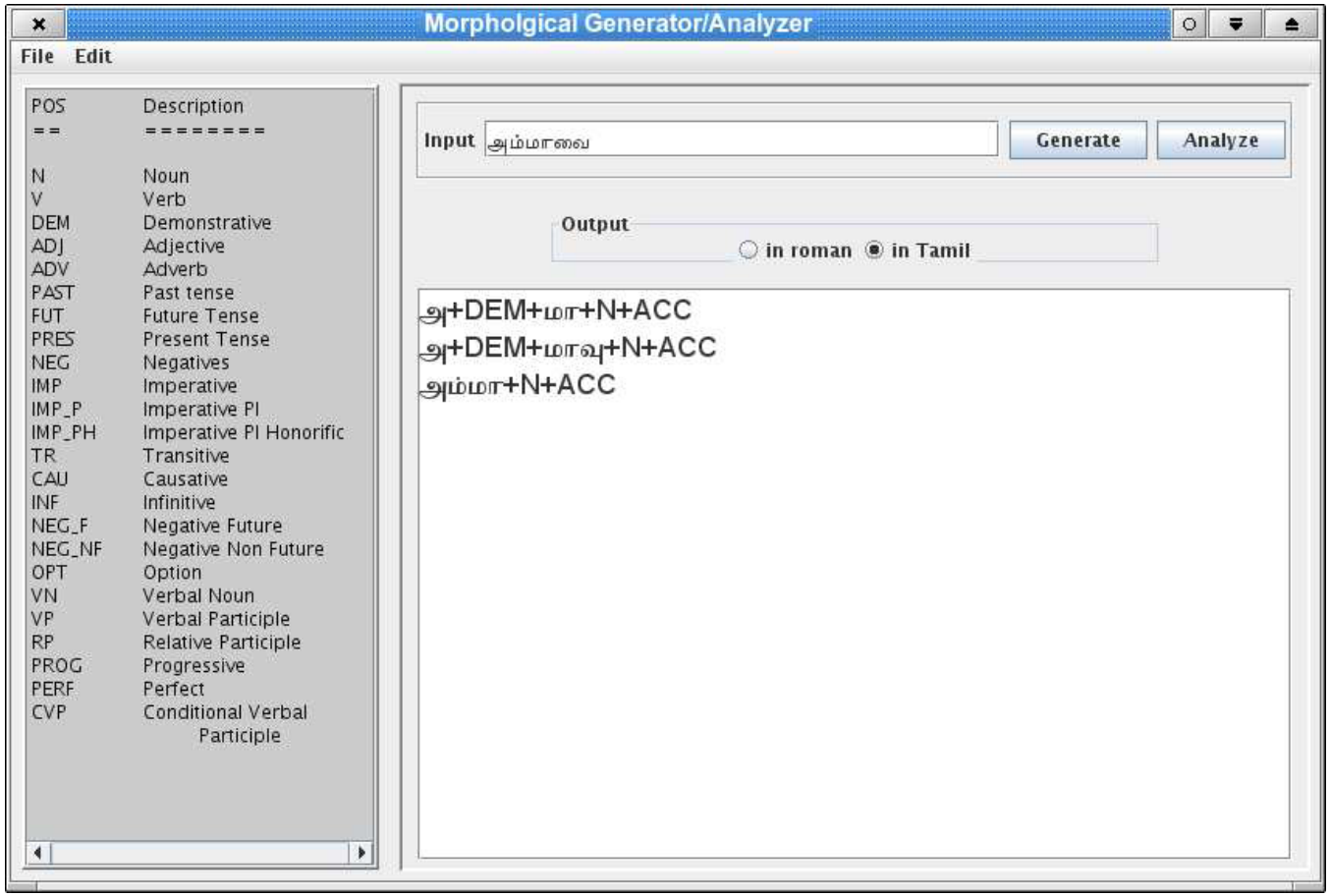



**Figure:** Morphological Analyzer screenshot

Thank you