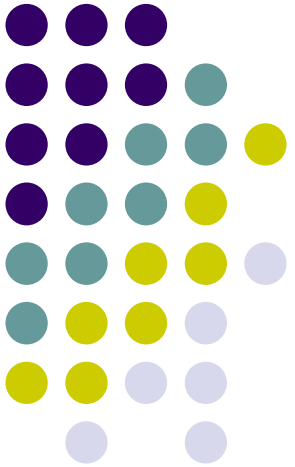
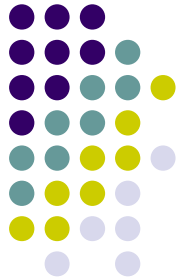
Morphological Generator for Tamil

-Menaka S, Vijay Sundar Ram and Sobha Lalitha Devi,

AU-KBC Research Centre

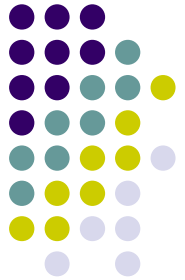


Overview



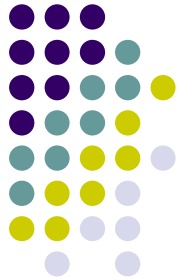
- Tamil Morphology Key ideas
- Morphosyntax and Morphophonemics
- Finite State Automata
- Morphological generator
- Evaluation

Morphological Generator



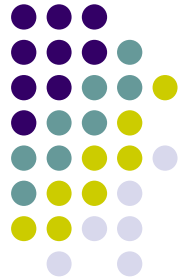
- What is it? Tool used in NLP
- What does it do? Root word -> Inflected form
- Who needs it? Inflecting languages
- Where is it used? MT, IR

Methods used



- Rule-based method (Ganapathiraju and Levin 2006)
- Corpus-based method (Lantin et al, Dasgupta and Ng, 2007)
- Finite-state method (Beesley and Karttunnen. 2003)

Tamil Morphology – key ideas



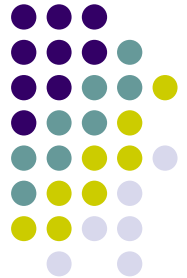
- Agglutinative – Suffixes attach in series to the root.

arapi + *katal* + *in* + *araci* => *arapikkatalinaraci*

'Arabian' + 'sea' + GEN + 'queen' => 'Queen of the Arabian Sea'

- Morphosyntax – Order in which suffixes attach to the root.
- Morphophonemics – Changes that take place during suffixation.

MorphoSyntax of Lexical Categories - Nouns



- Nouns (include pronouns)- Take Inflectional and Derivational Suffixes.
- Root + {number} + {case} + {DISJ/COOR/EMPH} + {PSP} + {EMP} + {INT/SUPP}

paiyan-kaL-ai-a: => paiyankaLaiya:

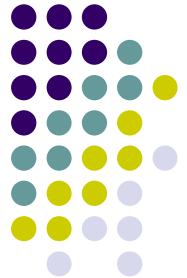
‘boy’-PL-ACC-INT => ‘the boys(OBJ)?’

- Derivation of verbs, adjectives, adverbs from nouns is possible.

azaku + a:na => azaka:na

'beauty' + ADJ => 'beautiful'

MorphoSyntax of Lexical Categories – Verbs…(1)



- Finite Verbs
- Root + Tense + PNG + {DISJ/EMPH/EMP/INT/SUPP}

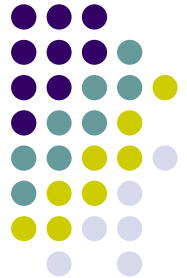
  pa:r-tt-a:n-a:m => *pa:rtta:na:m*

  'see'-PST-3SM-SUPP => 'it seems (he) saw'
- Root + INF + NEGVERB + {DISJ/COOR/EMPH/EMP/INT/SUPP }

  pa:r-a-illai-a:m => *pa:rkkavillaiya:m*

  'see'-INF-NEGVERB-SUPP => 'it seems (x) did not see'

MorphoSyntax of Lexical Categories – Verbs…(2)



- Relative participle
- Root + Tense/NEG + RP

pa:r-tt-a => *pa:rtta*

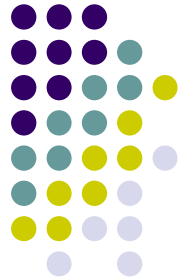
‘see’-PST-RP => ‘who saw’

- Pronominalisation

pa:r-tt-a-avan => *pa:rttavan*

‘see’-PST-RP-he => ‘he who saw’

MorphoSyntax of Lexical Categories – Verbs…(3)



- Non-Finite Verbs

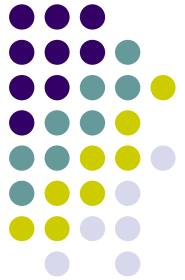
  root + {NEG} + INF/VBP/COND/CONC/HORT/OPT + {DISJ/COOR/EMPH} + {EMP} + {INT/SUPP}

  pa:Tu-a => *pa:Ta*

  'sing'-INF => 'to sing'

- Derivation of nouns, adjectives and adverbs.

Morphophonemics

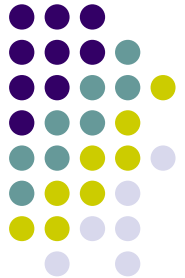


- Changes that occur when a suffix attaches to a root word.
- Change depends on
  - the nature of the end letter of the root word
  - the nature of the start letter of the suffix

ma:la:-a:l => ma:la:va:l pal-a:l => palla:l

‘Mala’-INS => ‘by Mala’ ‘tooth’-INS => ‘using tooth’

Paradigm-based approach.



- Follows from the morphophonemic changes.
- Those root words which behave similar are grouped.
- Paradigmatic classification for Tamil

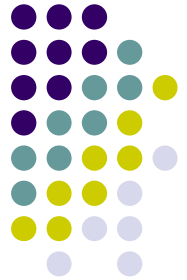
  – 36 noun paradigms and 34 verb paradigms

  ya:ci ‘beg’ takes *tt/kkiR/pp* as the three tense markers.

  viya 'wonder' takes *Ńt/kkiR/pp* as the three tense markers.

ya:ci-tt-a:n ‘beg’-PST-3SM	*ya:ci-kkiR-a:n* ‘beg’-PRE-3SM	*ya:ci-pp-a:n* ‘beg’-FUT-3SM
viya-Ńt-a:n ‘wonder’-PST-3SM	*viya-kkiR-a:n* ‘wonder’-PRE-3SM	*viya-pp-a:n* ‘wonder’-FUT-3SM

Finite State Automata…(1)



- A Finite-state automaton is a model of behavior consisting of a finite number of states, transitions from each state to another state and actions at each transition.
- Morphological generator moves from one state to another as each attribute is applied to the stem and the suffix is generated.

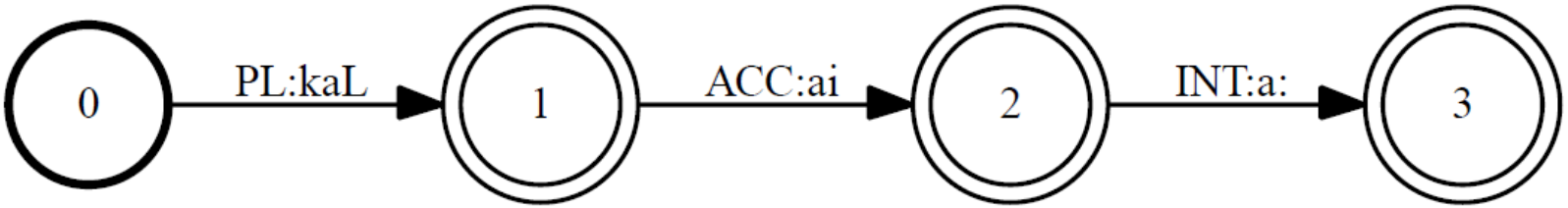
paiyan-kaL-ai-a: => *paiyankaLaiya:*

‘boy’-PL-ACC-INT => ‘the boys(OBJ)?’

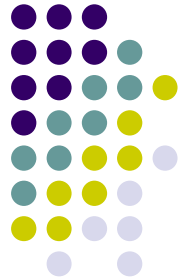
Finite State Automata…(2)

Input: *paiyan*, Plural, Accusative, Interrogative.

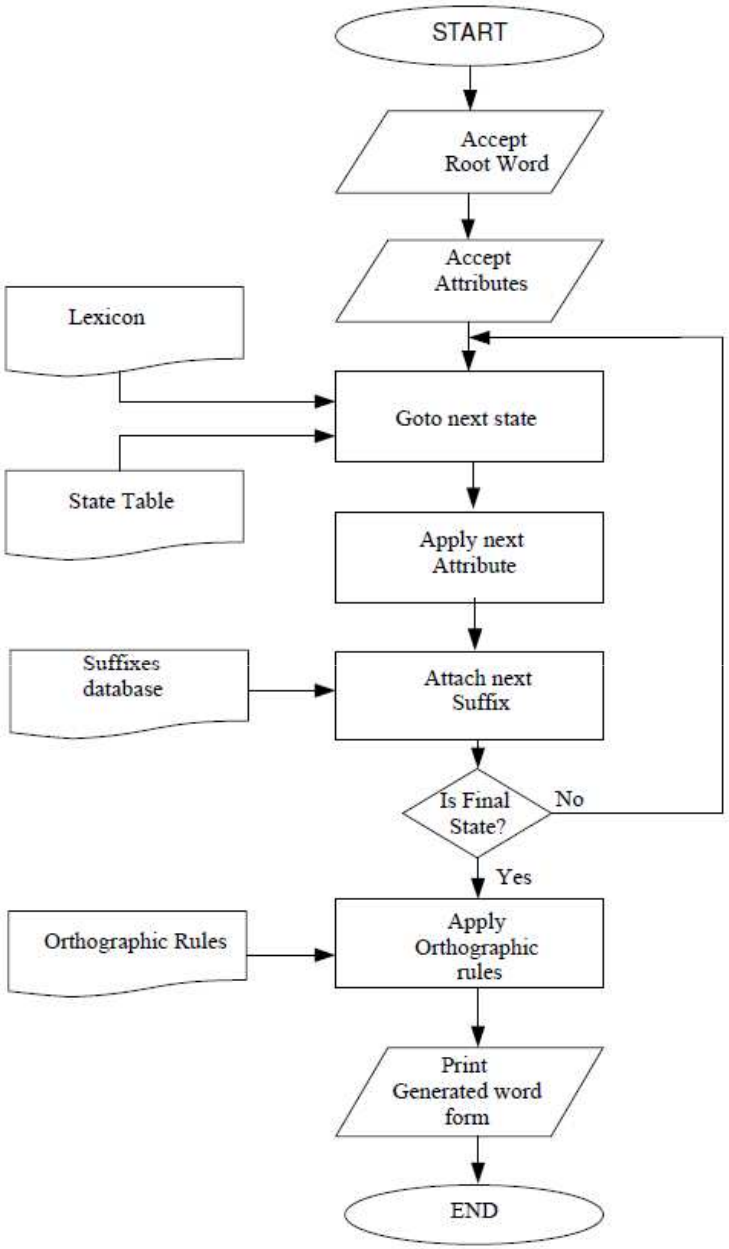
From State	To State	Attribute	Form Generated	Finalform
0	1	PL	*paiyankaL*	*paiyankaL*
1	2	ACC	*paiyankaLai*	*paiyankaLai*
2	3	INT	*paiyankaLaia:*	*paiyankaLaiya:*



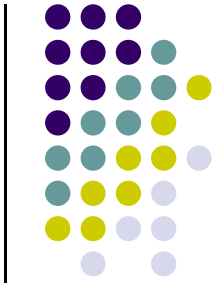
Design of MorphGenerator for Tamil



- A finite state automaton
- Moves from one state to another while attaching suffixes.
- End state produces the desired output
- Resource files
  - Lexicon
  - Suffix table
  - State table
  - Morphophonemic rules


START
Accept Root Word
Accept Attributes
Lexicon
Goto next state
State Table
Apply next Attribute
Suffixes database
Attach next Suffix
Is Final State?
No
Yes
Orthographic Rules
Apply Orthographic rules
Print Generated word form
END

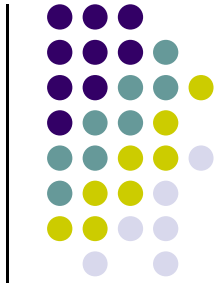
Evaluation – Experiment 1



2556 input words with noun roots spanning different paradigms and different attributes were tested.

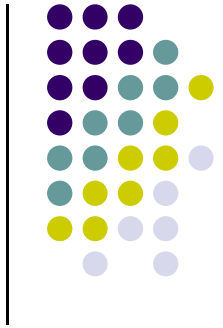
No. of True Positives (TP)	No. True Negatives(TN)	No. of False Positives (FP)	No. of False negatives (FN)	Precision TP/(TP +FP)	Recall TP/(TP + FN)	F-measure
2413	115	5	23	0.997	0.99	0.99

Evaluation – Experiment 2



19152 input words with verb roots spanning all the paradigms and various attributes were tested.

No. of True Positives (TP)	No. True Negatives(TN)	No. of False Positives (FP)	No. of False negatives (FN)	Precision TP/(TP +FP)	Recall TP/(TP + FN)	F-measure
17361	1451	38	302	0.997	0.98	0.99



Thank You!