Word Synthesizer Engine

Uma Maheshwar Rao[1], G. Christopher[2], M.

Center for Applied Linguistics & Translation Studies
University of Hyderabad
Hyderabad-46

[1]guraohyd@gmail.com, [2]efthachris@gmail.com

Content

- Introduction
- Organization of Data
  - *Lexicon*
  - *Feature Value Table*
  - *Synthesis Rule Set (Morphological Rules)*
- Computational Model of WordGen
  - Uni-planar
  - multi-planar
- Possible Combinations of Feature Values
- Input/ Output Specification
- Conclusion & Results

Introduction

- This Morphological Generator engine is based on word-and-paradigm method.
- A generic Engine which can be used for any language by plugging in a specific language data-base.
- This Generator synthesizes all and only the well-formed word forms.
- These word forms include both inflectional and productive derivational forms.
- The engine takes as input a root and along with it its inflectional categories (features) like gender, number, person and case in case of nouns and verbal categories in case of verbs and other relevant inflectional endings depending on the category.
- The Input and output are in Shakti Standard Form (SSF).

2. Organization of Data

- The current morphological generator requires the following basic resources or the morphological data base that is described below.

These resources are of three types:

- *Lexicon*
- *Feature Value Table*
- *Synthesis Rule Set (Morphological Rules)*

2.1 Lexicon

- The lexicon is a dictionary containing a list of roots/stems, each with its lexical category and paradigm type

- It is organized in the form of a simple linear, non-hierarchical sequence of the root, delimiter, lexical category, delimiter and paradigm type

S.No	Root/Stem form	Lexical Category (lcat)	Paradigm Type
1	*“winu”*	“v”	*“koVnu”*
2	*“maMwri”*	“n”	*“gaxi”*
3	*“welika”*	“adj”	*“lewa”*
4	*“appudu”*	“adv”	*“appudu”*
5	*“iwadu”*	“pn”	*“vAdu”*

2.2 Feature Value Table

- The Feature Value Table is essentially a list of affixes with their morpho-syntactic feature values like gender, number, person and the relevant morphological category information stored in the form of a table

- Feature value Table contains lcat, affix and case associated with nouns and pronouns; tense, aspect and modal categories with or without gender, number and person associated with verbs.

S.No	Rule No.	lcat	Affix	Gender	Number	Person
1	719	v	*iwi*	m	pl	2
2	653	n	*wopAtu*	null	sg	null
3	1649	pn	*lekuMdA*	null	pl	null
4	963	adj	*ti*	null	pl	null

2.3 Synthesis Rule Set

- The synthesis rule set is an exhaustive rule set, essentially a combination of concatenation processes which add the desired suffix to the given root/stem and which itself is appropriately modified by the relevant deletion rules.

- Both the add rule and deletion rule may apply vacuously in case the value of the character string to be added or deleted is null

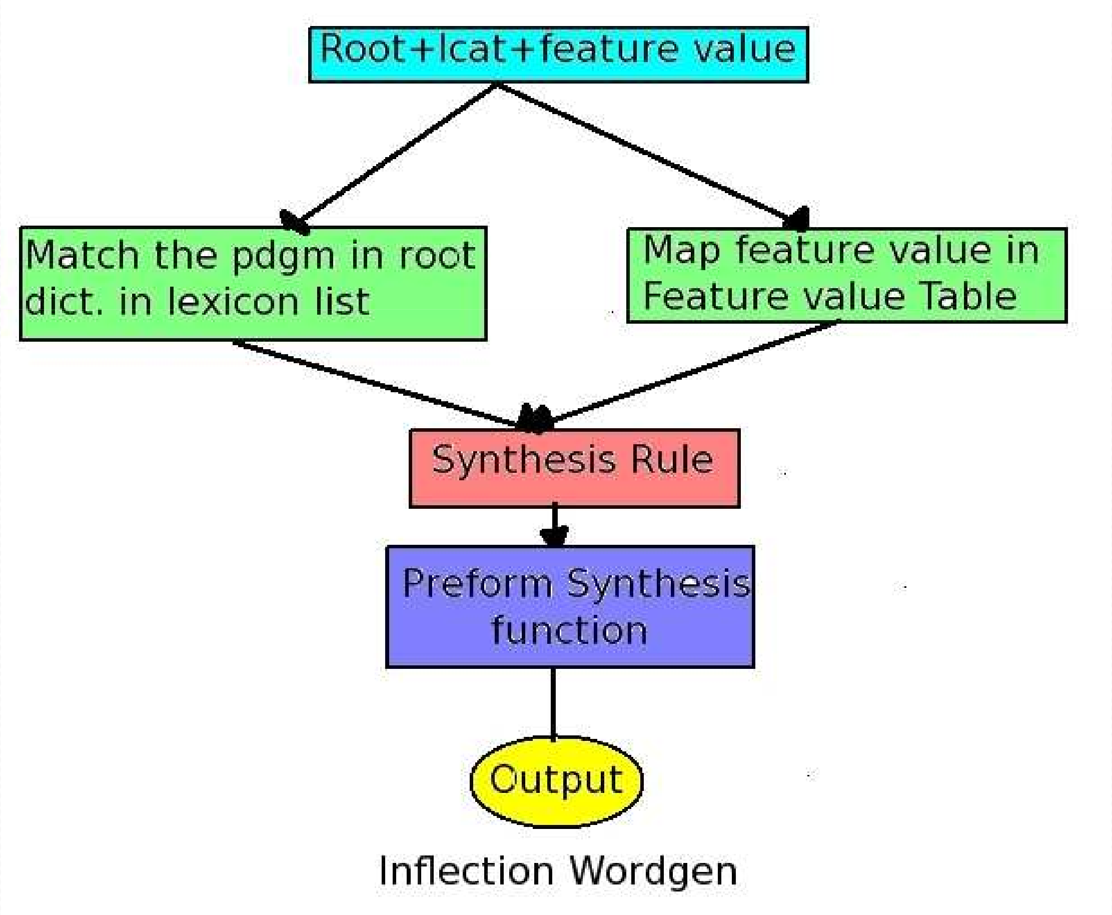
S.No	Concatenation of suffix by Add Rule	Stem/Root modification Delete Rule	Paradigm Type	Rule No.
1	*A*	*u*	*vAdu*	1649
2	*IsAdA*	*iyyi*	*wiyyi*	653
3	*akuMdA*	*u*	*poVg?du*	719
4	*NNilekuMdA*	*du*	*snehiwudu*	963

2.4 Morph Data organization

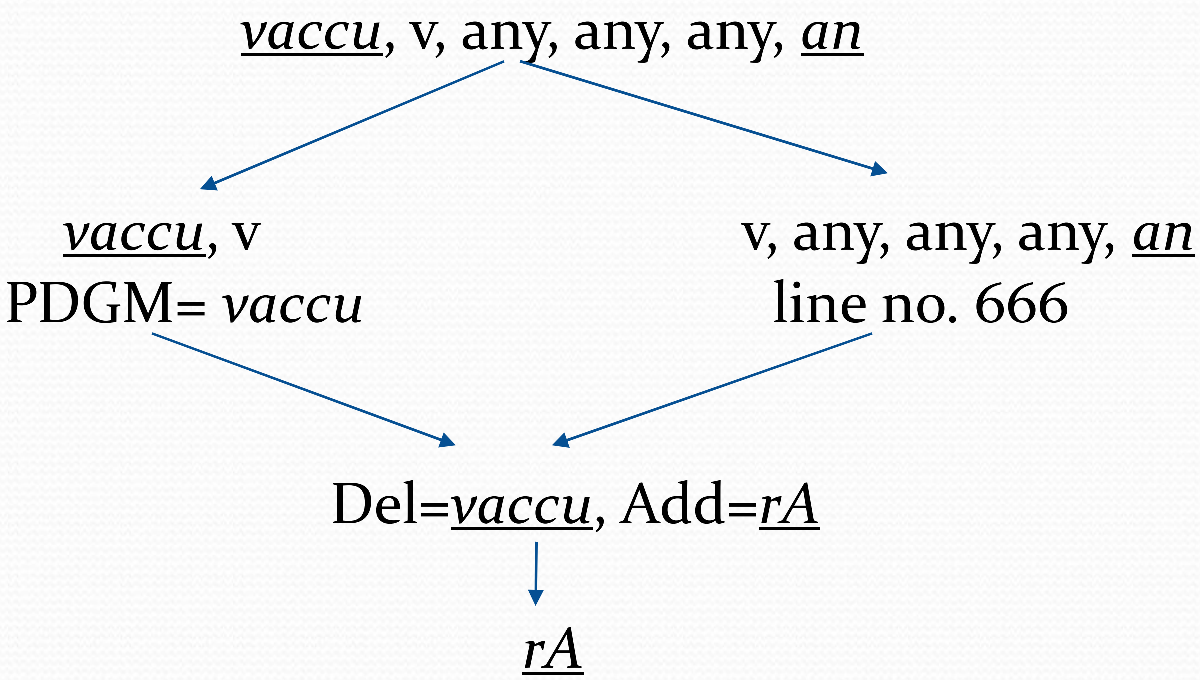
Paradigmatic Members/ wordforms	Common Maximal match	Functional/ Formative Element	Feature Values
winnAdu	*win*	*nAdu*	Past-m-sg-3
wiMtAdu	*wi*	*MtAdu*	np-m-sg-3
winadu	*win*	*adu*	neg-m-sg-3
wini	*win*	*i*	nf-past
wine	*win*	*e*	nf-adjl-ppl
winu	*winu*	o	imp-sg

4.1 Model of WordGen

-Uni-planar Morphological Generator


Root+lcat+feature value
Match the pdgm in root dict. in lexicon list
Map feature value in Feature value Table
Synthesis Rule
Preform Synthesis function
Output
Inflection Wordgen

Root word= *vaccu*, lexical category=v, gender= any, number=any, person=any and suffix=*an*



vaccu, v, any, any, any, *an*

vaccu, v
PDGM= *vaccu*

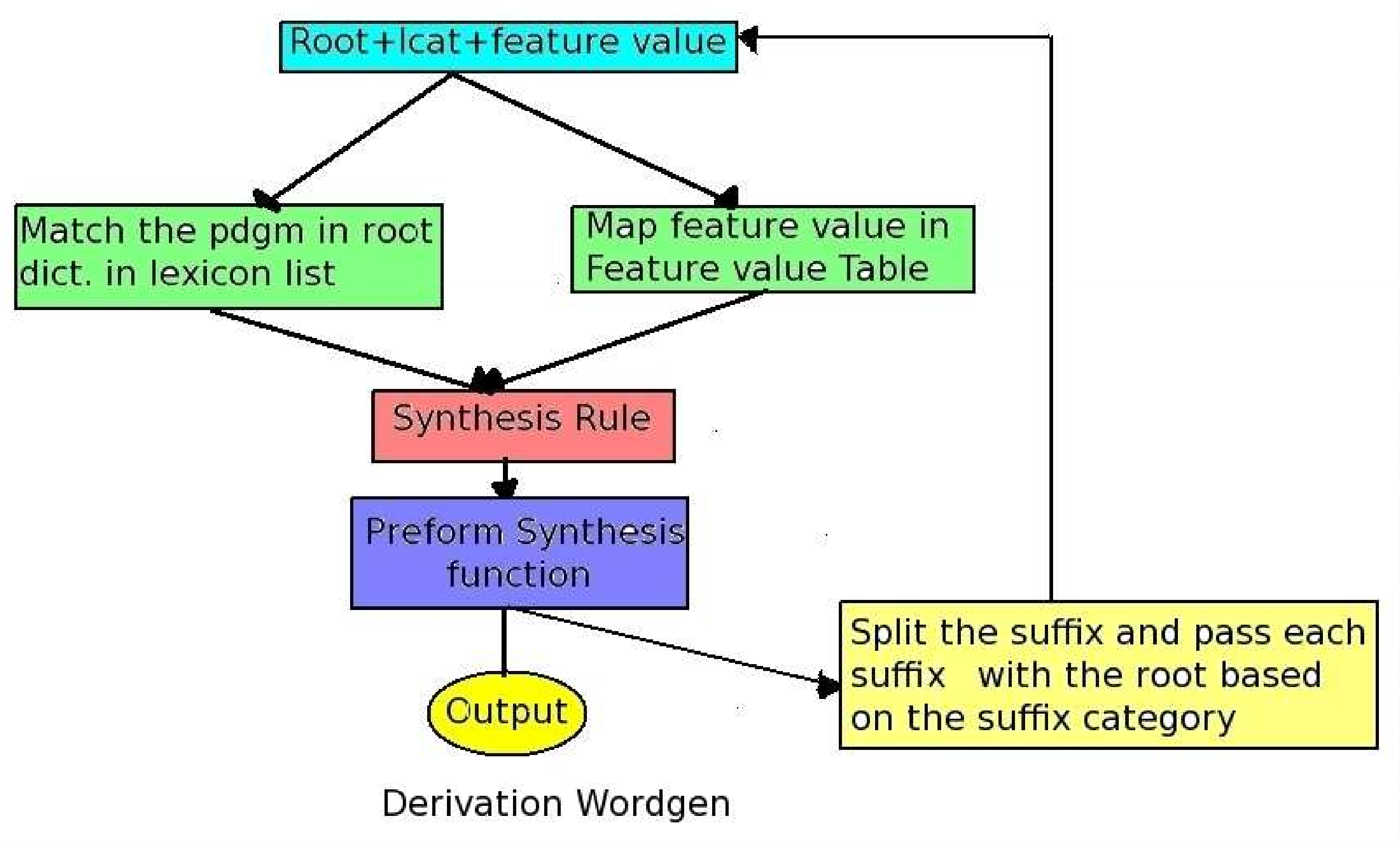
v, any, any, any, *an*
line no. 666

Del=*vaccu*, Add=*rA*

rA

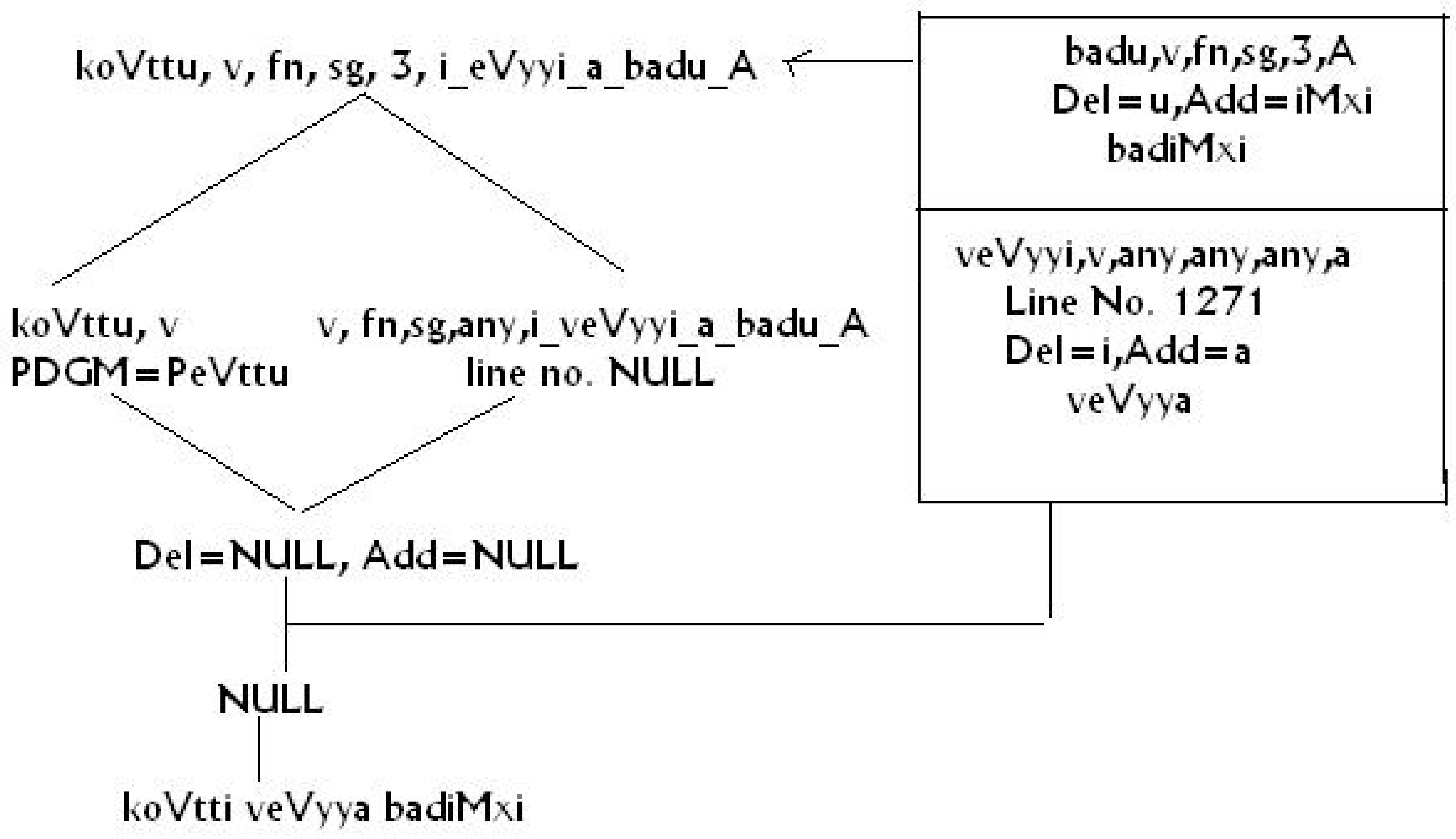
4.2 Model of WordGen

-Multi-planar Morphological Generator


Root+lcat+feature value
Match the pdgm in root dict. in lexicon list
Map feature value in Feature value Table
Synthesis Rule
Preform Synthesis function
Output
Split the suffix and pass each suffix with the root based on the suffix category

Derivation Wordgen

Root word = <u>*koVttu*</u>, lexical category = v, gender = fn, number = sg, person = 3 and suffix = *i_veVyyi_a_badu_A*



koVttu, v, fn, sg, 3, i_eVyyi_a_badu_A

badu,v,fn,sg,3,A
Del=u,Add=iMxi
badiMxi

veVyyi,v,any,any,any,a
Line No. 1271
Del=i,Add=a
veVyya

koVttu, v
PDGM=PeVttu

v, fn,sg,any,i_veVyyi_a_badu_A
line no. NULL

Del=NULL, Add=NULL

NULL

koVtti veVyya badiMxi

5. Possible Combination of Feature Values

NOUNS. root + mopho-syntatic functional categories

eg: noun + gender + number + person + d/o + case/other functional categories

PRONOUNS. root + mopho-syntatic functional categories

eg: pronoun + gender + number + person + d/o + case/other functional categories

ADJECTIVES. root + mopho-syntatic functional categories

eg: adjective + gender + number + person + d/o + case/other functional categories

VERBS. root + mopho-syntatic functional categories

eg: verb + gender + number + person + AuxVerb/other functional categories

NUMBER WORDS. root + mopho-syntatic functional categories

eg: number word + gender + number + person + d/o + case/other functional categories

NST (Nouns of Space and Time). root + mopho-syntatic functional categories

eg: locative + d/o + case/other functional categories

INDECLINABLES. They include the following.

Particles - look up in the root lexicon

eg: (Ewe, gAnI, EnA, etc.)

Adverbs - look up in the root lexicon

eg: adverb (bAgA, atlA, etc.)

Postpositions - look up in the root case marker list

eg: postposition (ni, ki, lo, va, etc.)

6. Input/ Output Specification

- Input for this computational model of Morphological Generator is in *Shaskthi Standard Format (SSF);*
- We have a token number, token, pos-tag and its morphological analysis.
- All these are in different fields (Columns).
- *Morph analysis*, is in fourth coloum in which

1st field is root, *2nd is lex.cat,*
3rd is gen, *4th is num,*
5th is per, *6th is case (d/o),*
7th is case marker/TAM, *8th is suffix.*

By using all the seven elements of the Morphological Analysis, the Generator generates the word forms and modifies the 2nd column i.e. *token* of the SSF format.

5.1 Input in SSF

```
<Sentence id=”1”>
1   ((      NP      <fs af='rAmudu,n,m,sg,3,d,ku,ku'>
1.1         rAmudu          NN      <fs af='rAmudu,n,m,sg,3,d,ku,ku'>
    ))
2   ((      NP      <fs af='Akali,n,m,sg,3,d,0,ku'>
2.1         Akali   NN      <fs af='Akali,n,m,sg,3,d,0,ku'>
    ))
3   ((      VGF     <fs af='veVyyi,v,m,sg,1,,A,A'>
3.1         veyyi   VM      <fs af='veVyyi,v,m,sg,1,,A,A'>
    ))
</Sentence>
```

5.2 Output in SSF

```
<Sentence id=”1”>
1   ((      NP       <fs af='rAmudu,n,m,sg,3,d,ku,ku'>
1.1         rAmudiki        NN       <fs af='rAmudu,n,m,sg,3,d,ku,ku'>
    ))
2   ((      NP       <fs af='Akali,n,m,sg,3,d,0,ku'>
2.1         Akali    NN      <fs af='Akali,n,m,sg,3,d,0,ku'>
    ))
3   ((      VGF      <fs af='veVyyi,v,n,sg,1,,A,A'>
3.1         vesiMxiVM        <fs af='veVyyi,v,n,sg,1,,A,A'>
    ))
</Sentence>
```

7. Conclusion & Results

- *WordGen* generates word forms for all the lexical classes where some sort of inflection is involved as in: nouns, pronouns, verbs, adjectives and locative nouns. This generator is designed to handle inflectional and productive derivational suffixes. The current version of the tool is integrated with IL-ILMT Hindi-Telugu, Telugu-Hindi, Telugu - Tamil and Tamil-Telugu systems (CALTS, University of Hyderabad).
- When tested with languages like Telugu, Hindi and Tamil their accuracy was 97.2%, 98% and 94% respectively.

Thank you !!!!