On the Description of Morphological Data for Morphological Analyzers and Generators: A Case of Telugu, Tamil and Kannada

Uma Maheshwar Rao G., Parameshwari K.

CALTS,University of Hyderabad

OVER VIEW

- Introduction
- Methodological Foundations for building MA and MG.
- Data Organization
- Interferences in Morphology
- Conclusion

Introduction

- Tamil, Telugu and Kannada, the three major and dominant Indian languages of Dravidian Family - regarded as morphologically rich and agglutinating.
- Words are formed not only from the concatenation of two or more morphological elements but also involve multilevel derivation.
- Morphological complexity demands sophisticated morphological analyzers and generators.

MA & MG :

- A Morphological Analyzer is a computational tool or a device which analyzes word into its root, its category and the relevant morpho-syntactic information in terms of its constituent morphemes.
- A Morphological Generator is an inverse of it i.e. given the root, the relevant category and the desired morpho-syntactic information, it synthesizes well-formed word forms.

METHODOLOGICAL FOUNDATIONS FOR BUILDING MA AND MG

- Requires an appropriate theoretical foundation of morphology.
- To Satisfy certain requirements as
  - Evolutionary System
  - Transparent System
  - Modular System
- An appropriate Organization of
  - Linguistic database
  - Meta-linguistic database
  - Computational modeling

Morphological Models

- Hockett (1954) distinguishes the morphological processes as,
  - Item and Arrangement (IA) model
    - Word forms as sequences of concatenated morphemes.
  - Item and Process (IP) model
    - Word form as the result of applying rules which modify a root or stem to produce a new one.
  - Word and Paradigm (WP) model
    - Word form as a functional projection of the root/stem (lexeme) according to the specifications of the formative elements in the paradigm.

Cont…

- The three models represent the combinatorial approach involving the segmentation of a string into a base and its constituent affixe(s).
- Ford and Singh(1985) propose a relational approach to morphology.
  - It focuses on the word and discards the notion of morpheme.
  - Alternatively known as whole-word morphology is essentially a list of exhaustive morphological relations expressed in the form of morphological strategies.

WORD AND PARADIGM BASED MORPHOLOGY FOR TAMIL, TELUGU AND KANNADA.

- WP : Better suited model for Computational Implementation.
- It assumed here involvement of an exhaustive collection of each and every word form relatable to a lexeme.
- Paradigm refers to an exhaustive set of morpho-syntactically related word forms of a given lexeme.

Cont…

- Each word form in a paradigm is considered as a formal expression of the root/stem (lexeme) associated with a morpho-syntactic function.
- Organized conveniently into tables, by classifying them according to the shared inflectional categories.
- Advantage of using paradigms : since it is easier to organize, modify and improve upon the data at a later stage

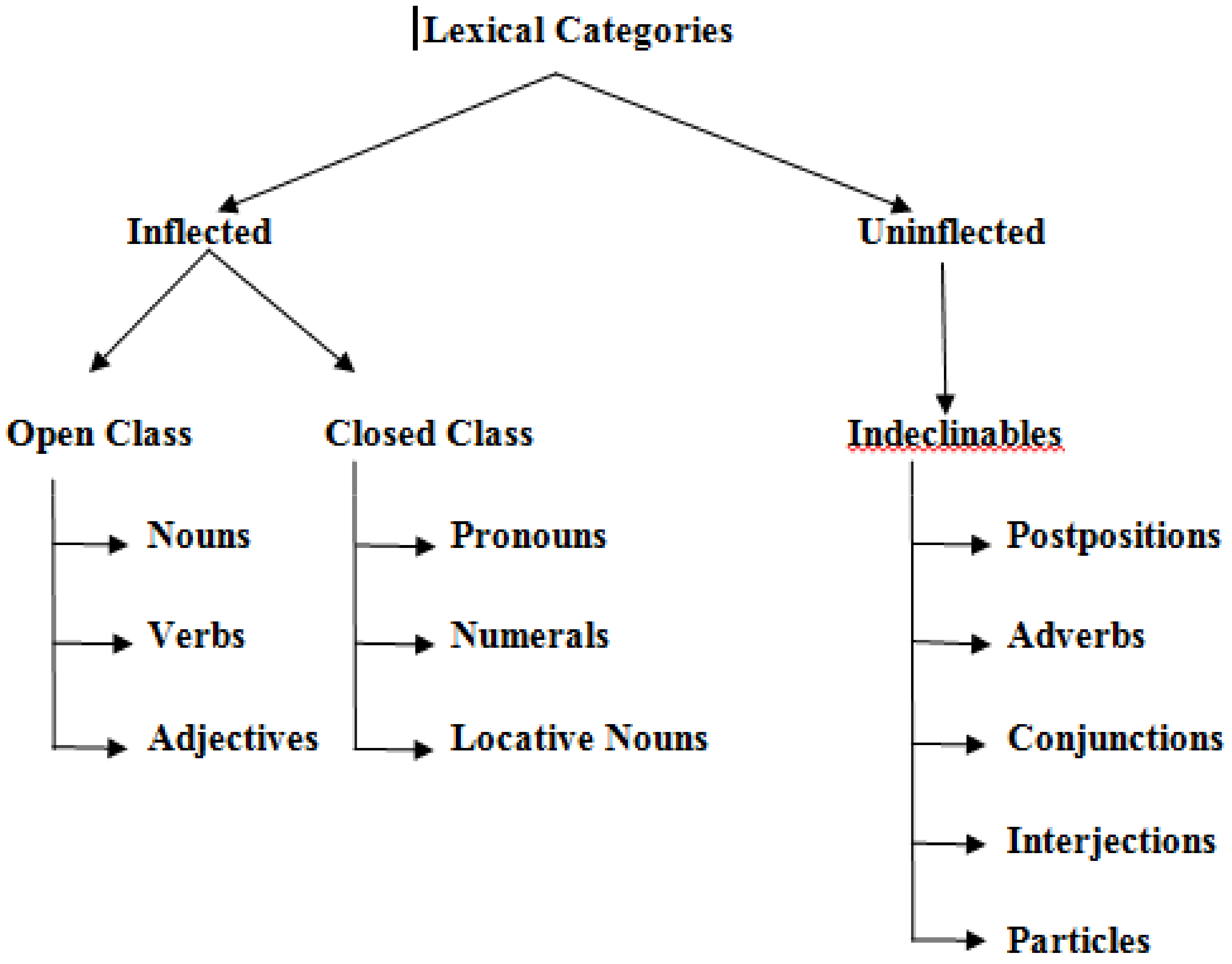
DICTIONARY OF ROOTS/STEMS:

- A root word dictionary in a MA and a MG differs from a conventional dictionary.
- In a Morphological dictionary, the root words are listed in such a way that each root/stem is identified by an appropriate paradigm type in addition to their lexical category.
- For instance,

  Tamil: mAtu,n,kAtu

DATA ORGANIZATION:

- Demand for Comprehensive but exhaustive analysis of Inflectional and productive Derivational.
- Morphological categories will be defined by their participation in inflection.
- There are two classes of roots/stems as far as the morphological phenomena are concerned.
  - A majority of roots/stems anchor various affixes or morpho-syntactic functional elements often called inflectional categories.
  - Conversely, a minority of roots incapable of receiving such inflection are often called as indeclinables.



Lexical Categories

Inflected

Uninflected

Open Class

Closed Class

Indeclinables

Nouns

Verbs

Adjectives

Pronouns

Numerals

Locative Nouns

Postpositions

Adverbs

Conjunctions

Interjections

Particles

MORPHOLOGICAL LAYERS.

- Word forms are considered to have one or more of the following layers.
  - root/stem + morphological category
  - root/stem + morphological category + category specific particles
  - root/stem + morphological category + (category specific particles) + syntactically relevant clitics.

Current Development

- The morphological analyzers and generators are developed for the three Dravidian languages viz. Tamil, Telugu and Kannada.
- Large corpora can be used as very effective and rich source for the identification and inclusion of new word forms into our morphological databases.
- The running texts of corpus are converted into groups of reverse sorted word forms enabling us to identify exhaustively the word forms of distinct paradigms.
- The well-organized data are implemented into the computational system with the adequacy of computational requirements, viz. exhaustiveness, coverage and precision

INTERFERENCE IN MORPHOLOGY

(i) EXTERNAL SANDHI

Tamil : *yAnYEk kutti* ‘small elephant’
anwac cattam ‘that law’
wamilYYw wAy ‘Mother of Tamil’
curYrYulAþ þayaNi ‘tourists’

This requires the deletion of the consonants before they are passed on to the Morphological Analyzer.

REDUPLICATIVES

- A number of word formation processes involving echo words, balancing words, complete reduplications etc. involve information from a neighboring word.
- Tam: *patam kitam*‘picture and the sort of’
- Tel : *kurcl gircl* ‘chair and the sort of’
- Kan: *puswaka giswaka* ‘book and the sort of’

WORD LEVEL VARIATION

- A number of word forms have variations wherein a single lexical item has two or more spellings. These includes,
- Orthographic Variations:

Tamil:	*anAwE*	‘orphan’
	anYAwE	‘orphan’
Telugu:	*puswakaM*	‘book’
	puswakamu	‘book’
Kannada:	*kiricu*	‘shout’
	kirucu	‘shout’

WORD LEVEL VARIATION

- Inflectional Variations:

Tamil:	*urYuppu-kkalY*	‘parts (of body)’
	urYuppu-kalY	‘parts(of body)’
Telugu:	*bledulu*	‘blades’
	bledlu	‘blades’
	blelYlu	‘blades’

WORD LEVEL VARIATION

- Dialectal Variation:

Tamil:	*paticcenY*	‘I read’
	patiwwenY	‘I read’
Telugu:	*vaccAnu*	‘I came’
	vaccinAnu	‘I came’

Nativized English loans

- Certain loan words from English are used very frequently.

Tamil	Telugu	Kannada	Gloss
pas	*bassu*	*bassu*	‘bus’
areVst	*areVstu*	*areVstu*	‘arrest’
potto	*Poto*	*Poto*	‘photo’

Such frequently used nativized English loans are added into the database so that they can be recognized

PROPER NOUNS

- Proper nouns such as names of persons, places or institutions etc. are an open class. Proper nouns increase in number, day by day and they may occur in any domain of language.

Ex. names

Tamil: *rAmacAmy, wevar, nAtAr etc.*

Telugu: *subbArAvu, rAvu, reVddi etc.*

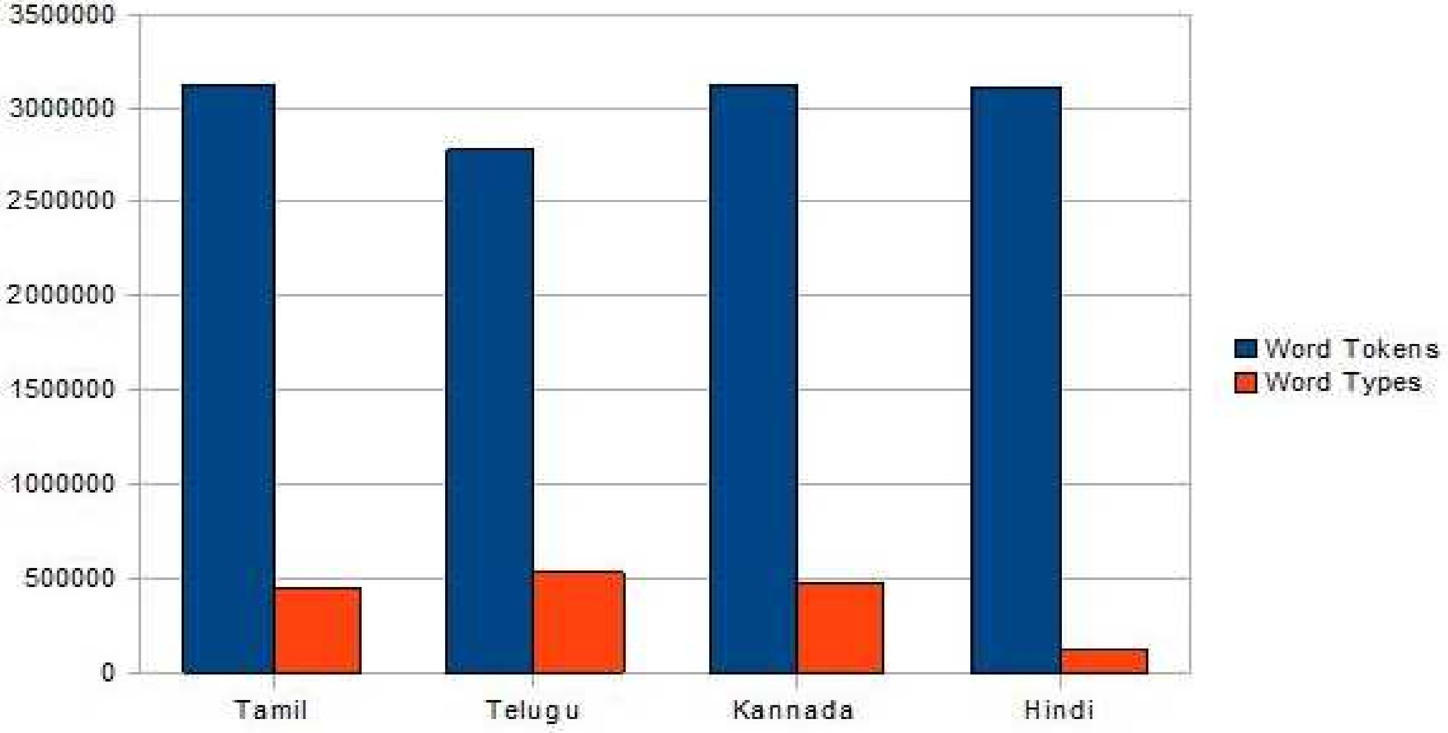
Kannada: *rAju, etc.*

Type Token Ratio of Tamil, Telugu,

kannada along with Hindi

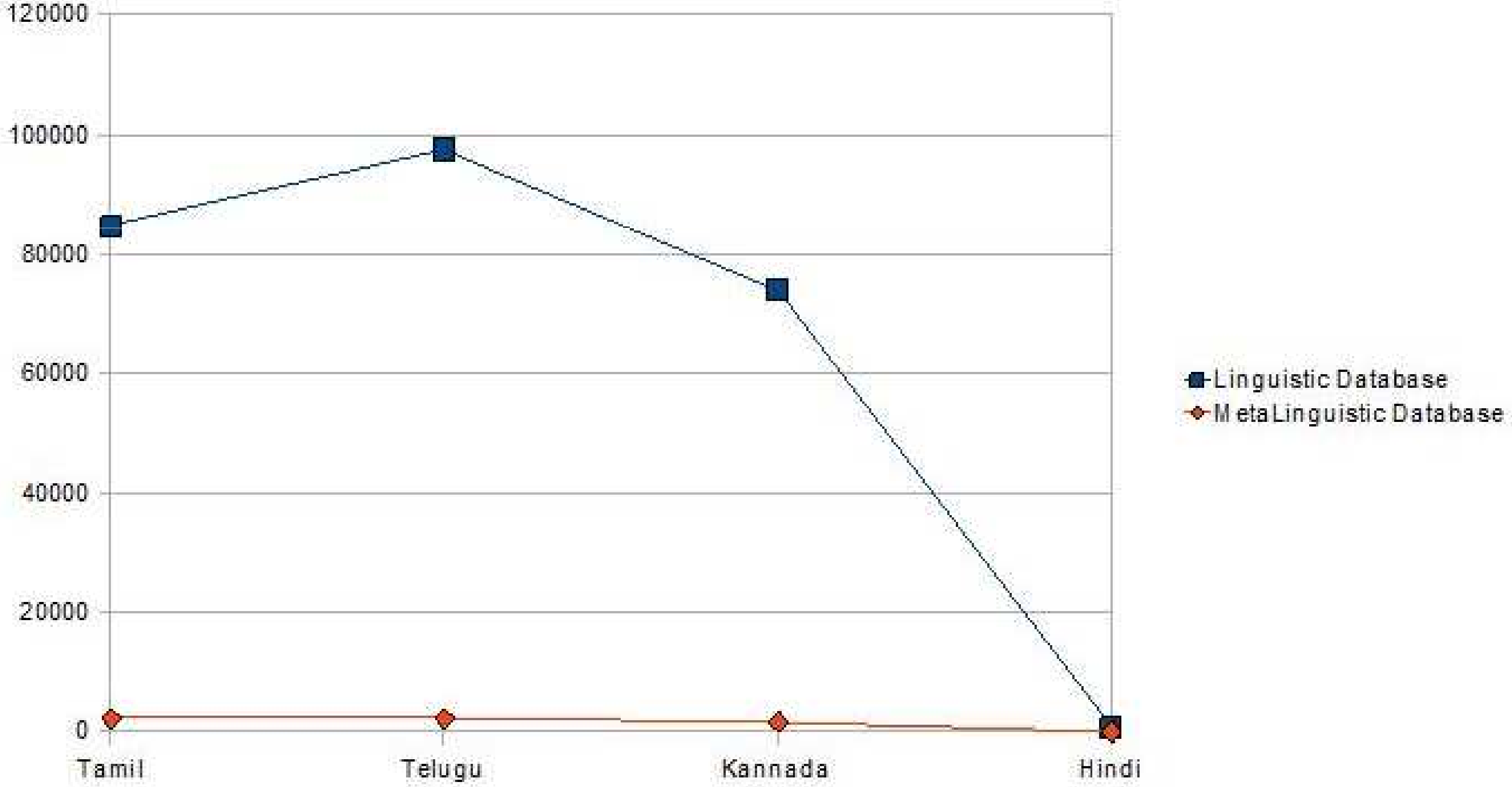
Corpus (CIIL)	Word Tokens	Word Types	Type Token Ratio
Tamil	3124447	445361	7.01
Telugu	2769797	534628	5.18
Kannada	3118987	474066	6.57
Hindi	3104668	120227	25.82

Type Token Ratio :



A Comparative Study on the Density of the Morphological Database

Language	Linguistic Database	MetaLinguistic Database	Density of the database
Tamil	84785	2258	37.54
Telugu	97485	2311	42.18
Kannada	74020	1794	41.25
Hindi	815	183	4.453



CONCLUSION

- The careful appraisal and study of the unrecognized words is conducted to identify and overcome the lapses by incorporating certain amount of data into the morphological database.

- The morphological data discussed above demonstrates that morphologically complex languages require exhaustively but enriched linguistic database for the purpose of developing a morphological analyzer/generator.