Introducing a measure of Morphological complexity in Indian Languages from the perspective of Morphological Analyzers

G. Uma Maheshwar Rao, Christopher M and Parameshwari K.
Center for Applied Linguistics and Translation Studies
University of Hyderabad

Introduction

A knowledge of the Morphological complexity of languages is crucial in the context of building morphological analyzers from the point of coverage, precision, and the amount of time required.

This paper focuses on the comparison of the morphological Complexity between different Indian Languages especially the Indo-Aryan and Dravidian family languages with respect to the measure of the complexity.

In order to compare the morphological complexity of these Languages vis-à-vis the building of the morphological analyzers we propose to show that there exists a relationship between the Type Token Ratio (TTR) and the Lexical Density Measure (LDM).

It can be stated that the relationship between the complexity of morphology and the corresponding analyzer can be obtained from the study of the two kinds of proportional relations expressed in terms of TTR and LDM.

This paper advocates that a considerably complex morphological data requires a well-sophisticated data organization techniques and morphological modelling for the building of Morphological Analyzers.

Questions:

what is a morphological complexity (MC)?

Morphological Complexity pertaining to language is described by various scholars as such, Morphological Complexity, according to Nichols (1992) is based on the number of points at which a typical sentence is capable of receiving inflection. She develops a measure to produce a comparative numerical index of complexity.

Max Bane (2008) defines Language's morphological complexity as the proportion of the lexicon’s total description length that is due to the description lengths of the affixes and signatures.

Berlin and Kay (1969) discover an elegant hierarchy of the number of basic colour terms in various languages. It is relatively simple to determine where languages fall on this

hierarchy and thus to compare complexity of their “color systems”.

The idea of efficient comparison of complex systems is implicit in the “Principles and Parameters” approach to language acquisition (Dorr, 1993).

According to McWhorter (2001) a language is more complex if it has more marked members in its phonemic inventory, or if it makes more extensive use of inflectional morphology. But the problem of practices is how do you compare two different inflectional paradigms, or how do you balance simple morphology with complex phonology.

Patrick Juola (2008) languages which are morphologically complex tend to have a wide variety of distinct linguistic forms, while languages which are morphologically simple have more word tokens, repeating a smaller number of types.

we try to describe Morphological Complexity with the help of Type Tokan Ration (TTR) and Lexical Density Measure (LDM). After analyzing the corpus available for langauges, TTR and LDM we generaise the Morphological Complexity is inversly propositional to TTR and directly propositional to LDM.

Morphological Complexity can also defined as the dot(.) product (scalar product) of Base form (

B f) and set of cross product of (vector product) of case (C) or Tense (T) with all other sets ranging

from Gender (A 1) to Clitic (A n).

What are the criteria that define MC?

Definition:

Morphological Complexity can be defined as the dot(.) product (scalar product) of Base form (

B f) and set of cross product of (vector product) of case (C) or Tense (T) with all other sets ranging from
Gender (A 1) to Clitic (A n).

Let Consider that

B_f = { Base form/ Root form/ Stem}
T = { Present, Past, ... }
C = { nom, acc, dat, ass, ... }
A_1 (GEN) = { m, f, n, fn, fm, ... }
A_2 (NUM) = { sg, pl, dual }
A_3 (PERSON) = { 1, 2, 3, ...}
..
..
..

.. ..
A_n (CLITIC) = { emp, dub, int .....}

Then we know that :

Verb Complexity = B_f . [T x { Π^n_{x-1} (A_1, A_2, ...,A_n) }]

Noun Complexity = B_f . [C x { Π^n_{x-1} (A_1, A_2, ...,A_n) }]

A Combination of these two is:

Morphological Complexity = $\mathbf{B_f}$. [(**C** | **T**) x { $\mathbf{\Pi^n_{x-1}}$ ($\mathbf{A_1, A_2, ...,A_n}$) }]

Complexity : Morphology vis-à-vis Morphological Analyzers

In order to compare the morphological complexity of Indian Languages the following languages are selected viz., Tamil, Telugu, Kannada and Malayalam from Dravidian Languages, Hindi,
Bengali, Marathi, Punjabi and Urdu from Indo-Aryan Languages.

The Complexity of languages can be obtained from the two kinds of proportional relations as
TTR and LDM.

Type Token Ratio :

Type Token Ratio (TTR) is the measure of the variation of vocabulary occurring in a text. The
'tokens' refer to the total number of word forms occurring in a text and 'types' refer to the total
number of unique or distinct word forms in a text. The TTR is interpreted as higher is the ratio
lower is the degree of complexity. The CIIL Corpus of Tamil, Telugu, Kannada along with Hindi are
taken for the analysis of Type Token Ratio. This analysis explicates the morphological complexity of
the languages. The TTR of Hindi is higher indicating lesser degree of complexity, whereas Telugu,
Tamil and Kannada exhibit lower TTR measures indicating a more complex Morphology.

Corpus (CIIL)	Word Tokens	Word Types	Type Token Ratio 1:
Telugu	2,769,797	534,628	5.18
Kannada	3,118,987	474,066	6.57
Tamil	3,124,447	445,361	7.01
Malayalam	2313855	542,657	15.58
Marathi	1784198	196916	9.06

Urdu	117240	7783	15.06
Oriya	2966417	192464	15.41
Bengali	2531295	162,454	15.58
Punjabi	2308030	104368	22.11
Hindi	3,104,668	120,227	25.82

Table 1: Type Token Ratio

Hence, we can state that there is an inverse proportional relation between the number of
distinct word forms (types) and the number of the total instances of wordforms (tokens).

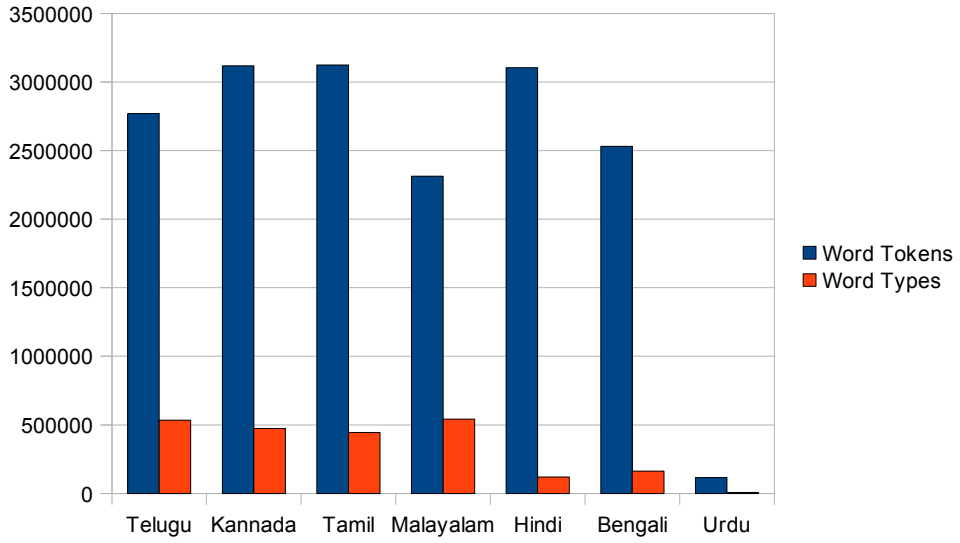

CHART 1: TYPE TOKEN RATIO

Lexical Density Measure :

Lexical Density Measure(LDM) here refers to the weightage of the linguistic database which is
used in building the Morphological Analyzer. The density measure is again a ratio of metalinguistic
database to the linguistic database. The Metalinguistic database involves the number of features or
morpho-syntactic categories of the relevant language and the Linguistic Database is a list of formal
expressions of the morpho-syntactic categories of the language as a set of paradigmatic forms along
with their variations. The table below depicts the Density of the databases of Tamil, Telugu and
Kannada which is considerably high when compared to that of Hindi.

Language	Meta Linguistic Database	Linguistic Database	Density of the database (LDM)
Telugu	2311	97,485	42.18
Kannada	1794	74,020	41.25
Tamil	2258	84,785	37.54
Malayalam	3756	43745	11.64
Punjabi	124	1306	10.53
Marathi	13983	106377	7.6
Hindi	183	815	4.453
Urdu	186	815	4.38
Bengali	5378	13162	2.45

Table 2: Lexical Density Measure

The higher scores of LDM in the above table display the morphological complexity of Dravidian Languages. Such a complexity require an appropriately designed Morphological Analyzer in order to ensure higher coverage and precision.

There is a direct proportional relation between the size of the metalinguistic database (number of morpho-syntactic categories) and the corresponding linguistic database (formal expressions in terms of word forms).

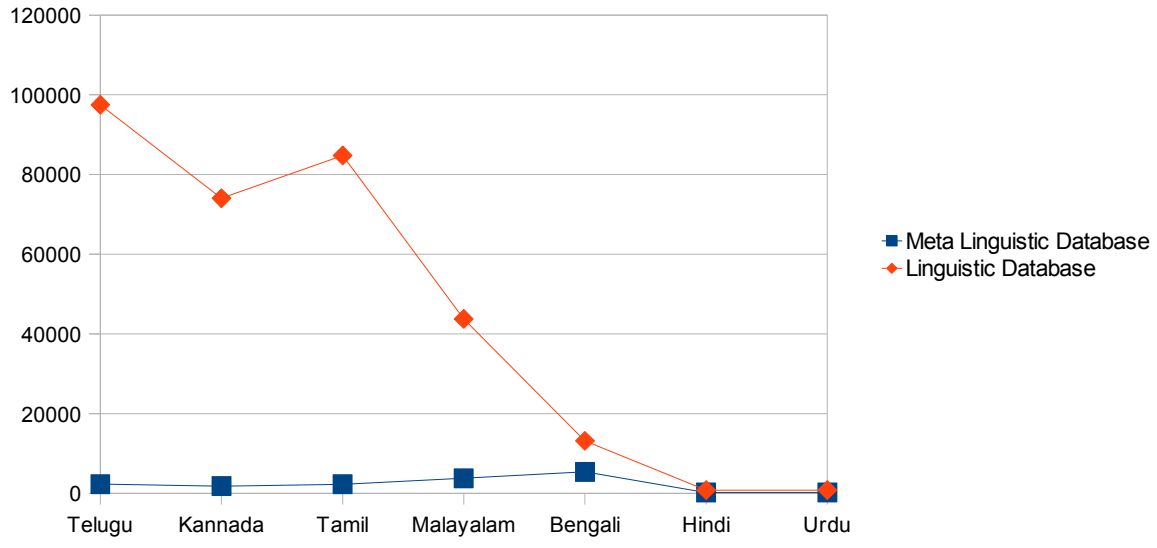

CHART 2: LEXICAL DENSITY MEASURE

The Morphological Complexity of languages vis-à-vis the morphological analyzers may be

expressed by the following statements:

$$\text{Morphological Complexity} \propto \frac{1}{\text{Type Toke Ratio (TTR)}}$$

$$\text{Morphological Complexity} \propto \text{Lexical Density Measure (LDM)}$$