# Unsupervised Morphological Analysis of Hindi Texts

**Aditi Krishn, Rabi Shanker Guha, Amitabha Mukherjee**
**Computer Science and Engineering**
**IIT Kanpur**

**February 20, 2012**

## Introduction

In this project we are trying to implement the algorithm described by Goldsmith[1] and further improved by Goldwater and Johnson[2] on Hindi text. These paper described a Bayesian procedure for the unsupervised learning of morpho-phonological rules from an unlabeled corpus. The primary goal is the determination of the location of the breaks between morphemes inside a word. The grammar returned by the algorithm would consist of a set of signatures alongwith phonological rules for insertion, deletion or replacement which further allow the grammar to collapse into a shorter signature.

## Motivation

There are several aspects of learning a language. One of them is learning the morphology of words. Ther is evidence that students who undertand how words are formed, by combining prefixes, sufffixes and root tend to have larger vocabularies and better reading comprehension and increased success in deciphering unfamiliar vocabulary[3]. This project aims at the same. Moreover unsupervised learning presents unusual challenges to the field of computation linguitics[2].

## Previous Work Done

Goldsmith in his paper An Automatic Morphological Analyzer,2004 describes a model for Bayesian model learning of morphemes. Linguistica[4] is a program designed to explore the unsupervised learning of natural language, with primary focus on morphology (word-structure).
This algorithm has been improved upon by Goldwater and Johnson[2] who describes a method for integration phonological modifiers in the output of Linguistica.

## Methodology

Linguistica uses three primitive types in its grammar: stems, suffixes, and sig-natures. Each signature is associated with a set of stems, and each stem is associated with exactly one signature representing those suffixes which it combines freely.
Sample Input

भारत, भारतीय
उड़ान, उड़, उड़ना, उड़ाना
दौड़, दौड़ान, दौड़ाना, दौड़ना
मार, मारा, मारना, मरवाया, मरना, मर, मरा

Analysed Text

σ1 = ({दौड़, उड़}×{$, आन , ना, आना})
σ2 = ({भारत}×{ $, ईय})
σ3 = ({मार}×{ $, आ, ना})
σ4 = ({मर}×{ $, आ, ना, वाया})

Goldwater and Johnson[1] add another primitive type to the grammar, the notion of a rule based on phonology.On adding phonological rules we expect to find a relation between σ3 and σ4.

**References**

[1] John Goldsmith, Linguistica: An Automatic Morphological Analyzer,2004.

[2] Sharon Goldwater and Mark Johnson,Priors in Bayesian Learning of Phonological Rules,

[3]http://www.uknow.gse.harvard.edu/teaching/TC102-407.html

[4]http://linguistica.uchicago.edu/