Magahi Verb Analyser and Generator

Ritesh Kumar & Dr. Girish Nath Jha
Jawaharlal Nehru University
New Delhi

Magahi

- Magahi appeared as a distinct language around 10th century like other New Indo-Aryan (NIA) languages.
- Grierson has classified Magahi under Eastern group of Outer sub-branch.
- Currently, Magahi speakers count up to 13,978,565 (Census, 2001).
- Ethnologue (1996) reports that Magahi is spoken mainly in Bihar and Jharkhand; but it is also spoken in some parts of West Bengal like Maldah District

Magahi

- Currently, three distinct varieties of Magahi could be recognized:
  - Central Magahi of Patna, Gaya, Hazaribagh;
  - South-Eastern Magahi of Ranchi and some parts of Orissa;
  - Eastern Magahi of Begusarai and Monghyr.
- Amongst these the Magahi spoken in and around Gaya and Patna is generally considered standard because of the obvious social and political reasons.

Verbs in Magahi

- In case of finite verbs, Magahi has three tenses —present, past and future.
- While present is unmarked, the past is marked by ‘-l-’ and ‘-b-’ functions as the marker for future.
- There are three aspects— progressive, stative and habitual.
- Also there are two moods—presumptive and subjunctive— represented morphologically on the verb

Verbs in Magahi

- Basically there are three types of verb stems in Magahi:
  - **Primitive**, monomorphic basic stems like /kʰɑ-/, /d̪ekʰ-/, /sʊn-/, etc.
  - **Derivative** stems. These are formed by adding various kinds of derivative suffixes to the verbal or non-verbal stem.
  - **Complex** verbs. These are formed by adding various kinds of models to the primitive and derived stems.

Complex verbs in Magahi

- The complex verbs in Magahi can be divided into two categories—compound verbals and conjunct verbals.
- Compound verbals involve combinations of two verb-stems.
- Conjunct verbals are those that involve the combination of a substantive (i.e., nouns and adjectives) and a verb stem.

Agreement in Magahi

- The most intriguing and unique feature of Magahi is its agreement system.
- The verb in Magahi agrees with both the subject and the object simultaneously.
- There is no gender and number agreement in Magahi.
- The verb agrees with the person and honirificity of both subject and object.

Agreement in Magahi

- Some examples:

(1)	həm	okərɑ	d̪ekʰə-l-i- əi
	I	him (-Honor)	saw 3P object (-Honour)
	I saw him; 3P Object, -Honour		
(2)	həm	ʊnkɑ	d̪ekʰə-l-i- əin
	I	him (+Honor)	saw 3P object (+Honour)
	I saw him; 3P Object, +Honour.		

Agreement in Magahi

- There is also this phenomenon of suspension of all agreements with object in certain construction, as in the following examples

(1) həm d̪ekʰəli/ d̪ekʰəlio

‘I saw’; Neutral object

(2) həm okərɑ d̪ekʰəliəi/ d̪ekʰəlio

I saw’; 3P Object, -Honour

Magahi as LRL

- According to the Census of India, 2001, Magahi is considered a dialect of Hindi.
- But the fact is that it is a completely different language, with closer relations with Bangla, Oriya, etc rather than Hindi.
- Literate, urban parents dissuade and forcefully stop children from using the language since it is considered 'uncouth' and the 'language of the illiterate'.

Magahi as LRL

- Consequently, Magahi does not have any online resources.
- And there is hardly any effort to develop these resources for the language, since neither the government nor the speakers are concerned or feel a need to develop the computationally useful resources.
- The basic aim of this analyser is to initiate some resource building and language processing for the language.

Needs for LRL

- Like any other LRL there are two basic needs of Magahi.
  - Need to standardise whatever little resources we have such that it could utilised for developing different tools, applications, etc.
  - Need to develop the language foundations (i.e., basic grammatical descriptions, dictionaries, etc.) and tools such that these standardised resources could be utilised.

Developmental Phases for LRLs

- There are four developmental phases for LRLs:
  - Initial Phase (Foundations): building of lexical data-base.
  - Second Phase: basic tools like morphological analysers, POS taggers, etc.
  - Third Phase: development of advanced tools and applications like web crawler and search engines.
  - Fourth Phase: development of general applications like those of information retrieval and extraction, question/answering systems, etc.

Phases in Magahi

- In case of Magahi, the foundational work has yet to be completed.
- There is no collection of corpus as such, since very little data is transferred on the computer, if any at all.
- However the primary job at the foundational stage, i.e., the grammatical and linguistic description of the language, is complete to a very large extent.

The Analyser/Generator

- In this paper we have also tried to take the work further to the second stage by developing a basic morphological analyser/generator for the verbs of Magahi.
- This tool analyses and gives the grammatical category of the given verb form and also generates the verb paradigm for that particular verb root.

The Analyser/Generator

- The users are provided with a GUI in which they are required to input a verb-root or verb-form and the system will give the verb-root, the grammatical category of the verb root and will generate all other forms of the verb.
- The data for developing this analyser is stored in three files in UTF-8 encoding.
- One file has all the lemmas and their English equivalent and the other two files have the inflections and the ECVs, along with the grammatical tags.

The Analyser/Generator

- The inputted form is searched through the list of the roots with the help of a lexicon reader and lexicon search engine.
- If it is there then it is attached with all the inflections and ECVs and finally returned which is displayed as output to the users with all the forms.
- If it is not found then the system checks whether it is a derived form of the verb.

The Analyser/Generator

- The output is displayed both in the Devanagari script and IPA.
- If it is not there then there is no output and the system prompts the user to enter another verb root form.
- The system is developed using Java/JSP as the programming language in the web domain.

A demo of the system

http://sanskrit.jnu.ac.in/student_projects/magahi-sea

The way ahead: Fixing up the Bugs

- As it is clear from the demo, the programme is not very clean.
- We need to fix up a few issues here and there like making the IPA transcriptions complete.
- Searching will be enabled through IPA and English equivalents also.

The way ahead : CL system

- We are planning to expand and make the system more robust by adopting the method of 'construction labelling (CL) system', for enhancing the argument structure specification.
- This system is especially designed for the LRLs and requires extensive linguistic expertise.
- It is a system of representing detailed morph-syntactic and semantic information in such a way that it is computationally useful.

The way ahead: CL system

- The main aim of this CL system is to identify and enumerate all the construction types (within the linguistic limits) of a particular language in a particular domain, down to a certain degree of detail.
- In this system the construction types are represented by strings of letters and hyphens which are called 'templates'.
- These templates are made up of 'labels'.

The way ahead: CL system

- Each construction is displayed from the top, first its properties as a whole are given, followed by properties of its main constituents, their syntactic properties and then finally their semantic properties.
- The area occupied by each type of the level is called 'slot'.
- Thus each slot consist of different labels like that for 'Parts of Speech' ,'valency', etc.

The way ahead : CL system

- This approach of construction labelling would be helpful in developing the morphological analysers/generators (and of course many other tools and applications also) which could analyse the morphemes of different words, if it is given a sentence or even a complete text.
- Later on it could be developed into a language generation tool also.

Open to Questions!