



*English-Bhojpuri SMT System: Insights  
from the Kāraka Model*

*Atul Kr. Ojha*

*[shashwatap9k@gmail.com](mailto:shashwatap9k@gmail.com)*

# *Objectives*

## ***(a) Primary***

- ▶ LT resources for Bhojpuri
- ▶ Improving the accuracy and fluency of the low-resource based MTs system (especially based on statistical method) such as the Indian languages

## ***(b) Secondary***

- ▶ Encoding of Kāraka model in the SMT model
- ▶ Suitability of PD and UD for English-Indian languages (E-ILs) using the SMT

# *Brief Overview of thesis*

There are five main contribution of this thesis:

- ▶ Studies the available E-IL MTs ([chapter 1](#)).
- ▶ Presents a feasibility study of Kāraka model for using the SMT between English-ILs with special reference to the English-Bhojpuri pair ([chapter 2 and 4](#)).
- ▶ LT resources for Bhojpuri ([chapter 3](#)).
- ▶ Towards developing an EB-SMT system using the Kāraka and the UD model for Dep-tree-to-string SMT model ([chapter 4](#)).
- ▶ A documentation of the problems has been secured that enlists the challenges faced during the EB-SMT system and another list of current and future challenges for E-IL MTs with reference of the English-Bhojpuri pair has been curated. ([chapter 5 and 6](#)).



# Bhojpuri Language: An Overview

Bhojpuri is an Eastern Indo-Aryan language, spoken by approximately 50,579,447 (Census of India Report, 2011) people, primarily in northern India which consist of the Purvanchal region of Uttar Pradesh, western part of Bihar, and north-western part of Jharkhand. It also has significant diaspora outside India, e.g. in Mauritius, Nepal, Guyana, Suriname, and Fiji. Verma (2003) recognises four distinct varieties of Bhojpuri spoken in India (shown in figure 3, it has adopted from Verma, 2003):

- ▶ Standard Bhojpuri (also referred to as Southern Standard): spoken in, Rohtas, Saran, and some part of Champaran in Bihar, and Ballia and eastern Ghazipur in Uttar Pradesh (UP).
- ▶ Northern Bhojpuri: spoken in Deoria, Gorakhpur, and Basti in Uttar Pradesh, and some parts of Champaran in Bihar.
- ▶ Western Bhojpuri: spoken in the following areas of UP: Azamgarh, Ghazipur, Mirzapur and Varanasi.
- ▶ Nagpuria: spoken in the south of the river Son, in the Palamu and Ranchi districts in Bihar.

# *LT Resources for Bhojpuri*

- Discusses the creation of language technological (LT) resources for Bhojpuri language such as monolingual, parallel (English-Bhojpuri), and annotated corpus etc.
- Methodology of creating LT resources for less-resourced languages.
- The initial resources created for the present study were a monolingual Bhojpuri corpus and a parallel English-Bhojpuri corpus.
- Both of these corpora were annotated with POS tag level.

- 
- For the parallel English-Bhojpuri corpus, the source language English was also annotated at the paninian dependency level.
  - Finally, it provides statistics of LT resources created and highlights issues and challenges for developing resources for less-resourced languages like Bhojpuri.
- 

# Details of Monolingual Bhojpuri Corpus

| Corpus source                                   | Corpus source information | Sentences | Words     | Characters |
|---|---------------------------|-----------|-----------|------------|
| Books   | bhojpuri nibandh          | 60,000    | 10,38,202 | 50,78,916  |
|   | tin nAtak                 |           |           |            |
|   | jial sikhiM               |           |           |            |
|   | rAvan UvAch               |           |           |            |
|   | bhojpuri vyakarana        |           |           |            |
| Magazines                                       | pAti                      | 40,029    | 5,77,878  | 28,06,191  |
|   | Parikshan                 |           |           |            |
|   | Aakhar                    |           |           |            |
|   | samkAlin bhojpuri sAhitya |           |           |            |
| Web-sources                                     | Anjoria                   | 40,029    | 5,77,878  | 28,06,191  |
|   | tatkaa Khabar             |           |           |            |
|   | bhojpuria BlogSpot        |           |           |            |
|   | Dailyhunt                 |           |           |            |
|   | Jogira                    |           |           |            |
|   | pandjiblogspot            |           |           |            |
|   | manojbhawuk.com           |           |           |            |
| Total number of sentences, words and characters |                           | 1,00,029  | 16,16,080 | 78,85,107  |



# *Statistics of English-Bhojpuri Parallel Corpus*

| Types of the Corpus     | Sentences | Words           | Characters       |
|-------------------------|-----------|-----------------|------------------|
| <b>English-Bhojpuri</b> | 65,000    | <b>4,40,609</b> | <b>23,29,093</b> |
|                         |           | 4,58,484        | 21,17,577        |







*Thank you*

*&*

*Open for*  *or ?*

