Core requirements for Language Research

Introduction

Bhojpuri is one of the Indo-Aryan languages spoken primarily in states of Uttar Pradesh and Bihar along with west-central part of Nepal. Devanagari is the primary script associated with Bhojpuri

CR-1: Description of Bhopuri

The word order is some what flexible in Bhojpuri and it is very similar to that in Hindi. The general word order in a Bhojpuri sentence is:
<Subject> <Object> <Verb>
As can be noticed it is different from the typical word order in an English sentence **(<Subject> <Verb> <Object>).**

For example, consider the following English sentence:
“ Please close the door.“

The Bhojpuri equivalent of this sentence would be:
“तनी दरवाज.ा ब॑द कर दा|“ {Tani darwaza band kar da |}
{Note: End of sentence is marked by a special character called “Poornaviram” written as “|“ instead of a full-stop(.)} Here the object **दरवाज.ा{darwaza}** is followed by the verb **ब॑द कर दा{band kar da}**.

As mentioned before, the canonical order of words in Bhojpuri is not necessarily fixed. The same sentence can be written and spoken in several ways. More generally, it is common in colloquial Bhojpuri to place a word or phrase that qualifies the preceding words at the end of the sentence.

Consider the following English sentence:
“I feel like going for shopping.“

The typical SOV Bhojpuri equivalent of this sentence would be :
“हमार बाज.ार जाव े क े मन करत बा|“ Humaar bazaar jaave ke man karat ba.}

The colloquial form of the same sentence can be:
“हमार मन करत बा बाज.ार जाव े क े|“ {Humaar man karat ba bazaar jaave ke.}
In accusative languages like Bhojpuri, the SUBJ(I) and SUBJ(T) have the same case-marking and induce agreement as shown below -

A. Intransitive
“I will sleep.“ translates to **“हम सोत बा ।“ {Hum sot ba}**

B. Transitive
“I will see a man.“ translates to **“हम मानक े द ेखी ।“{Hum manke dekhi}**

The accusativity in morphology can be explained by the structure based theory of case and agreement in the PP approach. Occupying the spec of a tensed Infl, SUBJ(I) and SUBJ(T) can have their nominative case properly licensed by Infl and induce morphological agreement by the mediation of Infl. The accusative case of OBJ is licensed by a transitive verb.

Noun - Adjective Relationship
The adjective generally follows the noun in the sentence unlike english where we have flexibility with respect to that.

“This flower is beautiful” AND **“It’s a beautiful flower”** both translate to **“अई फूल बड .ा सुनद्र र हौ ।“ {Ae phool bada sundar hau}**

Inflectional Morphology

Words in Bhojpuri take different forms in order to be part of a larger sentence. These forms lead to a number of inflectional categories. Some of the most important ones are : number, gender and person for nouns, pronouns, and mood, aspect,tense, and agreement for verbs.Some adjectives are also inflected for agreement. Suffixes are used for inflection.

1. Noun Morphology: Animate nouns (and some pronouns) are gender distinguished. Nouns referring to females are feminine, all others are maculine. Eg. babua-babuni (boy-girl). Nouns and pronouns also inflect for number. “-an” is the most common plural suffix. Eg. sonar-sonaran (goldsmith-goldsmiths).
2. Verb Morphology: Bhojpuri has an elaborate inflectional system to indicate tense, aspect and mode on the finite verb forms. Further agreement is required in terms of person and honorificity, gender and number. Eg. u- baith-al (he sat) , u-baith-al ba (he is seated).

Genders

Nouns in Bhojpuri fall under two genders: Masculine and Feminine. It does not have a neutral gender. The gender of inanimate objects in Bhohpuri is subjected to sound that is produced while pronouncing the word. A masculine tone indicates a male and feminine tone indicates a female.

Prepositions or Postposition

Bhojpuri has:
Possessive Pronouns:

1. 1st person singular - ham
2. 1st person plural - hamani kā
3. 2nd person singular - tu, tẽ, rauā
4. 2nd person plural - tohani kā
5. 3rd person singular - u, i
6. 3rd person plural - okani kā

Indefinite Pronouns:

1. Inanimate - Kuch
2. Non-honorific animate - Kauno
3. Honorific animate - Kehu

Demonstrative Pronouns:

1. This - Hei
2. That/Remote - Hau

Interrogative Pronouns:

1. Human (sing) - ke/kaun
2. Human (pl) - kinh ka/kaun
3. Non human - ka

Politeness

Bhojpuri syntax and vocabulary reflects a three-tier system of politeness. Any verb can be conjugated as per these tiers. For example, the verb “to come” in Bhojpuri is “aana” and the verb “to speak” is “bolna”. The imperatives “come!“ and “speak!“ can thus be conjugated five ways, each marking subtle variation in politeness and propriety. These permutations exclude a host of auxiliary verbs and expressions which can be added to these verbs to add even greater degree of subtle variation. For extremely polite or formal situations, the pronoun is generally ignored.
Similarly, adjectives are marked for politeness and formality. For example, “your” has several words (or synonym) but with a different tone of politeness: “tōr” (casual and intimate), “tōhār” (polite and intimate), “t’hār” (formal yet intimate), “rā‘ur” (polite and formal) and “āp ke” (extremely formal).

Bhojpuri honorifics are the most striking feature of the language. The honorific is dependent on person and proximity. Let us look at the occurrence of the honorifics.

Second person:

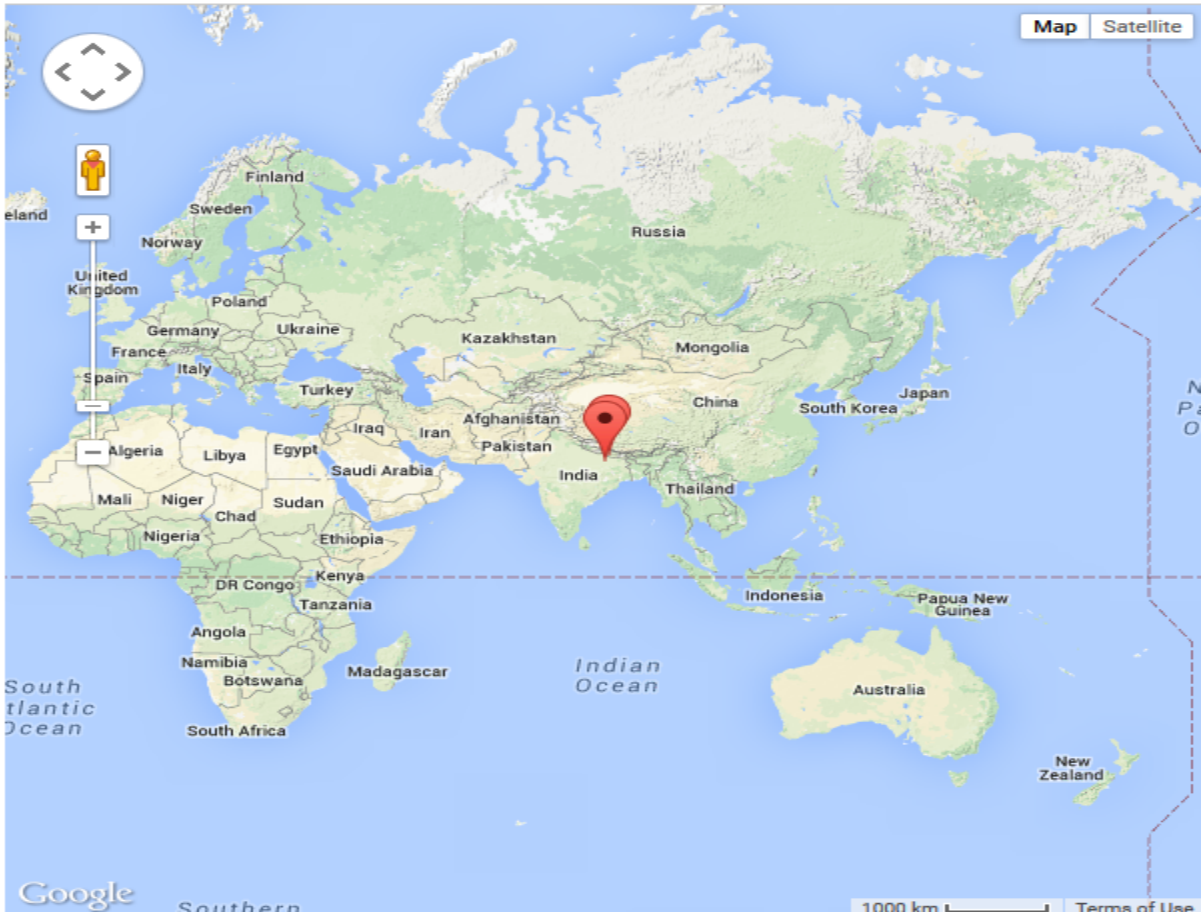
	Neutral	**-hon**	**hon**
Direct:	tu	te	reua
Oblique:	tohar	tor	raur

Third Person:

	Neutral	-hon,-prox	+hon,+prox	-hon,+prox
Direct:	u	u	i	i
Oblique:	unkara	okara	inkara	ekara

CR-2: Bhojpuri speaking regions and population

Bhojpuri is an Indo-Aryan language. It is closely related to Maithli and Magadhi, collectively known as Bihari languages.
It is spoken mostly in **North India** and **Nepal**. It is a common language in the following Indian states: Uttar Pradesh, Purvanchal, Bihar, and Jharkhand. It is an official language of Nepal, where almost 8% of the population speaks in Bhojpuri. A vernacular language of **Mauritius**, Bhojpuri is mostly used as a household language. Bhojpuri is also spoken is parts of Pakistan. In Suriname, Trinidad & Tobago, Guyana and Jamaica, a dialect of Bhojpuri, known as **Caribbean**



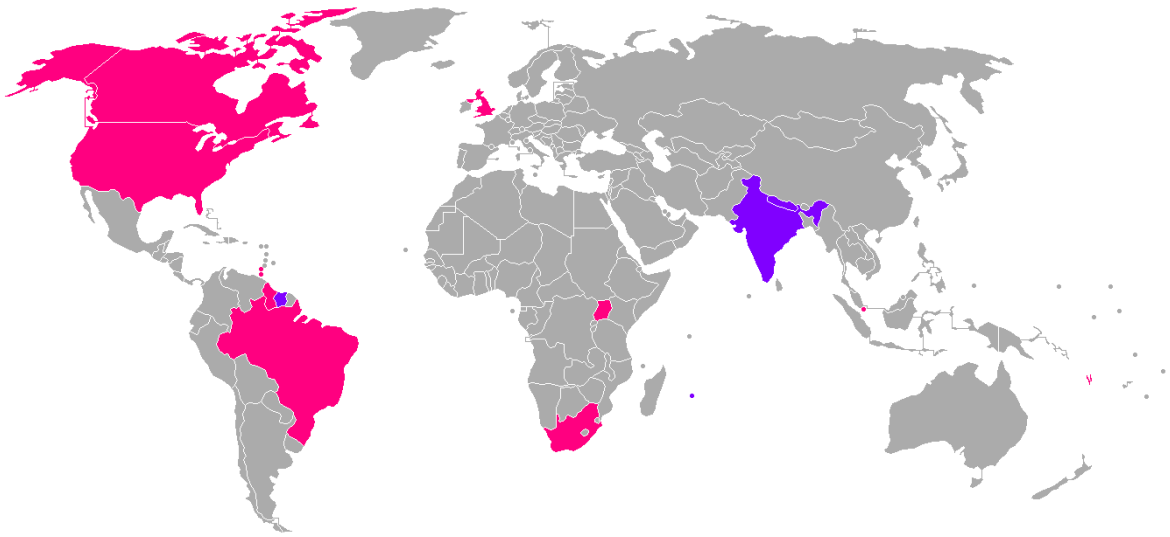
Hindustani is a widely spoken language.

Number of Bhojpuri speakers

According to the 2001 census, the number of Bhojpuri speakers in India is 33 million, though some people estimate the number to be 150 million in India, and a further 6 million elsewhere. These 6 million Bhojpuri speaking people lie outside the borders of India. Some of these countries are: Nepal, Mauritius, Fiji, Suriname, Guyana, Uganda, Singapore, Trinidad & Tobago, Saint Vincent and the Grenadines, Great Britain, and the United States.

Exposure to other languages

Being mostly a vernacular language, most of the Bhojpuri speakers across the world use it only for household purpose. Therefore,most of the Bhojpuri speakers are exposed to other native or non-native languages as well.
In North India, most of the Bhojpuri speaking population is well versed in Hindi, the official language of North India. The level of bi-lingualism and multi-lingualism is relatively high in Jharkhand, Bihar and Uttarakhand. Languages like Oriya, Bengali, Hindi, Santhli and Sadri are quite often used in these regions.
Bhojpuri is the third most popular language spoken in Nepal, with Nepali and Mathili being the top two languages. Bhojpuri is most popular in a region called Birgunj and most of the population here also speak in Nepali.
Similarly, in Mauritius French, English and Mauritian Creol is mostly widely used with Bhojpuri being only a household language.
After Dutch and Sranan Tongo, Caribbean Hindustani, a form of Bhojpuri, is the most widely spoken language of Suriname.
Bhojpuri has become an almost forgotten language in regions like Trinidad & Tobago and Guyana.



CR-3: Scripts associated with Bhojpuri

Bhojpuri was historically written in Kaithi scripts. But after 1894, Devanagari has served as the primary script. It is written left to right

Bhojpuri does not have many indigenous literatures. A few books have been printed in Bhojpuri, including , Lorikayan, or the story of Veer Lorik, a famous Bhojpuri folklore of Eastern Uttar Pradesh. No portion of the Bible has been translated into any of the dialects of Bhojpuri.

Bhojpuri uses the same script as the official language of India, Hindi, and so can reuse the infrastructure present for typing in Hindi. The main impediment for writing in Bhojpuri is the low literacy rate in the regions where it is spoken owing to economic backwardness. Writing in the Devanagari script is now possible in many online editors like (Quillpad)[www.quillpad.in], and offline software.

CR-4: Presence in Machine Translation

Currently no online machine translation systems exist for translating Bhojpuri to any other language, and human translators are employed for this task.
There aren't any research papers on the language in the Machine Translation archive either but this is a good comparison of some Indian languages and their structure Pronominals:A Comparative Study of the Languages of Bihar and West Bengal. Due to the low amount of research done on translation from Bhojpuri to other languages, no NLP tools exist either.