Hindi Dialects Phonological Transfer Rules for Verb Root Cǝlǝ

Diwakar Mishra  
Special Centre for Sanskrit Studies  
Jawaharlal Nehru University, New Delhi  
diwakarmishra@gmail.com

Kalika Bali  
Microsoft Research Lab India, Bangalore  
kalikab@microsoft.com

Abstract

Most Natural Language Processing (NLP) applications need to account for synchronic variations in a language as represented by its major dialects. However, most corpora available for the training and development of such systems tend to be dialect neutral. A framework that models synchronic variation can make NLP and Speech technology systems more robust to dialect variations. In this paper we present basic phonological transfer rules from standard Hindi to a number of its prominent dialects. We believe that this can be the first step towards a more general model for dialect variation in Hindi. The rules here describe morphophonemic change in simple verb forms between dialects taking the example of verb root cǝlǝ.

1. Introduction

Language change both synchronic and diachronic may be viewed as a complex dynamic system. Many believe that synchronic sound variation can provide us with interesting insights into diachronic variation and language evolution [1], [2], [3], [13]. While most NLP applications and especially, speech technology applications (recognition and generation both) need to be robust to dialect variation, the actual solution to this problem is largely ignored or worked at a superficial level. The main reason for this remains the dependency on large corpus of data for statistical NLP which views language as represented by the corpus. As corpus is more easily procured or collected for the dominant dialect/s, many dialect variations are not dealt with in such systems. This becomes more of a challenge for languages like Hindi and other Indian languages, where there is a scarcity of adequate databases even in the dominant or standard dialects. A holistic framework that would redress the issues posed by dialect variation therefore, continues to elude us.

Hindi is the primary official language of India and official language of ten of her states. According to the 2001 Census [4], 41% of the Indian population speaks Hindi as its first language, and more than 70% Indians can understand and speak Hindi to a certain level. Hindi is the lingua franca in many non-Hindi speaker states, such as the north eastern Indian state of Arunachal Pradesh, and is second most spoken language after Bengali in Andaman Islands and north eastern states [5].

The actual number of the dialects of Hindi is not finitely counted, but there are given individual data of 49 dialects of Hindi in census of India 2001 [6]. The most spoken of them are Bhojpuri, Rajasthani, Chhattisgarhi, Magahi, Pahari, Bundeli, Bagheli, Awadhi, Marwari, Mewari etc. [6]. Area-wise distribution of the major dialects is as follows: in Haryana, Haryanvi; in Uttarakhand, Garhwali and Kumauni; in Rajasthan, Rajasthani, Mewari, Harauti, Mewati, Marwari and Dhundhari; in Uttar Pradesh, Awadhi and Bhojpuri; in Madhya Pradesh, Bagheli, Nimadi, Malvi and Bundeli; in Bihar, Maithili and Magahi; in Jharkhand, Sadani; and in Chhattisgarh, Chhattisgarhi and Surgujiiya [7]. Via 92nd constitutional amendment 2003, Maithili is added to the 8th schedule of Indian constitution as a separate language.

In this paper, we present the phonological transfer rules from the Standard Dialect of Hindi, Khari Boli, to four of the major dialects of Hindi, viz., Bundeli, Bagheli, Kanaui and Awadhi. In the next section we define the scope of the work and describe the speech corpus on which the work is based. Transfer rules are presented in Section 3, and we conclude with discussion and future direction in Section 4.

In its spoken forms, Hindi encompasses a wide range of dialects. Roughly speaking, these varieties can be divided into western and eastern groups [12]. The standardized form of Hindi, commonly referred to as khaRI boll (literally 'standing language'), has a somewhat complex history. The modern standard language (as opposed to regional vernacular or literary dialects) arose through the infusion of considerable external (i.e., non-Hindi) vocabulary into a
grammatical skeleton based on vernacular dialect spoken in the Delhi area [12].

Figure 1: Map of Hindi speaking region indicating the regions of the selected dialects

2. Defining the Scope of the Work

The phonological change between two languages (or dialects), is very complex consisting of many simple or linear rules. There exist several categories in which the rules are applicable and the change patterns across the categories may vary a lot. Even within a category the change patterns may vary based on the division of the bases of that category or suffixes into different classes. However, these changes are not random but are regular phonological processes that can be captured systematically across word categories. Hence, at the initial stage of our work, we may begin with a single grammatical category (in this case, a regular verb) or one class of that category to map the forms across dialects.

Choudhary et. al. [8] in their attempt to derive synchronically different forms or dialects of Bangla from a single diachronic “parent” form show that the present day verbal inflections across two dialects of Bangla can be derived through phonological transfer rules (there named ‘rewrite-rules’) from Classical Bangla. They take 18 of 19 classes of verb roots with one root as representative, and 33 of 52 forms, which are relatively more regular, of one dialect and derive the forms in the other (synchronous) dialect.

In this paper, we have selected one verb root cola as a representative of verb roots having most regularity and least variation in their forms. A different set of transfer rules is applied to standard Hindi for each of the four dialects (Bundeli, Bagheli, Kannauji, and Awadhi). The standard dialect here is the form of Khari Boli accepted as the standard as described in [12].

Appropriate corpus is a must for any systematic study and development of language technology. For Hindi the state of annotated corpora availability is not encouraging; however there is tremendous text available in the form of news papers, magazines and literature. There are also corpora development efforts going on for Indian languages, like the Indian Languages Corpora Initiative (ILCI) project [9] funded by Department of Information Technology (DIT), Government of India, and running under the leadership of Jawaharlal Nehru University, that aims at collecting corpora in 12 languages including English. Written text for Hindi remains largely in the standard form with dialects mostly limited to literary use. The situation is even more challenging for speech corpora as very little is available for Indian languages, and that which includes dialects is negligible. For our research we had access to the speech corpus compiled and designed by Appen Technologies in 2006. The corpus was collected from seven different locations of Hindi speaking area each representative to one dialect. The map in Figure 1 below shows the areas of the dialects collected in the speech database. The database is collection of bi-way telephone conversations. The target was to collect the sample of speech of 1000 speakers, out of which 700 are males, and 300 are females, again out of which 700 are mobile phone conversations and 300 are landline conversations. The per-dialect target was 143 speakers. The actual collection (996 speakers) is almost near the target and constitutes over 60 hours of speech sample. The speech data is transcribed both in Devanagari and roman scripts. Unfortunately, this database is not available in the public domain.

There are certain issues however, with the speech database used. Though it claims to represent seven different dialects, Hindustani, Haryanvi, Bundeli, Bagheli, Kanaui, Awadhi and Chhattisgarhi, it was observed that a number of speakers actually used the standard form across several dialects. Some Awadhi marked speakers spoke Bhojpuri dialect. Pure Chhattisgarhi is almost absent and there is very little data for pure Bhojpuri and Haryanvi as well. Thus, we left out these dialects for analysis and formulating transfer rules.

The morphological and syntactic structure shows enough symmetry in these dialects, and it would seem reasonable to assume that mapping of these structures can be handled with phonological and morphological transfer rules.
3. The Transfer Rules between Dialects

Grammatical gender is marked in most forms of the verb in standard Hindi; other dialects mark gender in lesser forms. However, this varies from dialect to dialect. Hence, the source dialect – standard Hindi – requires gender and number information for almost all the verb forms while other dialects can generate many of the forms only with number. Most commonly such forms are first person forms, and all future tense forms in the non-standard varieties of Hindi do not mark gender. Due to this difference in morphology, standard to dialect transfer requires more rules as compared to dialect to dialect transfer. Similar observation is also made by Choudhury et. al. [8] for Bangla that the transfer between two dialects requires a very less number of rules as compared to transfer from Standard Colloquial Bangla to another dialect of Bangla.

The standard dialect has more morphological information than other dialect, so transfer from standard to other dialect is lossy. Such transfer leads to information loss and does not require any other extra information; while in reverse, i.e., transfer from other dialect to standard, might require additional rules for determining gender and number and inserting such information appropriately. What is more purposeful from the view of the need of Hindi language technology research is standard to dialect transfer rather than dialect to standard, because the need is to generate a dialect forms in the absence of dialect specific corpora. For the sake of convenience, the rule set is divided into two – one, for the past and present base and expectation mood (colata) “he would have walked”, and second, for the rest of verb morphology – other tenses and auxiliary verbs.

Key signs for the rules are given below:

In the rules, # = word boundary, ## = phrase boundary, R = root or part of root, (AUX) = of auxiliary verb, V(G,N) = vowel ending according to gender and number, V(ms.sg) = a, V(fm.sg) = i, V(ms.pl) = e, V(fm.pl) = i±nasal

3.1 Standard Hindi verb forms of colo

The root colo “to walk” is selected it assuming the representative of most regular and simplest verb forms. The forms here are before schwa deletion which is a common phenomenon in Hindi [14]. The forms included in the study, on which these transfer rules are applicable, are – present tense, present continuous tense, past tense, past continuous tense, past perfect tense, future tense, imperative and expectation mood.

The following tables present these forms of the root colo in the source dialect – the standard Hindi. The second person pronouns tu and ap behave like third person, so in second person forms listed here are only those having agreement with singular pronoun tUm.

### Table 1: Present tense and present continuous forms of colo in standard Hindi

<table>
<thead>
<tr>
<th>PNG</th>
<th>Present tense</th>
<th>Present continuous</th>
</tr>
</thead>
<tbody>
<tr>
<td>3p ms.sg</td>
<td>calata hæ</td>
<td>colo raha hæ</td>
</tr>
<tr>
<td>3p fm.sg</td>
<td>calati hæ</td>
<td>colo rahi hæ</td>
</tr>
<tr>
<td>3p ms.pl</td>
<td>calate hæ~</td>
<td>colo rahi hæ~</td>
</tr>
<tr>
<td>3p fm.pl</td>
<td>calati hæ~</td>
<td>colo rahi hæ~</td>
</tr>
<tr>
<td>2p ms</td>
<td>calate ho</td>
<td>colo rahi ho</td>
</tr>
<tr>
<td>2p fm</td>
<td>calati ho</td>
<td>colo rahi ho</td>
</tr>
<tr>
<td>1p ms.sg</td>
<td>calata hu~</td>
<td>colo raha hu~</td>
</tr>
<tr>
<td>1p fm.sg</td>
<td>calati hu~</td>
<td>colo rahi hu~</td>
</tr>
<tr>
<td>1p ms.pl</td>
<td>calate hæ~</td>
<td>colo rahi hæ~</td>
</tr>
<tr>
<td>1p fm.pl</td>
<td>calati hæ~</td>
<td>colo rahi hæ~</td>
</tr>
</tbody>
</table>

### Table 2: Past tense, past continuous and past perfect forms of colo in standard Hindi

<table>
<thead>
<tr>
<th></th>
<th>Past tense</th>
<th>Past continuous</th>
<th>Past perfect</th>
</tr>
</thead>
<tbody>
<tr>
<td>3p ms.sg</td>
<td>calata tha</td>
<td>colo raha tha</td>
<td>cola</td>
</tr>
<tr>
<td>3p fm.sg</td>
<td>calati thi</td>
<td>colo rahi thi</td>
<td>cali</td>
</tr>
<tr>
<td>3p ms.pl</td>
<td>calate the</td>
<td>colo rahi the</td>
<td>care</td>
</tr>
<tr>
<td>3p fm.pl</td>
<td>calati thi~</td>
<td>colo rahi thi~</td>
<td>cali~</td>
</tr>
<tr>
<td>2p ms</td>
<td>calate the</td>
<td>colo rahi the</td>
<td>care</td>
</tr>
<tr>
<td>2p fm</td>
<td>calati thi~</td>
<td>colo rahi thi~</td>
<td>cali~</td>
</tr>
<tr>
<td>1p ms.sg</td>
<td>calata tha</td>
<td>colo raha tha</td>
<td>cola</td>
</tr>
<tr>
<td>1p fm.sg</td>
<td>calati thi</td>
<td>colo rahi thi</td>
<td>cali</td>
</tr>
<tr>
<td>1p ms.pl</td>
<td>calate the</td>
<td>colo rahi the</td>
<td>care</td>
</tr>
<tr>
<td>1p fm.pl</td>
<td>calati thi~</td>
<td>colo rahi thi~</td>
<td>cali~</td>
</tr>
</tbody>
</table>

### Table 3: Future tense, imperative and expectation mood forms of colo in standard Hindi

<table>
<thead>
<tr>
<th></th>
<th>Future tense</th>
<th>Imperative</th>
<th>Expectation mood</th>
</tr>
</thead>
<tbody>
<tr>
<td>3p ms.sg</td>
<td>calega</td>
<td>cale</td>
<td>calata</td>
</tr>
<tr>
<td>3p fm.sg</td>
<td>calegi</td>
<td>cale</td>
<td>calati</td>
</tr>
<tr>
<td>3p ms.pl</td>
<td>cale~ge</td>
<td>cale~</td>
<td>calate</td>
</tr>
<tr>
<td>3p fm.pl</td>
<td>cale~gi</td>
<td>cale~</td>
<td>calati~</td>
</tr>
<tr>
<td>2p ms</td>
<td>caloge</td>
<td>cale</td>
<td>calate</td>
</tr>
<tr>
<td>2p fm</td>
<td>calogi</td>
<td>cale</td>
<td>calati~</td>
</tr>
<tr>
<td>1p ms.sg</td>
<td>calu~ga</td>
<td>calu~</td>
<td>calata</td>
</tr>
<tr>
<td>1p fm.sg</td>
<td>calu~gi</td>
<td>calu~</td>
<td>calati</td>
</tr>
<tr>
<td>1p ms.pl</td>
<td>cale~ge</td>
<td>cale~</td>
<td>calate</td>
</tr>
<tr>
<td>1p fm.pl</td>
<td>cale~gi</td>
<td>cale~</td>
<td>calati~</td>
</tr>
</tbody>
</table>
3.2 Standard Hindi to Bundeli

The schwa deletion rule is common in Hindi and its dialects with a little variation [10]. The rules mentioned here apply on the form before schwa deletion, and then general schwa deletion rules apply after application of transfer rule. A discussion on the schwa deletion rules is beyond the scope of this paper.

**First set transfer rules** – present/past base (or present/past form without auxiliary verb) and expectation mood form

**Exception**: These rules do not apply on feminine plural form in present tense and expectation mood.

**General**

\[ a \rightarrow o / C \_ ## \]

\[ V+length \rightarrow o / t \_ # \]

**Second set (general) transfer rules** – other forms including auxiliary verbs

**Exceptional rules**

\[ \theta \rightarrow g(V,G,N) / \ae(aux)\pmnasal \_ # \]

**General**

\[ o \rightarrow a_{o} / C \_ # \]

\[ a \rightarrow o / C \_ # \]

\[ e \pmnasal \rightarrow a_{e}\pmnasal / C \_ # \]

\[ u \rightarrow o_{-} / -- \_ # \]

\[ V+length \rightarrow V-length / R \_ V-length \]

\[ V\pmnasal+tense \rightarrow \alpha h \_ V\pmnasal-tense / R \_ g \]

\[ V (ex. calhaoge, calhaoge) \]

\[ \theta h(AUX) \rightarrow h a t(AUX) \_ / _ _ V \]

**Optional rule:**

\[ gV \rightarrow o / hV \_ # (ex. calhae, calhae~) \]

\[ ah \rightarrow o / R \_ \_ V \_ # (ex. calae~) \]

3.3 Standard Hindi to Bagheli

These rules also apply on the word before the application of schwa deletion rules.

**First set transfer rules** – present/past base (or present/past form without auxiliary verb) and expectation mood form

**Exception**: These rules do not apply on feminine plural form in present tense and expectation mood.

**General**

\[ a \rightarrow o / C \_ # \]

\[ V+length \rightarrow o / t \_ # \]

**Second set (general) transfer rules** – other forms including auxiliary verbs

**Exceptional rules**

\[ \theta \rightarrow g(V,G,N) / \ae(aux)\pmnasal \_ # \]

**General**

\[ o \rightarrow a_{o} / C \_ # \]

\[ a \rightarrow o / C \_ # \]

\[ e \pmnasal \rightarrow a_{e}\pmnasal / C \_ # \]

\[ u \rightarrow o_{-} / -- \_ # \]

\[ V+length \rightarrow V-length / R \_ V-length \]

\[ V\pmnasal+tense \rightarrow \alpha h \_ V\pmnasal-tense / R \_ g \]

\[ V (ex. calhaoge, calhaoge) \]

\[ \theta h(AUX) \rightarrow h a t(AUX) \_ / _ _ V \]

**Optional rule:**

\[ gV \rightarrow o / hV \_ # (ex. calhae, calhae~) \]

\[ ah \rightarrow o / R \_ \_ V \_ # (ex. calae~) \]

3.4 Standard Hindi to Kanauji

In the previous dialects the forms which are schwa ending, their parallel in Kanauji and Awadhi are I-ending. But this is also many times muted in pronunciation in the similar manner as schwa. Even then, the difference of o and I can be observed in the speech.

**First set transfer rules** – present/past base (or present/past form without auxiliary verb) and expectation mood form

**Exception**: These rules do not apply on feminine plural form in present tense and expectation mood.

**General**

\[ o \rightarrow a_{o} / C \_ # \]

\[ a \rightarrow o / C \_ # \]

\[ e \pmnasal \rightarrow a_{e}\pmnasal / C \_ # \]

\[ u \rightarrow o_{-} / -- \_ # \]

\[ V+length \rightarrow V-length / R \_ V-length \]

\[ V\pmnasal+tense \rightarrow \alpha h \_ V\pmnasal-tense / R \_ g \]

\[ V (ex. calhaoge, calhaoge) \]

\[ \theta h(AUX) \rightarrow h a t(AUX) \_ / _ _ V \]

**Optional rule:**

\[ gV \rightarrow o / hV \_ # (ex. calhae, calhae~) \]

\[ ah \rightarrow o / R \_ \_ V \_ # (ex. calae~) \]

3.5 Standard Hindi to Awadhi

Like other dialects, these rules apply on standard form before schwa deletion rules, and like Kanauji, many cases where vowel is changed to schwa in other dialects, here changes into I.
First set transfer rules – present/past base (or present/past form without auxiliary verb) and expectation mood form.

Exception: These rules do not apply on feminine plural form.
V+length → /i / t _ #

Second set (general) transfer rules

Exceptions:

If 2p
V+length → V-length / C _ #
Ø → u → V _ #

If 1p
i~ → en / -- # (ex. fm.pl pst.prf)
V+nasal → i / R _ # (ex. cali - imp)
V-nasal → en / R _ # (ex. calen - pst.prf)
V g V → Iba / R _ # (ex. callba)

General
ø → U / C _ # (ex. calU -2p.imp)
e+nasal → æ+nasal / C _ #
V+length → V-length / R V-length
V+nasal + C+stop + velar V+tense → i C-stop + velar+fricative æ+nasal / R _ # (ex. callhæ~)
Lh æ+nasal → i / R _ # (ex. cali)
th(AUX) → rah(AUX) / -- _ V
V(AUX) → æ / h _ #
if pl
ø → ∅ / a _ --

Optional rule
V → ∅ / Ib _ #

3.6 Comparative forms of standard Hindi and other dialects

The following are given some of the forms of standard Hindi with their respective other dialect forms. In the selected dialects other than standard, first person singular is used as plural, so their forms are not given. Here only present tense, past perfect, future tense and imperative forms are given for instance.

Table 4: Comparative verb forms of standard Hindi and other dialects

<table>
<thead>
<tr>
<th>Standard Hindi</th>
<th>Bundeli</th>
<th>Bagheli</th>
<th>Kamaaji</th>
<th>Awadhi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present tense</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>calata hæ~</td>
<td>calata hæ~</td>
<td>calata hæ~</td>
<td>calati hal~</td>
<td>calati hal~</td>
</tr>
<tr>
<td>calata hæ~</td>
<td>calata hæ~</td>
<td>calata hæ~</td>
<td>calati hal~</td>
<td>calati hal~</td>
</tr>
<tr>
<td>calata hæ~</td>
<td>calata hæ~</td>
<td>calata hæ~</td>
<td>calati hal~</td>
<td>calati hal~</td>
</tr>
<tr>
<td>calata hæ~</td>
<td>calata hæ~</td>
<td>calata hæ~</td>
<td>calati hal~</td>
<td>calati hal~</td>
</tr>
<tr>
<td>calata hæ~</td>
<td>calata hæ~</td>
<td>calata hæ~</td>
<td>calati hal~</td>
<td>calati hal~</td>
</tr>
<tr>
<td>calata hæ~</td>
<td>calata hæ~</td>
<td>calata hæ~</td>
<td>calati hal~</td>
<td>calati hal~</td>
</tr>
<tr>
<td>calata hæ~</td>
<td>calata hæ~</td>
<td>calata hæ~</td>
<td>calati hal~</td>
<td>calati hal~</td>
</tr>
<tr>
<td>calata hæ~</td>
<td>calata hæ~</td>
<td>calata hæ~</td>
<td>calati hal~</td>
<td>calati hal~</td>
</tr>
<tr>
<td>calata hæ~</td>
<td>calata hæ~</td>
<td>calata hæ~</td>
<td>calati hal~</td>
<td>calati hal~</td>
</tr>
</tbody>
</table>

4. Conclusion

In this paper, the authors have presented phonological transfer rules for the verb forms of root
The root `cǝlǝ` is selected as the representative of most regular verb roots which have least modification in base. This research is first of its type for Hindi and its dialects, so the scope has been narrowed to most popular verb forms, and that of most regular type of verb roots. The forms the rules are applicable on are – present tense, present continuous, past tense, past continuous, past perfect, future tense, imperative and expectation mood – for all persons, numbers and genders.

We believe that the transfer rules are going to be very useful in speech and language technology systems for making them more robust to dialect changes, and addressing the issues related to sparsity of dialect specific data. The importance of dialect sensitive NLP applications cannot be overly emphasized in a county like India, where they have an important role in the expansion of the reach of e-governance.

In future, these transfer rules need to be extended to all forms of all grammatical categories of Hindi to develop dialect-transducers for Hindi. Such Dialect transducers in a Hindi NLG system have the potential to automatically generate forms in several synchronic dialects expanding the scope of such a system. Phonological transfer rules will alone not be sufficient to develop a complete dialect transducer, but syntactic and semantic study will also play an important role in it. We hope that this first step in a relatively under explored area of dialects-transfer rules and dialect sensitive speech technology will prove challenging and interesting for future research.

References


