Hindi Dialects Phonological Transfer Rules for Verb Root *Cələ*

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 cələ.

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 Uttaranchal, 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 Surgujiya [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, Kanauji 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 bolI (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].

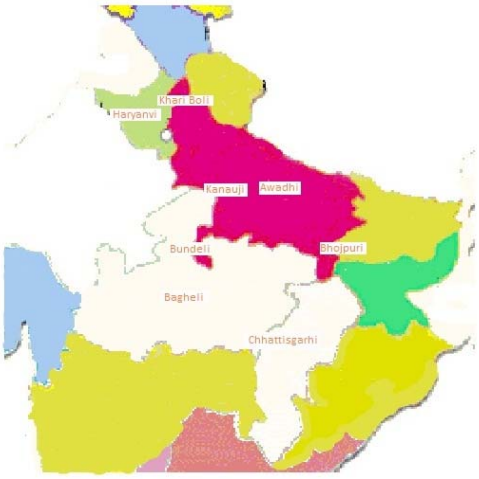

Figure 1: Map of Hindi speaking region indicating the regions of the selected dialects

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 take18 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 (synchronic) dialect.

In this paper, we have selected one verb root *cələ* 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, Kanauji, 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 (*cələta*) "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 *cələ*

The root *cələ* "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 *cələ* 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 *cələ* in standard Hindi

PNG	Present tense	Present continuous
3p ms.sg	cələta hæ	cələ rəha hæ
3p fm.sg	cələti hæ	cələ rəhi hæ
3p ms.pl	cələte hæ~	cələ rəhe hæ~
3p fm.pl	cələti hæ~	cələ rəhi hæ~
2p ms	cələte ho	cələ rəhe ho
2p fm	cələti ho	cələ rəhi ho
1p ms.sg	cələta hu~	cələ rəha hu~
1p fm.sg	cələti hu~	cələ rəhi hu~
1p ms.pl	cələte hæ~	cələ rəhe hæ~
1p fm.pl	cələti hæ~	cələ rəhi hæ~

Table 2: Past tense, past continuous and past perfect forms of *cələ* in standard Hindi

	Past tense	Past continuous	Past perfect
3p ms.sg	cələta tha	cələ rəha tha	cəla
3p fm.sg	cələti thi	cələ rəhi thi	cəli
3p ms.pl	cələte the	cələ rəhe the	cəle
3p fm.pl	cələti thi~	cələ rəhi thi~	cəli~
2p ms	cələte the	cələ rəhe the	cəle
2p fm	cələti thi~	cələ rəhi thi~	cəli~
1p ms.sg	cələta tha	cələ rəha tha	cəla
1p fm.sg	cələti thi	cələ rəhi thi	cəli
1p ms.pl	cələte the	cələ rəhe the	cəle
1p fm.pl	cələti thi~	cələ rəhi thi~	cəli~

Table 3: Future tense, imperative and expectation mood forms of *cələ* in standard Hindi

	Future tense	Imperative	Expectation mood
3p ms.sg	cəlega	cəle	cələta
3p fm.sg	cəlegi	cəle	cələti
3p ms.pl	cəle~ge	cəle~	cələte
3p fm.pl	cəle~gi	cəle~	cələti~
2p ms	cəloge	cəlo	cələte
2p fm	cəlogi	cəlo	cələti~
1p ms.sg	cəlu~ga	cəlu~	cələta
1p fm.sg	cəlu~gi	cəlu~	cələti
1p ms.pl	cəle~ge	cəle~	cələte
1p fm.pl	cəle~gi	cəle~	cələti~

3.2 Standard Hindi to Bundeli

The schwa deletion rule is common in Hindi and its dialects with a little variation [10] [14]. 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.

a → o / C _ ##

V+length → ə / t _ #

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

Exceptional rules

Ø → g V(G,N) / əe(AUX)±nasal _ #

General

o → ao / C _ #

a → o / C _ #

e ±nasal → ae±nasal / C _ #

u~ → əe~ / -- _ #

V+length → V-length / R _ V-length

V±nasal+tense → əhə V±nasal-tense / R _ g V (ex. cələhəoge, cələhəegi)

th(AUX) → hat(AUX) / -- _ V

Optional rule:

gV → ø / hV _ # (ex. cələhəe, cələhəe~)

əh → ø / R _ əV # (ex. cələe~)

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.

a → o / C _ ##

V+length → ə / t _ #

Second set (general) transfer rules

Exceptional rules

Ø → g V(G,N) / əe(AUX)±nasal _ # (ex. həega, həe~ge)

General

o → u / C _ #

a → o / C _ #

e ±nasal → əe±nasal / C _ #

u~ → əe~ / -- _ #

V+length → V-length / R _ V-length

V±nasal+tense → əhə V±nasal-tense / R _ g V (ex. cələhəoge, calahəegi)

th(AUX) → hət(AUX) / -- _ V

Optional rule:

gV → ø / hV _ # (ex. cələhəe, cələhəe~)

əh → ø / R _ əV # (ex. cələe~)

This optional rule applies more frequently in this dialect.

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 ə 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.

a → o / C _ ##

V+length → I / t _ #

Second set (general) transfer rules

Exceptions:

If 1p

V → en / -- _ # (ex. fm.pl pst.prf)

V+nasal → i / R _ # (ex. cəli - imp)

V-nasal → en / R _ # (ex. cəlen - pst.prf)

General

o → əU / C _ #

a → o / C _ #

e ±nasal → əI±nasal / C _ #

u~ → i / -- _ #

a → ø / a _ --

V+length → V-length / R _ V-length #

V±nasal C+stop+velar V+tense → i C-stop+velar+fricative əI±nasal / R _ # (ex. cəlIhəI~)

V+length → V-length / R _ V-length

th(AUX) → hət(AUX) / -- _ V

Optional rule:

əI±nasal → i±nasal / Ih _ # (ex. cəlIhi~)

if 1p

hV → Ibo / R _ #

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. cəli - imp)

V-nasal → en / R _ # (ex. cəlen - pst.prf)

V g V → Iba / R _ # (ex. cəlIba)

General

ə → U / C _ # (ex. cəlU -2p.imp)

e ±nasal → əe±nasal / C _ #

V+length → V-length / R _ V-length

V±nasal C+stop+velar V+tense → i C-stop+velar+fricative əe±nasal / R _ # (ex. cəlIhəe~)

Ih əe-nasal → i / R _ ~ (ex. cəli)

th(AUX) → rah(AUX) / -- _ V

V(AUX) → əe / h _ #

if pl

V(AUX) → V+nasal / h _ #

ə → ø / a _ --

Optional rule

V → ə / 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

Standard Hindi	Bundeli	Bagheli	Kanauji	Awadhi
Present tense				
cələta hæ	cələtə hæ	cələtə hæ	cələtI həI	cələtI həI
cələti hæ	cələtə hæ	cələtə hæ	cələtI həI	cələtI həI
cələte hæ~	cələtə hæ~	cələtə hæ~	cələtI həI~	cələtI həI~
cələti hæ~	cələtə hæ~	cələtə hæ~	cələtI həI~	cələtI həI~
cələte ho	cələtə ho	cələtə ho	cələtI həU	cələte həU
cələti ho	cələti ho	cələti ho	cələti həU	cələti həU
cələta hu~				
cələti hu~				
cələte hæ~	cələtə hæ~	cələtə hæ~	cəlItI hən	cəlItI hən
cələti hæ~	cələtə hæ~	cələtə hæ~	cəlItI hən	cəlItI hən
Past perfect				
cəla	cəlo	cəlo	cəla	cəla
cəli	cəli	cəli	cəli	cəli
cəle	cəle	cəle	cəle	cəle
cəli~	cəli~	cəli~	cəli~	cəli~
cəle	cəle	cəle	cəleU	cəleU
cəli~	cəli~	cəli~	cəlIU	cəlIU
cəla				
cəli				
cəle	cəle	cəle	cələn	cələn
cəli~	cəli~	cəli~	cələn	cələn
Future tense				
cəlega	cələhæ	cələhæ	cəlIhæ	cəli
cəlegi	cələhæ	cələhæ	cəlIhæ	cəli
cəle~ge	cələhæ~	cələhæ~	cəlIhi~	cəlIhəI~
cəle~gi	cələhæ~	cələhæ~	cəlIhi~	cəlIhəI~
cəloge	cələhəO	cələhəO	cəlIhəU	cəlIhəU
cəlogi	cələhəO	cələhəO	cəlIhəU	cəlIhəU
cəlu~ga				
cəlu~gi				
cəle~ge	cələhæ~	cələhæ~	cəlIhi~	cəlIba
cəle~gi	cələhæ~	cələhæ~	cəlIhi~	cəlIba
Imperative				
cəle	cəlæ	cəlæ	cələI	cələI
cəle	cəlæ	cəlæ	cələI	cələI
cəle~	cəlæ~	cəlæ~	cələI~	cələI~
cəle~	cəlæ~	cəlæ~	cələI~	cələI~
cəlo	cələO	cəlu	cələU	cələU
cəlo	cələO	cəlu	cələU	cələU
cəlu~				
cəlu~				
cəle~	cəlæ~	cəlæ~	cəli~	cəli
cəle~	cəlæ~	cəlæ~	cəli~	cəli

4. Conclusion

In this paper, the authors have presented phonological transfer rules for the verb forms of root

cələ ("walk") from standard Hindi to four of its dialects. The root *cələ* 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

[1] J. Ohala, "Sound change is drawn from a pool of synchronic variation". In L.E. Breivik & E.H. Jahr (Eds.), *Language change: Contributions to the study of its causes*. Mouton de Gruyter, Berlin, 1989.

[2] P. Niyogi, and R. C. Berwick, "The Logical Problem of Language Change: A Case Study of European Portuguese", *Syntax: A Journal of Theoretical, Experimental, and Interdisciplinary Research*, Vol. 1, 1998.

[3] D. Lightfoot, *The development of language: Acquisition, change and evolution.* Blackwell, Oxford, 1999.

[4] Census of India website – Census 2001, Statement 5, http://censusindia.gov.in/Census_Data_2001/Census_Data_Online/Language/Statement5.htm

[5] Census of India website – Census 2001, Statement 3, http://censusindia.gov.in/Census_Data_2001/Census_Data_Online/Language/Statement3.htm

[6] Census of India website – Census 2001, Statement 1, http://censusindia.gov.in/Census_Data_2001/Census_Data_Online/Language/Statement1.htm

[7] Asher, R. E., "Language in Historical Context", in Ed. Kachru, Braj B., Kachru, Yamuna and Sridhar, S. N., *'Language in South Asia',* Cambridge University Press, Cambridge, 2008

[8] Choudhury, Monojit, Alam, Monjur, Sarkar, Sudeshna and Basu, Anupam, "A Rewrite Rule Based Model of Bangla Morpho-Phonological Change", *Proceedings of the International Conference on Computer Processing of Bangla (ICCPB)*, Dhaka, Bangladesh, 2006, pp. 64-71

[9] Jha, Girish Nath, "The TDIL Program and the Indian Languages Corpora Initiative", *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC' 10)*, ELRA, Malta, 2010

[10] Choudhury, Monojit, Basu, Anupam and Sarkar, Sudeshna, "A Diachronic Approach for Schwa Deletion in Indo-Aryan Languages" in *Proc. of the Seventh Meeting of the ACL Special Interest Group in Computational Phonology* (SIGPHON 2004). ACL, Barcelona, Spain, 2004, pp. 20-26.

[11] Kachru, Yamuna, *Hindi,* John Benjamins North America, Philadelphia, 2006.

[12] Shapiro, Michael E., *A Premier of Modern Standard Hindi,* Motilal Banarsidass, Delhi, 1989.

[13] M. Dras, D. Harrison, and B. Kapicioglu, "Emergent Behavior in Phonological Pattern Change", *Proceedings of Artificial Life VIII*, 390 -393, Sydney, Australia, 2002.

[14] Ohala, Manjari, '*Aspects of Hindi Phonology'*, Motilal Banarsidass, Delhi, 1983