

LANGUAGE IN INDIA

Strength for Today and Bright Hope for Tomorrow

Volume 11 : 10 October 2011

ISSN 1930-2940

Managing Editor: M. S. Thirumalai, Ph.D.

Editors: B. Mallikarjun, Ph.D.

Sam Mohanlal, Ph.D.

B. A. Sharada, Ph.D.

A. R. Fatihi, Ph.D.

Lakhan Gusain, Ph.D.

Jennifer Marie Bayer, Ph.D.

S. M. Ravichandran, Ph.D.

G. Baskaran, Ph.D.

L. Ramamoorthy, Ph.D.

Comparative Study of Hindi and Dogri Languages with Regard to Machine Translation

Preeti Dubey, Shashi Pathania, and Devanand

Abstract

Hindi, the Official language of India, is regarded as one of the widespread languages after English and Mandarin. Hindi belongs to the Indo-Aryan (Indic) languages, and Dogri language also belongs to the same subgroup of the Indo-European family, i.e., Indo-Aryan family of languages. The script of Hindi language is Devanagari and Dogri has its own script namely, $\text{ḍogəra əkk}^h\text{ər}$ or ḍogər ; but now the younger generation has adopted nagəri script for Dogri. As such, almost all the modern literature has been and is being written in devanagəri. For the development of Machine Translation system from Hindi to Dogri Language, there is a need to find the similarity between both the languages. It is found that both the languages are closely related to each other. This paper explains the closeness between the Hindi and Dogri Language scripts.

Keywords: Machine Translation, International Phonetic Alphabets (IPA), Orthography

Introduction

Hindi and Dogri are closely related languages and have originated from Sanskrit. Hindi is the official language of India and is spoken and used by the people all over the country, the main regions being Haryana, Uttar Pradesh, Rajasthan, Bihar and Chhattisgarh. On the other hand, Dogri is mainly spoken in the Jammu region of J&K State and adjoining areas of Punjab, Himachal Pradesh and in the borders of Sialkot & Shakar Ghar tehsils in Pakistan. Both Hindi and Dogri use the same script i.e. Devanagri but the dialect is different. In this paper, closeness between Hindi and Dogri languages is explained which includes script, consonants, vowels, conjunct consonants, numerals, punctuation, abbreviations of the two languages.

Language in India www.languageinindia.com

11 : 10 October 2011

Preeti Dubey, Shashi Pathania, and Devanand

Comparative Study of Hindi and Dogri Languages with Regard to Machine Translation

298

Introduction: Devanagari Script

Devanagari script is used to document Hindi language and in writing the Dogri language. Devanagari script has evolved over a period of more than two thousand years. Devanagari emerged around 1200AD out of Siddham script, gradually replacing the earlier, closely related Sharda script (which remained in parallel use in Kashmir). Both were descendants of the Gupta script, derived from the Brahmi script attested from the 3rd century BC; Nagari appeared in the 8th century as an eastern variant of the Gupta script, contemporary to Sharada, its western variant.

Nāgarī is in Sanskrit the feminine of *nāgara*. The feminine form is used because of its original application to qualify the feminine noun *lipi* "script".

There were several varieties in use, one of which was distinguished by affixing *deva* "divine, deity" to form a tatpurusha compound meaning the "divine urban (e) [script]". Dogri also belongs to the Indo-Aryan group of Indo-European language family and typologically it is an isolative (analytic inflectional) language. The word *ḍuggar* is taken as derivative of the words 'ḍuḡgərtə', 'ḍuḡgər', 'ḍuḡgər' etc. but in view of law of phonetic change 'ḍuḡgər' has been taken as the most appropriate and scientific source of this word. Under the influence of regressive assimilation 'ḍuḡgər' became 'ḍuggər' (name of a particular region), which is "a region of difficult traverse", "invincible", "difficult to be subdued" etc. The earlier mention of "ḍuḡgər" occurs in Chamba Copper Plates of 11th Century A.D. and it refers to the community inhabiting the region between Ravi and Chenab. The earliest reference of Dogri (*ḍuggər*) is found in Nuh-Siphir, a Masnavi written by Amir Khusaro in 1317 A.D. ("Sindhi O Lahori O Doggar"). Here Duggar refers to the language of *ḍuggər* (the country between Lahore and Kashmir). Dogri written in this script was official language of J&K State during the reign of Maharaja Ranbir Singh (1857-1885 A.D.). But now the younger generation has adopted nagari script for Dogri, as such, almost all the modern literature has been and is being written in Devanagari.

Vowels and Consonants

Devanagari has 54 alphabets in all, including the fourteen vowels and thirty three consonants; whereas Dogri is written using Devanagari script and has thirty eight segmental and five supra segmental phonemes. Segmental phonemes have been divided into two broad groups i.e. vowels and consonants. It has ten vowel phonemes and twenty eight consonant phonemes.

Vowels

Devanagari Script possesses two different forms for each of the vowels - Full form and short form.

Full Form: In Devanagari, a full form is employed for a vowel that does not immediately follow a consonant or consonant cluster, i.e. in word-initial position or when the second of a sequence of vowels.

Short form (or *matra*):

Language in India www.languageinindia.com

11 : 10 October 2011

Preeti Dubey, Shashi Pathania, and Devanand

Comparative Study of Hindi and Dogri Languages with Regard to Machine Translation

In Devanagari, short form is used when the vowel immediately follows a consonant or consonant cluster. These short forms consist of lines, hooks or combination of both above, below or to the side of the consonantal characters. These vowels are written around (that is, below, above, to the right, and to the left) the consonant signs. The following table shows the both the full form and the short form of the vowels in Devanagari. Table: 1 shows the full form and short form of vowels in Devanagari.

Devanagari vowels		
Full Form	Short Form	IPA
अ	No Sign	ə
आ	ा	a
इ	ि	I
ई	ी	i
उ	ु	U
ऊ	ू	u
ए	े	e
ऐ	ै	ε
ओ	ो	O
औ	ौ	ɔ
ऋ	ृ	ɾ

Table 1: shows short and full form of Devanagari Vowels

Other symbols in
 अं an
 vowel
 अँ am
 – nasalized vowel
 अः ah
 voiceless breath

Devanagri

anusvaara – nasalized

chandrabindu / anunāsika

visarga – which adds after vowel

- The use of Chandrabindu 'ँ' and visarga 'ः' is not prevalent in Dogri orthography.
- 'ऋ' is not used in Dogri except for transliteration of some Sanskrit words e.g. ऋषि

Consonants: The consonants in Devanagari are presented in the table below C refers to Devanagari consonant and I refers o its IPA (International Phonetic Alphabet) :

C	I	C	I	C	I	C	I	C	I
क K a	k	ख Kha	k ^h	ग Ga	g	घ Gha	g ^h	ङ ṅa	ŋ
च ca	c, tʃ	छ cha	c ^h tʃ ^h	ज ja	ʃ, dʒ	झ jha	ʃ ^h dʒ ^h	ञ ña	ɲ

ट ṭa	ṭ	ठ ṭha	ṭ ^h	ड ḍa	ḍ	ढ ḍha	ḍ ^h	ण ṇa	ṇ
त ta	t̪	थ tha	t̪ ^h	द da	d̪	ध dha	d̪ ^h	न na	n
प pa	p	फ pha	p ^h	ब ba	b	भ bha	b ^h	म ma	m
य ya	j	र ra	r	ल la	l	व va	v		
श śa	ɕ, ʃ	ष ṣa	ʃ	स sa	s	ह ha	ɦ		

Table 2: shows the Devanagari consonants with their IPA

A dot below is used to supplement the alphabet to express additional sounds i.e. these consonants are formed with a dot diacritic. The additional consonants are as below are as below:

Additional Consonants

क qa ख kha ग ga ज्ञ za इ ra ढ rha
फ fa

Peculiarities of the Dogri Language

1. Dogri has the same basic consonants as Devanagari; but घ, झ, ढ, ध, भ, i.e. gha, jha ḍha ,dha and bha exist in Dogri orthography only.
 - i) Phonetically, they change into क, च, ट, त, प if used as the initial alphabet of a word.
 - ii) If these alphabets occur in the middle of a word, they are changed to ग, ज, ड, द, ब,
 - iii) if they occur at the end of a word then they are pronounced as the consonant that occurs just before them
2. Some Words in Dogri use इ and ज्ञ as first alphabet of a word e.g. डूर, इठा, जारा, डैका; which is not the case in Hindi.
3. Chandrabindu 'ँ' and visarga 'ः' are not used in Dogri.
4. ह (h) occurs rarely in Dogri vocabulary. In the words with ह as the initial letter, it is replaced with अ, आ, ओ etc. depending on the tone of the next alphabet e.g. हिरण is written as इरन , हैल becomes ऐल and हार is आर .

5. इ and उ are also used differently in Dogri. for translation of hindi words starting with इ and उ to dogri; they are used as short form to the preceding consonant e.g. उदास becomes दुआस and उखाड़ becomes खुआड़.
6. The use of ऐ, र, ज, ब, is very frequent in Dogri
- i) अ and आ are usually used as ऐ e.g. मंत्र of Hindi is written as मैतर in Dogri
 - ii) An extra “(र) /r “ is added in certain hindi words to change them to Dogri e.g. in certain words of Hindi language is given importance in Dogri e.g. नींद of Hindi is written as नींदर in Dogri ; similarly विलाप becomes बरलाप in Dogri
 - iii) ज is more prevalent as compared to य e.g. यजमान becomes जजमान and योगी becomes जोगी when य is placed in the middle of a word, it becomes ए e.g. बनाया becomes बनाएआ but when it is used at the final position of the word, it changes to the vowel or consonant just before य e.g. समय becomes समा, शिवालय is शवाला सराय becomes सरा, सिवाय is सवा in the Dogri language.
In words like माया, काया; the use of ‘य’ is same as in Hindi.
 - iv) Similarly ‘ब’ is used in place of ‘व’ e.g. वीणा of Hindi is written as बीन in Dogri and वकील is written as बकील; when used in the middle of a word it is replaced by आ e.g. बनवाया is written as बनोआया. Also ताव becomes ता and पाव is पा are examples of व’s placement at the final position. Its usage is retained in words like देवी, सेवा etc.
7. In Dogri orthography apostrophe comma serves double purpose.
- i) To express high falling tone after short vowel and to indicate high falling tone, single apostrophe comma is very frequent in the language. कु‘न, ब’न्न etc
 - ii) To indicate syncopated forms, apostrophe is very common in Dogri too. For Example: स ‘जा’ लै (This word is a combination of (साज + ऐलै) meaning ‘in the evening’. The first apostrophe indicates high falling tone and the other shows the syncopated form.

Dead and Live Consonants

Devanagari employs a sign known in Sanskrit as the *virama* or vowel omission sign. In Hindi it is called *hal* or *halant*, and that term is used in referring to the virama or to a consonant with its vowel suppressed by the virama. The virama sign (◌̣) nominally serves to cancel the inherent vowel of the consonant to which it is applied. When a consonant has lost its inherent vowel by the application of virama, it is known as a *dead consonant*; in contrast, a *live consonant* is one that retains its inherent

vowel or is written with an explicit dependent vowel sign. क्, ख् are examples of Dead consonants and क, ख are live consonants.

Consonant Conjuncts

Two or more letter forms are called Consonant Conjuncts. Devanagari has four consonant conjuncts namely: ज (ज्+ञ), क्ष (क्+श), श्र (श्+र), त्र (त्+र)

These consonant conjuncts are not used in Dogri except for transliteration of some Sanskrit words. 'ष' and 'ऋ' are also used for the same purpose in Dogri e.g. ऋषि, धनुष etc

In the second type of conjunct, a form of र is subjoined to certain consonants.

- i) When it is the second member of a cluster, it is indicated by a small diagonal slash under the sign of the first member of a conjunct: पन्द्रां, भ्राएं
- ii) When it is served as the first member of a conjunct the sound is indicated by a small hook placed on the top of the rekha for the second conjunct; This hook is deferred until after any matra written to the right side of the conjunct हांग

Consonant Sequences: Consonant sequences in Devanagari are classified in two classes: geminates and clusters.

Geminate (Doubled Consonants): It is written by writing the first component of the consonant cluster as the truncate form of the consonant. Dogri has fourteen geminate consonants that occur word medially. And word finally they occur only in those environments where these are not released. The following table shows the geminates in Dogri.

Geminate	Example of medial occurrence	Example of Final occurrence
Kk	पक्का (pəkka, "strong")	मक्क (məkk, "maize")
Gg	लग्गी (ləgg, "got hurt")	लग्ग (ləgg, "dateofmarriage")
Cc	कच्चा (kəcca, "unripe")	सच्च (səcc, "truth")
Jj	सुज्जना (sujjəna, "to swell)	अज्ज (əjj, "today")
tt	सुट्ठना (suəttna, "to throw")	सुट्ठ (suətt, "throw")
dd	हड्डी (əddi, "bone ")	उड्ड (udd, "fly")

Language in India www.languageinindia.com

11 : 10 October 2011

Preeti Dubey, Shashi Pathania, and Devanand

Comparative Study of Hindi and Dogri Languages with Regard to Machine Translation

Tt	कत्तना (kəttəna, “to spin”)	सत्त(sətt, “seven”)
Dd	सद्दना(səddəna, “to invite”)	रद्द (rədd, “cancel”)
Pp	टुप्पा (tuppa, “measuring pot”)	चुप्प (cup, “silent”)
Bb	डब्बा (dəbba, “box”)	बब्ब (bəbb, “father”)
Nn	बन्नना (bənnəna, “to tie”)	चन्न (cənn, “moon”)
Mm	लम्मा (ləmma, “tall”)	कम्म (kəmm, “work”)
Ss	रस्सा (rəssa, “thick rope”)	सस्स /səss/ “mother in law”)
Ll	कल्ला (kəlla, “alone”)	मुल्ल (mull, “price, value”)

Table 3: shows geminates used in Dogri

Five geminate aspirated stops and affricates also occur in both the positions. But in these geminations first aspirate tends to be converted into its un-aspirated counterpart e.g.

Geminate	Example of medial occurrence	Example of Final occurrence
kkh	मक्खन(məkkhən, “butter”)	अक्ख (əkkh, “eye”)
cch	मच्छी (məcchi, “fish”)	कच्छ (kəcch “near by”)
tth	मिट्ठा (mitṭha, “sweet”)	अट्ठ(əṭṭh, “eight”)
Tth	मत्था (mətṭha, “forehead”)	कत्थ(kəttṭh, “tale”)
Pph	लफ्फड (ləpphəṛ, “slap”)	झड़फ्फ(cəṛəpph, “encounter”)

Table 4: shows some more geminates formed by combination of two different consonants

Clusters: Vowel clusters are also very frequent in the language viz. two member vowel clusters, three member vowel clusters e.g. ai (आई/feminine. come), ae (आए/ Plural come), au (आऊ/ willing to come).

Some limitations of vowel clusters in Dogri are as below:

- i) ɔ (औ) never participates in vowel cluster
- ii) ε (ऐ) usually occurs as a last member of vowel cluster and in word final position.
- iii) Sequences of əa, aə, ai, iI, Ii, iu and ou do not occur.

In Dogri, consonant clusters with two members are very frequent-word initially and word medially. A few word final clusters also occur in the language. E.g. कटोरा /kɔra/ “bowl”, माल्क /malk/ “owner”, पनीर /pnir/ “cheese.”

Some limitations of consonant clusters are:

- i) H (ह) never participates in cluster formation in any position.
- ii) R (र) does not occur as initial letters in the cluster.
- iii) Word medially three member consonant clusters can be seen in certain syllable patterns, viz. word medially nasal +stop + trilled ;e.g. मंत्री /mentri/ “ minister”, पन्द्रां / pəndrā/ “ fifteen”, संस्कार /sənskar/ “ rite”.

Numerals: Numerals in Dogri are written in Persian.

Dogri	0	1	2	3	4	5	6	7	8	9
Hindi	०	१	२	३	४	५	६	७	८	९

Table 5: shows numerals as written in Dogri and Hindi Languages

Nasalized vowels

The two signs are used for nasalization. In Devanagari, *anusvara* (ँ) and *anunasika* (ँ) also called *chandrabindu* (ँ); whereas the use of chandrabindu is not prevalent in the orthography of Dogri.

Abbreviations

A dot (·) is used to mark abbreviation in Hindi as well as Dogri. A dot is used after the first syllable of the word to be abbreviation such as डॉ., पूं.

Supra-Segmental Features

Dogri has 5 supra segmental features, i.e. Length, Stress, Nasalization, Tone and Juncture.

1. Length

Length in Dogri has its important role in phonetic as well as phonological structure of the language. In Dogri there are ten vowels 3 short vowels and seven long vowels and all the ten vowels have been individually recognized as separate phonemes so the length feature hardly has any importance to prove their phonemic status on its own basis, but for extra long vowels length plays an important role to prove their phonemic status in word final position between long vowels and extra long vowels. For example:

Long Vowels	दा (of)	बी (also)	लै (have)	गलो (get melted)
Extra long Vowels	दाऽ (tric k)	बीऽ (seed)	लैऽ (rhyme)	गलोऽ (a creeper)

Table 6: shows the effect length.

On the level of consonants the feature of length is proven phonemic, e.g.

Short	कचा (bud)	बोली (deaf feminine)
Long	कच्चा (unripe)	बोल्ली (dialect)

Table 7: shows the length feature on the level of consonants

2. Stress

Although stress is not a prominent feature of Dogri like English, yet sometimes it is also proven phonemic at the level of syllables. The following pairs are distinguished by means of the stress, for example रला (rə'la, mixture); रला (rəla', mix imperative). In the first example, stress is on the 1st syllable/ rə' /and it means 'mixture' but in the second one rəla' stress is on the last (second) syllable, which means 'mix' in imperative sense.

Stress on 1st Syllable
रला /tə'la/ "sole of the shoe"

Stress on 2nd Syllable
रला /təla' / "tank"

3. Nasalization

Language in India www.languageinindia.com

11 : 10 October 2011

Preeti Dubey, Shashi Pathania, and Devanand

Comparative Study of Hindi and Dogri Languages with Regard to Machine Translation

Nasalization is also phonemic in Dogri and its phonemic status is established on the basis on the following oppositions:

Non Nasalized	Nasalized
बाग / bag / “garden”	बांग / bāg / “a crow of cock”
खा / kha / “eat” imperative	खां / khā / “may I eat ”

4. Tone

The major supra-segmental feature which widely appears in Dogri phonology is Tone. There are many levels of Tone but three of them stand in contrastive distribution. Therefore, it can be said that Dogri has three distinctive tones, namely (i) level-tone (ii) low or low rising tone (iii) high or high-falling tone. The phonemic status of tone in Dogri can be established on the basis of following oppositions, when one word can have three different meanings with the help of different tones. For example –

Level Tone	High-Falling	Low-Rising
पी / pi / “drink”	पीह / piə / “grind”	प्ही / pi/ “again”
कोड़ा / koṛā / “whip”	कोढ़ा / koṛa / “leper”	कोड़ा / kòṛa / “horse”

5. Juncture

There is one more suprasegmental feature – Juncture. The following examples show the existence of juncture in Dogri.

Continuous Flow / Close-Juncture	with the Pause / (Discontinuous) Open Juncture
बन्नना “to tie”	बन्न ना “please tie”
ओहदे “his, her”	ओहदे “give that

Table 8: shows examples of Juncture in Dogri Language

Juncture in Dogri has been proved phonemic on the syntactic level also. For example:-

In continuous flow: खाद्दी नेई पीती (neither eaten nor drunk)

With a break after: खाद्दी नेई, पीती (not eaten but drunk)

Similarly, with a break after: खाद्दी, नेई पीती (not drunk but eaten)

Conclusion

Language in India www.languageinindia.com

11 : 10 October 2011

Preeti Dubey, Shashi Pathania, and Devanand

Comparative Study of Hindi and Dogri Languages with Regard to Machine Translation

Indian languages are inflectional with a rich morphology, relatively free word order, and default sentence structure as SOV (Subject Object Verb). It is believed that Machine Translation systems can be developed with less effort using direct approach. Hindi and Dogri belong to the same group of Indo-European Family. Both the languages use the same script (Devanagari) but with some significant differences discussed in the paper.

Dogri is a highly inflectional language. Most interesting difference is inflection and the Juncture in Dogri

This comparative study is beneficial for the development of machine translation system from Hindi to Dogri and according to this study Direct approach of Machine translation shall be used; since direct approach is most suitable choice for language pair that are closely related to each other. A major challenge for the development of Hindi-Dogri Machine Translation system is the lack of language resources. Language resources refer to data resources such as digital bilingual dictionary, corpora, morphological analyzer and generator, etc.

References

1. Development of A Hindi to Punjabi Machine Translation System, V Goyal, Ph.D. thesis, Department of Computer Science, Punjabi University, Patiala ,2010
2. V Goyal and G S Lehal “Comparative Study of Hindi and Punjabi Language Scripts”, Nepalese Linguistics, Journal of the Linguistics Society of Nepal, Volume 23, pp. 67-82 (2008).
3. Sanjay chatterji et.al, “A Hybrid Approach for Bengali to Hindi Machine Translation ” in Proceedings of Natural language processing (ICON-2009), Hyderabad, India , pp(83-91)(2009).
4. Dr. Veena Gupta, “Dogri Vyakaran”, Jammu and Kashmir Academy of Art, Culture and Language, Jammu- Feb 1995.
5. www.lisindia.net
6. www.wikipedia.org

Preeti Dubey, **you may like to add your degrees if you prefer.**

Research Scholar

Department of Computer Science & IT

University of Jammu

Jammu

Jammu & Kashmir

India

preetidubey2000@yahoo.com

Language in India www.languageinindia.com

11 : 10 October 2011

Preeti Dubey, Shashi Pathania, and Devanand

Comparative Study of Hindi and Dogri Languages with Regard to Machine Translation

308

Shashi Pathania, **you may like to add your degrees if you prefer.**

Professor and Head

Dogri Department

University of Jammu

Jammu

shashipathania@gmail.com

Devanand, **you may like to add your degrees if you prefer.**

Professor and Head

Department of Computer Science & IT

University of Jammu

Jammu

Jammu & Kashmir

India

dpadha@rediffmail.com