ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(27), DOI:10.17485/ijst/2015/v8i27/81615, October 2015* ISSN (Online) : 0974-5645

Hindi-Kannada Named Entity Transliteration: Issues and Possible Solutions

Annarao Kulkarni*, B. R. Srivatsa and Chetan Baji

Centre for Development of Advanced Computing (C-DAC), Bengaluru - 560100, Karnataka, India; annaraogk@yahoo.com

Abstract

Indian languages belong to four language families, namely, the Indo-Aryan, Dravidian, Tibeto-Burman and the Austro-Asiatic. Hindi and Kannada belong to Indo-Aryan and Dravidian family respectively and are evolved from the ancient Brahmi script and have a common phonetic structure. But the Named Entity writing convention is different due to dialectic influence, language specific rules, and other factors. Due to this, the Named Entity Transliteration from Hindi to Kannada and vice versa is not one to one character mapping. This introduces many problems in Machine Translation (MT), Cross Lingual Information Retrieval (CLIR) and Parallel corpus creation between Hindi and Kannada. The paper discusses the Named Entity Transliteration issues encountered between Hindi and Kannada during the parallel corpora creation from Hindi to Kannada for the Indian Language Corpus Initiative (ILCI) project. In this paper, we discuss cases of no exact equivalence character between Hindi and Kannada, multiple mappings, diacritic marks, loan words and language specific transliteration issues in detail and propose the possible solution to resolve the problem. At implementation level, one may make use of either Finite-State Transducers (FST) or Regular Expressions

Keywords: Hindi, Kannada, Named Entity, Regular Expressions, Transliteration

1. Introduction

Named Entities (NE) are union of subsets of person name, place name, organization name, monetary expressions, dates, numerical expressions. Parallel pair of Named Entities between two languages of homogeneous and heterogeneous family and their classification plays an important role in Natural Language Processing (NLP) applications. Extracting NE is a challenging task. Various methods such as Hidden Marko Model (HMM), Conditional Random Fields (CRF), rule-based and hybrid approaches are in practice for Named Entity Extraction. Typically, NE of person name and place name are transliterated and organization name is translated for CLIR and MT.

Transliteration is not translation; the text remains the same, only the script in which it is rendered is changed[3]. Transliteration is usually done so as to preserve the accuracy of pronunciation as much as practically feasible without distorting the target language rules. In reality, it is a bit difficult in case of Named Entity Transliteration between Hindi and Kannada. Practically, we came across many issues while translating tourism domain ILCI corpora from Hindi to Kannada. In this paper, we address various issues we encountered during transliteration and propose possible solutions to resolve the issues. Rule based solution is better compared to other techniques. Techniques such as use of NE annotated parallel corpora and use of comparable corpora for NE transliteration may suffer due to lack of coverage and constrained resources. Methods based on phonetic information and statistical methods are also not feasible because they are computing incentive and language family dependent. Hence, designing a rule to address Named Entity transliteration and then implement those using either FST or regular expressions is more feasible.

* *Author for correspondence*

2. Related Works

Suggests finite state Hindi Urdu Machine Transliteration using Finite-State Transducers based on language specific characteristics, FST and UIT[1]. Problems have been faced at various levels like no equivalence of character set between two languages, diacritic marks and language specific issues for Urdu-Hindi Machine translation. Have come up with rule base for Hindi-Punjabi transliteration[8]. Suggest transliteration among Indian Languages using WX notation, by making use of common representation to transliterate among Indian languages[6]. In Hindi, syllabic vowels occur in the middle and end e.g., आइहोले (AihOlE), नकिारागुआ (nikArAguA). In Kannada, sequence of syllabic vowel is not allowed and pure vowel can occur only as a first syllable. This makes WX notation methodology difficult to implement.[5] have proposed statistical method for transliteration system across Indian languages using parallel corpora and algorithm similar to Soundx. However, Indian languages are resource constrained and due to lack of resources and NE aligned parallel corpora, rule based approach is suitable for Named Entity transliteration. ILCI parallel corpus is one such effort to create parallel corpora for 17 Indian languages with Hindi as source language. We have used ILCI parallel corpora for Hindi-Kannada to arrive at rule base for transliteration of NE from Hindi to Kannada.

3. Hindi and Kannada

Hindi is one of the official languages of India and is written in Devanagari Script. Devanagari is a script shared by ten other official languages of India and more than 70% Indians can understand and speak Hindi to a certain level. Hindi language has 49 dialects as per census of India 2001. The most spoken of them are Bhojpuri, Rajasthani, Chhattisgarhi, Magahi, Pahari, Bundeli, Bagheli, Awadhi, Marwari, Mewari etc[6]. Various phonological transformations take place and deviation is observed from Standard Hindi in written script due to dialectic influence thus resulting in different writing conventions. This is one of the issues while transliterating Named Entities from Hindi to other Indian languages for NLP applications. Guidelines of Central Hindi Directorate are considered as authentic for writing in Hindi. Both Hindi and Kannada languages evolved from the ancient Brahmi script and have a common phonetic structure with some differences in vowels, consonants, diacritic marks. This makes transliteration difficult. Kannada is one of the four major Dravidian languages. Kannada is rich in morphology and is agglutinative in nature. Kannada has many dialects and the major dialects are the old Mysore dialect, the coastal or Mangalore dialect and the northern or Dharwad dialect[7]. Kannada is a diglossic language. Literary variety differs in several respects from the spoken variety in phonology, morphology, lexicon and syntax. But, literary variety is the same across dialects.

4. Hindi-Kannada NE Transliteration

In the following section, we discuss issues with respect to no equivalent characters, multiple mapping, diacritic marks, language specific issues, loan words and suggest suitable solutions for NE transliteration from Hindi to Kannada.

4.1 Vowels

Table 1. Hindi-Kannada Vowels

Hindi	Kannada	ITRANS
अ	ಅ	A
आ	ಆ	A
इ	ಇ	I
ई	ಈ	I
उ	ಉ	U
ऊ	ಊ	U
ऋ	ಋ	R
-	ಎ	E
ए	ಏ	E
ऐ	ಐ	ai
-	ಒ	o
ओ	ಓ	O
औ	ಔ	au

Hindi contains 11 vowels and 10 dependent vowel symbols. Examples for them are ा, ी. These dependent vowel symbols are also called matras. Apart from these 11 vowels, Devanagari ऍ [U+090D] and ऑ [U+09011] are also considered as vowels for transliterating loan words borrowed from Urdu and English with their corresponding matras ॅ[U+0945] and ॉ [U+0940]. This

issue is discussed in section 4.6. Kannada contains 13 vowels and 12 dependent vowel symbols. If we compare Hindi-Kannada vowel set, Hindi doesn't have short [e] and [o]. Kannada contains both long and short [e] and [o]. Kannada language doesn't have ऍ and ऑ equivalents.

Case-1 Diphthongs: Sequence of syllabic vowels does not occur in Kannada language (S.N. Sridhar, 2007). But, there is a tendency to write Kannada diphthongs ಐ [ai] and ಔ [au] as sequence of syllabic vowels in Hindi like आइ [a|A+i|I] and आउ [a|A+u|U]. In such cases, replace Hindi sequence of syllabic vowels आइ [a|A+i|I] with Kannada dipthong ಐ [ai] and replace sequence of Hindi syllabic vowels आउ [a|A+u|U] with Kannada diphthong ಔ [au]. For example

ಐ [ai] – आइहोले - ಐಹೊಳೆ (aihoLe)
आइजोल - ಐಜೋಲ (aijOla)
ಔ [au] – आउटडोर – ಔಟಡೋರ್ (auTDOr)
आउटलुक – ಔಟ್‌ಲುಕ್ (auTluk)

Diphthong matras: Replace ा + इ|ई with Kannada ೈ [ai] matra. Replace ा + उ|ऊ with Kannada ೌ [au] matra. For example,

कराईकुडी – ಕರೈಕುಡಿ (karaikuDi)
ताइवान – ತೈವಾನ (taivAna)
राउडी – ರೌಡಿ (rauDi)

Case-2 Long Vowels: Long vowels (and matras) tend not to occur word finally, except as grammatical morphemes[7]. In Kannada, long vowel matra occurs only in initial and middle positions for Named Entities. Replace long vowel matra for the word finally syllable with short [a],[i],[e],[u] and [o] matra respectively. For example

भविानी – ಭೌವಾನಿ (bhivAni)
मनाली – ಮನಾಲಿ (manAli)
मसूरी – ಮಸೂರಿ (masUri)
पणजी – ಪಣಜಿ (paNaji)
रायबरेली – ರಾಯಬರೇಲಿ (rAyabarEli)
गनिी – ಗನಿ (gini)
पेरू – ಪೆರು (peru)
जम्मू – ಜಮ್ಮು (jammu)
बसिाऊ – ಬಿಸಾಮ (bisAvu)
कन्याकुमारी – ಕನ್ಯಾಕುಮಾರಿ (kanyAkumAri)
पुरी – ಪುರಿ (puri)
बेल्लारी – ಬಳ್ಳಾರಿ (baLLAri)

Case-3 Glide Insertion: When two vowels come together, they are replaced with glides [y] and [v] in Kannada. If the first vowel is front vowel [i, I] then [y] is inserted and [v] is inserted if the first vowel is back or low[7]. This rule is not applicable to diphthongs [ai] and [au]. Kannada words do not end with pure vowel and sequence of syllabic vowels does not occur in Kannada. If the final syllable of the NE ends with [A], [i, I] and [u, U] then replace with [vA], [yi] and [vu] respectively. For example,

बहराईच – ಬಹರಾಯಿಚ (baharAyica)
तरिुअनन्तपुरम् – ತಿರುವನಂತಪುರಂ (tiruvanaMtapuraM)
नोइडा – ನೋಯಿಡಾ (nOyiDA)
हसिुआ – ಹಿಸುವಾ (hisuvA)
गोआ – ಗೋವಾ (gOvA)
नकिारागुआ – ನಿಕಾರಾಗುವಾ (nikArAguvA)
हरदोई - ಹರದೋಯಿ (haradOyi)
राऊ – ರಾಮ (rAvu)
बसिाऊ – ಬಿಸಾಮ (bisAvu)

Case-4 Multiple Vowel Mapping: Hindi does not contain short [e] and [o]. For the syllable final [E] and [O] word, case 3 will be used, otherwise replace Hindi [E] and [O] with Kannada short vowel [e] and [o] respectively. This is a case of multiple mapping. However, both these possibilities are more suitable for data mining and Kannada data searching task. For example

राऊ – ರಾಮ (rAvu)
हरदोई – ಹರದೋಯಿ (haradOyi)
एर्नाकुलम – ಎರ್ನಾಕುಲಂ (ernAkulaM)
क्रोएशयिा – ಕ್ರೊಯೇಶಿಯ (kroyEshiya)
वेनेजुएला – ವೆನೆಜುವೆಲ (venejuvela)

4.2 Vowel Modifiers

Table 2. Hindi-Kannada Vowel Modifiers

Vowel modifiers	Hindi	Kannada	ITRANS
Chandrabindu	ँ	-	.N
Anuswara	ं	ಂ	M
Visarga	ः	ಃ	H

After a consonant, vowel or matra, a character can be used which modifies the vowel sound and is called vowel modifier[9]. This can be Chandrabindu, Anuswar or Visarga. Kannada doesn't have Chandrabindu. Following table shows the list of vowel modifiers.

Case-5 Chandrabindu ँ[.N]: Chandrabindu denotes nasalization of the preceding vowel. It is substituted with anuswara in Kannada. For example

राँची - ರಾಂಚಿ (rAMci)
हाँगकाँग – ಹಾಂಗ ಕಾಂಗ (hAMga kAMga)

Case-6 म [ma] as word final syllable: Word final syllable [ma] is replaced by anuswara [M] in Kannada. For example

मज़ोरम – ಮಿಜೊರಂ (mijoraM)
सक्किमि – ಸಿಕ್ಕಿಂ (sikkiM)
भद्राचलम – ಭದ್ರಾಚಲಂ (bhadrAcalaM)
चदिंबरम – ಚಿದಂಬರಂ (cidaMbaraM)
श्रीकाकुलम – ಶ್ರೀಕಾಕುಲಂ (shrIkAkulaM)

4.3 Diacritic Marks

A diacritic mark called Nukta is used in Devanagari for some languages to represent sounds from other languages. It takes the form of a dot placed below a character.

Case-7 Nukta: In Hindi, Nukta is used to modify consonants क, ख, ग, ज, ड, ढ, फ as क़, ख़, ग़, ज़, ड़, ढ़, फ़ respectively. However, there is no equivalent Nukta character in Kannada. It is necessary to drop Nukta while transliterating from Hindi to Kannada. Replace क़ with क ख़ with ख, ज़ with ज, ढ़ with ढ, फ़ with फ respectively. For example

धारवाड़ – ಧಾರವಾಡ (dhAravADa)
चंडीगढ़ – ಚಂಡೀಗಢ (caMDigaDha)
फ़्रान्स – ಫ್ರಾನ್ಸ (phrAnsa)
ज़िम्बाब्वे – ಜಿಂಬಾಬ್ವೆ (jiMbAbve)
अज़रबैजान – ಅಜರಬೈಜಾನ (ajarabaijAna)

4.4 Consonants

Hindi and Kannada share the same consonant set except for the non-varga consonant [La – ಳ] i.e., Devanagari consonant ळ. Devanagari ळ character is not used in Hindi. Consonants are categorized according to their phonetic properties into 5 Vargas and the remaining ones fall under the Non-Varga category. The last consonant in each Varga is called nasal consonant.

Case-8 Nasal consonants: If the nasal consonant is followed by any consonant belonging to that Varga, then, that nasal consonant is replaced by anuswar [.m] in Kannada. For example

अङ्कोला – ಅಂಕೋಲ (aMkOla)
केन्द्र – ಕೇಂದ್ರ (kEMdra)

ङ followed by क | ख | ग | घ is replaced by [M]
ञ followed by च | छ | ज | झ is replaced by [M]
ण followed by ट | ठ | ड | ढ is replaced by [M]
न followed by त | थ | द | ध is replaced by [M]
म followed by प | फ | ब | भ is replaced by [M]

Table 3. Hindi-Kannada Varga Consonants

क / ಕ [ka]	ख / ಖ [kha]	ग / ಗ [ga]	घ / ಘ [gha]	ङ / ಙ [~Na]
च / ಚ [ca]	छ / ಛ [cha]	ज / ಜ [ja]	झ / ಝ [jha]	ञ / ಞ [~na]
ट / ಟ [Ta]	ठ / ಠ [Tha]	ड / ಡ [Da]	ढ / ಢ [Dha]	ण / ಣ [Na]
त / ತ [ta]	थ / ಥ [tha]	द / ದ [da]	ध / ಧ [dha]	न / ನ [na]
प / ಪ [pa]	फ / ಫ [pha]	ब / ಬ [ba]	भ / ಭ [bha]	म / ಮ [ma]

Table 4. Hindi-Kannada Non-Varga Consonants

य / ಯ [ya]	र / ರ [ra]	ल / ಲ [la]	व / ವ [va]	श / ಶ [sha]
ष / ಷ [Sha]	स / ಸ [sa]	ह / ಹ [ha]	ಳ [La]	

Case–9 [ಳ - La] consonant in Kannada: In most of the cases (not always), Hindi ल [la] is substituted with ಳ [La] in Kannada. This is a case of multiple mapping for transliterating from Hindi to Kannada. For data mining, this rule is more useful. In such cases, search with both [la] and [La] replacements in Kannada. For example

केरल – ಕೇರಳ (kEraLa)
तमलिनाडु – ತಮಿಳುನಾಡು (tamiLunADu)
बेल्लारी – ಬಳ್ಳಾರಿ (baLLAri)
बेंगलूरू – ಬೆಂಗಳೂರು (beMgaLUru)
भुसावल – ಭುಸಾವಳ (bhusAvaLa)

4.5 Language Specific

Case-10 Conjunct formation for ल [la]: ल followed by consonant is often replaced by ल + halant to form a conjunct cluster. However, this is not a general rule. For example

कोलकाता – ಕೊಲ್ಕತ್ತ (kOlkatta)
सुलतानगंज – ಸುಲ್ತಾನಗಂಜ (sultAnagaMja)
जैसलमेर – ಜೈಸಲ್ಮೇರ (jaisalmEra)
तरुनिलवेली – ತಿರುನಲ್ವೇಲಿ (tirunalvEli)

Case–11 र [ra] followed by Consonant: If र [ra] is followed by a consonant, then, replace र with र + halant.

In Kannada, repha is inserted to form a conjunct cluster. This is not a general rule, though. For example

कारगिल – ಕಾರ್‌ಗಿಲ್ (kArgil)
इटारसी – ಇಟಾರ್‌ಸಿ (iTArsi)
करनूल – ಕರ್‌ನೂಲ (karnUla)
बरमूडा – ಬರ್‌ಮುಡ (barmuDa)

Case-12 Replace Hindi लैं and बैं with ಲ್‌ಯಾಂ [lyAM] and ಬ್‌ಯಾಂ [byAM] in Kannada. For example

फनिलैंड - ಫೆನ್‌ಲ್‌ಯಾಂಡ್ (phinlyAMD)
आइसलैण्ड – ಐಸ್‌ಲ್‌ಯಾಂಡ್ (aislyAMD)
न्यूज़ीलैंड – ನ್‌ಯೂಜಿಲ್‌ಯಾಂಡ್ (nyUjilyAMD)
बैंकाक – ಬ್‌ಯಾಂಕಾಕ್ (byAMkAk)
बैंड – ಬ್‌ಯಾಂಡ್ (byAMD)
बैंक – ಬ್‌ಯಾಂಕ್ (byAMk)

4.6 Loan Words

In Hindi, two vowels ऍ and ऑ with their corresponding vowel modifiers ॅ and ॉ are used for writing loan words from Urdu and English e.g., ऑल इंडिया रेडियो (All India Radio), ऑक्टोपस (Octopus), ऑमलेट (Omelette), ऑयल ऍण्ड गैस (Oil and Gas), ऑफ (of). In case of Kannada, loan words end with explicit halant [್] i.e., consonant without vowel part in it e.g., ಕೆನರಾ ಬ್‌ಯಾಂಕ್ (kenarA byAMk), ರೆಸಿಡೆನ್ಸಿ ರೋಡ್ (resiDensi rOD), ಹೋಟೆಲ್ ತಾಜ್ (hOTel tAj). When the case marker is attached to loan words, the explicit halant is replaced with enunciative vowel [u] e.g., bank–ಬ್‌ಯಾಂಕು (byAMku), car–ಕಾರು (kAru)

Case-14 While transliterating loan words from Hindi or English to Kannada, replace ऑ with [A] and ऍ with [a|e].

5. Conclusion

The rules discussed in this paper are applied to translation activities for creating parallel corpora from Hindi to Kannada for the ILCI project and these rules will help to get better quality of transliteration and translation output. Creating parallel dictionary of NE between two languages is a difficult task. Techniques, such as, use of parallel corpora for transliteration is not feasible due to constrained resources and lack of coverage. However, parallel corpora help in creating rule base for effective Named Entity transliteration and translation for CLIR and MT applications. This work can be extended to other Dravidian languages. However, there is no one-to-one character set mapping among Dravidian languages but the rules are the same. This work can be used as preprocessor and can be used with other existing transliteration techniques like phonetics based and parallel corpora based transliteration techniques.

6. Acknowledgments

While writing this paper, we referred Hindi-Kannada Tourism Domain ILCI parallel corpora with Hindi as source language. Examples of place name for this work have been taken from ILCI parallel corpora of Tourism domain. We would like to thank ILCI project team for this.

7. References

1. Abbas Malik MG, Boitet C, Bhattacharyya P. Hindi Urdu Machine Transliteration using Finite-State Transducers. COLING; 2008. p. 537-44.
2. Jha GN. The TDIL program and the Indian Languages Corpora Initiative (ILCI). Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC); 2010. p. 982-5.
3. Murthy KN, Badugu S. Roman Transliteration of Indic Scripts. Proceedings of the Tenth International Conference on Computer Applications (ICCA); 2012. p. 19-26.
4. Zhang M, Duan X, Pervouchine V, Li H. Machine Transliteration: Leveraging on Third Language. Proceedings of the 23rd International Conference on Computational Linguistics; 2010. p. 1444–52.
5. Srivastava R, Bhat RA. Transliteration System across Indian Languages Using Parallel Corpora. Proceedings of the 27th Pacific Asia Conference on Language, Information and Computation; 2013. p. 390-8.
6. Gupta R, Goyal P, Diwakar S. Transliteration among Indian Languages Using WX Notation. KONVENS; 2010. p. 147-50.
7. Sridhar SN. Modern Kannada Grammar. New Delhi: Manohar Publishers; 2007.
8. Goyal V, Lehal GS. Hindi to Punjabi Machine Transliteration System. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies; 2011.
9. Indian Script Code for Information Interchange (ISCII). Bureau of Indian Standards; 1991.