Digraphia of Maithili: Transliteration of Speech Corpus Transcripts to Tirhuta Script

Rajesha N[1], Narayan Choudhary [2]

[1]Central Institute of Indian Languages
n.rajesha@yahoo.co.in
[2]Central Institute of Indian Languages
choudharynarayan@gmail.com

Abstract

Maithili is a scheduled language in India that is widely represented by the Devanagari script and has an ethnic script called Tirhuta. The renaissance of the Tirhuta script for the Maithili took advancement when Unicode allocated a block for the Tirhuta and compliant fonts were prepared. The aspirations of the language community to assert the revival of their ethnic script and to be able to get the technological services in it will make the transliteration an essential step. The motivation of the proposed work is to provide transliteration between Devanagari and Tirhuta of Maithili. This paper presents the brief account of the digraphia in Maithili, and pioneering effort of LDC-IL to develop Maithili Sentence Aligned Speech dataset in Tirhuta using transliteration.

Keywords: Devanagari, Digraphia, Maithili, Tirhuta, Transliteration, LDC-IL

1. Introduction

A script is a collection of symbols or letters used to document language. The knowledge and culture can be transferred to generations while documenting and recording the information through language. Languages can be expanding beyond their initial demographics since writing allowed for long-distance communication and fostered trade and cultural interchange. Scripts like Roman, Cyrillic, Arabic, Devanagari use to represent many languages as well as a single language might be represented in multiple scripts. Sanskrit is represented in multiple scripts because of antiquity and geographical spread. According to Pollock (2009:274), the Brāhmi script in South Asia evolved through regional adaptations that reflected local identities and aesthetic preferences, which led to the development of various alphabets that were distinct from each other, while Sanskrit continued to be written in these scripts. Rosen, R. (2017:48-49) declares, 'There is not now nor has there ever been an "official" Sanskrit script.'

Scripts not only share the function of instrumentality to record the spoken language but also function as a symbol of identity carrying historical, cultural and political meanings of the subscribing language community. A language in the course of time may change its predominant script, a phenomenon known as diachronic digraphia, or it might manifest itself in two scripts concurrently known as synchronic digraphia. Elti di Rodeano (2019) states digraphia has two understandings: '*One, specular to the notion of diglossia, interprets it as a linguistic change, the use of two scripts in privative opposition for the same language, where variations can be identified variations, like dysphasic, diastratic, or diamesic ones; the other provides the notion of prestige as the single determining factor for the choice between two scripts.*'

2. Script Switching Trends

Script switching of any language merely on instrumental grounds of lack of fit between the phonemic composition and the writing system cannot provide necessary motivation. Script reforms can be carried out to better fit rather than changing the predominant script. The extreme reformers argue the script switching was needed to make a 'break with' or 'revive' the past. The script change occurs to politically or culturally distance the present from the immediate past. Howard (2001) explains the Arabic script used for Turkish language by the Ottoman Empire and its successor state the Republic of Turkey, switched it to Roman script to give such a break. Konkani language moving from Roman (Rōmi) to Devanagari in Goa[1] is also a similar act of administration, triggering diachronic digraphia. Such administrative moves are always limited by the jurisdiction of the authority. Hence beyond the political borders of Goa, Konkani exhibit synchronic digraphia. It uses Kannada script in Karnataka and Malayalam script in Kerala by the official recognition[2] and patronage[3] of the respective governments.

The script change occurs in order to assert the revival of ethnic script. The historical script of Meiteilon (Manipuri), which had been discontinued, was restored. A Cabinet resolution on May 18, 2005, brought Meiteilon written in Meitei Mayek into schools. Later the Manipur Official Language Act was amended in 2021 in line with the policy, to replace Bengali Script with Meitei Mayek Script in official usage[4]. In a similar endeavour reported by Marzluf and Myagmar (2019), Mongolian language law 2015 mandates all general education schools to teach the traditional 'national' script, Mongol Bichig. The law maintains the biscript policy, reaffirms the Cyrillic Mongolian script, and declares that Mongol Bichig would be used for all official and legal purposes in the future. The statute requires the president, parliament speaker, and

[1] The Goa, Daman and Diu Official Language Act, 1987
[2] Karnataka Government order ಸಂಕಇ 19 ಕರಅ94, Dated 20-April-1994 and ಇಡಿ 207 ಡಿಜಿಓ 2004, Dated 13-03-2007
[3] Kerala Government Order G.O.No.524/2012/CAD, Dated 8-11-2012
[4] Manipur Gazette No.427, Government of Manipur, Imphal, 2021

Rajesha N., Narayan Choudhary, *Digraphia of Maithili: Transliteration of Speech Corpus Transcripts to Tirhuta Script*. In: Z. Vetulani, P. Paroubek, I. Kardava (Eds.), Human Language Technologies as a Challenge for Computer Science and Linguistics – 2025, pp. 67–71. Poznań: Adam Mickiewicz University Press, 2025. © Authors, 2025. https://doi.org/10.14746/amup.9788323245414.13. Open Access chapter, distributed under the terms of the CC licence (BY-NC-ND, https://creativecommons.org/licenses/by-nc-nd/4.0/).

prime minister to utilise Mongol Bichig in all correspondence with international counterparts.

Script switch are made to demonstrate indigenous dominance in favour of one indigenous script like Devanagari even when it is marginalizing the other indigenous scripts like Kaithi, Mahājani etc. over a foreign dominating script like Arabic to provide a break in the historical continuum of the past. The Indian constitution, designating Hindi in Devanagari as the official language[5], can be interpreted as such a shift. This phenomenon has been documented by Dasgupta, J. (1970). Similarly, Sebba, M. (2007) and Brezianu, Andrei, Vlad Spânu (2010) points out, when Moldova was a part of the Soviet Union, the Soviet designated Cyrillic script, created by a missionary for Slavonic languages, as the official alphabet on the language spoken by the majority of Moldova, and named the language Moldovan. While the country was still a Soviet republic, Moldova's Supreme Soviet passed the adoption of Roman script by legislation in August 1989; Moldova promulgated Law No. 52 in March 2023, naming the national language as Romanian. At times, it may simply be a matter of administrative convenience. Sohoni, P. (2016), documented that Marathi language was represented both in Mōḍi and Devanagari (a variant known as Bālabōdh) scripts and it was made to move towards Devanagari[6].

3. Transliteration

Transliteration is a process of rendering a language from one writing system to another. Rejitha K. S. (2018) pointed out that transliteration is useful for those who know the language but not the script. It removes the barriers of learning multiple scripts. Automatic transliterator is beneficial for the languages that use multi-script. Languages use multiple scripts for several reasons. A language spoken across vast regions may develop different script usages due to historical, geographical, religious or cultural divergence make a language to adopt different scripts. A language might have an official script, but people continue using older scripts due to tradition or convenience.

Transliterator enhances the chance of survival and rejuvenation of ethnic scripts like Tirhuta, Śārada, Mōḍi over a dominating script like Devanagari. It even facilitates the natives who speak languages that are glottonyms but use different scripts to keep up identity. It also helps a second language learner to learn the intended language through their own script. Jayashree Nair et al. (2022), described Transliteration is used extensively in natural language applications such as information retrieval.

4. Maithili Digraphia

Maithili, an interesting example of digraphia, is recognised as a language by the 8th schedule of Indian constitution is spoken in the eastern part of India and some adjoining parts of Nepal. It has a literary history dating back to the 14th century. The language is traditionally written in script named Tirhuta, which means the script of the land bounded by Gaṇḍaki River in the west, Kōśī River in the east and Gaṅgā River in the south. It is also called as Mithilakshar and Pūrva-Vidēha. Mithila, the name of the capital of the land, Videha, appears frequently in Vedic and epic texts. Mithila and Videha eventually became synonymous. From ancient times it is renowned as the land given to intellectual pursuits. Rājendralala Mitra (1882), mentioned that the *Lalitavistara*, an ancient Sanskrit Buddhist text of great importance noted the pūrva-vidēha script in its 10th chapter.

In the Mithila region, Tirhuta was in use of ritual and socio-cultural practices as well as writing Sanskrit texts. Devanagari is widely recognised because it was primarily used for writing Sanskrit in northern India. Sir Ashley Eden, the then lieutenant-governor of Bengal introduced Hindi in the courts of Bihar from the 1st of January 1881. This move was welcomed by people. Mithilesh K.J., (2017) reasons, Hindi was welcomed because of the incomprehensibility of Urdu by most of the population and the cultural elite of the region are among Sanskrit scholars who already had the knowledge of Devanagari. Thus, the process of replacing Urdu and Persian with Hindi was enthusiastically taken up by the colonial administration. The efforts of Nagari Prachar Sabha of Kashi, Devanagari got an elevated position. The introduction of print technology Devanagari got momentum, and the usage of other scripts declined. The Tirhuta script saw the renaissance once the modern fonts are introduced in Tirhuta with the recognition of Unicode standards 7.0 in 2014 (Range U+11480 to U+114DF). Maithili become the second official language in the Indian state of Jharkhand in 2018[7]. Devanagari is the officially recognized script for Maithili in primary and secondary education, while Tirhuta is optionally recognized alongside Devanagari in higher education[8]. Maithili literature is published in the Tirhuta script that creates a need for transliteration to convert content prepared in Devanagari into Tirhuta.

4.1. LDC-IL's Approach to Maithili Digraphia

Linguistic Data Consortium for Indian Languages[9] (LDC-IL) is a scheme of the Ministry of Education, Government of India, with a mandate to create linguistic resources in Indian languages needed for natural language processing. This scheme works under the supervision of the Central Institute of Indian Languages, Mysore.

The consortium treats all scripts of any language equally. It strives to achieve balance among those scripts so that the language community may benefit from the rise of language technology in any and all the scripts that it subscribes. Since Maithili is a language of digraphia, the consortium decided to publish the sentence aligned speech dataset in both Devanagari and Tirhuta scripts. Both scripts originated from Brāhmi, and have a similar structure, thus as Garfield, E. (1977) denotes, the transliteration can ideally be achieved, by one-for-one character-by-character replacement. It should be a simple mechanical process.

4.2 Devanagari-Tirhuta Character Mapping

Devanagari script is a superset of characters that caters to multiple languages. Each subscribing language of Devanagari derives its own subset of alphabets for use.

[5] Constitution of India Article 343

[6] The Maharashtra Official Languages Act, 1964

[7] The Bihar Rajbhasha (Jharkhand Sanshodhan) Adhiniyam, 2018 Act 20 of 2018

[8] http://lisindia.ciil.org, Retrieved 24 October 2025.

[9] www.ldcil.org

Hence, the Devanagari block of Unicode accommodates alphabets of more than one language. While LDC-IL preparing transliteration schema, Non-Maithili alphabets of Devanagari are ignored as foreign character. The Tirhutha character Anji (U+11480) is not transliterated being a distinct cultural symbol and not present in Devanagari. The Tirhuta character Gvang (U+114C5; vēdika anusvāra) is also omitted from the schema not only because of non-availability of exact match but also as it was not necessary in the dataset in question. Tirhuta Unicode repertoire has only vowel signs for Short E and Short O, but not their corresponding vowels forms as they never occur in word initial position[10].
LDC-IL developed the character mapping as in Table 1 as a pre-requisite for transliteration.

Maithili Alphabet	Dēvanāgari	Tirhutā
Om - *a sacred syllable*	ॐ	𑒝
Letter A	अ	𑒁
Letter AA	आ	𑒂
Letter I	इ	𑒃
Letter II	ई	𑒄
Letter U	उ	𑒅
Letter UU	ऊ	𑒆
Letter Vocalic R	ऋ	𑒇
Letter Vocalic RR	ॠ	𑒈
Letter Vocalic L	ऌ	𑒉
Letter Vocalic LL	ॡ	𑒊
Letter E	ए	𑒋
Letter AI	ऐ	𑒌
Letter Candra O	ऑ	𑒍
Letter O	ओ	𑒎
Letter AU	औ	𑒏
Letter KA	क	𑒐
Letter KHA	ख	𑒑
Letter GA	ग	𑒒
Letter GHA	घ	𑒓
Letter NGA	ङ	𑒔
Letter CA	च	𑒕
Letter CHA	छ	𑒖
Letter JA	ज	𑒗
Letter JHA	झ	𑒘
Letter NYA	ञ	𑒙
Letter TTA	ट	𑒚
Letter TTHA	ठ	𑒛
Letter DDA	ड	𑒜
Letter DDHA	ढ	𑒝
Letter NNA	ण	𑒞
Letter TA	त	𑒟
Letter THA	थ	𑒠
Letter DA	द	𑒡
Letter DHA	ध	𑒢
Letter NA	न	𑒣
Letter PA	प	𑒤
Letter PHA	फ	𑒥
Letter BA	ब	𑒦
Letter BHA	भ	𑒧
Letter MA	म	𑒨
Letter YA	य	𑒩
Letter RA	र	𑒪
Letter LA	ल	𑒫
Letter VA	व	𑒬
Letter SHA	श	𑒭
Letter SSA	ष	𑒮
Letter SA	स	𑒯
Letter HA	ह	𑒰
Letter DDHA	ऴ	𑒱
Letter RHA	ढ़	𑒲
Letter FA	फ़	𑒳
Vowel Sign AA	ा	𑒰
Vowel Sign I	ि	𑒱
Vowel Sign II	ी	𑒲
Vowel Sign U	ु	𑒳
Vowel Sign UU	ू	𑒴
Vowel Sign Vocalic R	ृ	𑒵
Vowel Sign Vocalic RR	ॄ	𑒶
Vowel Sign Vocalic L	ॢ	𑒷
Vowel Sign Vocalic LL	ॣ	𑒸
Vowel Sign Short E	ॅ	𑒹
Vowel Sign E	े	𑒺
Vowel Sign AI	ै	𑒻
Vowel Sign Candra O	ॉ	𑒼
Vowel Sign Short O	ॊ	𑒽
Vowel Sign O	ो	𑒾
Vowel Sign AU	ौ	𑒿
Sign Candrabindu	ँ	𑓀
Sign Anusvara	ं	𑓁
Sign Visarga	ः	𑓂
Sign Virama (Halant)	्	𑓃
Sign Nukta	़	𑓄
Sign Avagraha	ऽ	𑓅
Abbreviation Sign	॰	𑓆
Digit Zero	०	𑓐
Digit One	१	𑓑
Digit Two	२	𑓒
Digit Three	३	𑓓
Digit Four	४	𑓔
Digit Five	५	𑓕
Digit Six	६	𑓖
Digit Seven	७	𑓗
Digit Eight	८	𑓘
Digit Nine	९	𑓙

Table 1. Devanagari-Tirhuta Character Mapping

[10] Proposal to Encode the Tirhuta Script in ISO/IEC 10646 https://www.unicode.org/L2/L2011/11175r-tirhuta.pdf

4.3 Implementation

A computer program in Javascript is developed to implement the Character Mapping. The tool is bidirectional, which can transliterate Tirhuta to Devanagari and Maithili alphabets of Devanagari to Trithuta. It swaps only those characters that are mapped between Devanagari and Tirhuta, while retaining all other characters such as Hindu-Arabic numerals, Roman characters, and punctuation marks in the output that are not mapped.

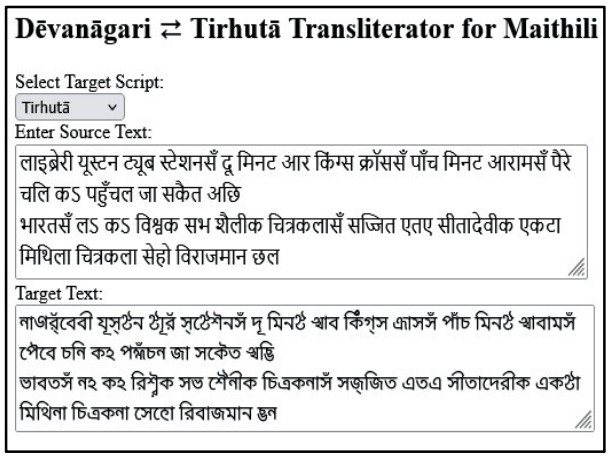

Fig. 1: Devanagari Tirhuta Transliterator

4.4 Testing

The transliterator program is tested on sentence list extracted from the LDC-IL Maithili text corpus (Ramamoorthy et al., 2019), which is in Devanagari. The Tirhuta sentences that contained untransliterated Devanagari characters were subjected to human validation. The Vowel Sign Chandra O (U+0949), a Devanagari character representing a borrowed sound is found in Tirhuta output. Since Devanagari 'Vowel Sign Chandra O' has no corresponding character in Tirhuta, It was then mapped to the 'Vowel Sign AA' as a surrogate as a normalisation measure. Some residual errors containing non-Maithili Devanagari characters due to typos in the testing corpus were found in the test output and were ignored.

4.4 Application

LDC-IL published the sentence aligned speech dataset for Automatic speech recognition systems, in both Devanagari (Shantanu Kumar et al., 2023) and Tirhuta (Dinesh Mishra et al, 2025) scripts. The initial transcriptions were created in Devanagari. Later it was transliterated to Tirhuta using a computer program that followed the aforementioned schema. Human validation of transliterated transcripts is performed to avoid the composition fallacy, to correct/debug any defects in the transliteration schema or computer program, and to adhere to the consortium's policy of human data validation prior to distribution.

4.4 The Resultant Dataset

The dataset is aligned, and segmented at sentence level using Praat software, the dataset contains Phonetic Transcription which are as it is pronounced in the audio, and Orthographic Transcriptions which is as the audio represented in standard orthography. The dataset is of 41:54:30 hours in duration, containing 21,412 audio segments collected from 300 native speakers. The dataset is available along with detail documentations to commercial users for a nominal cost and to non-commercial users free of charge. The dataset can be obtained from LDC-IL following the instructions provided in its data portal[11]. It can be used in the similar way to build the ASR systems as Devanagari transcript dataset released earlier.

5. Conclusion

LDC-IL, India's largest human-validated repository of Indian linguistic resources, took the initiative to publish language resources needed by the NLP research community in multi-script languages. For Maithili, a program has been developed to map Maithili in Devanagari script to Tirhuta script to support linguistic resources for ASR systems in Tirhuta.

References

Brezianu, Andrei, Vlad Spânu (2010), *The A to Z of Moldova*, Scarecrow Press, pp.110-111

Chaudhary, Radhakrishna, (2010) *A Survey of Maithili Literature*, Shruti Publications, ISBN : 978-93-80538-36-5, page 21

Choudhary, Narayan (ed.). 2019. *Linguistic Resources for AI/NLP in Indian Languages*. Central Institute of Indian Languages, Mysore. ISBN: 978-81-7343-295-8.

Choudhary, Narayan (ed.). 2021. *Compendium of Linguistic Resources in Indian Languages*. Central Institute of Indian Languages, Mysore. ISBN: 978-81-948885-6-7.

Dasgupta, J. (1970). *Language Conflict and National Development: Group Politics and National Language Policy in India*. India: University of California Press.

Elti di Rodeano, Sveva. (2019) *Digraphia: the Story of a Sociolinguistic Term*, Y. Haralambous (ed.), Proceedings of the "Graphemics in the 21st Century" Conference, Grapholinguistics Series, vol. 1, Fluxus Editions, Brest, 2019, p. 31-45 ISBN 978-2-9570549-0-9, doi: //10.36824/2018-graf-clti

Garfield, E. (1977). *Essays of an Information Scientist*. Vol:2, p.254-256, 1974-76 United States: ISI Press.

Howard, Douglas Arthur, *The History of Turkey*, Greenwood Press 2001, pp 98-99

Jayashree Nair, Riyaz Ahammed, and Anakha Shaji (2022), A study on Transliteration Techniques and Conventional Transliteration Schemes for Indian Languages, Edited by P. Karrupusamy, Valentina Emilia Balas, Yong Shi, Sustainable Communication Networks and Application: Proceedings of ICSCN 2021. (2022). Singapore: Springer Nature Singapore. pp 103-117, ISBN 9789811666056

Marzluf, Phillip, Myagmar Saruul-Erdene, (2019) "Mongolia Language education policy", The Routledge International Handbook of Language Education Policy in Asia, Edited by Andy Kirkpatrick, Anthony J. Liddicoat, Routledge, Oxon, doi 10.4324/9781315666235

[11] data.ldcil.org

Mithilesh K.J., (2017). Language Politics and Public Sphere in North India: Making of the Maithili Movement. India, OUP India, pp.85-90

Pollock, S. (2009). The Language of the Gods in the World of Men: Sanskrit, Culture, and Power in Premodern India. United Kingdom: University of California Press. Page 274

Rājendralala Mitra (1882), The Lalitavistara tranlsated from the orignal sanskrit, Calcuta, Asiatic Society of Bengal,

Ramamoorthy, L., Narayan Choudhary, Arun Kumar Singh & Dinesh Mishra (2019). A Gold Standard Maithili Raw Text Corpus. Central Institute of Indian Languages, Mysore. ISBN: 978-81-7343-236-1.

Rejitha K.S. (2018), Automatic Transliteration - A Malayalam Case Study in Sadakath: A Research Bulletin. Vol. VI, July. pp. 61-67. ISSN 2347-7644.

Rosen, R. (2017). Yoga FAQ: Almost Everything You Need to Know about Yoga-from Asanas to Yamas. United States: Shambhala.

Sebba, M. (2007). Spelling and Society: The Culture and Politics of Orthography Around the World. (n.p.): Cambridge University Press. pp81-82

Sohoni, P. (2016), Marathi of a Single Type: The Demise of the Modi Script, Modern Asian Studies, Cambridge University Press, vol. 51, issue 3, pp. 662-685, doi:10.1017/S0026749X15000542