DS-IASTConvert: An Automatic Script Converter between Devanagari and International Alphabet of Sanskrit Transliteration

[1]Subhash Chandra, [2]Vivek Kumar, [3]Anju

[1]Assistant Professor, [2]Doctoral Research Scholar, [3]Doctoral Research Scholar
[1]Department of Sanskrit,
[1]University of Delhi, New Delhi, India

Abstract: Sanskrit is considered as one of the oldest language in the world and belongs to Indo-Aryan language family. Ancient Sanskrit texts are typically written, interpreted and available in Devanagari Script (DS). Most of cases it is also being written in DS only in India. Not only the Sanskrit language, even most of the Indian languages also write in the DS. The DS is an abugida (alphasyllabary) used in India and Nepal. It writes from left to right, has a strong preference for symmetrical rounded shapes within squared outlines and is recognizable by a horizontal line that runs along the top of full letters. It is also originally very different from other Indic scripts. Deeply it is very similar except for angles and structural emphasis. Due to very strong knowledge system in the field of Science and Technology in Ancient India, western scholars had taken the interest to study the Sanskrit language. But they were unable to read and understand DS easily therefore, an effort were initiated by the scholars to map the DS in Roman Scripts. Finally, in 19th century the International Alphabet of Sanskrit Transliteration (IAST) was proposed by Charles Trevelyan, William Jones, Monier Monier-Williams and other scholars. It is presented and formalized in the Transliteration Committee of the Geneva Oriental Congress in September, 1894. The committee has approved IAST for use. This paper presents DS-IASTConvert: An Automatic Script Converter between Devanagari Script and International Alphabet of Sanskrit Transliteration. It is an online tool to convert Unicode DS text to IAST and IAST to Unicode DS text. The transliteration is based on an algorithm with the set of rules developed and stored in text file in tabular format. The innovation of this tool is the speed of conversion, online 24*7 availability, simplicity, easy to use, user friendly, automatically detection of script and also helps to learn the conversion scheme for manual conversion. This tool is very useful for Sanskrit researchers. It is available online for public use at no cost on http://cl.sanskrit.du.ac.in/ transliteration.

IndexTerms: **Devanagari to IAST, IAST, Devanagari, Transliteration, Diacritical Marks, DS-IASTConvert, Devanagari to Roman.**

I. BACKGROUND

Transliteration is a system that convey the same or as nearly as possible by means of one set of letters or characters the pronunciation of the words in languages written and printed in a totally different script (Karakos, 2003). Most of major Indian languages are written DS, it includes Hindi, Sanskrit, Marathi, Konkani, Nepali, Maithili, Sindhi, Bodo, Dogri, Santhali, Bhojapuri, Awadhi etc. (Bright, 1996). There are several methods of transliterations from DS to the Roman Script (RS). The process of transliterations from DS to theRS is known as Romanization of DS. It shares similarities, although no single system of transliteration has emerged as the standard (Sharma, 1972). IAST is a subset of the ISO 15919 standard, used for the transliteration of Sanskrit, Prakrit and Pāḷi into Roman script using diacritics. It is widely used standard for Romanization of Sanskrit, Prakrit and Pāḷi. It uses diacritics to disambiguate phonetically similar but not identical Sanskrit glyphs. Dental and retroflex consonants are disambiguated with an under dot symbol. An important feature of IAST is that it is reversible without any loss. IAST transliteration may be converted to DS without any ambiguity and with correct DS spelling. Many Unicode fonts fully support IAST display and printing. IAST is a transliteration scheme that allows the lossless Romanization of Indic scripts as employed by Sanskrit and related Indic languages.

IAST makes it possible for the reader to read the Indic text unambiguously, exactly as if it were in the original Indic script. It is this faithfulness to the original scripts that accounts for its continuing popularity amongst scholars. University scholars commonly use IAST for publications that cite textual material in Sanskrit, Pāḷi and other classical Indian languages. IAST is also used for major e-text repositories such as SARIT, Muktabodha, GRETIL, and sanskritdocuments.org. The IAST scheme represents more than a century of scholarly usage in books and journals on classical Indian studies. By contrast, the ISO 15919 standard for transliterating Indic scripts emerged in 2001 from the standards and library worlds; it includes solutions to problems such as representing Old Indo-Aryan and New Indo-Aryan languages side by side in library catalogues, etc. In IAST letters are modified with diacritics: long vowels are marked with an over line, vocalic (syllabic) consonants and retroflexes have an under dot.

A list of the letters in DS and IAST with phonetic values in International Phonetic Alphabet (IPA), an alphabetic system of phonetic notation based primarily on the Latin alphabet is shown in table 1. The list is valid for Sanskrit but can be fit for Hindi and other modern languages that use Devanagari script, with few phonological changes. This tool allows you to easily change the transliteration of single words or even entire texts. This is particularly very helpful when you want to produce Devanagari script or Roman script with diacritical marks. With the copy & paste function, you can then enter the generated words or texts in a word processing program of your choice.

II. OTHER ROMAN TRANSLITERATION METHODS

There are several other transliteration schemes from Devanāgarī scripts to the Roman script are available and being used for transliteration from DS to RS. Two major schemes are very popular for Indic script transliteration other than IAST. First schemes with diacritics that uses diacritics to map the letters. And other scheme is Indian languages TRANSliteration (ITRANS) which is an American Standard Code for Information Interchange (ASCII) transliteration scheme for Indic scripts and widely uses for DS. ASCII is a character encoding standard for electronic communication. ASCII codes represent text in computers, telecommunications equipment, and other devices. Most modern character-encoding schemes are based on ASCII, although they support many additional

characters. Under the schemes with diacritics includes two schemes first the National Library at Kolkata Romanization and second ISO 15919 schemes. And under ASCII schemes there are Harvard-Kyoto (HK), ITRANS scheme, Velthuis, Sanskrit Library Phonetic (SLP1), WX Notation are included.

Table 1 Mapping Alphabet List in DS and IAST

Vowels				
DS	**IAST**	**DS**	**IAST**	
अ	a	ऌ	ḹ	
आ	ā	ए	e	
इ	i	ऐ	ai	
ई	ī	ओ	o	
उ	u	औ	au	
ऊ	ū	अं	ṃ	
ऋ	ṛ	अः	ḥ	
ॠ	ṝ	ऽ	'	
ऌ	ḷ			

Consonants									
Velars		**Palatals**		**Retroflexes**		**Dentals**		**Labials**	
DS	**IAST**	**DS**	**IAST**	**DS**	**IAST**	**DS**	**IAST**	**DS**	**IAST**
क	ka	च	ca	ट	ṭa	त	ta	प	pa
ख	kha	छ	cha	ठ	ṭha	थ	tha	फ	pha
ग	ga	ज	ja	ड	ḍa	द	da	ब	ba
घ	gha	झ	jha	ढ	ḍha	ध	dha	भ	bha
ङ	ṅa	ञ	ña	ण	ṇa	न	na	म	ma
ह	ha	य	ya	र	Ra	ल	la	व	va
		श	śa	ष	ṣa	स	sa		

The National Library at Kolkata Romanization, intended for the Romanization of all Indic scripts, is an extension of IAST. It differs from IAST in the use of the symbols ē and ō for ए and ओ. A List of DS to National Library at Kolkata Romanization scheme is shown in table 2. ISO 15919 is a standard transliteration convention not only for Devanagari but for all South-Asian languages was codified in the ISO 15919 standard of 2001, providing the basis for modern digital libraries that conform to International Organization for Standardization (ISO) norms (Stone, 1998). ISO 15919 defines the common Unicode basis for Roman transliteration of South-Asian texts in a wide variety of languages/scripts. ISO 15919 uses diacritics to map the much larger set of Brahmic graphemes to the Latin script. The Devanagari-specific portion is very near to IAST. A List of DS to ISO 15919 transliteration scheme is shown in table 2.

Harvard-Kyoto (HK) is very similar to IAST. It does not contain any of the diacritic marks. Instead of diacritics, Harvard-Kyoto uses capital letters to represent the long vowel. The list of DS to HK transliteration scheme is shown in table 2. The Indian languages TRANSliteration (ITRANS) is an ASCII transliteration scheme for Indic scripts, particularly for DS. ITRANS scheme is an extension of Harvard-Kyoto (Chopde, 2009). ITRANS transliteration scheme for DS is shown table 2. The Velthuis system of transliteration is an ASCII transliteration scheme for the Sanskrit language from and to the Devanagari script. It was developed by Frans Velthuis, a scholar living in Groningen, Netherlands, who created a popular, high-quality software package in Latex for typesetting Devanāgarī. It does not use capital letters as compared to Harvard-Kyoto or ITRANS schemes (Velthuis & Pandey, 1991 and Pandey, 1998). The list of DS to Velthuis transliteration scheme is shown in table 2. The Sanskrit Library Phonetic Basic encoding scheme (SLP1) is an ASCII transliteration scheme for the Sanskrit language from and to the Devanagari script. It always uses a single character (Scharf, & Hyman, 2009). The list of DS to SLP transliteration scheme is shown in table 2.

WX notation is a transliteration scheme for representing Indian languages in ASCII is another scheme. This scheme originated at IIT Kanpur for computational processing of Indian languages, and is widely used among the natural language processing (NLP) community in India (Bharati, Chaitanya, Sangal & Ramakrishnamacharyulu, 1995). The list of DS to WX notation transliteration scheme is shown in table 2.

Table 2 Other Roman Transliteration Methods

Vowels							
DS	**IAST**	**ISO15919**	**HK**	**ITRANS**	**Velthuis**	**SLP1**	**WX**
अ	a	A	a	a	a	a	a
आ	ā	Ā	A	A/aa	aa	A	A
इ	i	I	i	i	i	i	i
ई	ī	ī	I	I/ii	ii	I	I
उ	u	u	u	u	u	u	u
ऊ	ū	ū	U	U/uu	uu	U	U
ए	e	ē	e	e	e	e	e

ऐ	ai	ai	ai	ai	ai	E	E
ओ	o	ō	o	o	o	o	o
औ	au	au	au	au	au	O	O
ऋ	ṛ	r̥	R	RRi/R^i	.r	f	q
ॠ	ṝ	r̥̄	RR	RRI/R^I	.rr	F	Q
ऌ	ḷ	l̥	lR	LLi/L^i	.l	x	L
ॡ	ḹ	l̥̄	lRR	LLI/L^I	.ll	X	LY
अं	ṃ	ṁ	M	M/.n/.m	.m	M	M
अः	ḥ	ḥ	H	H	.h	H	H
अँ		m̐		.N		~	az
ऽ	'	’	'	.a	.a	'	
Consonants							
DS	**IAST**	**ISO 15919**	**HK**	**ITRANS**	**Velthuis**	**SLP1**	**WX**
क	ka	ka	ka	ka	ka	ka	ka
ख	kha	kha	kha	kha	kha	Ka	Ka
ग	ga	ga	ga	ga	ga	ga	ga
घ	gha	gha	gha	gha	gha	Ga	Ga
ङ	ṅa	ṅa	Ga	~Na	"na	Na	fa
च	ca	ca	ca	cha	ca	ca	ca
छ	cha	cha	cha	Cha	cha	Ca	Ca
ज	ja	ja	ja	ja	ja	ja	ja
झ	jha	jha	jha	jha	jha	Ja	Ja
ञ	ña	ña	Ja	~na	~na	Ya	Fa
ट	ṭa	ṭa	Ta	Ta	.ta	wa	ta
ठ	ṭha	ṭha	Tha	Tha	.tha	Wa	Ta
ड	ḍa	ḍa	Da	Da	.da	qa	da
ढ	ḍha	ḍha	Dha	Dha	.dha	Qa	Da
ण	ṇa	ṇa	Na	Na	.na	Ra	Na
त	ta	ta	ta	ta	ta	ta	wa
थ	tha	tha	tha	tha	tha	Ta	Wa
द	da	da	da	da	da	da	xa
ध	dha	dha	dha	dha	dha	Da	Xa
न	na	na	na	na	na	na	na
प	pa	pa	pa	pa	pa	pa	pa
फ	pha	pha	pha	pha	pha	Pa	Pa
ब	ba	ba	ba	ba	ba	ba	ba
भ	bha	bha	bha	bha	bha	Ba	Ba
म	ma	ma	ma	ma	ma	ma	ma
य	ya	ya	ya	ya	ya	ya	ya
र	ra	ra	ra	ra	ra	ra	ra
ल	la	la	la	la	la	la	la
व	va	Va	va	va/wa	va	va	va
श	śa	Śa	za	sha	"sa	Sa	Sa
ष	ṣa	ṣa	Sa	Sha	.sa	za	Ra
स	sa	Sa	sa	sa	sa	sa	sa
ह	ha	Ha	ha	ha	ha	ha	ha
क्ष	kṣa	kSa	kSa/kSha/xa	k.sa	kza	kRa	
त्र	tra	Tra	tra	tra	tra	wra	
ज्ञ	jña	jJa	GYa/j~na	j~na	jYa	jFa	
श्र	śra	Zra	shra	"sra	Sra	Sra	

III. Features of DS-IASTConvert

DS-IASTConvert uses the method of mapping from Devanagari system of writing to IAST system based on phonetic similarity especially for Sanskrit Language.

Figure 1: Screen Shot of DS-IASTConvert User Interface

Through this tool, user need to type or copy/paste of Sanskrit Devanagari text in Unicode or IAST and after clicking on a button text will be converted into IAST or DS. The system is very useful for Sanskrit researchers to automatically convert DS text into IAST or IAST text in DS. System generates the result with the help of associated data files. The sample of the rule file is shown in Table 3. Special features of DS-IASTConvert are listed below:

3.1 Auto Script Detection

Auto Script Detection is one of the useful and unique features of the system. DS-IASTConvert automatically detects the writing script. The system uses a script detection module and rules to detect the script. When user gives the input in Devanagari Script, System automatically converts in IAST and if user give the input in Roman script then system converts the input into Devanagari script.

3.2 Showing the Conversion Rules

Second most unique features of the DS-IASTConvert is that the system also shows mapping rules converting the text to IAST to DS or DS to IAST. This feature is very helpful to learn the conversion rules and also helpful in reading the script. System has over mouse function where user can take the curser of the mouse on the converted text, the conversion rules appears in new popup windows. The tools is also useful for self-learning of transliteration.

3.3 User Friendly

Third most unique features of the DS-IASTConvert is that the system is very easy. User can just copy and paste any text written in DS or IAST to convert the script. And after converting, the converted text can be pasted and used in any text editors.

Table 3: Map Table for Unicode Devanagari <> IAST

maprtu:0=०,1=१,3=३,2=२,5=५,4=४,7=७,6=६,8=८,9=९,
maprtu:k=क्,kh=ख्,g=ग्,gh=घ्,ṅ=ङ्,'=ऽ,
maprtu:c=च्,ch=छ्,j=ज्,jh=झ्,ñ=ञ्,
maprtu:t=त्,th=थ्,d=द्,dh=ध्,n=न्,
maprtu:ṭ=ट्,ṭh=ठ्,ḍ=ड्,ḍh=ढ्,ṇ=ण्,
maprtu:p=प्,ph=फ्,b=ब्,bh=भ्,m=म्,
maprtu:y=य्,r=र्,l=ल्,v=व्,ś=श्,ṣ=ष्,s=स्,h=ह्,.=।,
maprtu:a=अ,ā=आ,i=इ,ī=ई,u=उ,ū=ऊ,ṛ=ऋ,ṝ=ॠ,ḷ=ऌ,ḹ=ॡ,e=ए,ai=ऐ,o=ओ,au=औ,aṃ=अं,aḥ=अः,
maprtu:ā=ा,i=ि,ī=ी,u=ु,ū=ू,ṛ=ृ,ṝ=ॄ,ḷ=ॢ,ḹ=ॣ,e=े,ai=ै,o=ो,au=ौ,ṃ=ं,ḥ=ः,
maprtu:ka=क,kha=ख,ga=ग,gha=घ,ṅa=ङ,
maprtu:ca=च,cha=छ,ja=ज,jha=झ,ña=ञ,
maprtu:ta=त,tha=थ,da=द,dha=ध,na=न,
maprtu:ṭa=ट,ṭha=ठ,ḍa=ड,ḍha=ढ,ṇa=ण,
maprtu:pa=प,pha=फ,ba=ब,bha=भ,ma=म,
maprtu:ya=य,ra=र,la=ल,va=व,śa=श,ṣa=ष,sa=स,ha=ह,.=।,ā=aā

IV. Conclusions

Sanskrit has an inventory of thirteen vowels and thirty-four consonants: all of which can be unambiguously encoded into Devanāgarī. As of recent history, Sanskrit is also often Romanized for simpler typesetting and wider scholarship. The dominant Romanization scheme is the International Alphabet of Sanskrit Transcription (IAST), which evolved out of various earlier Romanization schemes. Converting any text written in DS in to IAST form is very challenging and time consuming task because few letters cannot be typed directly by the keyboard. User needs to insert diacritics from the symbols list. Therefore, DS-IASTConvert: An Automatic Script Converter between Devanagari and International Alphabet of Sanskrit Transliteration has been developed and presented here. The system is available at http://cl.sanskrit.du.ac.in for public use.

V. Future Direction

India has 22 official languages writing in 12 scripts. Script convertor can fill the gap of learning and understanding the data and knowledge. The major objective of the authors is develop transliteration system between Indian Language's scripts to Indian Language's scripts. Therefore, the rules and module can be utilized to develop the system. The system is also very useful for information mining form the text available in various scripts.

References

1. Bharati, A., Chaitanya, V., Sangal, R., & Ramakrishnamacharyulu, K. V. (1995). *Natural language processing: a Paninian perspective* (pp. 65-106). New Delhi: Prentice-Hall of India.
2. Bright, William. "The devanagari script." *The world's writing systems* (1996): 384-390.
3. Chopde, A. (2009). Itrans Indian language transliteration package version 5.2 source.
4. *Daya Nand Sharma, Transliteration into Roman and Devanāgarī of the Languages of the Indian Group, Survey of India, (1972).*
5. Daya Nand Sharma, Transliteration into Roman and Devanāgarī of the languages of the Indian group, Survey of India, 1972.
6. Dhore, Manikrao, Shantanu Dixit, and Ruchi Dhore. "Hindi and Marathi to English NE Transliteration Tool using Phonology and Stress Analysis." *Proceedings of COLING 2012: Demonstration Papers* (2012): 111-118.
7. Shashir's Notes." *Modern Transcription of Sanskrit*, shashir.autodidactus.org/shashir_umich/sanskrit_transcription.html.
8. Karakos, Alexandros. "Greeklish: An experimental interface for automatic transliteration." *Journal of the American Society for Information Science and Technology* 54.11 (2003): 1069-1074.
9. Pandey, A. (1998). Romanized Indic and LATEX.
10. Scharf, P., & Hyman, M. (2009). Linguistic issues in encoding Sanskrit. *Motilal Banarsidass, Delhi*.
11. Stone, A. (1998). ISO Committee Draft 15919: Transliteration of Devanagari and Related Scripts into Latin Characters.
12. VELTHUIS, F., & Pandey, A. (1991). Devanagari for TEX.