ISSN: 2456 - 3145
International Journal of Interdisciplinary Research in Arts and Humanities
Impact Factor 5.225 / www.dkirf.org / dvpublication@gmail.com
Volume 3, Conference World Special Issue 1, 2018

LANGUAGE TECHNOLOGY AND NATURAL LANGUAGE PROCESSING: A DEVELOPMENTAL STUDY OF ROMAN (WX) NOTATION OF LADAKHI

Irfan ul Salam*, Humaira Khan & Afreen Nazir*****
Department of Linguistics, University of Kashmir, Srinagar, Jammu and Kashmir

Cite This Article: Irfan ul Salam, Humaira Khan & Afreen Nazir, "Language Technology and Natural Language Processing: A Developmental Study of Roman (WX) Notation of Ladakhi", International Journal of Interdisciplinary Research in Arts and Humanities, Volume 3, Conference World Special Issue 1, Page Number 51-55, 2018.

Abstract:

In the present world of Digitalization, Language Technology has become a vital medium to preserve, populate and modernize one's language and culture simultaneously. This is known that computers can understand the Roman Script, in processing and editing, it has become a very important task to develop a standard Roman Notation for non-roman languages to be able to use for various NLP (Natural Language Processing) applications. This motivates to develop such a notation called as WX-Notation schema for Ladakhi language. Ladakhi is a Tibetic language, classified under Tibeto-Burman language family. It is spoken in Leh district of Jammu & Kashmir state. Ladakhi uses Tibetan script called 'yige' or Western Tibetan Archaic. WX-Notation is a transliteration scheme for representing Indian languages in ASCII (American Standard Code for Information Interchange,) for computational processing of Indian languages, and is widely used among the natural language processing (NLP) community in India. this transliteration scheme is productive as every consonant and every vowel has a single mapping into Roman. Hence it is a prefix code, advantageous from computation point of view. It is high time for Ladakhi language, to come into the world of language computation in order to sustain the linguistic and cultural aspects. In this process there are other roman notations available, but how scientifically they are, raises some important questions.

- Can these notations be case sensitive?
- Are they anywhere near to the writing style of Ladakhi script?
- How complex are these to enter in computers?
- How easy they are for a new learner?
- Do we have technology to convert from Roman to the Target Script?

In this research paper, an attempt is made to answer these questions and a usable WX notation scheme is discussed. A demonstration of the converter will also be discussed.

Key Words: Computational Linguistics, Ladakhi Script, Language Technology, NLP & WX- Notation

Introduction:

Technology today, is very much part of everything that we do and our world depends on it for greater efficiency and reliability of our existence and continuity of our survival. In the present world, Language Technology has become a vital medium to preserve, populate and modernize once language and culture simultaneously. To be able to use all these technologies, Digitization is crucial as it is the prerequisite for data processing, storage and transmission. Language Technology is often called Human Language Technology (HLT) consists of language processing, popularly called as the field of NLP (Natural Language Processing), also known as Computational Linguistics. Therefore, when language comes in contact with Information Technology, it needs to be organized so that it can be processed by computational means. This often requires broad knowledge not only about Linguistics, but also Computer Science and other related fields. Everyone today as a matter of fact surrounded with computational applications related to Machine translation, Natural language interface, Document processing and information retrieval, Grammar and style checking, Computer-Assisted language learning and many more. All these tools are the bi-products of Computational Linguistics or NLP. Among the most Tibeto-Burnan Languages, Ladakhi is less explored from Computational linguistics perspective.

Unicode:

Unicode is a character encoding standard that has worldwide acceptance. For example, Microsoft softwares as we know, uses Unicode at its core. We understand the binary nature of computer processing and it just deals with numbers. They store letters and other characteristics by assigning a number for each one. It is rather important to know that, before Unicode was invented, there were hundreds of encoding systems for assigning these numbers. No encoding technique could compensate enough characters and that became the most challenging disadvantage and this has been the problem any computational language oriented task would come across. Thus, in order to overcome this limitation and to avoid font conflicts and make it compatible across the globe, Unicode resolved all these issues all together. Unicode is now a computing industry standard for consistent encoding, representation and handling of text expressed in most of the world's writing systems. The latest version of Unicode contains a repertoire of 136,755 characters covering 139 modern and historic scripts and it is handled by the Unicode Consortium.

Roman Notation/WX Notation:

WX-Notation is a transliteration scheme for representing Indian languages in ASCII (American Standard Code for Information Interchange). 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. The salient features of this transliteration scheme are: Every consonant and every vowel has a single mapping into Roman. Hence it is a prefix code, advantageous from computation point of view. Typically the small case letters are used for un-aspirated consonants and short vowels while the capital case letters are used for aspirated consonants and long vowels. While the retroflexed voiceless and voiced consonants are mapped to 't, T, d and D', the dentals are mapped to 'w, W, x and X'. Hence the name of the scheme "WX", referring to the idiosyncratic mapping and which makes WX- Notation case-sensitive in nature. Ubuntu Linux provides a keyboard support for WX notation.

Since it is understood that WX-Notaion is an effective transliteration tool, very compatible for computers and can understand the Roman Script easily, in processing and editing, it has become a very important task to develop a standard Roman

ISSN: 2456 - 3145
International Journal of Interdisciplinary Research in Arts and Humanities
Impact Factor 5.225 / www.dkirf.org / dvpublication@gmail.com
Volume 3, Conference World Special Issue 1, 2018

Notation for non roman languages to be able to use for various NLP applications and development of Tools like Morphological Analyzer, POS Tagger, Chunker etc. This motivates us to develop such a notation for *Ladakhi* as well. *WX-Notation* schema which is already used by many Indian Languages is also adopted for *Ladakhi*, which uses Western Tibetan Archaic Script also called as *'yige'*.

Ladakhi:

Ladakhi is classified under Tibeto-Burman language family. *Ladakhi*, also called *Bhoti* or *Bodhi*, is a Tibetic language spoken in the *Ladakh* region of India. It is the predominant language in the district of *Leh* of the *Jammu & Kashmir* state. Though a member of the Tibetic family, *Ladakhi* is not mutually intelligible with Standard Tibetan.

Ladakhi has approximately 200,000 speakers in India, and perhaps 12,000 speakers in the Tibet Autonomous Region of China, mostly in the *Qiangtang* region. *Ladakhi* has several dialects: *Lehskat* after Leh, where it is spoken; *Shamskat*, spoken in the northwest of Leh; *Stotskat*, spoken in the Indus valley and which is tonal unlike the others; *Nubra*, spoken in the north of *Leh*; *Purigi/Balti* spoken in the Kargil district. The significant difference in the dialects remain in the tone or way of speaking. The varieties spoken in Upper *Ladakh* and *Zangskar* have many features of *Ladakhi* and also western dialects of Central Tibetan.

It is high time that *Ladakhi* should be documented, preserved and digitized, so as to make it available to come into the world of language computation in order to sustain their linguistic and cultural heritage. Such research works will result into various scope for revitalizing the language all together and ensure its long sustenance. In this process there are other roman notations available, but how scientifically they are, raises some important questions.

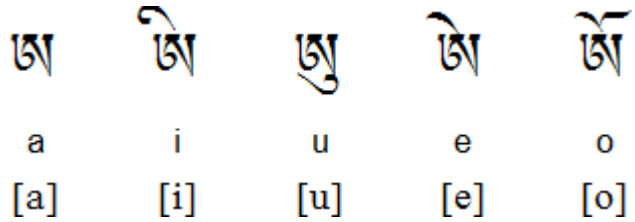
- Can these notations be case sensitive?
  - Yes, WX-Notation is case sensitive as it provides greater scope to encompass as many characters as possible.
- Are they anywhere near to the writing style of Ladakhi script?
  - Since WX-Notations have been specially designed for Indian languages and the scripts pertaining to Bhrami. The Tibetan script itself has been derived from Bhrami in the 7th century AD. Therefore it is probable to face lesser issues as expected.
- How complex are these to enter in computers?
  - Unicode system provides subtle processes to integrate the script information into computer readable format, which is a time consuming process, but achievable.
- How easy they are for a new learner?
  - It is the most convenient alternative to learn and understand a language and it very productive in nature.
- Do we have technology to convert from Roman to the Target Script?
  - WX-Convertor module is the tool that can convert a UTF (unicode text format) into WX (Roman Text) format and vice-verse.

Writing System of Ladakhi:

Ladakhi has total of 39 characters in its writing system and the script used is Western Tibetan Archiac.

Consonants:

Vowels:



Some Writing Features of Ladakhi Script:

- Type of writing system: syllabic alphabet or abugida. Each letter has an inherent vowel /a/. Other vowels can be indicated using a variety of diacritics which appear above or below the main letter.

ISSN: 2456 - 3145
International Journal of Interdisciplinary Research in Arts and Humanities
Impact Factor 5.225 / www.dkirf.org / dvpublication@gmail.com
Volume 3, Conference World Special Issue 1, 2018

- Direction of writing: left to right in horizontal lines.
- Syllables are separated by a dot.
- Consonant clusters are written with special conjunct letters.

Transliteration Schema:

S.No	Glyph	Code	IPA	Roman	WX	Word
1.	ཀ	0f40		ka	ka	/kalak/ Mud
2.	ཁ	0f41		kha	Ka	/Khar/ Palace
3.	ག	0f42		ga	ga	/golak/ Bald
4.	ང	0f44		nga	fa	/Ngamo/ Morning
5.	ཅ	0f45		ca	ca	/chipa/ Sparrow
6.	ཆ	0f46		cha	Ca	/chang/ Alcohol
7.	ཇ	0f47		ja	ja	/jiksten/ World
8.	ཉ	0f49		nya	Fa	/nya/ Fish
9.	ཏ	0f4a		tta	ta	/Tangmo/ Cold
10.	ཐ	0f4b		ttha	Ta	/Thagu/ kettle
11.	ད	0f4c		dda	da	/damba/ Cheek
12.	ཏ	0f4f		ta	wa	/tutu/ Throat
13.	ཐ	0f50		tha	Wa	/Thalba/ Dust
14.	ད	0f51		da	xa	/daman/ Musical instrument
15.	ན	0f53		na	na	/Namza/ Weather
16.	པ	0f54		pa	pa	/Pabu/ Shoe
17.	ཕ	0f55		pha	Pa	/phey/ Wheat flour
18.	བ	0f56		ba	ba	/Balang/ Cow
19.	མ	0f58		ma	ma	/marpo/ Red
20.	ཙ	0f59		tsa	taz	/Tsepo/ Basket
21.	ཚ	0f5a		tsha	Taz	/tsha/ Salt
22.	ཛ	0f5b		dza	daz	/Dzago/ Friend
23.	ཝ	0f5d		wa	va	/watse/ Fox
24.	ཞ	0f5e		zha	jaz	/Jara/ Blind
25.	ཟ	0f5f		za	Jaz	/zumo/ Sickness
26.	འ	0f60		-a	a	/thap/ Stove
27.	ཡ	0f61		ya	ya	/yar/ Summer
28.	ར	0f62		ra	ra	/ri/ Mountain
29.	ལ	0f63		la	la	/lam/ Road
30.	ཤ	0f64		sha	Ra	/shing/ Wood
31.	ས	0f65		ssa	sa	/sa/ Ground
32.	ཧ	0f67		ha	ha	/handang/ Retard
33.	ཨ	0f68		a	A	/Ama/ Mother

ISSN: 2456 - 3145
International Journal of Interdisciplinary Research in Arts and Humanities
Impact Factor 5.225 / www.dkirf.org / dvpublication@gmail.com
Volume 3, Conference World Special Issue 1, 2018

34.	ར	0f6c		rra	rY	/Ruspa/ Bone
35.	ཀ	0f6b		kka	kY	/Kalak/
36.	ི	0f72		kigu	i	/Rigu/
37.	ེ	0f7a		denbu	e	/Stebo/
38.	ོ	0f7c		naro	o	/Kore/
39.	ུ	0f74		Zhapshku	u	/thugu/

Design of Transliteration Tool:

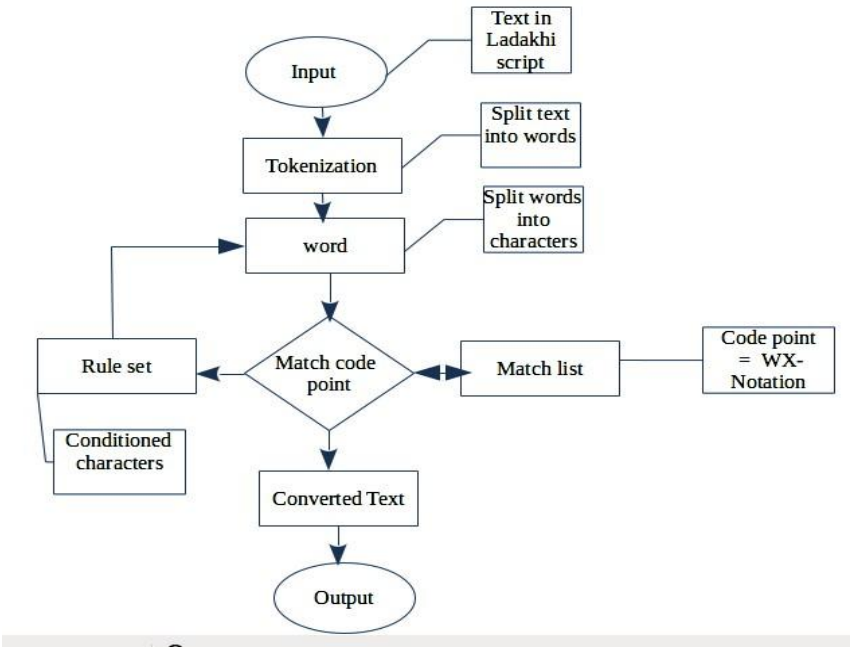


Figure 1: Above computational model describing the working of WX-Converter

Issues in Unicode Chart 10.0:

- Some sounds are not captured in the unicode chart.

Conclusion:

It is apparent that developing such tools will not only facilitate linguistic output incorporated with technology, but it will also lead to new research opportunities in various other domains of language studies. WX-Convertor is a tool of utmost significance as it opens the doorway to technology and allows us the following:

- Web compatibility
- Multilingual application
- Simplified application development
- Inter-operability
- Easy to deploy worldwide
- Easy to provide shared acess
- Ease of migration of existing code

References:

1. Vishal Goyal and Gurpreet Singh Lehal. 2009. Hindi-punjabi machine transliteration system (for machine translation system). George Ronchi Foundation Journal, Italy, 64(1): 2009.
2. Rohit Gupta, Pulkit Goyal, Allahabad IIIT, and Sapan Diwakar. 2010. Transliteration among indian languages using wx notation. Semantic Approaches in Natural Language Processing, page 147.
3. Kanika Gupta, Monojit Choudhury, and Kalika Bali. 2012. Mining hindi-english transliteration pairs from online hindi lyrics. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), pages 23–25.
4. Jagadeesh Jagarlamudi and a Kumaran. 2008. Cross Lingual Information Retrieval System for Indian Languages. In Advances in Multilingual and Multi modal Information Retrieval, pages 80–87. Springer.
5. Girish Nath Jha. 2010. The tdil program and the indian language corpora initiative (ilci). In Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010). European Language Resources Association (ELRA)
6. Kevin Knight and Jonathan Graehl. 1998. Machine transliteration. Computational Linguistics, 24(4):599–612.

7. Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology Volume 1, pages 48–54. Association for Computational Linguistics.
8. Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pages 177–180. Association for Computational Linguistics.
9. Bhadriraju Krishnamurti. 2003. The Dravidian Languages. Cambridge University Press.
10. Amba Kulkarni, Rahmat Yousufzai, and Pervez Ahmed Azmi. 2012. Urdu-hindi-urdu machine translation: Some problems. Health, 666:99–1.
11. Jin-Shea Kuo and Ying-Kuei Yang. 2004. Generating paired transliterated-cognates using multiple pronunciation characteristics from Web Corpora. In PACLIC, volume 18, pages 275–282.
12. Gurpreet S Lehal and Tejinder S Saini. 2010. A hindi to urdu transliteration system. In Proceedings of ICON-2010: 8th International Conference on Natural Language Processing, Kharagpur.
13. Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. In Soviet physics doklady, volume 10, page 707.
14. Abbas Malik, Laurent Besacier, Christian Boitet, and Pushpak Bhattacharyya. 2009. A hybrid model for urdu hindi transliteration. In Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration, pages 177–185. Association for Computational Linguistics.
15. David Matthews. 2007. Machine transliteration of proper names. Master's Thesis, University of Edinburgh, Edinburgh, United Kingdom.
16. Boris New, Veronica Ara´ujo, and Thierry Nazzi. 2008. Differential processing of consonants and vowels in lexical access through reading. Psychological Science, 19(12):1223–1227.
17. Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, pages 160–167. Association for Computational Linguistics.
18. Vincent Pagel, Kevin Lenzo, and Alan Black. 1998. Letter to sound rules for accented lexicon compression. Ar Xiv preprint cmp-lg/9808010.
19. Kalyani Patel and Jyoti Pareek. 2009. Gh-map-rule based token mapping for translation between sibling language pair: Gujarati–hindi. In Proceedings of International Conference on Natural Language Processing.
20. Vladimir Pervouchine, Haizhou Li, and Bo Lin. 2009. Transliteration alignment. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1- pages 136–144. Association for Computational Linguistics.
21. Alok Rai. 2001. Hindi nationalism, volume 13. Orient Blackswan. R Russell and M Odell. 1918. Soundex. US Patent, 1.
22. Hassan Sajjad, Alexander Fraser, and Helmut Schmid. 2012. A statistical model for unsupervised and semi-supervised transliteration mining. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume1, pages 469–477. Association for Computational Linguistics.
23. Muhammad Adeel Zahid, Naveed Iqbal Rao, and Adil Masood Siddiqui.2010. English to Urdu transliteration: An application of Soundex algorithm. In Information and Emerging Technologies (ICIET), 2010 International Conference on, pages 1–5. IEEE.
24. Min Zhang, Xiangyu Duan, Vladimir Pervouchine, and Haizhou Li. 2010. Machine transliteration: Leveraging on third languages. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 1444–1452. Association for Computational Linguistics.
25. Min Zhang, Haizhou Li, Ming Liu, and A Kumaran. 2012. Whitepaper of news 2012 shared task on machine transliteration. In Proceedings of the 4th Named Entity Workshop, pages 1–9. Association for Computational Linguistics.Wikipedia : https://en.wikipedia.org/wiki/Ladakhi_language