Building a Cross Script Kashmiri Converter: Issues and Solutions

Article · January 2010

CITATIONS
0

READS
301

3 authors:

Aadil Lawaye
Baba Ghulam Shah Badshah University
16 PUBLICATIONS **42** CITATIONS
SEE PROFILE

Aadil Kak
University of Kashmir
5 PUBLICATIONS **36** CITATIONS
SEE PROFILE

Nali Mehdi
SKEMA Business School
36 PUBLICATIONS **388** CITATIONS
SEE PROFILE

Building a Cross Script Kashmiri Converter: Issues and Solutions

Aadil Amin Kak, Nazima Mehdi and Aadil Ahmad Lawaye

aadilaminkak@yahoo.com; nazimamehdi@yahoo.com; aadillawaye@yahoo.com

University of Kashmir

Abstract

Kashmiri is a new entrant in the realm of Natural Language Processing. Efforts in this direction are only now taking place by developing different NLP tools. The paper in question talks about the development of a Persio-Arabic Devanagari converter. Here the main focus is on handling some issues which were faced while developing the converter.

1.0 Introduction

Kashmiri language is primarily spoken in the Kashmir province and some parts of the Jammu province of the state of Jammu and Kashmir State and by migrant populations in the rest of India and abroad. Various scripts such as Sharda, Devanagari, Roman, and Perso- Arabic have been used for Kashmiri. The earliest script used for writing Kashmiri is the Sharda script which is now only used by some Kashmiri pundits for writing horoscopes and a large number of Sanskrit literary works, and old Kashmiri works were written in this script. Presently, the official script of Kashmiri is the modified Persio-Arabic script with additional diacritic marks to represent Kashmiri specific sounds. Alternative scripts like the modified Devanagari script with additional diacritic marks is used by writers and researchers for representing the Kashmiri text related to language, literature, and culture in Hindi. In addition to these scripts Roman script is also used for writing Kashmiri. Regarding the modified Persio-Arabic script, it is written from right to left. It has two modes: *nasakh* or the type script, and *nastalikh*, the handwritten version.

1.1 Need for Developing a Converter

No manual work is needed i.e. people who are unaware of Persio-Arabic or the Devanagari script can automatically convert the text from Persio-Arabic to Devanagari and vice versa. It is very useful for mutually exclusive people who can use only one script and are unaware of the other.

The converter was built by developing different rules on the basis of character combinations and character positions (Initial, medial and Final).

1.2 Issues and solutions

There is no one to one mapping in both the scripts i.e. there are certain letters in Persio-Arabic which cannot be replaced by Devanagari equivalents, examples are given below.

In Kashmiri letter "و" when comes in combination with different letters like ا , و , ◌ُ is pronounced differently but in Devanagari it remains same as "व" with all the combinations. For example, if we have a word like اوٚل meaning 'nest' in Kashmiri it is not represented as अवॅल but as ओल. Thus here the problem was "و" cannot be simply replaced by "व" as it appears but depending upon the combination letter "و"gets replaced by ओ which was handled by building different rules.

Another example is when "و" comes in combination with ◌ُ in a word like کوٗل meaning 'dumb' in English, then "و" is replaced by "ो" not by "व" and the word in devanagari will be "कोल" not कवॅल.

From the examples above it is clear that all the Devanagari letters does not have one to one correspondence with Persio-Arabic characters. So in order to solve such an issue different rules were built on the basis of the character position and the combination of the characters was also taken into consideration.

1.2.1 Rules for Developing a Converter

Below give are the set of rules which were built in order to handle the above mentioned issues.

1. **Rules for Noon (ن)**

a. If word starts or ends with the noon **"ن"** then this is replaced by **"न".**

b. If word starts with **"اَنْ"** then this is replaced by **"अँ"**, else this combination is replaced by **"ॅा"**.

c. If Noon comes in combination with the اِ , اَ , اُ or اٗ then noon is replaced by ◌ं.

d. If noon comes in combination with this diacratic**"**◌ं**"**, then this combination is replaced by ◌ं.

2. **Rules for Aleef (ا)**

a. If word starts with **" ا "** then it is replaced by **"अ"**, and in all other places it is replaced by **"ा"**.

b. If aleef comes in combination with "◌ٔ ", then this combination is replaced by **"अ ॅ"**.

c. If aleef comes in combination with "◌ٍ ", then this combination is replaced by **"इ"**.

d. If aleef comes in combination with "ئِ ", then this combination is replaced by **"ई"**.

3. Rules for Aleef Above Wavy Hamza (أ)

a. If Word starts with " أ ", then it is replaced by **"ऑ"** else it is replaced by **"ॉ"**.

4. Rules for vaav

a. If word starts with " او " then this combination is replaced by "**ओ**", and in all other places it is replaced by **"ाव"**.

b. If "وٚ"comes after these characters (آ, ◌ٍ , ◌َ , ◌ُ ,◌ٍ), then"وٚ" is replaced by **"व"**.

c. If "وٚ" comes in combination with the following characters, then it is replaced by

Character Combination	Replaced By
وا	**वा**
وِی	**वी**
وی	**वी**
وۍ	**व्य**
یو	**ेव**
وٚ	**ौ**
وٗ	**ू**
وَ	**व**
وِ	**वि**
وُ	**वु**
وٍ	**वु**
وٖ	**वू**
ؤ	**वॅ**

d. If word starts or ends with " و " , then this is replaced by **"व"** else it is replaced by **"ो"**.

5. Rules for Yeh (ی)

a. If word ends with "**ی**", then it is replaced by "**ी**", or if a word starts with "**ی**" , then it is replaced by **"य"**, else it is replaced by "**े**".

b. If word starts with "**یٖ**", then it is replaced by **"ई"**, else it is replaced by "**ी**".

c. If Yeh "**ی**"comes in combination with "**◌ٚ**", then this combination is replaced by "**ॆ**".

d. If word starts with "**یُ**", then in this combination Yeh "**ی**" is replaced by **"य"**, else it is replaced by "**्य**".

e. If Yeh "**ی**"comes before or after these characters (ا**,** ◌ٚ**,** ◌ٍ**,** ◌َ **,** ◌ُ **,** وٚ **),** then Yeh"**ی**" is replaced by **"य"**.

f. If a word starts with Yeh "ی", then it is replaced by **"य"**, else it is replaced by "**े**".

6. Rules for (ے)

a. If words end with "◌َے" , "ےُ" or "◌ٍے", then it is replaced by **"य "**.

b. If word ends with "**ے**", then it is replaced by " **े** ",

c. If " **ے** " comes in combination with " ◌ٚ ", then it is replaced by "**ॆ**", else it is replaced by "**ए**".

7. Rules for (ھ)

a. " **ھ** " always comes in combination with the following characters and is replaced as shown in below table

Character Combination	Replaced By
پھ	**फ**
تھ	**थ**
ٹھ	**ठ**
چھ	**छ**
ژھ	**छ़**
کھ	**ख**
گھ	**घ**

8. Palatalization

Devanagari does not have any separate character representing Palatalization like Persio-Arabic script. In Devnagari '**य**' is used for palatalization. Generally it is represented by by shortening the character which comes before it, but in Persio-Arabic a separate character " **ۍ**" is used. Examples are given below.

Persio-Arabic Palatalization	Devanagari Palatalization
مٔنٛزۍ	मँज़्य
ناٗوۍ	नॉव्य
تُہۍ	तुह्य
پأٹھۍ	पॉठ्य
أکۍ	अॅक्य

1.3 Table Representing Persio-Arabic-Devanagari Characters

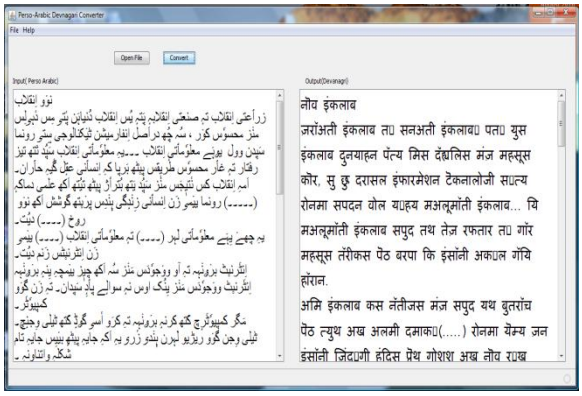
Below given is the list of other characters present in Persio-Arabic Script with their equivalents in Devanagari.

Persio- Arabic	Devanagari
آ	आ
ب	ब
پ	प
ت	त
ٹ	ट
ث	स
ج	ज
چ	च
ح	ह
خ	ख़
د	द
ڈ	ड
ذ	ज़
ر	र
ڑ	ड़
ز	ज़
ژ	ज़
س	स
ش	श
ص	स
ض	ज़
ط	त
ظ	ज़
ع	अ
غ	ग़
ف	फ
ق	क़
ک	क
گ	ग
ل	ल
م	म
ن	न
ہ	ह
ۄ	ॏ
ۍ	◌्य
◌ِ	◌ि
◌ُ	◌ु
◌ٗ	◌ू
◌ٛ	◌ॅ
◌ٚ	◌ॆ
◌ٕ	◌ॖ
◌ٔ	◌ॅ
◌ٟ	◌ॗ
١	१
٢	२
٣	३
۴	४
۵	५
۶	६
۷	७
٨	८
٩	९
٠	०

‘	,
۔	.
؟	?
؛	;

1.4 Snap shot of the Converter

A tool was developed for converting the text from Perso-Arabic to Devanagari using the java by implementing the above discussed rules. The snapshot of the tool is given below:



Snap shot showing the converted text from Persio-Arabic Script to Devanagari Script

1.5 Conclusion

The converter developed so far is the first automatic Persio-Arabic- Devanagari converter. This converter is developed by using a rule based approach. The converter is tested on 10000 words and more than 90% accuracy was found. work is also going on the reverse working of the converter as well that is from Devanagari to Persio-Arabic so that texts can be easily converted from both the scripts either from Persio-Arabic to Devanagari or vice versa.

References

[1] Ali, A. R., and M. Ijaz. "English to Urdu transliteration system." *Proceedings of Language and Technology*, 2009:15-23.

[2] Bhat, R. K. *A Descriptive Study of Kashmiri*. New Delhi: Amar Prakashan, 1987.

[3] Dhar, T. N. *Kashmir Language and Devanagari – Kashmiri Script.* http://vitasta.org/pdf/Vitasta2006SecIII.pdf

[4] Grierson, G. A. 1911. *Standard Manual of the Kashmiri Language*. 2 vols. Oxford: Oxford University Press. Reptd. Rohtak: Light and Life Publishers, 1973.

[5] Jurafsky, D., and H. J. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. London: Prentice Hall, 2000.

[6] Jung, S. Y., S. L. Hong, and E. Paek.. "An English to Korean Transliteration Model of Extended Markov Window." *Proceedings of the 18th Conference on Computational Linguistics*, 2000: 383-389.

[7] Kachru, B. B. *A Reference Grammar of Kashmiri*. Urbana-Champagne: U of Illinois Press, 1969.

[8] --- . *An Introduction to Spoken Kashmiri*. Urbana-Champagne: U of Illinois Press, 1973.

[9] Kak, A. A., and R. Talashi. *Keshur: Akh Grammari Vetshnai* (Kashmiri: A Grammatical Description). Jammu: Nihar Publications, 2004.

[10] Koul, O. N. *Spoken Kashmiri: A Language Course.* http://www.koshur.org/Kashmiri/introduction.html

[11] Raina, M. K. *How to Read and Write Kashmiri in Devanagari* ? http://iKashmir.net/mkraina/index.html

[12] Shant, R. L. "Devnagari Script for Kashmiri: A Study in its Necessity, Feasibility and Practical." *Kshir Bhawani Times.* Kashmiri Pandit Sabha: Jammu, 1997 http://www.koausa.org/languages/devan1.html

[13] Wali, K., and O. N. Koul. *Kashmiri: A Cognitive-Descriptive Grammar*. London: Routledge, 1997.