International Journal of Advanced Research and Development
ISSN: 2455-4030
Impact Factor: RJIF 5.24
www.advancedjournal.com
Volume 2; Issue 6; November 2017; Page No. 809-811

Variation of Maithili: A vital issue of script unification

Saroj Kumar Jha, Shwetangi Kumari, Dr. Piyush Pratap Singh
Department of Computational Linguistics, School of Language Mahatma Gandhi International Hindi University, Wardha, Maharashtra, India

Abstract
This paper attempts at studying the script variation causing the problem of text unification for Maithili language data collected from the various region of Bihar especially Maithili spoken area. Updating the *devanagari* script for developing the Maithili language corpora for Machine Translation (English-Maithili) and reverse. Many of the linguist & scholars (Maithili Literature) have discussed the technicality of text with different mode of writing, and the conclusion drawn lead us to the process of text unification towards all the data collected from the web and the other sources like newspapers, magazine and manuscripts of Maithili language. However, the alternations are found at various grapheme and pronunciation levels. Some problematic issues from the analysis have cropped up like how these varieties can be brought under unified approach and reconstruct the data for processing through technology. There are several examples which can justify not only the close class lexicons are added with huge numbers of scripting style but the open class as well. This reconstruction elaborates the issue hampering the way in unifying the script of Maithili Language. It's messed up due to a variety of forms of Maithili, given to the socio-political reason especially in northern Bihar (Jha, Govind. *et al.*; 1986).

Keywords: POS, scripts, corpora, mithilakshar

Introduction
A diachronic study has been adopted for the study of the script known as *Mithilakshar* to write Maithili language constantly since 7th century till date. The *Mithilakshar* script has evolved from the *Bramhi* and later the modifications kept on with politically motivated system. Therefore the script got its own recognition, ease of use and popularity among the eminent scholars and writers since its frequent used by the *Vidyapati* (a renound Maithili Scholar). A huge number of words borrowed from *Latin, Greek*, to *English*, same as from *Arabic, Farsi* and Sanskrit to *Hindi* and *Maithili* as well. The outstanding *Maithili* scholar *Vidyapati* and others argued about *Mithilakshar* origined from symbol and inscriptions of different temple and caves from the various parts of *Northern Bihar*. The oldest specimen of *Tirhuta* is a *Shaivite* temple inscription in *Tilkeshwarsthāna* near *Kusheshwarsthāna* in *Darbhangā* district of Bihar. Such inscription reflected as the direct evidence of ancient *Magadhi Prakrit* language in these temples which were built on *Kāttikasudi* in "Shake 125" (AD 203) on the next day of *Diwali*. The holiday was immensely regarded as an auspicious day for icon installation throughout the temple. These inscriptions are quite different from the modern *Tirhuta* whereas the research focused to the gradual development of *Mithilakshar* from the bottom to top whereas other language nurtured with full prop up.

As soon as *Mithilakshar* replaced with *Devnagari* due to socio-political reasons and a huge dominance of *Devnagari* script onto Maithili and other IL script, the essence (phonological utterance and lexical representation) of Maithili is wiped out. Several scholars started using it (*devnagari*) with the various morpho-syntactic aspects with the dilutive semantic notation like: *राति* (night) in *Maithili* is written in *Devanagari* *रात/रात्रि* but phonologically it's called "rAit" instead of "rAt/rAtri" which conventionally manipulated by the user having L_1 knowledge of Maithili dominated by Hindi. Hence *Maithili* language incorporated the deviated form of lexicons commonly used at current trend in the *blogs* and *social media* like W*hatsApp, Facebook, twitter* etc., by young age writers and the elite class writers as well, for the sake of ease of writings. Being a linguist and a part of Computational Linguist the research-group, have encountered the several issues related to textual corpora (orthographic from) and web crawled corpora from the ample of websites for analysis and research perspective. The challenges identified as the part of diachronic linguistics analysis of written text available to us at in form of written text & web drawn text. The research groups have opted the few challenging illustration from the text of various centuries started from 60s to 21st century. Therefore confronting this issue, the group have line up a parameter that how the group can appropriately identify the kinds of possible graphitic variation which hamper the text processing work in Maithili-Hindi/English MT System.

Primarily the collected corpora in slot of three decades like from 1961-1980, 1981-2000, and 2001-2015 respectively. The source of corpora is taken from the newspaper, e-book, web and some hand written text of *Maithili* from research scholars of all the three decades having Maithili as an academic subject in their Master/MPhil and PhD, from the renowned universities. The estimated collected corpora for this research are approximate 90,000K sentence corpus of different variety and the data is filtered decade wise from 1961-80, 1981-2000, and 2001-2015 of 32,458, 29,948 and 31,729 respectively and the text domains are from various sections like *Literature, Education, Health and History*. The following areas of

corpora have been selected to mark the token and its types. The fields of the acronyms are listed below:

Table 1

Evaluation Tags		
S.N.	**Category**	**Acro.**
1.	Postposition attachments	PSPatt.
2.	Long and Short Distinction	LSD
3.	Homograph	HOMO
4.	Blending	BLND
5.	Stylistics Change	STYchng.

After the corpus cleaning, the tagged corpora have been classified decade wise from (1960-2015) through MAT (Maithili POS Tagger) application to fix the appropriate POS tags to every word.

Therefore the tagged corpora are manually checked and evaluate the appropriate set GC tag to every word and therefore we tabled the detail the words count separately in the detail format. The format is clearly kept the distinction of TAGs of NN, NNP, VP, and PSP on the basis of the token. Suppose the system identified the PSP attached with "*रामक*" (NNP) & "*काजक*" (NN), and then counts of such tokens exists in the given corpora are also based on "Long and Short Distinction (LSD), Homograph (HOMO), Blending (BLND), Stylistics Change (STYchng)" including postposition attachments (PSPatt). Below the examples are mentioned:

Example

1. *गाछक आम निक लाइग रहल अछि।*

Gaachak aam nik laaig rahal achi.
Mango of the tree tastes nice.

2. *कवितियाक गाछ लाइग रहल अछि।*

kavitiyAk gAch lAig rahal achi.
Kavita's tree is planted.

In the above examples the PSP attached with NN and NNP both in italic word in the examples and the following entity counts are shown below.

Table 2

Evaluation in thousand (1961-80) in %				
S.N.		**NN/P**	**PRP**	**Total**
1.	PSP att.	31.29	23.84	55.13
2.	LSD	24.18	12.32	36.40
3.	HOMO	16.13	15.09	31.22
4.	BLND	07.25	13.33	20.58
5.	STY	03.08	17.03	20.11

In the above table, the data evaluation stats only 16141.36 words out of 32,458 which are PSP attached with NN, NNP.

Table 3

Evaluation in thousand (1981-00) in %				
S. N.		**NN/P**	**PRP**	**Total**
1.	PSP att.	27.41	19.18	46.59
2.	LSD	18.61	09.22	27.83
3.	HOMO	11.81	09.33	21.14
4.	BLND	05.55	10.23	15.78
5.	STY	02.75	11.03	13.78

In the above table 3, the data evaluation stats only 13,952.77 words out of 29,948 which are PSP attached with NN, NNP.

Table 4

Evaluation in thousand (2001-15) in %				
S. N.	**Item**	**NN/P**	**PRP**	**Total**
1.	PSP att.	24.81	16.78	41.59
2.	LSD	15.62	08.16	23.78
3.	HOMO	08.11	06.67	14.78
4.	BLND	04.55	10.23	14.78
5.	STY	02.68	09.85	12.53

In the above table 4, the data evaluation stats only 13,196.09 words out of 31,729, NN attached with PSP, and NNP.

Table 5

Gross Evaluation in %				
S. N.	**Item**	**1960-80**	**1981-00**	**2001-15**
1.	PSP att.	55.13	46.59	41.59
2.	LSD	36.40	27.83	23.78
3.	HOMO	31.22	21.14	14.78
4.	BLND	20.58	15.78	14.78
5.	STY	20.11	13.78	12.53

In the gross evaluation, we found the continuous decrement of every decade from 1980-2015 and the new increment of variety entered in the text with different perspective and arrangement. This graph doesn't reflect the fall of PSP words used in the text but the writing methodology of writers has acquired the variations of style and writing methodology. The other areas of text is also stating the distinction among the frequent use of new to old and the range of effect on text, is a point of question today. Moreover the graph representation is stating the same opinion as below.

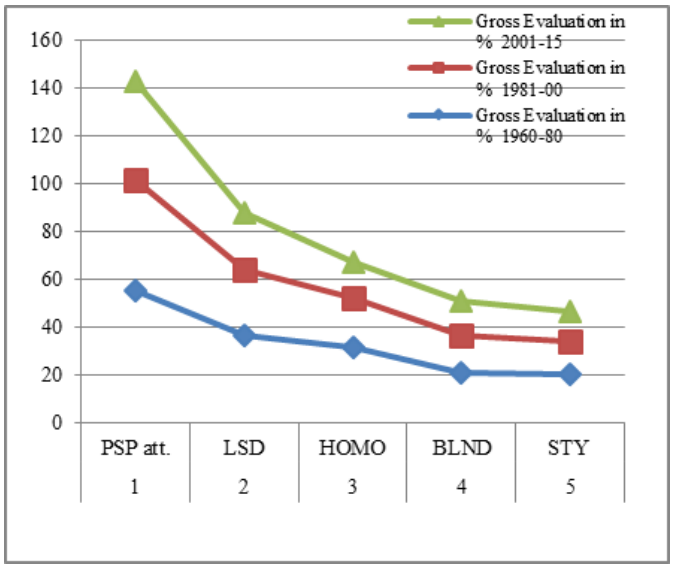

Fig 1

Result & Discussions

The throughout analysis has attempted against the huge amount of corpora to identify the possible linguistics factors as well as statistical and manual analysis of the corpora to

draw a line and parameter during writing the text keeping less possible changes which should be the less hurdle for corpus collection mechanism. After finding this statics of corpus variation a few guidelines are found and these are mentioned below:

- If NN/NNP are attached with PSP "*काजक /रामक*", count it as one and assigned the tag of first entity.
- Distinction of Long and Short between ***"आ"***and***"ओ"***, write it as first for the same meaning of "And".
- The Homograph of PSP "*क*" and VM ***"क"*** will be written separately like PSP "*क*" and VM ***"क'"***.
- Blending forms of VM "*बुझाइछ*" where ("*बुझ+अछि*") is clubbed and created a new form from the existing one.
- The autographic form of text variation in respect to free writing style, and HOMO & Hon (+,-) is followed in various cases.

Conclusion

The conclusion of this paper is to prepare a unified format of collecting data of various region and simultaneously treat all the autographic forms of Maithili language to be written in one style, which will allow machine to treat is as one form with confined meaning. So the process will lead to the guideline to collect corpus and create the machine readable text for easy processing which would be better for all the future work of MT & NLP.

Reference

1. Biber D, Conrad S, Reppen R. Corpus linguistics - Investigating Language Structure and Use. Cambridge: Cambridge University Press, 1998.
2. Aarts J, Meijs W Eds. Corpus Linguistics: Recent Development in the Use of Computer Corpora in English Language Research. Amsterdam Atlanta, GA.: Rodopi, 1984.
3. Baker M, Gill F, Tognini-Bonelli E. Eds. Text and Technology: In honour of John Sinclair. Philadelphia: John Benjamins, 1993.
4. Boguraev B, Pustejvsky J Eds. Corpus Processing for Lexical Acquisition. Cambridge, Mass.: MIT Press, 1996.
5. Botley SP, McEnery AM, Wilson A Eds. Multilingual Corpora in Teaching and Research. Amsterdam-Atlanta, GA: Rodopi, 2000.
6. Dash NS. Corpus Linguistics and Language Technology: With Reference to Indian Languages. New Delhi: Mittal Publications, 2005.
7. Dash NS. Language Corpora and Applied Linguistics. Kolkata: Sahitya Samsad, 2007.
8. Dash NS. Corpus Linguistics: An Introduction. New Delhi: Pearson Education-Longman, 2008.
9. Dash NS. Corpus Linguistics: Past, Present and Future. New Delhi: Mittal Publications, 2009.
10. Dash NS. Language Corpora and Applied Linguistics. Kolkata: Sahitya Samsad, 2007.
11. Dash NS. Corpus Linguistics: An Introduction. New Delhi: Pearson Education-Longman, 2008.
12. Dash NS. Corpus Linguistics: Past, Present and Future. New Delhi: Mittal Publications, 2009.
13. Dash NS. Corpus-based Analysis of the Bengali Language. Saarbrucken, Germany: Verlag Dr Muller Publications, 2009.