International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 03, March 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354

IDENTIFICATION OF MODIFIER COMPONENT IN UPPER AND LOWER ZONE OF DEVANAGARI CHARACTERS

Manoj Kumar Gupta[1], C. Vasantha Lakshmi[2], C. Patvardhan[3]

[1] Ph.D.(C. S.) Scholar, Dept. of Physics and C. S., Dayalbagh Educational Institute, Agra (India)

[2]Dept. of Physics and Computer Science, Dayalbagh Educational Institute, Dayalbagh, Agra (India)

[3]Dept. of Electrical Engineering, Dayalbagh Educational Institute, Dayalbagh, Agra (India)

ABSTRACT

Upper and lower zone of the Devanagari character contains the modifiers of the core character which are present in middle zone. Though these modifiers are very less in number but their proper identification is very important. As some modifier have their part in middle zone which required to be clubbed with the part of the connected component of the modifier in upper zone to correctly identify the modifier. Shirorekha provides a very good basis for separating modifier in upper zone. An attempt is made to identify all possible connected symbols in upper and lower zone of Devanagari character. The frequency analysis done on two documents with different contents and sizes shows that out of all upper zone modifier, the presence of े *(vowel* ए *) is highest with around 33%. And more than 90% of the symbols in the upper zone are coverd by 5 modifiers viz.* े ं ी ो ि*. The results also shows that out of all lower zone modifier, the presence of* ु *(vowel* उ*) is highest with around 55%. And more than 90% of the symbols in the lower zone are coverd by 3 modifiers viz.* ु ू ्*. It has been validated on an article of unknown font. The utility of the proper identification of possible upper and lower zone modifiers along with associated character is in the correct assembly of word after recognition.*

Keywords: Optical Character Recognition, Modifiers, Upper Zone, Lowe Zone

I. INTRODUCTION

Devanagari characters can be divided into three zones and a header line as shown below in Figure 1.

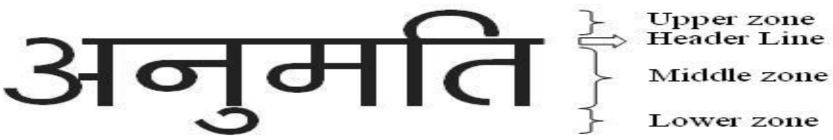


Fig. 1. Example of various zones in devanagari character

Upper and lower zone contains the optional vowel modifiers (matra) above to core character in the upper zone or the modifiers below to core characters in the lower zone. Some modifiers have their part in the middle zone also. A vowel following a consonant may take a modified shape depending on whether the vowel is placed to the left, right, top, or bottom of the consonant [1]. Matras may be attached or detached from the base characters. Hence, the detection of the zone boundaries of character is an important task [2].

Middle zone characters are the core character which can either be single or conjunct characters, though theoretically there can be 46656 triconsonantal conjunct but there are only 345 frequently used symbols in the middle zone. Various structural features viz. bar type, touching count [3], number of water bodies [4], number of left surface cavities [5], place of the touching point to the shirorekha and bar can further classify the connected symbols in the middle zone into manageable chunks of classes [6]. recognition accuracies are further enhanced by identification of the possible character set for single letter words [7].

Motivation for the present work: How do we handle the attached symbols that are attached in the top and bottom strip? How many are there and in how many ways are they attached? Is there a simple way to separate that out? These are the pertinent questions behind the motivation for undertaking this research work.

The paper is organized as follows: Section I describes the proposed approach. Section II describes the identification of the symbols in the upper and lower zone. Frequency Analysis is described in Section III. Validation of proposed scheme over unknown font is discussed in Section IV. Finally, Conclusions are given in Section V.

II. PROPOSED APPROACH

Though modifiers in upper and lower zone are very less in number but their proper identification along with associated character is very important for correct assembly of word after recognition. Moreover some modifiers have their part in middle zone. There are some modifiers or their portions which could be above the header line, below the character and some in the middle zone. The position and the way in which the marks are joined to the base character are required to be identified.

Shirorekha provides a very good basis for separating modifier in upper zone. Steps required to be performed to extract the upper zone symbols starting from reading a page of text are given below in Figure 2.

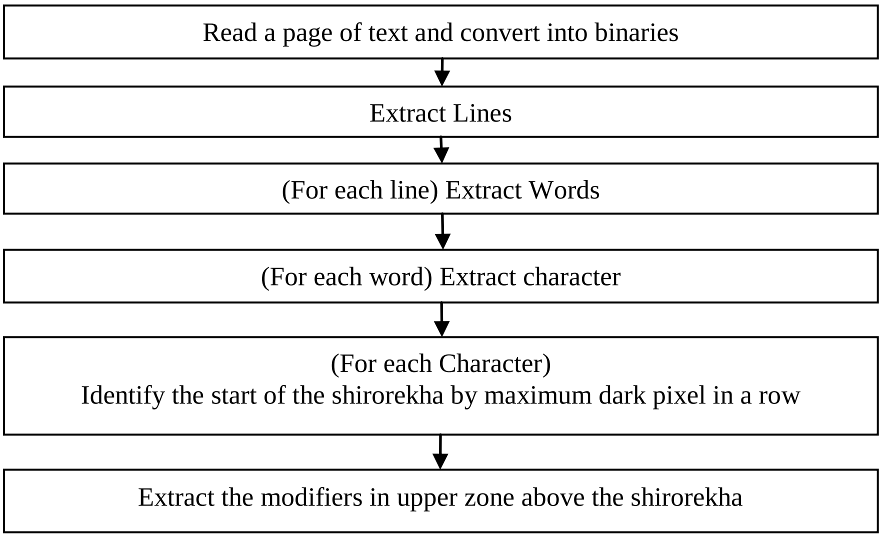

Read a page of text and convert into binaries
Extract Lines
(For each line) Extract Words
(For each word) Extract character
(For each Character)
Identify the start of the shirorekha by maximum dark pixel in a row
Extract the modifiers in upper zone above the shirorekha

Fig. 2. Steps for extraction of symbols in upper zone

The position and the way in which the marks are joined to the base character may vary from font to font. In some fonts the vowel or consonant modifier may touch the character it modifies and in others it may not. Hence the identification of start of lower zone symbols depends upon the identification of end of middle zone. Steps required to be performed to extract the lower zone symbols starting from reading a page of text are given below in Figure 3.

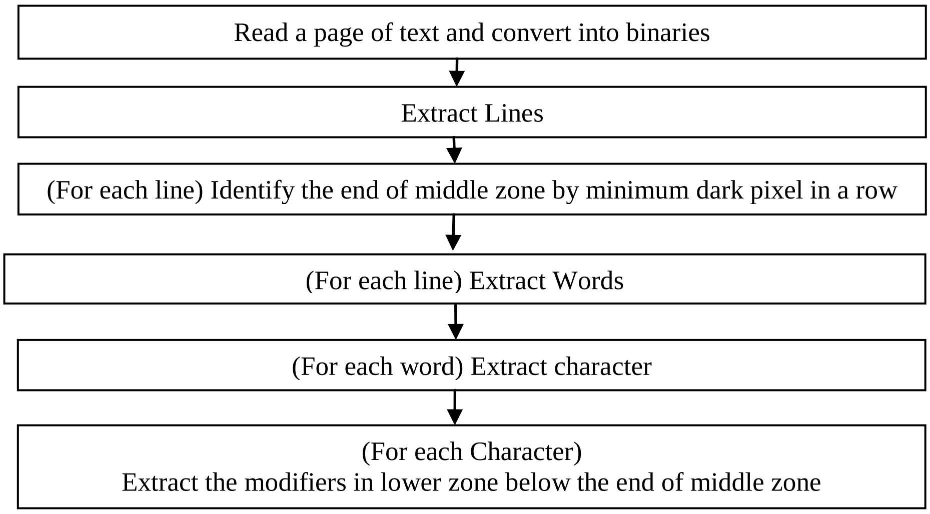
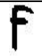
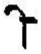
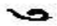
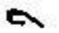

Read a page of text and convert into binaries
Extract Lines
(For each line) Identify the end of middle zone by minimum dark pixel in a row
(For each line) Extract Words
(For each word) Extract character
(For each Character)
Extract the modifiers in lower zone below the end of middle zone

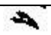
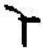
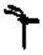
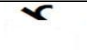
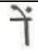
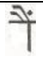
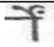
Fig. 3. Steps for extraction of symbols in lower zone

III. IDENTIFICATION OF THE SYMBOLS IN UPPER AND LOWER ZONE

Devanagari script has 13 vowels. The part of the modifiers could be above the header line, below the character or even in the middle zone. Apart from these, there are various other possible symbols formed due to the presence of multiple modifiers on a particular character. In addition to these, there are symbols which are having unconnected component in the lower zone. All these are depicted below in Table I.

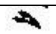
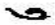
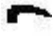
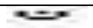
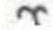
TABLE I. List of Modifiers of the Devanagari Script

Sno	Vowel/ Multiple Vowel/ Others	Modifier symbols attached with consonant क	Symbols in upper zone/ upper zone with portion in middle zone/ lower zone
1	अ	क्	
2	इ	कि	
3	ई	की	
4	उ	कु	
5	ऊ	कू	
6	ऋ	कृ	

7	ऋ	कृ	
8	ऋ	टृ	
9	ए	के	
10	ऐ	कै	
11	ओ	को	
12	औ	कौ	
13	अं	कं	
14	अं	कँ	
15	ए ऋ	कें	
16	इ ऋ	थिं	
17	ई ऋ	थीं	
18	ओ ऋ	खों	
19	ए अं	में	
20	ई अं	हीं	
21	ओ अं	यों	
22	ऐ अं	मैं	
23	ओ ऋ अं	तों	
24	ड़	ड़	

Some modifiers in the above list have their part in middle zone which required to be clubbed with the part of the connected component of the modifier in upper zone to correctly identify the modifier. A summary of distinct symbol in the upper and lower zone are shown below in the Table II.

TABLE II. List of distinct Symbols in the Upper and Lower Zone

Sno	Symbol in Upper Zone	Sno	Symbol in Lower Zone
1		1	
2		2	
3		3	
4		4	
5		5	
6		6	
7			
8			

IV. FREQUENCY ANALYSIS OF SYMBOLS IN UPPER ZONE AND LOWER ZONE

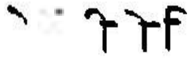
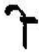
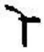
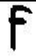
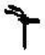
Frequency analysis is done manually to find out the most frequently used symbols in upper and lower zone in two articles of different contexts and sizes [8][9]. The first article contains the 3236 character and the second article contains the 1912 character. The results of upper zone are summarized in Tables III. The results shows that out of all upper zone modifier, the presence of (vowel ए) is highest with around 33%. And more than 90% of the symbols in the upper zone are coverd by 5 modifiers viz. 

TABLE III. Results of Frequency Analysis of Symbols in Upper Zone

Sno	Symbols in upper zone	Article I		Article II	
		Frequency	%age	Frequency	%age
1		409	32.72	244	33.70
2		222	17.76	107	14.78
3		197	15.76	91	12.57
4		158	12.64	123	16.99
5		137	10.96	80	11.05
6		63	5.04	35	4.84
7		33	2.64	26	3.59
8		27	2.16	16	2.20
9		2	0.16	-	-
10		1	0.08	-	-
11		1	0.08	1	0.14
12		-	-	1	0.14
	Total	**1250**	-	**724**	-

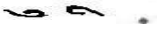
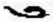
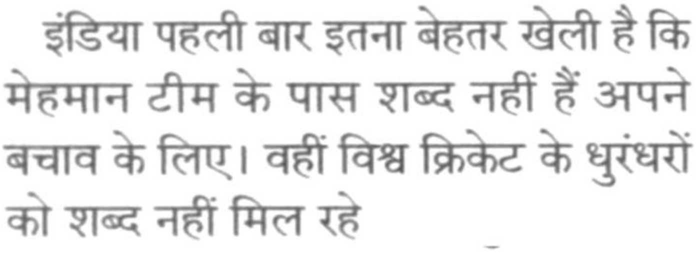
The results of the frequency analysis of most frequently used symbols in lower zone are shown below in Table IV. The results shows that out of all lower zone modifier, the presence of (vowel ੳ) is highest with around 55%. And more than 90% of the symbols in the lower zone are coverd by 3 modifiers viz. 

TABLE IV. RESULTS OF FREQUENCY ANALYSIS OF SYMBOLS IN LOWER ZONE

Sno	Symbols in lower zone	Article I		Article II	
		Frequency	%age	Frequency	%age
1		77	57.46	43	55.13
2		35	26.12	12	15.38
3		18	13.43	14	17.95
4		3	2.24	9	11.54
5		1	0.75	-	-
	Total	**134**	-	**78**	-

V. VALIDATION OVER UNKNOWN FONT

To validate the proposed scheme, a system is developed in JAVA. The Figure 4 shows the image of an article [10] of unknown font used for testing.



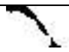
इंडिया पहली बार इतना बेहतर खेली है कि
मेहमान टीम के पास शब्द नहीं हैं अपने
बचाव के लिए। वहीं विश्व क्रिकेट के धुरंधरों
को शब्द नहीं मिल रहे

Fig. 4. Image of an article of unknown font used for testing

Program extracted the 30 words from the above test image. The details of the character identified in all three zones viz. middle, upper and lower zone of test image are summarized in Table V below. Some modifier has their part in middle zone which required to be clubbed with the part of the connected component of the modifier in upper zone to correctly identify the modifier.

TABLE V. List of Character Identified in Middle, Upper and Lower Zone in Test Image

Sno	Word	Zone	Char 1	Char 2	Char 3	Char 4	Char5
1	इंडिया	Upper Zone	ं	ि	े		
		Middle Zone	ड	ा	ड	य	ा
2	पहली	Upper Zone			ी	⌞	
		Middle Zone	प	ह	ल	ा	
3	बार	Middle Zone	ब	ा	र		
4	इतना	Middle Zone	इ	त	न	ा	
5	बेहतर	Upper Zone	े				
		Middle Zone	ब	ह	त	र	
6	खेली	Upper Zone	े	ी	⌞		
		Middle Zone	ख	ल	ा		
7	है	Upper Zone	ै				
		Middle Zone	ह				
8	कि	Upper Zone	ि	े			
		Middle Zone	ा	क			
9	मेहमान	Upper Zone	े				
		Middle Zone	म	ह	म	ा	न

10	टीम	Upper Zone					
		Middle Zone	ट	।	म		
11	के	Upper Zone					
		Middle Zone	क				
12	पास	Middle Zone	प	।	स		
13	शब्द	Middle Zone	श	।	ब्द		
14	नहीं	Upper Zone					
		Middle Zone	न	ह	।		
15	हैं	Upper Zone					
		Middle Zone	ह				
16	अपने	Upper Zone					
		Middle Zone	अ	प	न		
17	बचाव	Middle Zone	ब	च	।	व	
18	के	Upper Zone					
		Middle Zone	क				
19	लिए	Upper Zone					
		Middle Zone	।	ल	ए		
20	।	Middle Zone	।				
21	वहीं	Upper Zone					
		Middle Zone	व	ह	।		
22	विश्व	Upper Zone					
		Middle Zone	।	व	श्व		
23	क्रिकेट	Upper Zone					
		Middle Zone	।	क्र	क	ट	
24	के	Upper Zone					
		Middle Zone	क				

25	धुरंधरों	Upper Zone		ं		े	ं
		Middle Zone	ध	र	ध	र	ा
		Lower Zone	ु				
26	को	Upper Zone	े	·			
		Middle Zone	क	ा			
27	शब्द	Middle Zone	श	ा	ब्द		
28	नहीं	Upper Zone		ी	ं		
		Middle Zone	न	ह	ा		
29	मिल	Upper Zone	ि	ˋ			
		Middle Zone	ा	म	ल		
30	रहे	Upper Zone		े			
		Middle Zone	र	ह			

Middle zone connected component along with their various structural features of the middle zone component viz. Bar Type, Touching count [3], Number of water bodies [4], number of left surface cavities [5], place of the touching point to the shirorekha and bar which have been extracted by the program on the above image of unknown font are summarized in Table VI.

TABLE VI. List of Character Identified in Middle Zone with Structural Properties

SNO	Middle Zone Char	Bar Type	Touching Count	Number of Water Bodies	Number of Left Surface Cavities	First Shirorekha Touching Point from Left in End Bar Character	First Bar Touching Point from Bottom in End Bar Character
1	इ	No Bar	1	1	2		
2	ा	End Bar	1	1	0	After Mid Point	Above Mid Point
3	ड	No Bar	1	1	2		
4	य	End Bar	2	1	1	Before Mid Point	Below Mid Point
5	ा	End Bar	1	1	0	After Mid Point	Above Mid Point
End of Word							
6	प	End Bar	2	1	0	Before Mid Point	Below Mid Point
7	ह	No Bar	1	1	1		
8	ल	End Bar	1	2	1	After Mid Point	Above Mid Point
9	ा	End Bar	1	1	0	After Mid Point	Above Mid Point

End of Word							
10	ब	End Bar	1	1	1	After Mid Point	Below Mid Point
11	ा	End Bar	1	1	0	After Mid Point	Above Mid Point
12	र	No Bar	1	1	1		
End of Word							
13	इ	No Bar	1	1	2		
14	त	End Bar	1	1	1	After Mid Point	Above Mid Point
15	न	End Bar	1	1	1	After Mid Point	Above Mid Point
16	ा	End Bar	1	1	0	After Mid Point	Above Mid Point
End of Word							
17	ब	End Bar	1	1	1	After Mid Point	Below Mid Point
18	ह	No Bar	1	1	2		
19	त	End Bar	1	1	1	After Mid Point	Above Mid Point
20	र	No Bar	1	1	1		
End of Word							
21	ख	End Bar	2	1	1	Before Mid Point	Below Mid Point
22	ल	End Bar	1	1	1	After Mid Point	Above Mid Point
23	ा	End Bar	1	1	0	After Mid Point	Above Mid Point
End of Word							
24	ह	No Bar	1	1	1		
End of Word							
25	ा	End Bar	1	1	0	Before Mid Point	Above Mid Point
26	क	Mid Bar	1		1		
End of Word							
27	म	End Bar	2	1	1	Before Mid Point	Above Mid Point
28	ह	No Bar	1	1	1		
29	म	End Bar	2	1	1	Before Mid Point	Above Mid Point
30	ा	End Bar	1	1	0	After Mid Point	Above Mid Point
31	न	End Bar	1	1	1	After Mid Point	Above Mid Point
End of Word							
32	ट	No Bar	1	1	1		
33	ा	End Bar	1	1	0	Before Mid Point	Above Mid Point

34	म	End Bar	2	1	1	Before Mid Point	Above Mid Point
End of Word							
35	क	Mid Bar	1		1		
End of Word							
36	प	End Bar	2	1	0	Before Mid Point	Below Mid Point
37	ा	End Bar	1	1	0	Before Mid Point	Above Mid Point
38	स	End Bar	2	1	1	Before Mid Point	Above Mid Point
End of Word							
39	श	No Bar	2	1	1		
40	ा	End Bar	1	1	0	After Mid Point	Above Mid Point
41	ब्द	No Bar	1	1	1		
End of Word							
42	न	End Bar	1	1	1	After Mid Point	Above Mid Point
43	ह	No Bar	1	1	2		
44	ा	End Bar	1	1	0	After Mid Point	Above Mid Point
End of Word							
45	ह	No Bar	1	1	1		
End of Word							
46	अ	End Bar	2	1	2	Before Mid Point	Above Mid Point
47	प	End Bar	2	1	0	Before Mid Point	Below Mid Point
48	न	End Bar	1	1	1	After Mid Point	Above Mid Point
End of Word							
49	ब	End Bar	1	1	1	After Mid Point	Below Mid Point
50	च	End Bar	1	1	1	After Mid Point	Below Mid Point
51	ा	End Bar	1	1	0	After Mid Point	Above Mid Point
52	व	End Bar	1	1	1	After Mid Point	Below Mid Point
End of Word							
53	क	Mid Bar	1		1		
End of Word							
54	ा	End Bar	1	1	0	After Mid Point	Above Mid Point
55	ल	End Bar	1	1	1	After Mid Point	Above Mid Point
56	ए	No Bar	2	1	0		

End of Word							
57	।	End Bar	1	1	0	After Mid Point	Above Mid Point
End of Word							
58	व	End Bar	1	1	1	After Mid Point	Above Mid Point
59	ह	No Bar	1	2	2		
60	।	End Bar	1	1	0	After Mid Point	Above Mid Point
End of Word							
61	।	End Bar	1	1	0	After Mid Point	Above Mid Point
62	व	End Bar	1	1	1	After Mid Point	Above Mid Point
63	श्व	End Bar	3	2	1	Before Mid Point	Below Mid Point
End of Word							
64	।	End Bar	1	1	0	After Mid Point	Above Mid Point
65	क्र	Mid Bar	1		2		
66	क	Mid Bar	1		1		
67	ट	No Bar	1	1	1		
End of Word							
68	क	Mid Bar	1		1		
End of Word							
69	ध्	End Bar	3	1	1	Before Mid Point	Below Mid Point
70	र	No Bar	1	1	1		
71	ध	End Bar	3	1	0	Before Mid Point	Below Mid Point
72	र	No Bar	1	1	1		
73	।	End Bar	1	1	0	Before Mid Point	Above Mid Point
End of Word							
74	क	Mid Bar	1		1		
75	।	End Bar	1	1	0	After Mid Point	Above Mid Point
End of Word							
76	श	No Bar	2	1	1		
77	।	End Bar	1	1	0	Before Mid Point	Above Mid Point
78	ब्द	No Bar	1	1	1		
End of Word							
79	न	End Bar	1	1	1	After Mid Point	Above Mid Point
80	ह	No Bar	1	1	1		

81	ा	End Bar	1	1	0	After Mid Point	Above Mid Point
End of Word							
82	ा	End Bar	1	1	0	After Mid Point	Above Mid Point
83	म	End Bar	2	1	1	Before Mid Point	Below Mid Point
84	ल	End Bar	1	2	1	After Mid Point	Above Mid Point
End of Word							
85	र	No Bar	1	1	1		
86	ह	No Bar	1	1	1		
End of Word							

VI. CONCLUSIONS

Though Upper and lower zone modifiers are very less in number but their proper identification is very important. Some modifier has their part in middle zone which required to be clubbed with the part of the connected component of the modifier in upper zone to correctly identify the modifier. Shirorekha provides a very good basis for separating modifier in upper zone. The frequency analysis done on two documents with different contents and sizes shows that out of all upper zone modifier, the presence of े (vowel ए) is highest with around 33%. And more than 90% of the symbols in the upper zone are covered by 5 modifiers viz. े ं ी ो ि. The results also show that The results shows that out of all lower zone modifier, the presence of ु (vowel उ) is highest with around 55%. And more than 90% of the symbols in the lower zone are covered by 3 modifiers viz. ु ृ ़. It has been validated on an article of unknown font.

The utility of the proper identification of possible upper and lower zone modifiers along with associated character is in the correct assembly of word after recognition.

REFERENCES

[1] R. Jayadevan, S. R. Kolhe, P. M. Patil, U. Pal "Offline Recognition of Devanagari Script: A Survey" in IEEE Transaction on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Year: 2011, Volume: 41, No. 6 pp 782-796

[2] J. Dholakia, Atul Negi, S. Rama Mohan "Zone Identification in the Printed Gujarati Text" in Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05), Year 2005, Vol. 1, pp 272 – 276

[3] M. K. Gupta, C. Vasantha Lakshmi, M. Hanmandlu, C. Patvardhan "An Exhaustive Font and Size Invariant Classification Scheme for OCR of Devanagari Character" in International Journal on Natural Language Computing, Feb.2015, Vol. 4 No. 1 pp 1-21

[4] M. K. Gupta, C. Vasantha Lakshmi, C. Patvardhan "Classification of Devanagari Characters based on Water Bodies" in International Journal of Computer & Mathematical Sciences, Jan. 2016, Vol. 5 Issue. 1 pp 18-27

[5] M. K. Gupta, C. Vasantha Lakshmi, C. Patvardhan “Classification of Devanagari Character based on Left Surface Cavity” in International Journal of Recent Scientific Research, Feb. 2016, Vol. 7 Issue. 2 pp 8741 - 8746

[6] M. K. Gupta, C. Vasantha Lakshmi, C. Patvardhan “Identification and Use of Touching Point Structural Property for Piece wise Classification of Devanagari Characters” in International Journal of Computer & Mathematical Sciences, Feb. 2016, Volume 5, Issue 2, pp. 83 - 93

[7] M. K. Gupta, C. Vasantha Lakshmi, C. Patvardhan “Identification of Character Pattern in Devanagari Words for Enhancement of Recognition Accuracy” in Advances in Computer Science and Information Technology, Jan. 2016, Vol. 3 Issue. 1 pp 5-8

[8] Vivek Mohan, Tenaliram Ki Kathain, Jhole Main Katora, 1st ed., Raja Pocket Books,Delhi, 2004, pp 3-7

[9] Vishnu Sharma, Panchtantra Ki 101 Kahaniya,, 7th ed., Manoj Publications, Delhi, 2010, pp 9-12

[10] Hindi Magazine, Sukravar, 18 Dec. - 24 Dec., 2010, pp. 58