Effect of Transliteration on Readability

Sambhav Jain, Kunal Sachdeva, and Ankush Soni

Language Technologies Research Centre, IIIT Hyderabad, Iran
sambhav.jain@research.iiit.ac.in,
{kunal.sachdeva,ankush.soni}@students.iiit.ac.in

Abstract. We present our efforts on studying the effect of transliteration, on the human readability. We have tried to explore the effect by studying the changes in the eye-gaze patterns, which are recorded with an eye-tracker during experimentation. We have chosen Hindi and English languages, written in Devanagari and Latin scripts respectively. The participants of the experiments are subjected to transliterated words and asked to speak the word. During this, their eye movements are recorded. The eye-tracking data is later analyzed for eye-fixation trends. Quantitative analysis of fixation count and duration as well as visit count is performed over the areas of interest.

Keywords: Eye Tracking, Transliteration, Readability.

1 Introduction

Readability can be technically accounted as the ease with which the text can be read and understood. There are various factors that can be explored to measure the readability such as "speed of perception," "perceptibility at a distance," "perceptibility in peripheral vision," "visibility," "the reflex blink technique," "rate of work" (e.g., speed of reading), "eye movements," and "fatigue in reading" (Tinker, Miles A. 1963).

The early research on readability date back to 1880's when English professor L. A. Sherman, pointed out that average sentences length is getting shorter with time and attributed the fact to ease of reading shorter sentences than the longer one by common mass(Sherman, L.A. 1893). The first psychological study in the field was (Kitson, Harry D. 1921), which observed that each type of reader bought and read their own type of text and the respective text types differ in sentence length and word length trend, showing that sentence length and word length are the best signs of being easy to read.

The research on readability since then has been explored extensively in the field of psycholinguistics. Major research methodologies that are employed here include Behavioral Tasks, Language Production Errors, Neuroimaging and Eye Movements. A typical behavioral task would include presenting the subject with linguistic stimuli and ask to perform an action in response (e.g. articulating a given word). The response to the stimuli is recorded and measured (if required). Often this is also complemented by "priming effect" where the earlier linguistic stimuli is provided along with a supporting or disaccording linguistic stimuli and the effect of it is compared with the earlier observation. Language production error methodology analyzes error patterns and investigates a systematic process responsible for it. Neuroimaging take advantage of medical techniques like positron emission tomography (PET), functional magnetic resonance imaging (fMRI),

D. Harris (Ed.): EPCE/HCII 2013, Part I, LNAI 8019, pp. 49–57, 2013.

event-related potentials (ERPs) in electroencephalography (EEG) and magneto encephalography (MEG), and Trans cranial magnetic stimulation (TMS). Eye tracking make use of a device called eye-tracker which can determine and record the point of gaze (where one is looking) or the motion of an eye relative to the head.

Research on readability is not limited to single language but various studies engaging multiple languages have been performed with multilingual readers (Caramazza&Brones, 1979, Soares&Grosjean, 1984).

Language pairs with different writing scripts gave a new angle to the research on readability with an additional factor of orthographical complexity getting introduced. India, with more than a hundred languages and almost each having a different writing script from other, provides great research opportunities.

Kumar et. al. 2010 performed fMRI study of phrase reading for Hindi-English bilinguals and observed left putamen activation for the less fluent language (English). Das et al. 2011, also employed neuroimaging to reveal dual routes to reading in simultaneous proficient readers of English-Hindi orthographies. Rao, et al. 2011 have targeted Hindi – Urdu orthographies and did behavioral analysis on the readability of the two. They observed a relatively faster orthographic characteristics speed in Hindi word naming as compared to that in Urdu.

The advent of digital communication mediums have given rise to the use of transliterated text where the text is written in different script, generally English in most of the cases. However, there is no major readability study yet on the transliterated text. This has motivated us to study the effect of transliteration on human readability. We have tried to explore the effect by analyzing the eye movement of the subject while reading.

In past, eye-tracking has been explored by linguistic researchers to investigate the human reading patterns in language (Rayner 1998). However, there are no studies on the readability research of transliterated text using eye-tracking.

More recently, eye tracking has also been applied to experimental studies in translation process research (Jakobsen and Jensen 2008, Pavlovicand Jensen 2009, Alves, Pagano and Silva 2010, Hvelplund 2011, Carl and Kay 2011, Carl and Dragsted 2012, among others). In most of these works, eye-tracking data have provided input for quantitative analyses of fixation count and duration in areas of interest in texts.

Eye-tracking studies have also contributed to the investigation of the lexical retrieval of words and the processing of syntax, semantics, and discourse. In reading studies, the movements of the eyes are recorded as sentences are read. Typical dependent variables are word-based duration measures such as the time the eyes dwell on a word before proceeding to the next word or the probability to move backwards from a word. Increased in this times and rates of regressions on a specific word are commonly interpreted as posing difficulty to process that word or one of the previous words (Rayner, 1998; Clifton et al., 2007; Vasishth et al., 2013).

2 Span and Scope of Transliteration

In this research, we have studied the effect of transliteration on human readability by analyzing the eye-movement of the participants subjected to reading stimuli.

Transliteration is the process of converting a text from one writing script to another by substituting the alphabets. For example in Chinese, the text '母亲' means 'mother' and

pronounced 'mouchan'; to represent it in text as 'mouchan' instead of '母亲' is transliteration. Here the substitution is done from Chinese alphabets (source script) to Latin alphabets (target script). Across transliterations, the pronunciation of the lexicon however remains unaltered. Off late, transliteration is quite frequently seen especially in case of digital communication like email, chat, blogs etc. The target language in majority of the cases is observed to be English. This is due to that fact that there is an ease to type in English given Latin layout keyboard. The reverse is also seen in practice where an English word is observed in a different script other than Latin. This is majorly seen in case of borrowed vocabulary words. Globalized use of English as official language is accounted as the main reason for it.

The abundant use of transliteration in digital communication has introduced a need for better design of text input mediums and product designers are now considering factors effecting readability, to come up with better display devices. However these are challenging issues as investigating the factors that contribute to better reading or writing experience are not straight forward as writing and reading are not just physical but also a unique cognitive ability of humans, and cognitive aspects are tough to be directly articulated, identified or answered.

Here we have made an effort towards identify such factors, by exploring the eye-tracking technique. Eye-tracking has been extensively explored in past for readability research to investigate the human reading patterns (Rayner 1998). Except here we are having transliterated text instead of the regular text. We have chosen Hindi and English languages, written in Devanagari and Latin scripts respectively, due to high availability of Hindi-English bilingual speakers in the neighborhoods.

3 Experiment

3.1 Objective

The objective of the experiment is to report the changes in the human reading pattern when the text is transliterated. The independent variables in our experiment were as follows:

- Fixation Count
- Fixation Duration
- Visit Count
- First Fixation Duration

3.2 Participants

The experiment is conducted at IIIT Hyderabad which is a deemed university in South India, with two major streams of Computer Science and Electronics & Communication. The university has ample number of students from North India where Hindi is the majorly spoken language. Twenty-four proficient biliterate readers of Hindi-English volunteered from the campus. They included 17 male and 7 female students, aged 20–28 years approximately. All except one claimed Hindi as their native language, but all of them agreed to have received formal education in Hindi during schooling. The participants were given small incentive in form of chocolates for being part of the experiment.

3.3 Stimulus Material

Our stimuli consisted of 8 slide shows. Each slide show comprised of three kinds of slides (Fig1).

a) Instruction Slide.

This slide contains the general information about the experiment.

b) Fixation Slide.

This slide just had a star figure in the middle of the slide. This slide appeared between any two consecutive word slides so that the first fixation landing position could be captured more effectively. The fixation screen is important to minimize the chances of subject already looking at the position where a stimulus is to appear.

c) Word Slide.

This slide contains a single word in one of the quadrant of the slide which is selected randomly.

We establish a baseline with one pair of slideshow, having words in Hindi and English with their respective scripts. The other three pair of slides contains transliterated text with English words in Devanagari and Hindi words denoted in Latin. The words are automatically selected from 'Gyannidhi Corpus' (*Arora et. al. 2003*). We only considered words having more than three syllables (approximately greater than 8 characters) to have ample eye movement for a word. On the basis of their frequency of occurrence (Low, Medium and High) in the corpus, 10 words are randomly selected from each frequency class and for each language, giving us the three pair of slides mentioned earlier.

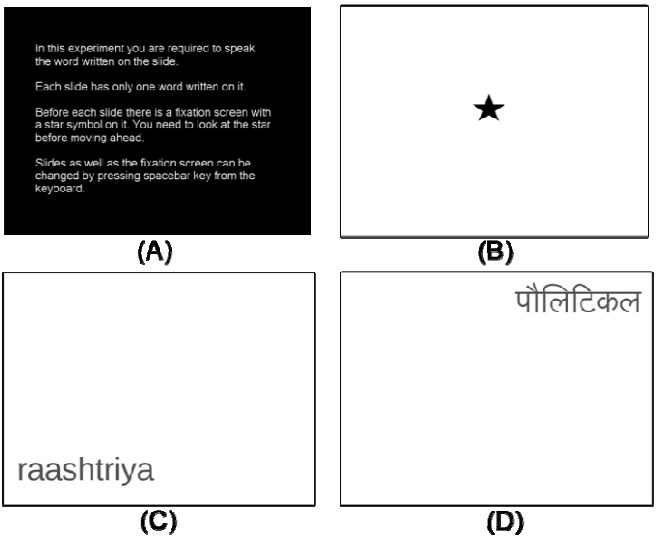

Fig. 1. (A) Instruction slide (B) Fixation Slide (C) Transliterated Hindi word in Latin (D) Transliterated English word in Devanagari

3.4 Procedure

We employ eye tracking machine Tobii X120. Tobii X120 eye tracker is widely used for research in the academic community, and to conduct usability studies and market surveys. The stimuli are loaded in the eye tracker and each participant's response is recorded on the same. We maintained a random order for the above mentioned slide shows so that any kind of bias can be avoided .The internal order of slides in a slide show is also randomized. For analysis, each word is enclosed in a rectangular boundary and the resulting rectangle is partitioned in four equal parts (fig.2) for area of interest analysis, over selected parameters viz. number of fixation points, duration of fixation and number of visits. It is important to divide the word in areas of interest as we want to know which part does subject focuses on. This approach provides an advantage of increased statistical power over a normal whole-volume analysis approach.

3.5 Task

Each participant was subjected to above mentioned eight slide shows .They were asked to enunciate the word occurring on the screen thereby ensuring that they read the complete word. During the process their eye movements are captured with the eye tracker.

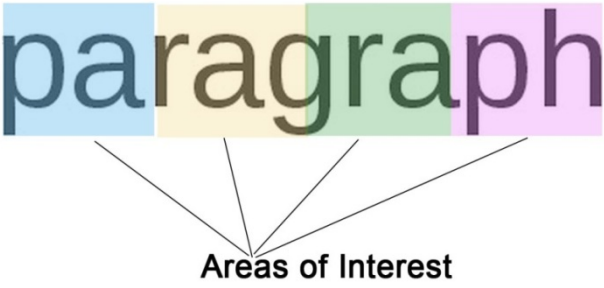

Fig. 2. Four rectangular regions of *Area of Interest* analysis

3.6 Results

We performed the Area of Interest (AOI) analysis for the independent variables in our experiments. The readings recorded by the eye-tracking apparatus are processed via Tobii AOI analysis application. The results for visit count and fixation duration are shown below in Table1 and Table2 respectively. The heat maps for fixation duration are shown in Fig 3. The detailed results for first fixation duration and fixation count are not shown due to space constraint.

4 General Discussions

Our experiment results show an increase in the average reading duration for the transliterated text over the baseline. The average fixation time in case of English written in Devanagari is noted to be greater than that in the case of Hindi written in Latin. This reinforces the observation by (Rao et. al. 2011) that the readability depends on the complexity of the orthography. A high concentration of fixations is seen in the second and the third partitions of the boundary as compared to the first and the fourth partition. The observation can be accounted for, from the study of peripheral vision and central vision by (Legge et. al. 2001). There is an increase in the number of visits in the AOIs over the baseline for both English and Hindi. Also, for both, the languages, the transliterated text, showed a subtle increase in the average fixation number and average fixation count for low frequency words as compared to their medium and high frequency counterparts.

Table 1. ExperimentResults for Visit Count per word averaged over 24 participants. (N=Visit Instances, Mean =Avg. visits, Sum=Total visits, Sdev= Standard deviation)

	Rectangle 1		Rectangle 2		Rectangle 3		Rectangle 4		Total	
Baseline English	N	8.7	N	18.9	N	17.6	N	7.7	N	529
	Mean	1.471	Mean	2.111	Mean	2.085	Mean	1.247	Mean	1.871
	Sum	12.8	Sum	39.9	Sum	36.7	Sum	9.6	Sum	990
	Sdev	0.468	Sdev	1.054	Sdev	0.944	Sdev	0.515	Sdev	1.05
Baseline Hindi	N	8.4	N	18.9	N	17.8	N	7.9	N	530
	Mean	1.595	Mean	2.101	Mean	1.949	Mean	1.354	Mean	1.858
	Sum	13.4	Sum	39.7	Sum	34.7	Sum	10.7	Sum	985
	Sdev	0.684	Sdev	0.964	Sdev	1.234	Sdev	0.493	Sdev	1.22
High English	N	10.5	N	19.6	N	19.1	N	8.7	N	579
	Mean	1.752	Mean	2.444	Mean	2.005	Mean	1.345	Mean	2.009
	Sum	18.4	Sum	47.9	Sum	38.3	Sum	11.7	Sum	1163
	Sdev	0.992	Sdev	1.123	Sdev	1.045	Sdev	0.471	Sdev	1.1
High Hindi	N	13.2	N	19.9	N	19.9	N	13.1	N	661
	Mean	1.538	Mean	2.724	Mean	2.523	Mean	1.634	Mean	2.21
	Sum	20.3	Sum	54.2	Sum	50.2	Sum	21.4	Sum	1461
	Sdev	0.673	Sdev	1.333	Sdev	1.274	Sdev	0.7	Sdev	1.33
Mid English	N	13.5	N	19.8	N	19.8	N	10.5	N	636
	Mean	2.252	Mean	3.096	Mean	2.495	Mean	1.638	Mean	2.489
	Sum	30.4	Sum	61.3	Sum	49.4	Sum	17.2	Sum	1583
	Sdev	1.26	Sdev	1.483	Sdev	1.199	Sdev	0.806	Sdev	1.61
Mid Hindi	N	13.5	N	19.3	N	19.1	N	12	N	639
	Mean	1.704	Mean	2.642	Mean	2.215	Mean	1.433	Mean	2.089
	Sum	23	Sum	51	Sum	42.3	Sum	17.2	Sum	1335
	Sdev	0.916	Sdev	1.261	Sdev	1.171	Sdev	0.561	Sdev	1.38
Low English	N	12.8	N	19.2	N	19.5	N	12.6	N	6
	Mean	2.008	Mean	3.089	Mean	2.944	Mean	1.802	Mean	2.5
	Sum	25.7	Sum	59.3	Sum	57.4	Sum	22.7	Sum	15
	Sdev	0.999	Sdev	1.562	Sdev	1.687	Sdev	0.756	Sdev	2.26
Low Hindi	N	12.6	N	19.3	N	19.3	N	11	N	622
	Mean	1.841	Mean	3.114	Mean	2.233	Mean	1.582	Mean	2.312
	Sum	23.2	Sum	60.1	Sum	43.1	Sum	17.4	Sum	1438
	Sdev	0.911	Sdev	1.688	Sdev	1.242	Sdev	0.687	Sdev	1.71

Table 2. Experiment Results for Fixation Duration per word averaged over 24 participants. (N=Avg. fixations, Mean =Avg. fixation time per fixation (in sec.), Sum=Total fixation time (in sec.), Sdev= Standard deviation (in sec.))

	Rectangle 1		Rectangle 2		Rectangle 3		Rectangle 4	
Baseline English	N	12,8	N	39,9	N	36,7	N	9,6
	Mean	0,216	Mean	0,269	Mean	0,3	Mean	0,309
	Sum	2,762	Sum	10,719	Sum	11,015	Sum	2,965
	Sdev	0,085	Sdev	0,116	Sdev	0,137	Sdev	0,138
	Rectangle 1		Rectangle 2		Rectangle 3		Rectangle 4	
Baseline Hindi	N	13,6	N	39,9	N	35,2	N	11
	Mean	0,291	Mean	0,292	Mean	0,339	Mean	0,332
	Sum	3,962	Sum	11,641	Sum	11,937	Sum	3,655
	Sdev	0,115	Sdev	0,166	Sdev	0,181	Sdev	0,169
	Rectangle 1		Rectangle 2		Rectangle 3		Rectangle 4	
High English	N	18,4	N	47,9	N	38,3	N	11,7
	Mean	0,283	Mean	0,274	Mean	0,311	Mean	0,322
	Sum	5,204	Sum	13,132	Sum	11,903	Sum	3,762
	Sdev	0,136	Sdev	0,137	Sdev	0,16	Sdev	0,158
	Rectangle 1		Rectangle 2		Rectangle 3		Rectangle 4	
High Hindi	N	20,3	N	54,2	N	50,2	N	21,4
	Mean	0,26	Mean	0,254	Mean	0,282	Mean	0,289
	Sum	5,283	Sum	13,763	Sum	14,154	Sum	6,187
	Sdev	0,127	Sdev	0,109	Sdev	0,126	Sdev	0,163
	Rectangle 1		Rectangle 2		Rectangle 3		Rectangle 4	
Mid English	N	30,4	N	61,3	N	49,4	N	17,2
	Mean	0,3	Mean	0,264	Mean	0,311	Mean	0,319
	Sum	9,119	Sum	16,169	Sum	15,35	Sum	5,486
	Sdev	0,17	Sdev	0,127	Sdev	0,155	Sdev	0,153
	Rectangle 1		Rectangle 2		Rectangle 3		Rectangle 4	
Mid Hindi	N	23	N	51	N	42,3	N	17,2
	Mean	0,289	Mean	0,255	Mean	0,299	Mean	0,292
	Sum	6,647	Sum	12,981	Sum	12,63	Sum	5,025
	Sdev	0,127	Sdev	0,113	Sdev	0,136	Sdev	0,163
	Rectangle 1		Rectangle 2		Rectangle 3		Rectangle 4	
Low English	N	25,7	N	59,3	N	57,4	N	22,7
	Mean	0,283	Mean	0,259	Mean	0,298	Mean	0,32
	Sum	7,273	Sum	15,342	Sum	17,111	Sum	7,263
	Sdev	0,121	Sdev	0,12	Sdev	0,166	Sdev	0,155
	Rectangle 1		Rectangle 2		Rectangle 3		Rectangle 4	
Low Hindi	N	23,2	N	60,1	N	43,1	N	17,4
	Mean	0,275	Mean	0,254	Mean	0,289	Mean	0,301
	Sum	6,379	Sum	15,255	Sum	12,449	Sum	5,234
	Sdev	0,118	Sdev	0,11	Sdev	0,135	Sdev	0,153

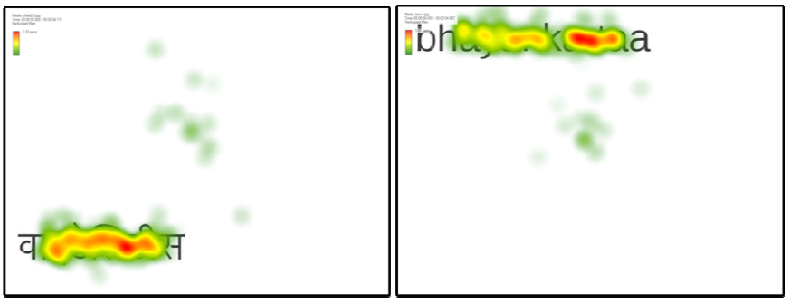

Fig. 3. Heat Maps for Fixation Duration

5 Conclusion and Future Work

The increase in the statistics of fixation duration and visit count indicates that there is an extra effort required on the part of a reader to process and to speak the transliterated text. Thus, our results conclude that the familiar word forms are quickly perceived by the mind rather than the unfamiliar forms and thus readability is not just the mere process of identifying the constituent alphabets but more of a cognitive process.

For simplicity, here we have assumed that each character takes an equal effort in reading on the part of the reader. In future we wish to relax this assumption and investigate the effect over transliterated texts. Also, existing quantitative readability measures (Sinha 2012, Benjamin, 2012) can be explored for selecting the stimuli text.

References

1. Rayner, K.: Eye movements in reading and information processing: 20 years of research. Psychological Bulletin Journal (1998)
2. Arora, K.K., Shukla, S.A.V.G.V.N., Agrawal, S.S.: Gyannidhi: A parallel corpus for Indian languages including nepali. In: Proceedings of Information Technology: Challenges and Prospects (ITPC-2003), Kathmandu, Nepal (May 2003)
3. Rao, C., Vaid, J., Srinivasan, N., Chen, H.C.: Orthographic characteristics speed Hindi word naming but slow Urdu naming: evidence from Hindi/Urdu biliterates. Springer Reading and Writing Journal (2011)
4. Legge, G.E., Mansfield, J.S., Chung, S.T.L., et al.: Psychophysics of reading. XX. Linking letter recognition to reading speed in central and peripheral vision. Citeseer Vision Research Journal (2001)
5. Tinker, M.A.: Legibility of print, vol. 1. Iowa State University Press, Ames (1963)
6. Sherman, L.A.: Analytics of literature: A manual for the objective study of English prose and poetry. Ginn (1893)
7. Kitson, H.D.: How to use your mind; a psychology of study. JB Lippincott Company (1921)
8. Soares, C., Grosjean, F.: Bilinguals in a monolingual and a bilingual speech mode: The effect on lexical access. Memory & Cognition 12(4), 380–386 (1984)

9. Alves, F., GonçAlves, J.L., Szpak, K.: Identifying Instances of Processing Effort in Translation Through Heat Maps: an eye-tracking study using multiple input sources. In: 24th International Conference on Computational Linguistics, p. 5 (December 2012)
10. Von der Malsburg, T., Vasishth, S.: What is the scanpath signature of syntactic reanalysis? Journal of Memory and Language 65(2), 109–127 (2011)
11. Kumar, U., Das, T., Bapi, R.S., Padakannaya, P., Joshi, R.M., Singh, N.C.: Reading different orthographies: an fMRI study of phrase reading in Hindi–English bilinguals. Reading and Writing 23(2), 239–255 (2010)
12. Das, T., Padakannaya, P., Pugh, K.R., Singh, N.C.: Neuroimaging reveals dual routes to reading in simultaneous proficient readers of two orthographies. Neuroimage 54(2), 1476–1487 (2011)
13. Carl, M., Jakobsen, A.L., Jensen, K.T.: Studying human translation behavior with user-activity data. In: Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science, NLPCS 2008, Barcelona, Spain, pp. 114–123 (June 2008)
14. Pavlović, N., Jensen, K.T.: Eye tracking translation directionality. Translation Research Projects 2, 93 (2009)
15. Hvelplund, K.T.: Allocation of Cognitive Resources in Translation: an eye-tracking and key-logging study (Doctoral dissertation, Københavns Universitet Københavns Universitet, Det Humanistiske Fakultet Faculty of Humanities, Institut for Engelsk, Germanskog Romansk Department of English, Germanic and Romance Studie) (2011)
16. Carl, M., Kay, M.: Gazing and Typing Activities during Translation: A Comparative Study of Translation Units of Professional and Student Translators. Meta: Journal des Traducteurs 56(4) (2011)
17. Carl, M., Dragsted, B.: Inside the monitor model: Processes of default and challenged translation production. Translation: Computation, Corpora, Cognition 2(1) (2012)
18. Clifton, C., Staub, A., Rayner, K.: Eye movements in reading words and sentences. Eye Movements: A Window on Mind and Brain, 341–372 (2007)
19. Vasishth, S., von der Malsburg, T., Engelmann, F.: What eye movements can tell us about sentence comprehension. Wiley Interdisciplinary Reviews: Cognitive Science 4(2), 125–134 (2013)
20. Benjamin, R.G.: Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty. Educational Psychology Review, 1–26 (2012)
21. Sinha, M., Sharma, S., Dasgupta, T., Basu, A.: New Readability Measures for {B}angla and {H}indi Texts. In: Proceedings of COLING 2012: Posters (2012)