Title :

Infrequent words are difficult to comprehend : Analysis using gaze tracking

Abstract :

Parameters like saccade length, fixation duration and regression frequency reveal a lot about the text in silent reading. Like whether the text is conceptually difficult or not, has difficult words or not and lot more.
This is in accordance with the view that eye's saccadic movements reflect the online cognitive processing of the text being read(Processing versus Oculomotor model).
Confusion in parsing information is reflected in reduction in saccade length, and increase in fixation duration and regression frequency.
Experiments have been conducted to show that as the frequency of word increases, the fixation time on that particular word decreases. Short within-word regression indicates difficulty in processing that word, hence frequency can also play a role here.
Morover, parafoveal preview(slight focus on first few letters of the word on right of currently fixated word) triggers preliminary parafoveal word analysis. Here also parafoveal preview does better if the parafoveal word has higher frequency. Richer the analysis, faster the reading rate.

This report deals with a similar experiment that analyses whether infrequent words are difficult to comprehend with words drawn from "hindi" language.

Theories to test with “hindi” language:

1. Infrequent words are more fixated upon as compared to frequent words[1].
2. With 3-4 occurences of the infrequent word in the same paragraph, the fixation duration decreases with each subsequent occurence[1].
3. With 3-4 occurences of the frequent word in the same paragraph, the fixation duration remains roughly the same with each subsequent occurence[1].
4. Short within-word regression takes place on infrequent words[2].
5. More the frequency of the parafoveal word, less will the fixation on it. Sometimes, it may even get skipped[4].
6. Reading and comprehension of the text takes place in parallel[5].

Methodology:

We conducted the experiment and the subjects were asked to read the following hindi paragraph.

हिंदी के आख्यान पढ़ने का हमेशा से मुझे बहुत शौक था। ये शौक मुझे धरोहर में अपनी माँ से लिया है, जो कि स्वयं हिंदी के आख्यान किताबी कीड़ो कि तरह चाट डालती थी। परन्तु हिंदी के आख्यान ढूंढ पाना बहुत कष्टदायक कार्य है। माँ ने कभी पुस्तके खरीदी ही नहीं थी, वे ग्रंथालय से पुस्तके ला ला के पढ़तीं और उन्हें लौटा देती थी। इस कारणवश पुस्तको का कभी संग्रह न हो सका। जैसे तैसे मैंने रबिन्द्रनाथ टैगोर की "काबुलीवाला" पढ़ी थी और आनंद से प्रफुल्लित हो उठी थी। परन्तु उसके बाद से कोई अवसर ही न मिल सका। लेकिन भगवान के आगे कहीं कुछ रुक पाया है? कुछ ही दिनो पहले मुझे एक ऐसा ही स्वर्णिम अवसर मिल गया। ऐसा जिसकी मुझे हमेशा से तलाश थी। लैंडमार्क में हिंदी पुस्तको का ढेर लगा था जिसे मैं उठा लायी ! अब एक एक कर के मैं वो सारी किताबें पढूंगी, जो पहले न

पढ़ सकी थी। इनमे से सब से ऊपर थी मुंशी प्रेमचंद की
"नैराश्य लीला" और हरिवंश राइ बच्चन की "बचपन के साथ"। इस उपन्यास के बारे में मैंने माँ से बहुत सुना था।

Design of the Paragraph:

We considered a hindi corpus[6] and found out some words with very less frequency. आख्यान (9.78531371192e-06) , नैराश्य (frequency – 1.39790195885e-06) , स्वर्णिम (frequency - 3.1452794074e-06) , ग्रंथालय (frequency – 3.49475489711e-07) are some of the infrequent words that have placed in the paragraph.
This was done to test theory (1) whether infrequent words are more fixated upon. आख्यान was repeated 3 times to test theory (2).

We also had some frequent words in the paragraph like भगवान (frequency – 0.000116724813564) and पुस्तके (frequency – 0.00013749823759).
पुस्तके was repeated 3 times to test theory (3).

To test theory 6

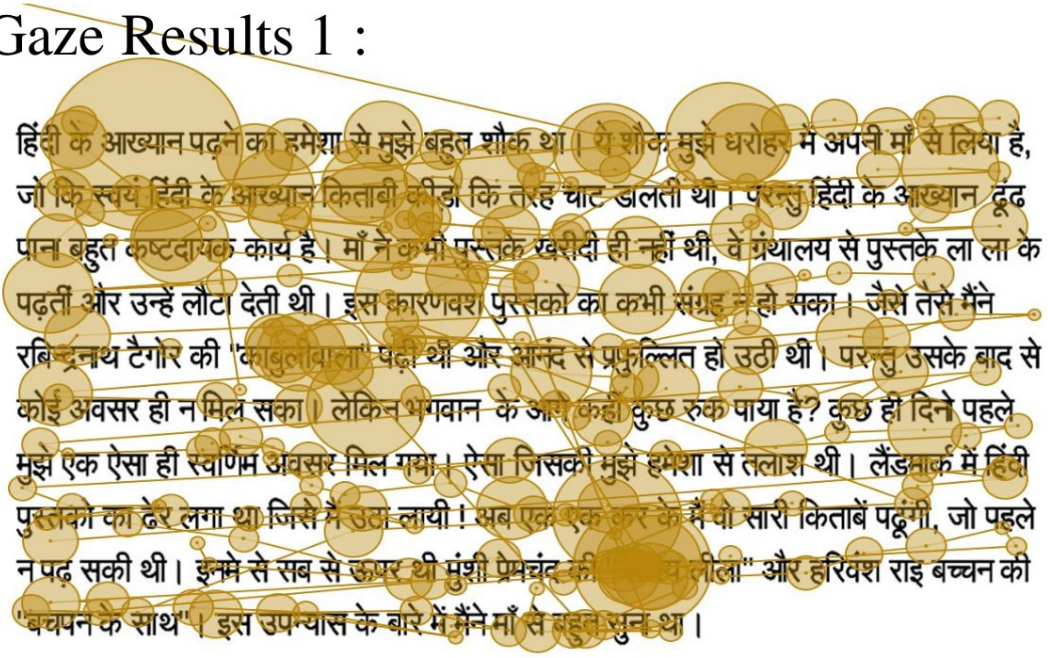
- We took an infrequent word - टै गोर with frequency 7.68846077365e-06
- **Though it is very infrequent but people are very much aware of Rabindranath Tagore, a famous writer in India. By theory (1), the fixation on टै गोर should be high.**
  But, on reading रबिन्द्रनाथ, people would be expecting टै गोर to follow if both reading and comprehension of text were to take place in parallel. Thus it is most likely to get skipped even though its infrequent.
  Hence, if theory (6) is correct, टै गोर would be skipped.
  Had it been serial processing of text, it would have high fixation according to theory (1).

Results :

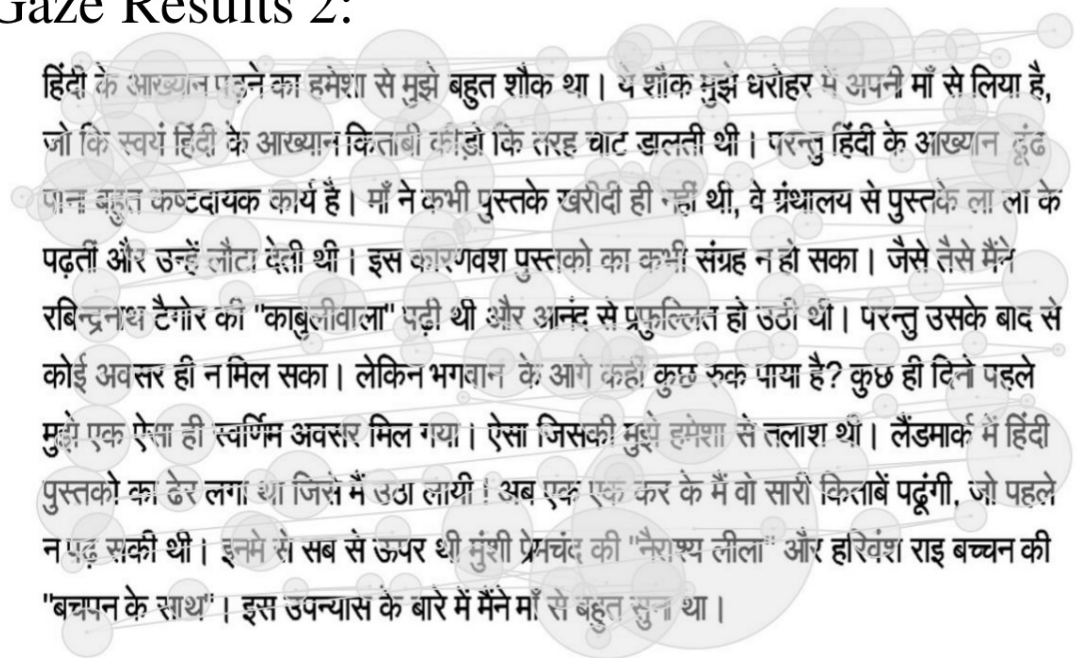
Below are the screenshots of two subjects reading the paragraph. The radius of the circle is directly proportional to the fixation duration. One can observe that :

1. Words like आख्यान are more fixated upon. (Theory 1)
2. The fixation on आख्यान decreases continously. (Theory 2)
3. The fixation on पुस्तके remains roughly the same. (Theory 3)
4. Short word regression on नैराश्य
5. टै गोर, बच्चन, प्रेमचंद have not been focussed, skipped even thus proving theory (6).
   Though they have less frequency(high fixation) but due to parallel processing of text and priming because of famous authors, they are **parafoveally processed.**

Gaze Results 1 :



Gaze Results 2:



References :

1. Becker, W., & Jürgens, R. (1979). Analysis of the saccadic system by means of double step stimuli. Vision Research, 19, 967-983
2. K. Rayner. Eye movements in reading and information processing:20 years of research. Psychological bulletin, 124:372-422, 1998.
3. Matthew S. Starr and Keith Rayner. Eye movements during reading: some current controversies. Trends in Cognitive Science, 5 (2001), pp. 156-163.
4. Inhoff, A.W. and Rayner, K. (1986) Parafoveal word processing during eye fixations in reading:effects of word frequency. Percept. Psychophys. 40, 431-439.
5. Duffy, S. A., Morris, R. K.,& Rayner, K. (1988). Lexical ambiguity and fixation times in reading. Journal of Memory and Language, 27, 429-446.
6. http://www.cfilt.iitb.ac.in/