IndiEn: A YouTube Comments based Indian English Corpus

Vivek Sikarwar (vivek.sikarwar@jaipur.manipal.edu)[1], Mohit Sharma[2], Anurag Khare[3], Ark Verma[4]

[1,2]Manipal University Jaipur, Jaipur, India, [1,3,4]Indian Institute of Technology Kanpur, Kanpur, India

Most of the visual word recognition was started with the language Englsih and continues to be. English language is used as official language in various countries including India. Indian English (IE) is recognized as a distinct, nativized variety of English. There are various corpora available for English language like SUBTLEX-US and SUBTLEX-UK. In this study, a new corpus has been created and validated its frequency estimates through an LDT.

Experiment:

Participants: N=30, Age range = 18 ~ 37 years (M=21.04, SD=3.97).

Stimuli: Corpus was created using YouTube comments. Only Indian origin channels were taken. Words selected were common between current corpus, English Lexicon Project, SUBTLEX-US and SUBTLEX-UK. A 2x2 design was chosen with 2 factors i.e. frequency and length with 2 levels each which were high and low for frequency and short and long for length. There were four conditions: i) High-short (high frequency and short length), ii) High-long (high frequency and long length), iii) Low-short (low frequency and short length) and iv) Low-long (low frequency and long length). Following criteria were chosen for selecting words for each condition: i) High frequency: zipf value greater than 4.5, ii) Low frequency: zipf value less than 3.5, iii) Long length: 7 to 9 letters, and iv) Short length: 3 to 5 letters. For selecting words for High-long condition, a list of words was created based on the criteria zipf value greater than 4.5 and number of letters 7 to 9; then, randomly 125 words were selected. Similarly, 125 words each were selected for the remaining 3 conditions also.

Procedure: It was a lexical decision task. Detailed procedure is shown in figure 1.

Results: Models: freq∗length+(freq+length|subject)+(1|target). Main effect of frequency was found for RTs (Est = 0.09, t-value = 17.90, $p < 0.001$) and accuracy (Est = -1.43147, t-value = -13.257, $p < 0.001$). Effect of length was found for accuracy only (Est = 0.56309, t-value = 6.257, $p < 0.001$). See figure 2 and 3. Correlation values can be found in table 1.

Discussion: Results show that freq estimates of current corpus correlates well with RTs of Indian population which makes it a useful tool for selecting words for carrying out studies on visual word recognition in English language on Indian population. Results further suggest that freq estimates of current corpus is better than those of English Lexicon Project, SUBTLEX-US and SUBTLEX-UK. Further, there is a need to compare the freq estimates with more corpora to confirm if the effect of frequency estimates of a corpus can vary with geographical variations.

Figures and Tables

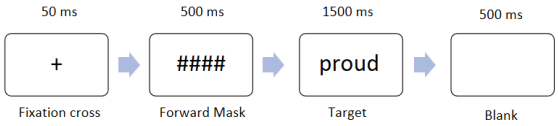

Figure 1: Procedure of Experiment.

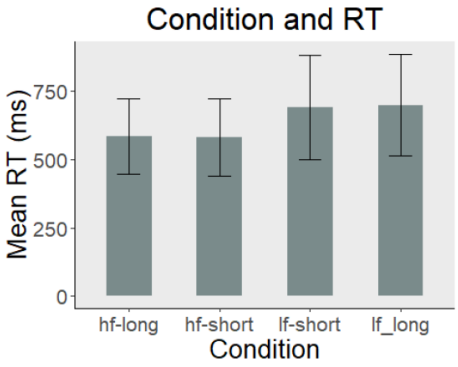

Figure 2: Relation between RT and conditions

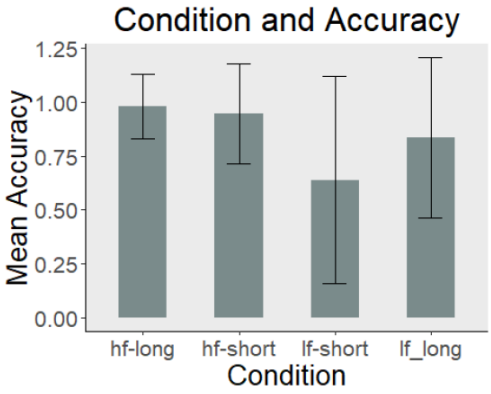

Figure 2: Relation between Accuracy and conditions

Table 1: Correlations between variables for word stimuli

	IndiEN_zipf	mean_rt_exp	mean_acc_exp	mean_rt_ELP	mean_acc_ELP	Sub_UK_zipf
mean_rt_exp	-0.73					
mean_acc_exp	0.55	-0.7				
mean_rt_ELP	-0.7	0.69	-0.65			
mean_acc_ELP	0.45	-0.58	0.79	-0.7		
Sub_UK_zipf	0.88	-0.73	0.55	-0.75	0.52	
sub_US_zipf	0.87	-0.72	0.55	-0.73	0.49	0.96