Abstract

Text readability refers to all text properties that interact with a reader during reading and affect the extent of understanding of the text and the cognitive load. Readability or text difficulty depends on the lexical, syntactic, and semantic and discourse dimensions of the textual content along with the background of the reader and the concerned language. In spite of having many practical significances and huge user base, there are few attempts for text readability Bangla. In this thesis, I have studied text readability in Bangla in three textual levels: full text, sentence and word. A Bangla dataset in Unicode encoding has been prepared to make up for the unavailability of automatically accessible data in Bangla that is annotated according to the reading difficulty. User feedback has been recorded through a number of empirical user surveys using diverse methodologies. Two target user groups have been considered at every step of the work to study the subjective effect user background on text readability and to validate the final models. For full text readability, the research establishes the inapplicability of English readability formulae such as Flesch Reading Ease index for Bangla. Estimation models and binary classification models of Bangla text readability have been developed using regression analysis and support vector machines respectively. A short study has also been performed on the relation between the readability and latent semantic analysis in Bangla. In sentence level readability, I have studied both the effect of constituents and organization of words in a sentence on the comprehension difficulty. Regression method has been used at first to model user feedback on sentence reading difficulty with sentence attributes. In the next step, an entropy approach based on dependency grammar of Bangla has been used to model the user response with successive reading of words in a sentence. A computational model has been proposed based on depth of the dependency structure and number of unprocessed dependencies to measure processing difficulty of sentence surface forms. At word level, I have studied the effect of orthography through feature based complexity, phonology, semantics and word familiarity on written word recognition in Bangla.

Keywords: Text readability, Sentence comprehension, Word recognition, Computational models, Correlation, Regression, Support Vector Machine, Entropy, User study