Investigating Chinese Text Readability: Linguistic Features, Modeling, and Validation

Yao-Ting Sung[1], Ju-Ling Chen[1], Yi-Shian Lee[2], Jih-Ho Cha[2], Hou-Chiang Tseng[2], Wei-Chun Lin[1], Tao-Hsing Chang[3], and Kuo-En Chang[4]

[1]Department of Educational Psychology and Counseling, National Taiwan Normal University

[2]Research Center for Psychological and Educational Testing, National Taiwan Normal University

[3]Department of Computer Science and Information Engineering, National Kaohsiung University of Applied Sciences

[4]Graduate Institute of Information and Computer Education, National Taiwan Normal University

This study aims to (a) develop readability indicators based on the textual factors that influence reading comprehension; (b) construct the readability model for Chinese text; and (c) validate the proposed readability models. This study constructs readability models employing step regression and SVM, using 24 readability indicators as its predictive variable and the grade level of 386 textbook articles as the criteria. The proposed models are then validated according to an additional 96 texts. The results show that in step regression, the critical predictors are the number of complex words, proportion of simple sentences, average logarithm of content word frequency, and number of personal pronouns. In the SVM model, the critical predictors selected by using the *F*-score include the number of complex words, number of two-character words, number of characters, and number of intermediate-stroke characters. The accuracy rates of step regression and SVM are 55.21% and 72.92%, respectively. Both models predict the texts more accurately at the lower grade levels than at the higher grade levels.

Keywords: *accuracy, readability, stepwise regression, support vector machine*