The Factorial Structure of Written Hebrew and its Application to AES

Anat Ben-Simon & Yael Safran

- Abstract -

In 2000, NITE launched the Hebrew Language Project (HLP), the goal of which is to develop computational tools for the analysis and evaluation of Hebrew texts. The present paper summarizes the initial development, analysis and organization of machine-generated statistical and NLP text features and mapping of the underlying structure of written Hebrew through analysis of the structure of these features. To this end, the paper reports the results of two successive studies.

The purpose of the first study was to examine the characteristics of 133 machine-generated quantified features, to identify the ones most relevant to text difficulty and writing quality and to combine them into empirically based and theoretically meaningful linguistic categories. The study also examined the effect of the text-feature clustering model on the accuracy of the automated score. To attain these goals, a three-stage analysis was carried out using two text corpora and two essay corpora.

The second study focuses on analysis of the factorial structure of writing ability and the validation of machine-generated text features used for its prediction. A factor analysis applied to the selected AES features using five essay-corpora, revealed three AES dimensions: lexical complexity (fluency), topical analysis (content) and vocabulary. However, the AES dimensions failed to align with raters' scores on compatible or close dimensions.