**A new Dataset for Automatic Recognition of Complex Sentences in Hebrew Texts**

SE-3-6

By: Shira Malka; shirma3@ac.sce.ac.il  
 Ron Bar Zvi; ron.bzeve@gmail.com

Advisors: Dr. Natalia Vanetik1, Dr. Marina Litvak2

1Shamoon College of Engineering, Beer-Sheva

State institutions websites are usually defined with elevated Hebrew language, which can cause reading and understanding issues, especially for disadvantaged groups, such as immigrants, elders and mentally retarded people. Sentence simplification aims to modify a sentence in order to make it easier to comprehend. Because Hebrew is a low-resourced language, we built our own dataset comprising of 1000 sentences collected from governmental websites and annotated with binary labels related to sentence complexity. This dataset can be employed in the future for training machine learning models for automatic detection of complex sentences. We reached 72.4% agreement on sentence complexity labels using Kappa score.

**Keywords:** Machine learning, Natural Language Processing, Hebrew, Sentence Simplification, Text Simplifier