Word Frequency Distribution in English and Hungarian texts

Mária Csernoch[1], László Hunyadi[2]

[1]Institute for English and American Studies, [2]Department of General and Applied Linguistics,
University of Debrecen, 4010 Debrecen, Hungary
mcsernoch@hotmail.com

Models based on the frequency of word-types assume that tokens in a text occur randomly. Even though, many different strategies can be followed in the process of selecting words randomly. The best models proved to be those that assume that word-types are binomially distributed. Based on this binomial distribution a model was created first by Baayen [1] than by Hoover [2]. In these models a constant was calculated for each type which occurred m times in the original text. This constant served as the predicted number of types occurring in the text. Summing up these constants for each type a predicted number of word-types can be calculated. However, the model created this way is static, thus it always provides a constant for a selected M (M ≤ N, where N is the number of tokens in the text). In the era of the new generation of personal computers dynamic models can also be built based on the same assumptions.

The ultimate goal of our studies was to build such a dynamic model. The question was whether this new model can help us to explain the regularities of the introduction of word-types in a text. Furthermore, can we see what the sources are which force the authors to use new types in their works, and can we find any significant sign to use this method to recognize the author of a selected work?

Hungarian and English literary works and English textbooks were analyzed the find regularities in the introduction of word-types in these works. To carry out the study the texts were divided into short, constant-length intervals with a usual length of 100 words. One of the advantages of this method was that the short intervals allowed us to follow minor changes in the texts. Based on the frequency of the word-types in the original text a model, an artificial text was created. Comparing the original and the artificial text we were able to find intervals in the original text which made a kind of stand out. These jumps were found to be responsible for something unpredicted, sometimes illogical events in the discourse of the text. Analyzing the textbooks, we learned that the introduction of word-types in these books showed resemblance to randomly chosen and then concatenated short stories. It seemed that the authors of the textbooks ignored that not only the number of word-types should be increased, but the words should be repeated a certain times in these books.

References

1. Baayen R. H.: The Effect of Lexical Specialization on the Growth Curve of the Vocabulary. Computational Linguistics 22. (1996) 455-480.
2. Hoover D. L.: Another Perspective on Vocabulary Richness. Computers and the Humanities 37 (2003) 151-178.