THE DEFINING OF SAMPLE SIZE FOR STATISTICAL ANALYSIS OF TEXTS
Ihor Kulchytskyy
Applied Linguistics Department, Lviv Polytechnic National University, 12 S. Bandera Str., Lviv, 79013, UKRAINE, E-mail: bis.kim@gmail.com
Lviv Polytechnic National University

The exploring of methods and techniques of quantitative study in linguistics is not going to lose topicality. One of the important areas of quantitative study is the examining of informational and statistical properties of the text. They are commonly used in the text attribution and deciphering. On the other hand, for any statistical analysis of the text it is important to correctly choose the method and quantity. This study is an attempt to establish the sufficient percentage of works of art by Marko Cheremshyna to determine the probable relative frequencies of symbols in them as well as to examine the stability of these frequencies.

The material of the research is the complete works by Marko Cheremshyna published in 1937. The main object is the relative frequency of letters in the Ukrainian alphabet. Since hyphens, apostrophes, and space (the last divides the text into words) are also used in Ukrainian texts, during calculating Marko Cheremshyna’s works are interpreted as the set of symbols of the extended Ukrainian alphabet including apostrophe, hyphen and space. As the creation of the text selections are made with the means of the computer programme, the length of the paragraph is equal to the length of text passage. The completed works are converted into electronic form and normalized. For each of them the relative frequencies of symbols of the expanded alphabet were defined. From the received samples there were created 5 identical texts for research length of about 470 000 symbols that differ only in paragraph length (about 100, 200, 300, 400 and 500 symbols respectively). The paragraph length was chosen arbitrarily, all small letters were changed into capital, text symbols that were not included in the extended alphabet were replaced by a space, only one space was left between words, the text was divided into paragraphs of fixed length to within the accuracy of a word, that if you add the word to the next paragraph and its length becomes greater than the predetermined, the word is not cut and the paragraph is left bigger to several symbols; during calculations the end symbol of the paragraph is considered as a space.

With the use of Pearson’s chi-squared test the optimal length of text passages and the amount of text works by the author was determined when the relative frequency of symbols coincides with the frequency counted for all works and each symbol rank of the extended alphabet in a frequency distribution was defined. In total 965 000 experiments were conducted. Thus, for Marko Cheremshyna’s works the optimal passage length is 100 characters and the sample length is about 42 % of the total one.

Keywords: quantitative study, sample, sample length, frequency, Marko Cheremshyna.