УДК 009

COMPUTER-ASSISTED TEXT ANALYSIS EXEMPLIFIED BY THE WITCHER BOOKS

Arina S. Petrova (ITMO University), **Iuliia A. Iusma** (ITMO University),

Annotation. The report presents the results of the computer analysis of the Russian translation of the text "The Witcher". The results of the study are the marked-up text corpus, the frequency analysis of words for each book and for the description of the main characters, the most common connotations for each book, the frequency of the appearance of characters' names throughout the book series. The work also contains the visualization of the results and the graph of the relationship between the characters.

Introduction. Computers have learned how to work with structured information, but most of the information in the world is presented as unstructured texts. An example of this type of information could be works of fiction. Computer analysis of literary texts of various styles and genres is now becoming one of the necessary components of linguistic research, as linguist Georgy Vekshin notes that computer methods are widely used in modern literary studies. Computer analysis greatly simplifies the work with huge amounts of texts and makes it possible to draw conclusions based on accurate information rather than on subjective human opinion. Data obtained through automated analysis can be used to refute or confirm research hypotheses. There are many off-the-shelf tools that can be used for analysis, but many researchers still prefer to work with texts without using any of them at all. The first reason is the lack of ready-made tools for analyzing Russian text. The second reason is the inability to get correct analysis results without manual markup of the text. The third reason is the genre and vocabulary of the work. For example, to analyze fantasy fiction it is necessary to understand the meaning and vocabulary that the author puts into certain terms. Therefore, with the existing ready-made tools for text analysis, there is still no ready-made solution for every text.

Main part. We propose the following algorithm for the study of the literary text. The first stage is to determine the desired results. Our expected results are a frequency analysis of the entire text, a frequency analysis of the characters' descriptions, an analysis of the use of the characters' names throughout the book series, connotations, a sentiment analysis of the characters' speech, a network analysis of the characters' relationships, and a map of the characters' movements. The second step is the manual markup of the text. The results of this stage are the dataset of characters' descriptions, the dataset of direct speech of the characters, the table with the movement of the characters on the map, and the table with the relationship of the characters. The third step is text analysis using the Python programming language and its *nltk* library in particular. At this stage, firstly it is necessary to do tokenization - the selection of individual words, and to remove all the unnecessary for the analysis symbols. Next, it is needed to do lemmatizing and to make a list of stop words, then analysis can be done. The fourth step is the visualization of the results.

Conclusion. The analysis of the Wicher book let us see the inner sights of literary texts and formulated recommendations for working with the text. Moreover, sentiment analysis and topic modeling help us make all the vital topics, which are presented in the books, visible and highlighted with the key words. To simplify the presentation of the results we made the visualization. In the future, we plan to analyze the English-language translation as well as the original itself for comparison purposes, and to create a website on which the results of the analysis will be presented.

Arina S. Petrova (author)

Iuliia A. Iusma (author)