ABSTRACT

Title: " The development of automated database replenishment system of stable combinations (on the basis of natural language)"

Author: Mariya I. Butuzova

Research supervisor: Associate Professor, Ph.D. Nikolai A. Knyazev

Research initiator: State budgetary institution of additional professional education "Center for Continuous Development"

Topical importance: creation of all sorts of word combinations. Such dictionaries are necessary primarily for teaching Russian to foreigners, for automating translation, for resolving lexical homonymy in algorithms for automatic text analysis, for identifying paronymic errors, etc. The creation of such dictionaries is a very laborious process, requiring the joint efforts of many qualified specialists. Any, even partial, automation of this process is of undoubted practical interest.

Goal: creation of a suitable algorithm for further development of the automated search system for stable combinations (on the material of a natural language) and updating the database for further use.

Tasks: the study of all parts of speech in the Russian language for the presence of specific characteristics. The study of combinations of some parts of speech with other parts of speech. Study of word combinations of a certain type for the presence of specific characteristics. Creation of an algorithm and its block diagram based on the identified characteristics. Creation of a database of finite symbols of the most frequency parts of speech. Writing a visual shell and code to solve the tasks. Testing the program for syntactic and morphological errors. Create exceptions in the code for certain parts of speech.

Theoretical value: consists in justifying the need to use purely linguistic methods for solving the tasks posed.

Practical applicability: is to use the created software to compile glossaries, dictionaries, as well as for a more extensive analysis of the text by linguists.

Results: a large amount of information and material was studied, the goal was achieved and the tasks were accomplished. System of automated replenishment of the base of stable combinations was developed. Also, software for this system was implemented using a high-level C # programming language. As the extracted information, stable word combinations were chosen, which make it possible to extract the semantic core from the text and which will later be used as databases of dictionaries and glossaries. After testing on 20 scientific papers of various directions, the result was achieved in 80% of the accuracy of the developed program (automated system).

Implementation advice: to obtain the best, more accurate results of the developed automated system, it is necessary to check on more articles of different subjects. Based on the problems identified during the work, it should be noted that inaccuracies in the algorithm will be corrected and improved in the future. It is important to add that not every combination of parts of speech is appropriate. For correct operation it is necessary to conduct the most detailed analysis of the database of tokens for incompatible finite symbols of one part of speech with the final symbols of the other part of speech. This will identify the most frequent word combinations. At the moment, 20% of errors occur due to two adjacent parts of speech, which are not combinations by grammatical features.