Impact Factor:					
	ISRA (India) = 4.971	SIS (USA) = 0.912	ICV (Poland) = 6.630
	ISI (Dubai, UAE) = 0.829	РИНЦ (Russia) = 0.126	PIF (India) = 1.940
	GIF (Australia) = 0.564	ESJI (KZ) = 8.997	IBI (India) = 4.260
	JIF = 1.500	SJIF (Morocco) = 5.667	OAJI (USA) = 0.350

SOI: 1.1/TAS DOI: 10.15863/TAS

International Scientific Journal

Theoretical & Applied Science

p-ISSN: 2308-4944 (print) **e-ISSN:** 2409-0085 (online)

Year: 2020 **Issue:** 11 **Volume:** 91

Published: 07.11.2020 **http://T-Science.org**

QR – Issue QR – Article

A.N. Abdullayev
Samarkand State University
Associate Professor
Department of Information technology

Laziz Shukurov
Samarkand State University
Master student
Republic of Uzbekistan, Samarkand

DEVELOPMENT OF FULL-TEXT DOCUMENTS USED IN THE ACTIVITIES OF UNIVERSITIES WITHOUT DICTIONARY ALGORITHMS FOR CONTROLLING SPELLING ERRORS

Abstract: *The basis of the article is the definition of the main approaches, principles and methods of building an information processing system for monitoring and correcting errors in electronic texts in natural languages, as well as the development of methods for assessing and analyzing probabilistic and quantitative indicators of the effectiveness of the system under study.*

Key words: *Algorithm, synthesis, analysis, method, text, information, coding.*

Language: *English*

Citation: Abdullayev, A. N., & Shukurov, L. (2020). Development of full-text documents used in the activities of universities without dictionary algorithms for controlling spelling errors. *ISJ Theoretical & Applied Science, 11 (91),* 43-47.

Soi: http://s-o-i.org/1.1/TAS-11-91-9 ***Doi***: https://dx.doi.org/10.15863/TAS.2020.11.91.9

Scopus ASCC: *1700.*

Introduction

In accordance with the purpose of the article, the following theoretical and practical tasks are set:

- selection of areas of research, determination of the basic principles of building a computer system for monitoring and error correction, analysis of the probability of distortions at the stages of input, transmission and processing of electronic texts;

- development of a methodology for determining the amount of redundancy, which would provide the required information reliability when applying methods and software tools for monitoring and error correction based on the use of redundancies of various nature,

- development of methods for the synthesis of probabilistic processes, determining the amount of information: taking into account the ongoing probabilistic processes during the usual reception of a message and when using a correcting code.

- development of a methodology for determining the required rational memory volume of a computer information processing system with built-in means of monitoring and correcting errors in electronic texts.

This paragraph outlines the main approaches to the creation of methods, algorithms for monitoring and correcting errors in texts when processing data in electronic document management systems (EDMS) of enterprises.

One of the important criteria for the functioning of enterprise EDMS is reliable data exchange. However, in real conditions, the reliability of information is very low and is equal to approximately 3.4 · 10-2 osh / sign. It has been established that about 85% of errors in the total volume of distortions belong to the human operator, scanning and recognition processes. Moreover, for the normal functioning of the system, it is required to increase the reliability of the processed information up to 10-5-10-6 osh / sign, which emphasizes the urgency of solving the problem

of building a software control (detection) and correction (automatic correction) of errors in texts.

Methods for entering and transmitting information. Text information in the form of typewritten texts, documents, graphs, etc. can be entered by a human operator, through scanning devices, including recognizing software systems, and also transmitted as files by machine media, by e-mail, etc. decoding of text information. In this regard, in terms of this work, along with others, we also solve problems related to the development of methods, algorithms and programs for monitoring and correcting errors based on effective methods of encoding, compressing and decoding information.

Information coding.

The input text presented for solving problems of control and information processing is usually encoded. Various methods can be used to encode text information, in particular Shannon-Fano, Huffman codes, arithmetic coding; vocabulary methods: Ziv-Lempel, Lempel-Ziv-Welch and algorithms using the Burrows-Wheeler transformation, ASCII machine code, etc. It has been proven that due to the simplicity and high efficiency in the created IC it is possible to give preference to the use of an arithmetic code.

The peculiarity of this method is that the transmitted text, as a rule, is encoded with real numbers, or you can use decimal numbers. This makes it possible to construct an algorithm for adapting the boundaries for decoding, which can be used to control the validity of real decoded texts. In this regard, for the control and correction of errors in texts, a task has been posed and is being solved related to the development, research and application of software methods based on the use of features, schemes and rules of the recommended arithmetic coding.

Software methods of information control.

In foreign and domestic practice, software methods for monitoring, detecting and automatically correcting errors in text information have not been studied enough, and so far there are no developments that are effectively used in practice. The results of the research prove the possibility of using software methods for controlling digital information to detect single and double spelling errors in texts, for example, through the use of linear, modular and plane summation algorithms.

It was noted above that human operator, scanning and recognition errors account for a significant proportion of the total amount of distortion. At the same time, such errors are characterized as errors of large multiplicity (k-grams). Consequently, the tasks of control and correction of errors in texts should be solved in a new setting, which will take into account the noted conditions for processing information. Moreover, it is required that the software system built into a computer data processing system provides comfortable conditions for detecting and correcting errors based on the use of modern computer technologies.

Along with the application of software methods that use artificial redundancy, it is also effective to use natural redundancy to control and correct errors in texts.

When setting the problem of information control on the basis of natural redundancy, the error correction software can be implemented on the basis of the development of the following error control methods: along the boundaries of code sets, recoded text information; according to special reference books of word forms of a natural language. In addition to these, you can use methods that take into account logical connections between sequences of phrases, words, letters or the specifics of the coding system; control over the boundaries of valid codes, etc.; methods that take into account statistical relationships and data correlations; semantic methods that take into account the properties of the language and the structure of the formation of word forms; methods for detecting grammatical errors based on morphological analysis; non-morphological methods (dictionary and non-dictionary). Dictionary methods include methods based on the use of an unstructured list of all admissible word forms, and dictionary-free methods include methods that test part of word forms (methods of diagrams, trigrams, n-grams) and the method of hash codes; methods based on the application of algorithms for the classification and recognition of texts of the studied language, etc.

It should be noted that among these methods, the method of morphological analysis of word forms can be distinguished as the main method for checking spelling errors in texts. In this regard, within the framework of this work, it is planned to conduct detailed studies related to the development of a software data processing system for controlling the spelling of the Uzbek language (Latin, Cyrillic).

As you know, natural languages, in accordance with the principles of their structure, are divided into three groups: analytical, inflectional and agglutinative. Agglutinative languages, to which most of the Turkic languages belong, in particular, Uzbek, are characterized by an intermediate position between analytical and inflectional languages. On the one hand, they retain a very rich system of inflectional and word-generating affixes, but, on the other hand, this system is characterized by significant constructiveness and simplicity. However, despite the relative simplicity and constructiveness of the Turkic languages, the problems of developing spell-checkers for them are poorly studied by linguists, focusing mainly on European languages. We know the list of research teams engaged in the development of automated speech and text processing in natural languages.

The most advanced research in this area is being

carried out by Kemal Oflazer's research group in Turkey as part of a NATO project.

Numerous works of American linguists are devoted to the problem of dividing the flow of text into component parts. Xerox offers tools for automated processing of words in natural languages, based on the use of multi-level grammar description. Describing the vocabulary of the Uzbek language using the two-level grammar proposed by American linguists is a legitimate task within the framework of constructing a spelling control system for the modern literary Uzbek language.

The next problem is taking into account the current state of the Uzbek spelling - the simultaneous presence of two different graphic systems (Cyrillic and Latin) in it. The Uzbek spell-checker must be able to check spelling in both views and, if necessary, translate from one view to another. The most difficult and time-consuming task is to prepare a fairly representative dictionary of the roots of words of the modern Uzbek literary language. Here, the recently released AS of Uzbekistan spelling dictionaries in the Latin alphabet are of great help.

Thus, the study of the main aspects of theoretical and practical problems, on the results of which a computer system based on the morphological analysis of texts in the Uzbek language can be built, is an interesting and unsolved problem.

Assessment of the quality of texts. Along with the above, the task is related to checking the quality of the text, the solution of which will require the development of specific algorithms and programs for comparing the original and the checked text. You can offer several options for solving this problem: letter-by-letter, symbolic comparison; comparison of words; comparison of words with extraction of roots; comparison of strings; comparison of frequency characteristics of words, arrays, strings and pages; comparison of intervals of appearance of letters, symbols, words, lines in the page.

Thus, in this section, the main approaches to the creation of a computer system for monitoring and correcting errors in texts are proposed, tasks are formulated that determine the following areas of theoretical and practical research:

development of methods for analyzing probabilistic processes occurring in various conditions of transmission and processing of information and determining the optimal volume of the dictionary of word forms of the Uzbek language;

development of methods and algorithms for control and correction of texts based on: algorithms for optical text recognition; probabilistic model of making mistakes; arithmetic coding method; software methods of information control by linear, modular, planar summation and methods of morphological analysis based on a multilevel model of representation of word forms of the Uzbek language.

Models without dictionary morphological analysis

At the present stage of development of information technologies, the morphological component has become an integral part of intelligent information retrieval systems (ISS). In the 60-70s. XX century, all experimental research in the field of machine morphology began with the creation of a machine dictionary. There was no single generally accepted format and structure of such a dictionary. These circumstances had two consequences: firstly, all algorithms automatically became dictionary-dependent, and secondly, each algorithm was developed for a specific dictionary format.

The main problem in developing a machine-oriented algorithm for linguistic processors is the amount of initial data used by the program, that is, the amount of dictionaries that have to be compiled manually. Research in this area is aimed at minimizing baseline data. Works on morphology can be roughly divided into two categories:

- theoretical, in which descriptions of morphological laws and formal models of the morphology of natural language are presented;
- applied, dedicated to the description of software-implemented systems with a morphological module.

In theoretical works, multilevel formal models of morphology are built, mostly intended for synthesis. Such models of morphological synthesis imply the presence of large dictionaries with a complex structure. They describe a wide range of morphological phenomena. Many components of these models are redundant for machine analysis tasks (phonetic realization of a word, accent paradigm, a large number of derivational affixes).

Models that use a vocabulary are able to give a more complete analysis of a word form (i.e. operate with a large number of grammatical features).

The degree of accuracy of such an analysis is higher than in models that do not use a dictionary; however, in the space of real texts, systems that use a dictionary often fail. This is due to the fact that no complete dictionaries exist.

The vocabulary of the language is constantly updated, new words appear. Each subject area has its own terminology, its own subset of the language vocabulary, and it is impossible to include all existing terminology in the general dictionary, just as it is impossible to list all existing names and surnames that have regular declination.

The algorithms of programs that work without a dictionary use probabilistic-statistical methods and lexicons of suffixes or quasi-suffixes, bases or quasi-bases constructed empirically. The paper describes a working model of morphological analysis that does not require voluminous dictionaries of the foundations of open classes of words. The model was developed in line with engineering linguistics. The model uses the

following lexicons:

1. Lexicon of endings and reflexives;
2. Lexicon of suffixes;
3. Lexicon of quasi-roots;
4. Lexicon of prefixes;
5. Lexicon of bases;
6. Lexicon of basics.

Each unit of such a lexicon is assigned all possible grammatical characteristics of word forms, of which this unit can be a part. An example of a lexicon unit for quasi-roots:

-ni-
noun, 11, -e,
noun, 8, th,
verb, -th;
where 11, 8 is the type of declination.

Models of morphological analysis based on a dictionary of word forms

The analysis of word forms in non-vocabulary models is based on the rules of searching and combining units of different lexicons, which leads to the unification of hypotheses. Such an analysis does not use the capabilities of texts entering the system. In this regard, a method is proposed that is reduced to empirical compression of the original dictionary of word forms. For this, common letter chains in a variety of word forms are identified, and all possible meanings of the grammatical categories of these word forms are assigned to each chain of letters. Empirical compression of the grammatical dictionary of the Russian language leads to the creation of a large number of scattered lexicons of different structures, each of which requires a separate data reading procedure. This approach to morphological analysis cannot be called, in full measure, wordless.

A similar method is based on the description of probabilistic-statistical methods for creating auxiliary lexicons based on the original text corpus.

All algorithms of this kind have the same disadvantages:

• precise linguistic methods of analysis are not used;

• a large volume of lexicons;

• probabilistic statistical methods do not work well with a small sample.

The accuracy of this analysis is much lower than for systems that work with a dictionary. These algorithms do not allow you to choose unique grammatical characteristics, although in most cases they allow you to build a common basis or quasi-basis for a set of word forms and lemmatize a word form.

The freest form of analysis developed at the University of Chicago. The model allows, by statistical processing of a large array of texts, by analyzing the frequency of occurrence of a sequence of characters in word forms, to select a set of affixes and root morphemes that are relevant for a given language. The program works with most European languages, including Russian. The work was carried out within the framework of scientific research and did not receive applied implementation.

Algorithms of morphology are based on self-learning of the program on open arrays of real texts and combine two approaches: linguistic

- a formalized grammar for constructing morphological hypotheses and a mathematical one - a correlation method that allows unifying a morphological hypothesis.

Morphological analysis without a dictionary is a central component of the automatic text database (DB) indexing system implemented in Oracle DBMS. However, the output of the system is an automatically constructed grammatical dictionary of fundamentals and an associated document index intended for full-text search in the database. The essence of intelligence is the ability to make intelligent decisions in the absence of completeness of data and facts. The intelligence of the system increases with the decrease in the amount of static information used in the analysis.

In our case, we are talking about the use of linguistic information in morphological analysis in the tasks of automatic indexing of text databases. In this regard, let us highlight the main criteria that distinguish the majority of intelligent systems, which the projected text auto-indexing processor adheres to:

• The ability of the system to explain each step of the decisions made. The analysis does not use probabilistic and statistical methods.

• Using the rules and properties that characterize the given subject of analysis. To construct morphological hypotheses of word forms, a formalized grammar and the property of the Russian language are used that most of the grammatical categories in Russian are calculated from inflection.

• Modularity of the system, which provides effective change and replenishment of rules and properties, and also makes it possible to tune the analyzer to other natural languages with advanced morphology.

• Multiple interpretations. The analyzer leaves all the homonyms of the meanings of inflectional categories.

• Self-learning and a mechanism for correcting previously made incorrect decisions. The volume of the read texts replenishes the number of word forms used in the analysis process, thereby increasing the accuracy of morphological analysis and making it possible to correct incorrectly constructed bases and the meanings of their grammatical categories.

• Modeling of human intellectual behavior. In this case, we are talking about an attempt to emulate the thoughts of a person studying a foreign language, who is faced with the task of classifying the words of a given language, in conditions when he has at his disposal a large array of texts, some knowledge of the

morphology of the language and there is no dictionary of the language in which texts.

References:

1. Abdurokhmonov, F., et al. (1979). *Khozirgi Uzbek adabiy tili*. ed. Teacher. (pp.54-99). Tashkent.
2. Azlarov, E., et al. (1993). *Textbook of the Uzbek language*. ed. Teacher. (pp.45-100). Tashkent.
3. Anoshkina, Zh.G. (1995). *Morphological processor of the Russian language*. (pp.17-23). Almanac "Govor", Syktyvkar.
4. Apresyan, Yu. D., et al. (1989). *Linguistic support of the ETAP-2 system* ... (pp.65-87). Moscow: Nauka.
5. Akhatov, A.R., & Khimmatov, I.K. (2015). *Algorithms for controlling the reliability of electronic text processing in distance learning systems*. "Modern information and communication technologies in education: problems and solutions" ilmiy-amaliy konferentsi materialari, TATU Samaragand branches, 2015 yil 15-16 April. (pp.150-153). Samaraand.
6. Akhatov, A.R., & Himmatov, I.K. (n.d.). *Technology of control over the accuracy of processing electronic texts in automated linguistic systems.*
7. Burlak, S.A., & Starostin, S.A. (2001) *Introduction to linguistic comparative studies*. (pp.32-89). Moscow: Editorial URSS.
8. Wirth, N. (2001). *Algorithms and data structures*. (pp.35-150). SPb..
9. Gryaznukhina, T.A., et al. (1999). *Syntactic analysis of a scientific text on a computer*. (pp.15-86). K. // Scientific Thought.
10. Zhumanov, I.I., & Dzhuraev, M.K. (2004). A text correction method based on a probabilistic model of making mistakes. *In RZh "Vestnik TSTU"* No. 1, Tashkent, pp. 38-44.