DIVIDING UZBEK TEXTS INTO ELEMENTARY UNITS

Akhmedova Husniya Khusanovna
Lecturer, Tashkent University of Information Technologies named after Muhammad al Khorazmi, Tashkent, Uzbekistan
khusniya150586@gmail.com

Annotation: *Along with the development of information and communication technologies, the demand for means of storing, processing and transmitting data in the form of electronic text is also increasing. This article discusses issues of structural analysis and grouping of electronic texts in the Uzbek language. An algorithm for dividing electronic text into structural units has been developed and described.*

Аннотация: *Наряду с развитием информационно-коммуникационных технологий возрастает и спрос на средства хранения, обработки и передачи данных в виде электронного текста. В данной статье рассматриваются вопросы структурного анализа и группировки электронных текстов на узбекском языке. Разработан и описан алгоритм разделения электронного текста на структурные единицы.*

Аннотация: *Axborot-kommunikatsiya texnologiyalarining rivojlanishi bilan birga ma’lumotlarni saqlash, qayta ishlash va elektron matn ko‘rinishida uzatish vositalariga talab ham ortib bormoqda. Ushbu maqolada o‘zbek tilidagi elektron matnlarni tizimli tahlil qilish va guruhlash masalalari muhokama qilinadi. Elektron matnni strukturaviy birliklarga ajratish algoritmi ishlab chiqilgan va tavsiflangan.*

Key words: *electronic text, Uzbek language, algorithm, paragraph, sentence, elementary unit.*

Ключевые слова: *электронный текст, узбекский язык, алгоритм, абзац, предложение, элементарная единица.*

Kalit so‘zlar: *elektron matn, o‘zbek tili, algoritm, abzas, gap, elementar birlik*

INTRODUCTION

One of the most important tasks is the correct organization of the process of preliminary text processing when developing linguistic solutions that fully comply with the norms of the Uzbek literary language. Therefore, software that performs text preprocessing analyzes and groups the structure of electronic text and its semantic components according to their characteristics. Presentation of data in the form of text is one of the most widely used methods, and text is the most convenient form of information for computer processing. A text is an ordered collection of words that convey a specific meaning. Texts used and processed by computers are called electronic text. A lot of research has been carried out in the world on the organization of word processing and continues today. For example, in the book "125 Problems in Text Algorithms: With Solutions" by Maxime Crochemore, Thierry Lecroq and Wojciech Rytter, about the characteristics of elements that can be found in text files, their classification, approaches used in the organization of preprocessing, models and algorithms detailed information is provided [1]. V.V. Dikovitsky, M.G. Shishaev's research work entitled

"Обработка текстов естественного языка в моделях поисковых систем" talks about the computer linguistic model and methods of text preliminary processing and organization of information search [2].

Also, the textbook by E.I. Bolshakova, E.S. Klyshinsky, D.V. Lande et al. naimed (available only in Russian) "Автоматическая обработка текстов на естественном языке и компьютерная лингвистика" covers the main issues of computational linguistics: from the theory of linguistic and mathematical modeling to technological solutions [3]. Classification and clustering of textual information, fundamentals of fractal theory of textual information are considered. The list of such studies can be continued for a long time. The results of conducted and ongoing research can be seen in the capabilities of modern text editors used in computers. Encoding systems are used for electronic representation of text elements in computing machines. Such systems represent the binary code of any symbol that can be found in the electronic text [4,5].

Texts in the Uzbek language are expressed using letters and symbols of Latin graphics. Table 1 below lists the letters of the Latin alphabet.

Table-1. Latin alphabet of the uzbek language

№ tupe	1	2	3	4	5	6
letter	Aa	Bb	Dd	Ee	Ff	Gg
	Hh	Ii	Jj	Kk	Ll	Mm
	Nn	Oo	Pp	Qq	Rr	Ss
	Tt	Uu	Vv	Xx	Yy	Zz
letter+sign	O‘o‘	G‘g‘				
letter+letter	Shsh	Chch	ng			
sign	'					

Electron word processing refers to the input, editing, formatting and printing of text and documents using computers. This process can be complex depending on the creation, design, structure and other technical features of the text. Today, using modern technologies, it is possible to obtain useful information from electronic text, check it and process it accordingly. The basis of such technologies are the methods, models and algorithms used in the organization of text processing. [6,7].

Researchers such as A.Norov, Sh.Muradov, B.Akmuradov, U.Khamdamov, Dj.Elov, Dj.Sultanov, I.Narzullayev, M.Mukhiddinov have discussed a number of problems of organizing the processing of Uzbek texts and their computer-linguistic solutions in their works [5-10]. However, there are many issues that arise in the implementation of Uzbek word processing and are waiting for their solution. Small texts can be read and analyzed without using special tools. However, it takes a lot of time and resources to extract and analyze the necessary information from a large amount of text.

ALGORITHM FOR DIVIDING UZBEK WORDS INTO ELEMENTARY UNITS

Parsing the incoming text first begins by checking whether it is written in Latin. Since the system is designed only for texts in the Uzbek language written in the Latin alphabet, the process is not carried out for texts in which elements of the Latin alphabet are not observed. In order to clearly and efficiently organize the process, the composition and structure of the text is studied, that is, the number of headings, paragraphs, sentences and elementary units is determined [1].

Editing text paragraph by paragraph is carried out using a developed program for analyzing and dividing the text into blocks. A block diagram representation of this program is shown in Figure 1.[8-10]

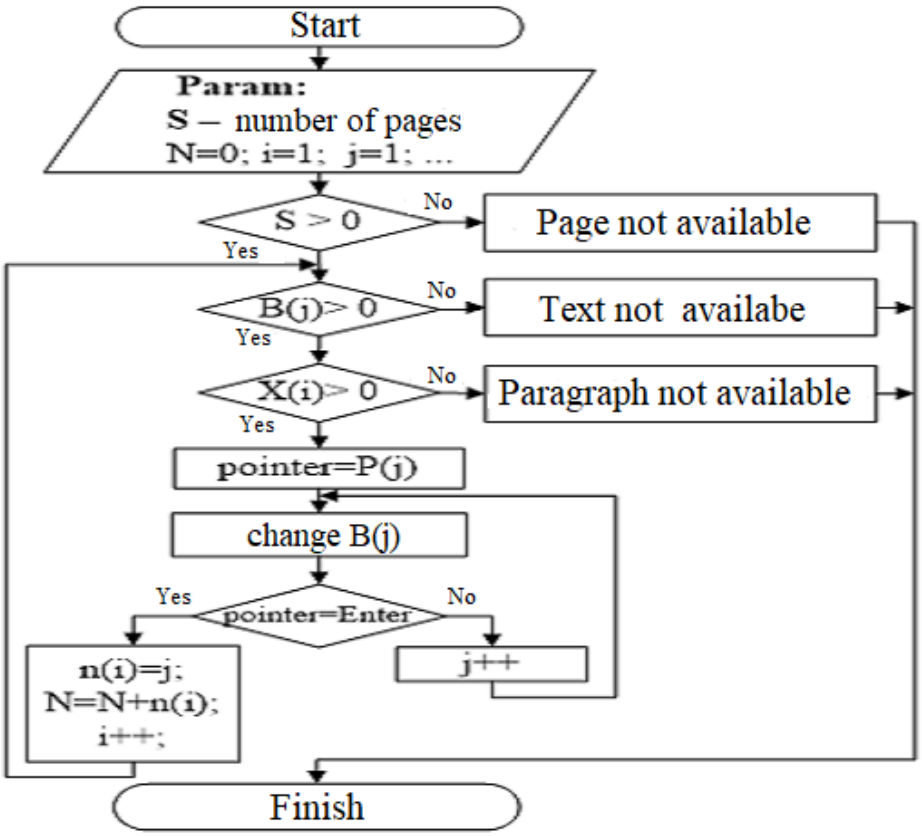

Figure 1. Algorithm analysis of electronic text and dividing edinits

where: X(i) – paragraphs; n(i) – number of units in the paragraph;

N – total number of units; P(j) – probes; B(j) represents units.

As you can see from the flowchart in Figure 1 above, the text pre-processing process starts from the first page of the file. The process flow can be described as follows.

- - the parameters necessary for editing the text are initially determined and announced;
- - at the next stage, the presence of pages in the file is checked;
- - at the next stage, the presence of text elements on the current page is checked;
- - at the next stage, the presence of a paragraph on the current page is checked;
- - the pointer separates the first unit after itself;

As mentioned above, the term "unit" is defined indicatively and can be any text element or sequence of elements[11].

Text units can be defined as a set of text elements located between any two paragraph marks, space and input characters, as well as additional dash, parallelism, proportionality, etc. symbols, which are considered structural spacing. In turn, such additional symbols are counted independently [3]. At this stage, the pointer then splits the character string from the current point to the first character or space after it and passes it as a unit to the next stage.

- B(i) unit received at the next stage is edited;

- at the next stage, the equality of the pointer to the Enter symbol is checked:

- at the last stage, statistical data is saved and the transition to the keying paragraph is made; Only one paragraph is processed in this last step of the paragraph-editing algorithm. By repeating the algorithm cyclically, it is possible to process all the paragraphs in the file [12,13].

One line consists of a sequence of several units. Then the total number of units (B) in one paragraph can be determined by the following expression 1:

(1)

where: Pl – l is the number of units in the paragraph, n is the number of lines.

Using this, it will be possible to determine the total number of units in the text using the second formula below.

(2)

where: P_{jami} is the total number of units in the text, m is the number of paragraphs

Based on the same rule, it is possible to determine the number of all units in the file.

CONCLUSION

Given the large number of pages and units contained in a large electronic text file, it is appropriate to implement the processing of such text editing algorithms on small components. According to this approach, i.e., after editing a part of the text, by switching to the keying part, it is possible to transfer the considered part to the next processing stage. This feature is especially important in modern real-time systems working with electronic text.

About this article, it is possible to divide all Uzbek words into syllables using the proposed algorithm. Because dialects are distinguished by the different pronunciation of words. However, the rules of syllable translation in literary language can be applied to all dialects. It is also possible to divide words learned from other languages into syllables using the proposed algorithm. In that case, it is important to remember that such words can lead to a loss of pronunciation in the language in which they are learned.Currently, created electronic dictionary is not to generalize and standardize all the words and terms used in the speech of ethnic Uzbeks living or working in Uzbekistan and other countries. Because, Uzbek language is a complex historical language with many dialects. In addition, changes have been made to the language standards of Uzbek peoples living outside the territory of Uzbekistan, depending on the ethnic composition of the population. This means that the electronic dictionary of Uzbek words created in this work includes only words and terms that comply with the standards of literary language, which are legally maintained in the Republic of Uzbekistan.

Finally, it can be concluded that many words can be expressed in Uzbek language using a small number of syllables, using the feature of articulation of words. This solution allows us to avoid the problem of memory, which is the main drawback of the concatenative method in the development of the Uzbek language text-to-speech synthesizer. Future work includes expanding the dictionary created above on demand and supply and forming a database of syllables accordingly, as well as the gradual improvement of the developed algorithm.

LITERATURE:

1. Y. Ning, S. He, Z. Wu, C. Xing, and L.-J. Zhang. A Review of Deep Learning Based Speech Synthesis, Applied Sciences, vol. 9, no. 19, pp. 4050, 2019. https://doi.org/10.3390/app9194050
2. J. A. E. Nogra. Text Analysis on Instagram Comments to Better Target Users with Product Advertisements, International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), vol. 9, no. 1.3, pp. 175-181, 2020. https://doi.org/10.30534/ijatcse/2020/2691.32020
3. J. Sotelo, S. Mehri, K. Kumar, J.F. Santos, K. Kastner, A. Courville, and Y. Bengio. Char2wav: End-to-end Speech Synthesis, In Proceedings of the International Conference on Learning Representations Workshop, Toulon, France, April 2017, pp. 24–26.
4. C. Pappas. Top 10 Text to Speech (TTS) Software for eLearning, 2019. Available online: https://elearningindustry.com/top-10-text-to-speech-ttssoftware-elearning (accessed on 22 June 2020).
5. Akmuradov B., Khamdamov U., Elov Dj., Sultanov Dj., Narzullayev I. Organization of initial text processing in the Uzbek language synthesizer // International Conference on Information Science and Communications Technologies (ICISCT 2021). 4-6 November, Tashkent – 2021. 5p.
6. Braunschweiler N., Buchholz S. Automatic sentence selection from speech corpora including diverse speech for improved HMM-TTS synthesis quality // INTERSPEECH 2011. 28-31 August 2011, Florence, Italy. P 1821-1824
7. Боярский К. К., Введение в компьютерную лингвистику // Учебное пособие, СПб: НИУ ИТМО, 2013. – 72 с
8. Цитульский А. М., Иванников А. В., Рогол И.С. Интеллектульный анализ текста // StudNet 2020. №6. С. 476-483.
9. Y. Li, T. Jianhua, H. Keikichi, X. Xiaoying, and L. Wei. Hierarchical stress modeling and generation in mandarin for expressive Text-to-Speech, Speech Communication, 72, pp. 59-73, 2015.
10. M.B. Akçay, and O. Kaya. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, Speech Communication, 116, pp. 56-76, 2020. https://doi.org/10.1016/j.specom.2019.12.001
11. V. R. Reddy, and K. Sreenivasa Rao. Prosody modeling for syllable based text-to-speech synthesis using feedforward neural networks, Neurocomputing, 171, pp. 1323-1334, 2016.

12. N. P. Narendra, and K. Sreenivasa Rao. Optimal weight tuning method for unit selection cost functions in syllable based text-to-speech synthesis, Applied Soft Computing, vol. 13, no. 2, pp. 773-781, 2013. https://doi.org/10.1016/j.asoc.2012.09.023

13. A.C. Pugach. Comparative analysis of speech synthesis methods, Young scientist, 26, pp. 154-156, 2016.