Natural Language Modeling Issue

Guli Toirova Ibragimovna

Associate Professor of the Department of Uzbek Linguistics, Bukhara State University PhD in philosophy, associate Professor. Uzbekistan

E-mail: tugulijon@mail.ru

Nargiza Jaxonova

1st year master's degree from Bukhara State University, Uzbekistan

ABSTRACT

Relevance: *In Uzbek linguistics, a number of studies have been carried out on automatic translation, the development of the linguistic foundations of the author's corpus, the processing of lexicographic texts and linguistic-statistical analysis. However, the processing of the Uzbek language as the language of the Internet: spelling, automatic processing and translation programs, search programs for various characters, text generation, the linguistic basis of the text corpus and national corpus, the technology of its software is not studied in any monograph. Discusses the transformation of language into the language of the Internet , computer technology, mathematical linguistics, its continuation and the formation and development of computer linguistics, in particular the question of modeling natural languages for artificial intelligence. The Uzbek National Corps plays an important role in enhancing the international status of the Uzbek language.*

Objective: *To emphasize the importance of linguistic modules, such as phonology, morphology and spelling, in the formation of the linguistic base of the national corpus of the Uzbek language.*

Methods: *The article uses rational-typological, comparative, meaningful, discursive methods of analysis.*

Results: *The article is scientifically substantiated by the need to create an algorithm for phonological, morphological and spelling rules for the formation of a lexical-grammatical code, one of the independent components of linguistic programs, a linguistic module and an algorithm and its types are analyzed. The need for algorithms for phonological, morphological and spelling rules for the formation of the lexical and grammatical code is scientifically substantiated. The importance of such linguistic modules as phonology, morphology and spelling in the formation of the linguistic base of the national corpus of the Uzbek language is emphasized.*

Conclusions: *Given the fact that raising the international status of the Uzbek language, bringing it to the level of the world language of communication, studying and teaching the Uzbek language abroad, expanding and honing the capabilities of our national language will be carried out directly through the national corpus, the practical significance of the work will be a key development factor and survival.*

Key words: *corpus, spelling module, morphological module, linguistic module, word-combination modules, word algorithm, formula algorithm, tabular algorithm, graphical algorithm*

I. Introduction

The term "linguistic module" plays an important role in the field of computer linguistics. For example, the conversion of natural language into a machine language , i.e. the development of ways to process text via a computer system. In this end, linguistic programs in other languages are being created. The linguistic module is an integral part of these linguistic programs. For example, if the lexical module is surrounded by a dictionary layer (words), the grammatical module edits symbols, punctuation, letters and other characters, the spelling rules of the spelling module, the morphological module analyzes words (from word to lexeme analysis) and the synthesis process (lexeme formation), the supersyntactic unit in the syntactic module-the interconnecti phenomenon.

II. Literature review

Analysis of the relevant literature. In her research, M. Abzalova notes: "In order to obtain realistic results in the development of a linguistic framework of word classes, first of all, the affixes that form them and their combinations are attached to words and are the best way to reach the linguistic base." We recommend using the following linguistic modules suggested by M. Abzalova in the formation of the Uzbek Language National Corps:

“The affixes added to the key words in the modulation of the noun category are defined as follows:

affix of affiliation: q_a= *-niki;*

affix of place : u_j= *-dagi;*

affix of limiting: ch_q[3]= {*-gacha, -kacha, -qacha*}*;*

affix of plural: Pl_a= *-lar;*

consonant affixes (with variants): k_a [7] = {-ning, -ni, -ga, -ka, -qa, -da, -dan}; possessive affixes: e_a [9] = {- m, -im, -ng, -ing, -lari, -miz, -imiz, -ngiz, -ingiz};

noun-forming affix: sh_y = -lik;

1st type affix of person-number category: sh_s1 [-man, -san, -miz, -siz; -simiz, -sisiz]

affixes: -mi, -chi, -gina, -kina, -qina, -dir, -u, -yu, -da, -a, -ya.

The following examples can be given to the module of attaching the given affixes to the core (A = base, N = derivative):1. N=A□□q_a; боланики= бола□□ники

2. N=A□□u_j; boladagi = bola□□dagi

3. N=A□□ch_a[1]; bolagacha= bola□□gacha

4. N=A□□ Pl _a; bolalar= bola □□lar

5. N=A□□k_a[7]; bolaning= bola □□ning

6. N=A□□e_a[6]; bolam= bola □m

7. N=A□□ k_a□□ e_a[6]; bolalarim= bola □□lar□□im

8. N=A□□ k_a[6]; bolamga = bola □□m□□ga

9. N=A□ Pl_a□e_a[6] □ k_a[6]; bolalarimga=bola□ Pl_a; lar□□e _a[6];m□k_a[7];ga

10. N=A□□ e_a[6] □□u_j; bolamdagi=bola□□m□□dagi.

The modulation continues in this order” [2].

In the process of creating a national corpus in the Uzbek language, an optimum version of M. Abzalova is being used. The algorithm of phonological , morphological and orthographic rules shall be established in order to form a lexical-grammatical code in the linguistic norms module of the Uzbek language phrases.

III.Methodology of research

What's the [6] algorithm? Algorithm, algorithm-a clear rule (program) for the execution of actions in a certain order that are used to solve problems of a particular type. One of the basic concepts for cybernetics and mathematics. The rule that performed four arithmetic operations on a decimal number system was called an algorithm in the Middle Ages. [15] The computer with its computing power is fast, clean, accurate and at the same time "completely incomprehensible"[7]. The idea that when we use it to solve a number of problems, the computer invents something on its own is a mistake, and a clear and complete instruction is needed for the computer to work. An algorithm is a rigidly set order that performs the action needed to produce the final result. This may sound strange, but we're always confronted with an algorithm in real life. An example of this is the use of a payphone, which includes a sequence of actions required for a successful phone call. The rules for the use of home appliances, etc., in a short, understandable way, tell us what to do in one way or another, and determine the algorithm of our actions. According to historians and mathematicians,[21] the word "algorithm" is derived from the name of our great ancestor Abu Abdullah Muhammad ibn Musa al-Khwarizmi, and his famous book "Kitab al-jabr wa al-muqabala" has given rise to another popular term "algebra." It is fair to say that the basic algorithm for the production of instructions is controlled in the process of computer-assisted activities. We can not, however, transfer our records directly from the algorithm to the computer, because they are written in a language that the computer does not understand, only people understand. For a computer to understand an algorithm, it is translated into a machine language, just as algorithms written in a machine language are called programs or computer programs. Important features of the optional algorithm: the accuracy of the algorithm - the value of each step, discreteness - the process of solving the problem can be divided into several simple steps (execution steps) so as not to cause difficulties for the computer or person, the publicity - usefulness of the algorithm - the end of the actions of the algorithm, which allows to obtain the desired result with the initial data in the final steps [20].

IV. Analysis and results

The national corpus of the Uzbek language is the lexical unit that exists in the Uzbek language, such as synonyms , antonyms, homonyms, assimilation words, hierarchies of words; it is necessary to be able to automatically analyze the morphological structure of the word, the construction of the word, the meaning of the word, its morphological features. In other words, in the process of composing, lemming, marking the corpus, it is necessary, on the basis of individual searches, to find and interpret those words which form part of the corpus in the texts. In order to do this, the above-mentioned algorithm, linguistic modeling, must be carried out. M. Abzalova 's research "Linguistic modules of the program for editing and analyzing texts in the Uzbek language"[2], A. Eshmominov 's research" Synonymous database of the Uzbek national corpus"[17], automatic analysis of the morphological characteristics of words. It is necessary to use some parts of Sh. Khamroeva 's research on "Linguistic bases for the creation of the author's corpus of the Uzbek language"[18], N. Abdurahmanova 's research on" Linguistic support for the program for the translation of English texts into Uzbek"[1].

“Dictionary of synonyms of Uzbek language”, “Explanatory dictionary of Uzbek words”, “Dictionary of obsolete words of Uzbek language”, “Dictionary of synonyms of Uzbek language”, “Dictionary of Uzbek words”, “Dictionary of synonyms of Uzbek language” "Dictionary of contradictory words of the Uzbek language", "Dictionary of word classification of the Uzbek language", "Educational etymological dictionary of the Uzbek language", "Educational toponymic dictionary of the Uzbek language" can serve as a linguistic support. Only such dictionaries are reworked, lemma words;

depending on the nature of the words, it is necessary to delimit their series and connect the members of the lemma series with each other. Only then can the revised dictionary form the basis of the software for the programmer.

In the final stage, texts prepared with meta-metric and morphological markings undergo several more automatic transformations. The following programs written in “Perl” language are used:

1) The converter converts the working format of the socket to the final format. The converter converts the morphological analysis in parentheses to the correct format <w lex =… ..gr =….>. It also checks for some spelling errors in order to further improve the quality of the search engine, translates the name into Latin, adds insufficient characters, identifies different forms of the verb;

2) **Semantic markup program (Semmarkup).** The program adds basic semantic characters to words using a special semantic dictionary. This method makes semantic search in the corpus much easier. The semantic dictionary is formalized in the form of a table, the first column contains a lexeme and a phrase, and the remaining columns contain semantic symbols. After the program compares the morphological characters of the word with the dictionary and finds similarities, it copies the semantic characters in the sem attribute of the <w> tag. In multi-character words, however, certain errors may occur in the semantic search;;

3) **Statistical programs (Gramstat, Metastat).** These programs are designed to collect statistics on the distribution of grammatical and metamaterial characters in texts. This method allows you to quickly find errors in the characters. The **gramstat** program allows distribution in morphological analysis (lexeme, word group, lexeme, and grammatical features of word form) for individual parts.

Most modern layout languages are based on SGML / XML, in which the defined text covers two parallel data layers: visible (text itself) and hidden (tagged or marked) [11]. In this case, the hidden part of the information is placed inside the text, but special markers <…> are included, which, in turn, separate it from the visible text. Unlike external methods of annotation writing (e.g. comments), the markup is always incorporated into the text and is an integral part of it. Subsequent levels of structural analysis are used by some corporations. In particular, some small corpuscles will be connected on the basis of a complete syntactic analysis. Such cases are usually characterized by a profoundly interpreted or syntactic structure. For example, a syntactic markup is like a large tree in itself. We know that manual analysis of texts is a valuable and time-consuming task. Currently , various software analysis tools are available on Russian and foreign sites, which are open (directly) accessible. They are individual, i.e. independent and subdivided into websites. In this case, it should be noted that in recent years, developers have focused on web applications. These systems have several advantages: the ability to analyze (mark) a single document by multiple users at once does not require the installation of additional software, but with the exception of the browser, access rights are limited, and the marking process can be monitored. In particular, let's pay attention to the process of analyzing the text from the story "Speech" by A.Qahhor. Text goes as following: “*You don't love me, you 're not happy with our marriage, I've been waiting until this hour, this minute, you haven't said a word, it's been a year since we put our heads on a pillow ...*

The speaker really forgot about it, but he was talking.”

The text mentioned above is distinguished by the following features:

1-table

№	**Type according to the sentence structure**

<table>
<tr><td>1.</td><td>[simple sentence]</td><td colspan="3">&lt;СГ&gt;, &lt;/СГ&gt;</td></tr>
<tr><td>2.</td><td>[уюшган гап]</td><td colspan="3">&lt;УГ&gt;, &lt;/УГ&gt;</td></tr>
<tr><td>3.</td><td>[complex sentence]</td><td colspan="3">&lt;ҚГ&gt;, &lt;/ҚГ&gt;</td></tr>
<tr><td><b>№</b></td><td colspan="4"><b>The type of sentence used for the purpose of expression</b></td></tr>
<tr><td>1.</td><td>[дарак гап]</td><td colspan="3">&lt;дг&gt;</td></tr>
<tr><td>2.</td><td>[сўроқ гап]</td><td colspan="3">&lt;сг&gt;</td></tr>
<tr><td>3.</td><td>[буйруқ гап]</td><td colspan="3">&lt;бг&gt;</td></tr>
<tr><td><b>№</b></td><td colspan="4"><b>Depending on whether or not the owner is represented in the linguistic construction of the speech</b></td></tr>
<tr><td>1</td><td>[эгали гап]</td><td colspan="3">&lt;Е+&gt;</td></tr>
<tr><td rowspan="2">2</td><td rowspan="2">[эгасиз гап]</td><td rowspan="2">&lt;Е-&gt;</td><td>[шахси номаълум гап]</td><td>&lt;ш.н.г&gt;</td></tr>
<tr><td>[атов гап]</td><td>&lt;а.г&gt;</td></tr>
<tr><td></td><td></td><td></td><td>[семантик-функционал шаклланган гап]</td><td>&lt;с.фш.г &gt;</td></tr>
<tr><td><b>№</b></td><td colspan="4"><b>According to the participation of the primary and secondary segment</b></td></tr>
<tr><td>1</td><td>[йиғиқ гап]</td><td colspan="3">&lt;йг&gt;</td></tr>
<tr><td>2</td><td>[ёйиқ гап]</td><td colspan="3">&lt;ёг&gt;</td></tr>
<tr><td><b>№</b></td><td colspan="4"><b>According to the presence of parts that do not make grammatical connection with the sentence</b></td></tr>
<tr><td>1</td><td>[ундалма]</td><td colspan="3">&lt;у&gt;, &lt;/у&gt;</td></tr>
<tr><td>2</td><td>[киритма]</td><td colspan="3">&lt;к&gt;, &lt;/к&gt;</td></tr>
</table>

The morphological marking system includes word, lemma, and tag. A word form is a morphological unit in a selected text. The first step in marking a word is to lemma it, that is, to bring out the lexeme form of the word. The most difficult step in marking inflected languages is lemmatization, that is, attaching the lexeme form of a word to a word as a tag. Because we know that in inflected languages the grammatical meaning of the word is mixed with the core of the word. Unlike inflected languages, the process of lemma in agglutinative language is much easier [4]. Initially, the analysis options for word forms are given in the form of a list, by selecting the correct option or editing the existing option. The editor makes it easy to navigate the text and make global changes and alterations. Thus, the marking application falls into a familiar environment and makes effective use of all the features of this editor. For the purpose of visual separation, different elements of the text are decorated in different colors and styles. Particularly,

— Analysis of the layout and the command variant is formalized in the form of hidden text and is usually not visible in normal mode;

— word forms are formalized in different colors depending on the number of analysis options: zero, one or more.

The grammatically impersonal part of the word is the same as the stem or base lemma. The mark is given in the character <*> of the lemma. If the lemma in all the word categories is based on this principle, that is, the principle that "the root part of the word is equal to the lemma," the verb lemma II in the verb group is given in the form of an imperative mood. In dictionary articles, the verb is given in the form of an action name: <go>. However, this form is not appropriate for the corpus because the text in the corpus is searching for the <bar> form, not the <go> form of the word. The verb lemma is therefore given as <taught>, not <be>, shown as <blind>, received as <received> [17]. The marking process requires writing 5 to 10, sometimes even more, morphological tags (comments) for each word.

The main advantage of SGML / XML compared to other layout languages (TEX, RTF) is that it has strict syntax of markup commands, differentiating attributes and elements, clear indication of element boundaries, self-documentation, automatic verification of grammatically correct entry.

The most authoritative standards for corpus data encoding are: TEI (Text Encoding Initiative)[5], CES (XML Corpus Encoding Standard)[8], EAGLES (European Advisory Group on Language Engineering Standards)[10]. In particular, TEI is recognized as a well-developed standard, defining the rules for the expression of different types of texts and textual information elements, with particular emphasis on: structure, title, style of speech (prose, poetry , drama), pages, quotations, footnotes or links (footnotes, comments), corrections, tables, formulas, specific characters (characters), linguistic annotations, etc. The special title of the standard shall be subject to the rules for the coding of the case. Although TEI is not specifically tailored for corpus applications, it often works in conjunction with similar standards. For example, the British National Corpus (BNC), the Czech National Corps, the Hungarian National Corps, etc. The XCES standard is an advanced application of TEI, designed solely for the corpus and intended to identify specific labels specific to the corpus.

But when we studied the TEI and XCES universal standards in detail, we found that they were too complex, unnecessary, and inconvenient for text mass marking. The full provisions of the TEI are very broad and not always reasonable, and it is therefore difficult enough to comply with all the requirements of this standard. The format is not compact, and the size of the content is usually increased. The format loses its clarity function, for example, it is suggested that meta-attributes be written in the form of text in the tag, so that when the markup is removed, the original text returns to its original state, error occurs.

You can also restrict yourself to TEI applications by rejecting "redundant" tags. The minimum set of tags is selected from the TEI to represent the body: <text> -text, <p> -header, <s> -word, <w> -word, and morphological analysis is written in the form of <w ana = ...> attribute. However, such an appearance does not fully comply with the standard of the housing layout. This view is reminiscent of a simplified HTML version.

The corpus format has a number of HTML languages, with some special tags attached for linguistic units. This format specifies the coding requirements for important text information and includes:

1) meta text attributes;

2) text structure elements (title, paragraph, poems, footnote or link (footnotes, comments) and tables at the bottom of the page);

3) linguistic units (sentences, words);

4) lexical information (grammatical, semantic signs);

5) text formatting parameters, special characters, etc [20].

Meta text attributes are written in texts in different situations, so that steps 2 and 3 can be done in parallel or arbitrarily. But the text must have the name of the file identified and recorded. It does not perform any actions, such as renaming a single connection or file, as such actions could disrupt the operation of the entire system. For the purpose of storing metadata, simple Excel spreadsheets with a predefined structure are used, with the first column containing the name of the file (clearly specified path) and the other columns with metamata attributes and process information. This allows you to use Excel's built-in tools effectively and makes the search engine much easier. For example, search, filtering, analysis and data processing (to-do list, auto-filling, statistics). In this case, the tables must be stored in a text format, and this format must be understood by Excel. This allows the file stored in the spreadsheet view to accept not only Excel but also other spreadsheet programs and increase the runtime efficiency.

Theoretically, metadata can be stored separately from each text, but according to the HTML rules, the data must be stored in the file header so that the Yandex-server can index the data. When storing metadata in separate memory, there is always a problem of synchronization, meta-tables, and text interactions with each other.

V. Suggestions

The following methods are used to store metadata in separate memory:

1) The *metas* table creates meta-table headers by collecting meta-text attributes from the file headers. In Excel, it can be modified manually. At the initial processing stage, some metadata can be added to the text, such as the author's name, title and date of creation. At the final stage, the Metas.bat program collects all attributes and completes the verification phase.

2) Meta.txt takes the meta text attributes from the modified meta-tables and transfers them to the existing text. This program checks the availability of the file and updates the title. In the tables, most attribute actions are separated by a" "symbol. When the text is changed, each action will appear as a separate attribute. Metamata attributes can therefore move freely between text and meta-tables. Meta-metric, on the other hand, will need to be carried out interactively with several cycles of verification.

3) MetaTest checks the accuracy of the meta-table. In this case, the actions of the attribute in the normative table are compared with those shown in the templates. The program identifies incorrect actions with a "#" character and can be checked and corrected manually.

All the above programs are done in Perl.

At the final stage of processing, texts prepared with meta-metric and morphological markings undergo several more automatic transformations. The converter checks for some markup errors in order to further improve the quality of the search engine by converting the morphological analysis in parentheses to the correct format <w lex =… ..gr =….>.

VI. Conclusion / Recommendations

In conclusion, it should be noted that the role of linguistic modulation in the formation of the national body 's linguistic base is incomparable. It is therefore necessary to create an algorithm as a basis for the production of controlled instructions in the computer process. It is important to develop specific linguistic module forms by marking each word group in the development of a morphological marking algorithm.

Given that increasing the international status of the Uzbek language, raising it to the level of a world language of communication, learning and teaching Uzbek abroad, and expanding and polishing the capabilities of our national language directly through the national body, the practical significance of the work will be a key factor for development and survival.

References:

1. Abduraxmonova N.Z. Linguistic support of the program for translating English texts into Uzbek (on the example of simple sentences): Doctor of Philosophy (PhD) il dis. aftoref. - Tashkent, 2018.
2. Abjalova M. Linguistic modules of the program of editing and analyzing texts in the Uzbek language (for the program of editing texts in official and scientific style): Doctor of Philosophy (PhD)… dis. – Fergana, 2019. – P.22.
3. Avliyokulov N.X. Technology of modular teaching of professional sciences. - T.: Yangi asr avlodi, 2004. –106 p Stepanov A.N. 6.3. Archiving of file objects // Informatics: basic course: for students of humanities specialties of universities. - Peter, 2010. - 719 p.
4. Vanyushkin A. S., Grashchenko L. A. Assessment of algorithms for the selection of key words: tools and resources // New information technologies in automated systems. - 2017. - № 20. - S.. 95–102.
5. Zakharov V.P. Corpus Linguistics: Uchebno-metod. posobie. - SPb., 2005. - 48 p.
6. Kasyanov V. N., Kasyanova E.V. Introduction to programming. - http://pco.iis.nsk.su/ICP
7. Kasyanova E.V. Yazyk programming Zonnon for platforms .NET // Programmnye sredstva i matematicheskie osnovy informatiki. - Novosibirsk: ISI SO RAN, 2004. - P.189–205.
8. Kutuzov A.B. Corpus linguistics. - (Electronic resource): License Creative commons Attribution Share-Alike 3.0 Unported (Electronic resource) - //lab314.brsu.by/kmp-lite/kmp-video/CL/CorporeLingva.pdf
9. Manturov O.V. and dr. Explanatory dictionary of mathematical terms. –M .: Prosveshchenie, 1965. - 509 p.
10. Melchuk I.A. Poryadok slov pri avtomaticheskom sinteze russkogo slova (predvaritelnыe soobshcheniya) // Nauchno –texnicheskaya informatsiya. 1985, №12. -S.12-36.
11. Nedoshivina E.V. Programs for working with corpus texts: a review of the main corpus managers. Uchebno-metodicheskoe posobie. - St. Petersburg. - 2006. 26 p.
12. Safarova R.G. and b. Classification of pedagogical technologies used in the process of modular teaching in general secondary schools. / Methodical manual. - T .: State Scientific Publishing House "National Encyclopedia of Uzbekistan", 2016. –176 p.
13. Stepanov A.N. 6.3. Archiving of file objects // Informatics: basic course: for students of humanities specialties of universities. - Peter, 2010. - 719 p.
14. Explanatory dictionary on theoretical mechanics. –M .: MFTI. 2007. – 68 p.
15. Toirova G. About the technological process of creating a national corps. // Foreign languages in Uzbekistan. Electronic scientific-methodical journal. - Tashkent. 2020, № 2 (31), –B.57– 64. https://journal.fledu.uz/uz/ 2-31-2020
16. National encyclopedia of Uzbekistan. 5 volumes. Volume 1 - Tashkent: State Scientific Publishing House of the National Encyclopedia of Uzbekistan, –2006. – B.201.

17. Eshmo'minov A. Dictionary of synonyms of the National Corps of the Uzbek language: Doctor of Philosophy (PhD) in Ph.D. aftoref. - Karshi, 2019.
18. Тоирова Г. Важность интерфейса в создание корпуса. International Scientific Journal «Internauka», *// Международный научный журнал «Интернаука». – 2020. – №7.* Онлайн *журнал.* https://doi.org/10.25313/2520-2057-2020-7-5944
19. 19Toirova G. The Role Of Setting In Linguistic Modeling. //International Multilingual Journal of Science and Technology. ISSN: 2528-9810 Vol. 4 Issue 9, September – 2019,-P.722-723 http://imjst.org/index.php/vol-4-issue-9-september-2019/
20. Fries Ch.C. The structure of English. An introduction to the construction of English sentences. - L., 1969. – S.98.
21. Zemanek H. Lecture Notes in Computer Sciece 122 (1981), 1-81 [elek.res.] Http://elganzua124.github.io/ taocp / OEBPS / Text / ch01.html