International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor (2018): 7.426

The Linguistic Norms of the Grapheme Editing Module

Abzhalova Manzura Abdurashetovna

Head of the Scientific – Methodical Department of the Navoi State Mining Institute

Abstract: ***Text processing systems in Uzbek are currently one of the most perspective areas of informational technology. From such kind of systems as automatic analysis system is a complex of multistage process. This article discusses the first phase of the graphic arithmetic. The article describes the functions of the graph paper and focuses on the basis of the graphics modification modules.***

Keywords: automatic analysis, grapheme analysis, linguistic module, word processing, linguistic base.

1. Introduction

In recent years automatic processing of texts is one of the important and fast developing directions of computer linguistics. [1] In computer technology, which has become an integral part of our daily lives, text is essential. Therefore, the members of practical linguistics from the beginning started creating new programs which automatically translates texts from one or several languages into another ones analyzing from morphological, syntactic and semantic point of view.

Currently in order to enrich and to perfect linguistic base of programs connected with texts, scientific and practical researches are on the process.[2]

In text processing automatic edition and analysis are very important. This type of computer based analysis is characterized by a time consuming linguistic analysis and material savings comparing to human based analysis. In addition, automated analysis will also help to increase the user's readiness during the same process of operation. The orthographic, syntactic or semantic error analysis application that is typed by the logger will instantly identify and write the correct writing options. The word *"хоким"* must be inserted, but the user of the program writes the word in the form of the "*хоким",* so that the program lets the user judge and replace the word as a result of the linguistic modules it has. As a result, the user realizes that the word "xokim" does not exist in the Uzbek dictionary. Also, the analysis of application not only checks the correctness of the words, but also monitors the signs being used in their place. Let's say, *Dear friend, do not waste your luck!* The sentence is grammatically correct, but the comma is a space blank after the word *of my friend*. The linguistic edtiting program is built on these minor errors.

The following linguistic modules should be developed for the perfect text editing program [3],[4]
1) Graphical analysis modules.
2) Morphological analysis modules.
3) Syntax analysis modules.
4) Semantic analysis modules.

The graphical analysis is considered as the first step to analyze text linguistically. At this stage, the paragraphs in the text, word, phrase, number, punctuation, and other special symbolic characters are identified.

Grapheme is written text (letters, punctuation). It is called token, tokenization in the English language. The purpose of the graphical analysis phase is to identify and classify the smallest units in the text. Such units include: word, paragraph, punctuation marks, dates symbols of monetary units, word combination numbers, IP addresses, and file names, phone numbers.

Graphics analysis performs the following functions:
1) Classify text flow into words and their classification. These units will be commented on for further analysis during the process. For example, AA-word consists only of the initials. Aa – word, as the head begins with the capital letter. To do this, capital and lower case letters are included in the application module;
2) Frequency editing of words outlined in the text;
3) Combines some combinations of words into larger units- "persuasive phrases" (and signs the beginning and the end of the phrases IB1 …IB2);
4) Gives a special character to words(hidden, vague words);
5) Calculates and controls the places taken;
6) Identifies the paragraph and puts the numbers;
7) Defines abbreviations, acronyms.

To carry out these tasks, it is required to develop a formal approach based on linguistics methods.
The following serve as grapheme edit sources
- Glossary of abbreviations;
- Dictionary of graphic characters;
- The rules of the use of linguistic punctuation and their place in the text;
- Personal names and animal nicknames dictionary;
- Glossary of terms written in a foreign language.

Thus, the basis of the graphics analysis modules consists of the following signs and combinations:
1) The numbers from 0-9 (the rest complex numbers are included in the algorithm) the roman numbers.
2) Capital and lower case letters of the English, Latin and Cyrillic alphabet.
3) Punctuation marks: dot, comma, quotes, brackets, hyphens, semi colon, dotted comma, suspension points, question and exclamation mark.

4) Mathematical expressions: Add (+), subtraction (-), division (:), multiplication(x; *), bracket and its form (),[]; infinity sign (∞), elevation (x^2), umbilical(√) equality (=) and inequality (≠), large (>) and small (<)symbols. The algorithm is justified by the fact that they are just numbers.
5) The hyphen (–) and dash (-) are signed as a different characters
6) In the Uzbek language writing based on Latin graphics the function of "tutuq belgisi (‘)" hard sign is introduced.
7) Diacritical marks. While using the Latin alphabet in writing, the algorithm is replaced by an asterisk (‘) on the "o‘ " and " g‘ " letters which is a mistake and warns text writers to of this error.
8) Attribution signatures: digital ($^{2;3;\cdots}$) and stars (*)
9) Special characters are: ∧, ∨, %, &, \, {}_ , №
10) Currency symbols:
11) Abbreviations: *НДКИ, ТошМИ, ЮНЕСКО, СамДУ(* NSMI, TMI, UNESKO SSU) and etc .

Rule: the writing of the abbreviations:
1) Are written only by capital letters, after capital letters the punctuation (.) dot and space are not put: USA, NSMI, UNN and etc;
2) The first syllable of the word and initials of the next words are taken : *ЎзМУ, СамДУ, БухДУ, ЎзФА and etc.*;
3) Part of the first word and full form of the next word are taken: *мединститут, сантехник, агитпоезд*.
4) The first part of the word is taken*:* химфак, ижрокўм.

Main part of the signs in the text includes the alphabet letters of a particular natural language. When inserting words in characters of a certain alphabet (Cyrillic, Latin or any other) other elements might be included in the words. For example, *maktab*, *makr*, *garb*, *akuiya*. Grapheme analysis searches for the mistakes and transfers them to the next (morphological) stage. In some cases there are special signs in texts, such as ©, *, &, №, &, <, >. Such signs are stored in the special data store called "Special signs - SpS" ("махсус белгилар – МахБ") where every sign has its own mission. They even include 'Space' and 'Paragraph', and the sign showing the end of the line. For example, the sign system ("lexeme"), - can be expressed as follows:
(- opening parentheses
" – opening quotation mark
Lexeme- lexical unit
" – closing quotation mark
) – closing parentheses
, - comma
- hyphen

Such notes are essential for the place and importance of the signs in syntactic or semantic analysis. Moreover, there are such cases in writing, which cannot be neglected. For instance, words might be written with a space between each other like in *A P И З A* (A P P L I C A T I O N). If grapheme analysis modules do not include any information on the above given case, the application will accept the letters in *A P И З A* as individual characters rather than constituents of a word.

Likewise, comma or point used to separate whole numbers in decimals might not serve to compart but accept it as a whole.

1.25 } whole numbers
5,5 }

We can observe the diacritic marks in written in Latin alphabet Uzbek texts be used differently.

Correct	Wrong
O`o` G`g`	O‘ o‘
O′o′ G′g′	G‘ g‘
O'o' G'g'	
fe`l	fe'l
a‘lo	a'lo

It's a pity that these cases are often met in mass media and advertisements. Every mark has its meaning and role in a text. The proper use of the diacritic marks (‘- the reverse of the hard sign) to express the letters o‘ and g‘ in Latin alphabet is explained in the "Orthographic rules of the Uzbek language based on Latin alphabet". The tasks of the hard sign (one of them is to prolong a vowel when used after a vowel, as in she'r, ma'no; another is to compart syllables when used after a consonant, as in sur'at, qit'a) show that it cannot be used interchangeably with diacritic mark.

Every case, as in the above given example, is formed as the types of text elements in creating modules of the grapheme analysis stage of a language processing application. Below are their types and elements:

ЛекБ – lexical unit, a lexeme made up of letters of a particular language (as in *volida, buyruq, yurak, hayot, oldin*);
ЧетЛ – a lexeme of a foreign language (*приказ, бюро*);
РБ – numeric unit (*1986, 18/04/2012, 5.05, 19,25*)
ХРБ – letter-numeric unit (*Боинг*-767, СУ-26, "Келажак овози – 2018", "Йил аёли – 2019")
АББР – abbreviation (*БМТ, МДҲ, ЎзР, НДКИ, ЎзМУ*)
КБ – shortenings (*ва ҳок., ва бошқ., м-н:*)

The following signs are included in language processing application database:

Concept or term	Sign
Upper case	Â
Lower case	Ă
Space	#
Indent / Paragraph	¶
Punctuation marks	

Signs correction, seemed unimportant at first sight, has a great role in writing text orthographically. Any negligible linguistic error may distort the meaning of the text. Therefore, in creating linguistic supply for the application analyzing and correcting Uzbek texts a profound attention is paid to linguistic signs making up grapheme analysis modules that scan the graphemes and shortened words written in Cyrillic and Latin. Grapheme analysis is considered to be the initial stage of multilevel analysis application, which serves for the next stages to be effective and full-fledged.

2. Conclusion and Recommendations

The following conclusion is made as a result of the study of stepwise work and setting-up guidance of foreign linguistic analysis application:

- The role and importance of linguistic module is great in creation of linguistic processor. Therefore, in Uzbek computer linguistics the establishment of thorough modules serves as foundation for the formation of text processing linguistic application.
- At the initial stage of the introduction of the Uzbek language, for the ideal and effective work of lingua system analyzing and correcting texts of official and scientific style the system is modeled by dividing it into grapheme, morpheme and syntax analyzing stages.
- At the grapheme analyzing stage elements are labeled.
- The formation of grapheme analysis module gives an opportunity to achieve higher results in forming language system.

References

[1] *Peter Jackson, Isabelle Moulinier*. Natural Language Processing for Online Applications. — John Benjamins Publishing, 2002.
[2] *Bolshakov I.A., Gelbukh A.* Computational Linguistics. Models, Resources, Applications. — IPNUNAM-FCE, 2004, -PP.186.
[3] *Большакова Е.И. и др.* Автоматическая обработка текстов на естественном языке и компьютерная лингвистика (учеб. пособие) — М., 2011. -С.106.
[4] *Леонтьева Н.Н.* Автоматическое понимание текстов: системы, модели, ресурсы. – М. 2006.