ANALYSIS OF SENTENCES AND WORDS USED IN AZERBAIJANI TEXTS

Sardar Gasimov[1], Ibrahim Ibrahimov[2]

[1]Azerbaijan State Oil Academy, Baku, Azerbaijan, *qs-114@mail.ru*
[2]Azerbaijan State Oil Academy, Baku, Azerbaijan,ibrahim757-6@mail.ru

The work presented is devoted to the development of software and to the statistical investigation of the lengths of sentences and words used in arbitrary Azerbaijani texts.

The major purpose of the work is to find out words of what length are used most of all in the works of various authors by analyzing words and sentences. And this in turn allows us to determine the peculiar "Writing style" of each author on the basis of the lengths of words. The purpose of studying this "Writing style" as a feature is to make use of it in recognition systems which will be developed in the future [1].

We often come across such a case when, suppose, some text was found, but the author of the text is unknown. Then by conducting such analysis we could say to which author the text belongs. This as well may be interesting to authors themselves to find out words of what length they use most of all. Thus the development of such software and its analysis is of grand practical importance.

If we conduct an analysis on the different works of same author, we will see that the diagrams obtained are either similar, or slightly different. Note that this difference manifests itself not in visual view, but only in the percentage ratio of words. For example, let us look at the analysis of words of M.F.Akhundov's works "Aldanmish Kevakib" and "Haji Gara" in the following diagrams:

Figure 1. Diagram 1. Analysis of words from M.F.Akhundov's work "Aldanmish Kevakib".

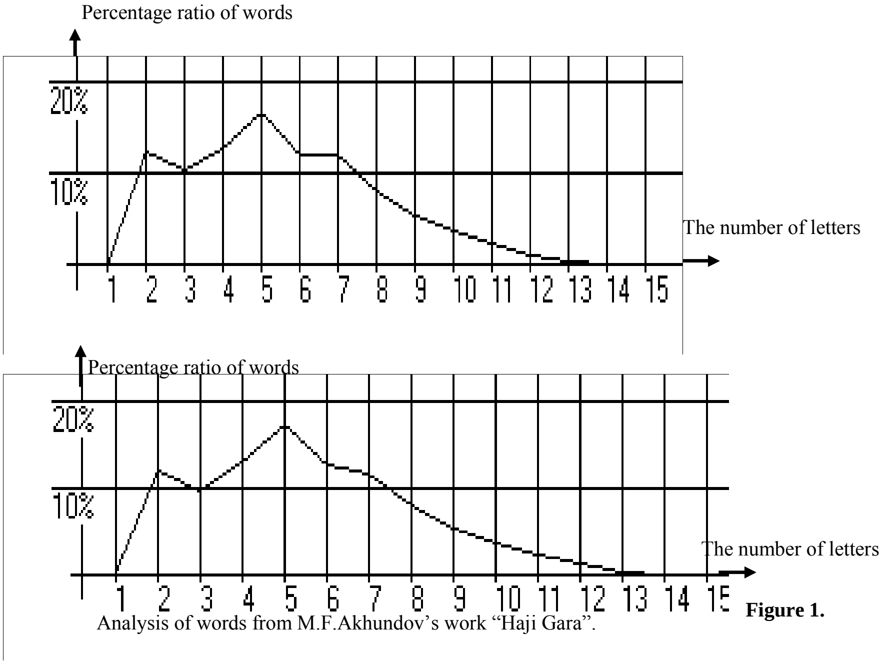
Y axis – Percentage ratio of words

X axis – The number of letters

Figure 1. Diagram 2. Analysis of words from M.F.Akhundov's work "Haji Gara".

Y axis – Percentage ratio of words

X axis – The number of letters



Analysis of words from M.F.Akhundov's work "Haji Gara".

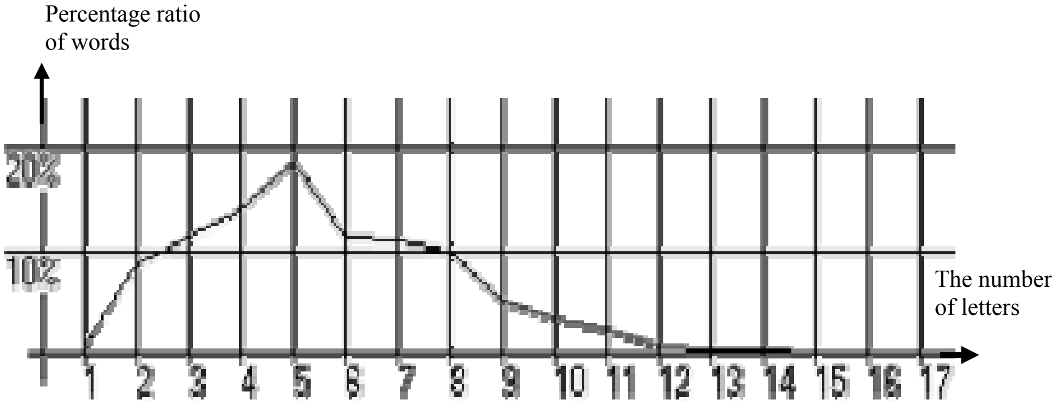
Figure 1.

In order to investigate the “Writing style”, we can make use of such notions from Statistical Analysis as expectation, dispersion, moments of the second and higher orders, etc. If we designate the length of words used in the author’s text by ξ, then the expectation and dispersion corresponding to fig. 1 will be as follows:

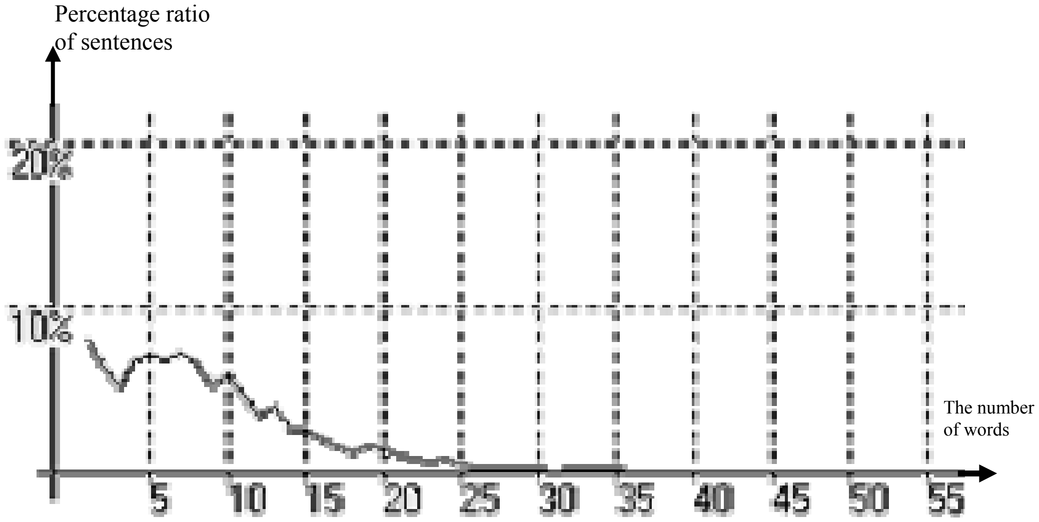
$$M\xi = 5{,}632 \text{ and } D(\xi) = 2{,}642 .$$

The same attributes with regard to Anar’s work “Dantenin yubileyi” are as follows (see fig. 2):

$$M\xi = 5{,}644 \text{ and } D(\xi) = 2{,}536 .$$



Analysis of words from Anar’s work “Dantenin yubileyi”.



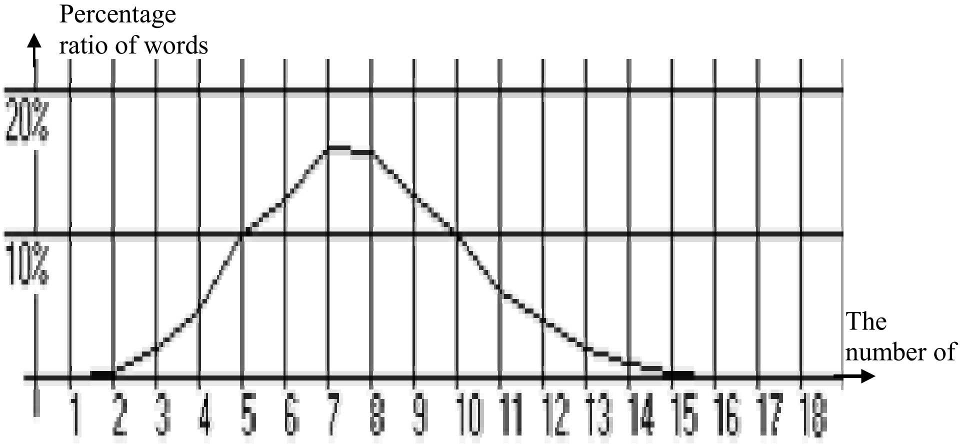
Analysis of sentences from Anar’s work “Dantenin yubileyi”.

Figure 2

Figure 2. Diagram 1. Analysis of words from Anar’s work “Dantenin yubileyi”.
Y axis – Percentage ratio of words
X axis – The number of letters
Figure 2. Diagram 2. Analysis of sentences from Anar’s work “Dantenin yubileyi”.
Y axis – Percentage ratio of sentences
X axis – The number of words

Note that the graphics of the analyses of Suleiman Sane Akhundov’s work “Gan Bulaghi” and Anar’s work “Dantenin yubileyi” are close to each other. Moreover, the two authors use words of five letters most of all. But there are sharp differences in the percentage ratio of words consisting of seven, eight and nine letters. These analyses are different on the lengths of sentences as well, i.e. the writer Anar most of all uses sentences of 5-10 words, but S.S.Akhundov most of all used sentences of 15-20 words. And this in turn shows once more that each author has their peculiar writing style.

If we look at the analysis of Azerbaijan language dictionary (see fig. 3), we will see that there are most of all words of seven letters. It is clear from fig. 3 that the percentage ratio increases steadily till words of seven letters and then decreases. But the results of the analyses conducted show that authors most of all use words of five letters.



Analysis of words from the Azerbaijan language dictionary.

Figure 3.

Figure 3. Analysis of words from the Azerbaijan language dictionary.
Y axis is percentage ratio of words,
X axis is the number of letters.

The software developed in Borland Delphi 7 programming environment allows analyzing arbitrary texts printed in Azerbaijan language using Latin alphabet [2].

The software was developed in such a way that it would be possible to change the alphabet; therefore it is possible to conduct analysis for any language and the corresponding alphabet.

The results of the analysis conducted can be used in neuro-fuzzy recognition systems of an author. At this time it is possible to enter into the system the words and set of words used most often, which is an important feature of "Writing style".

Currently, we are conducting works on the application of neuro-networks and the development of software for recognition systems.

Literature

1. Michael Huth, Mark Ryan. Logic in Computer Science: Modelling and Reasoning about Systems. Cambridge University Press, 2004, 427 p., www.cs.bham.ac.uk/research/projects/lics/
2. Сухарев М. - Основы Delphi. Профессиональный подход, Издательство: Наука и Техника, 2004, 603 с., www.booksgid.com/programmer/4355-.html