www.dergipark.gov.tr
ISSN:2148-3736
El-Cezerî Fen ve Mühendislik Dergisi
Cilt: 5, No: 3, 2018 (888-903)

El-Cezerî Journal of Science and Engineering
Vol: 5, No: 3, 2018 (888-903)
DOI : 10.31202/ecjse.442153

ECJSE

Research Paper / Makale

Mathematical and Information Models for Evaluating Readability of Texts in Azerbaijani Language

Ismayil SADIGOV

Institute of Information Technology of the Azerbaijan National Academy of Sciences, Baku/ AZERBAIJAN
ismayil.sadigov@gmail.com

Received/Geliş: .10.07.2018 **Revised/Düzeltme:** 28.08.2018 **Accepted/Kabul:** 29.09.2018

Abstract: The article describes software that calculates the quantitative characteristics of texts in the Azerbaijani language and determines the readability of texts based on the Flesch reading-ease formula and Flesch-Kincaid grade-level formula adapted for these texts. The problems and ways to solve them, related to the calculation of certain parameters of texts explain by the examples.

Keywords: Text complexity, readability formulas, Flesch reading-ease formula, Flesch-Kincaid grade-level formula.

Azerbaycan Türkcesindeki Metinlerin Okunabilirliğini Değerlendirmek için Matematiksel ve Bilgi Modelleri

Öz: Bu makalede, Azerbaycan türkcesindeki metinlerin niceliksel özelliklerini hesaplayan ve bu metinler için uyarlanmış olan Flesch okuma kolaylığı ve Flesch-Kincaid eğitim düzeyi formülleri esasında metinlerin okunabilirliğini tanımlayan yazılım anlatılmıştır. Metinlerin bazı parametrelerinin hesaplanmasında karşılaşılan problemler ve bunları çözme yolları örneklerle açıklanmıştır.

Anahtar kelimeler: Metinlerin zorluğu, okunabilirlik formülleri, Flesch okuma kolaylığı formülü, Flesch-Kincaid eğitim düzeyi formülü.

1. Introduction

The idea of automatic processing and analysis of texts has emerged in the early ages of the development of computational techniques. In the 50s of the last century, in the US, USSR, Great Britain, France and some other countries was founded computational linguistics. In the 1960s, in the United States was developed the first frequency dictionary of the English language using computers. [1]

The interdisciplinary field that studies the application of mathematical models to describe linguistic regularities is called computational linguistics. Computational linguistics can be divided into two major directions. One of them explores the use of computer technologies for linguistic researches, i.e. application of known mathematical methods (e.g. statistical processing) to identify regularities. The detected regularities are used in another direction – in understanding texts written in natural language, in other words, creating mathematical models for solving linguistic issues and studying applications that work on the basis of these models. This area of computational linguistics is closely

How to cite this article
Sadıgov I., "*Mathematical and information models for evaluating readability of texts in Azerbaijani language",* El-Cezerî Journal of Science and Engineering, 2018, 5(3); 888-903.

Bu makaleye atıf yapmak için
Sadıgov I., *"Azerbaycan Türkcesindeki Metinlerin Okunabilirliğini Değerlendirmek için Matematiksel ve Bilgi Modelleri",* El-Cezerî Fen ve Mühendislik Dergisi 2018, 5(3); 888-903.

related to the field of artificial intelligence, which is developing natural language text processing systems. [2]

Figure 1 shows the general scheme of text processing. Regardless of what language the given text is, its analysis goes through the same stages. The first two stages (splitting text into sentences and words) are almost identical to the majority of natural languages. The only difference may be related to abbreviations and punctuation marks that indicate the end of sentence. The next two stages (characteristics of separate words and syntactic analysis), on the contrary, depend heavily on the chosen natural language. [3]

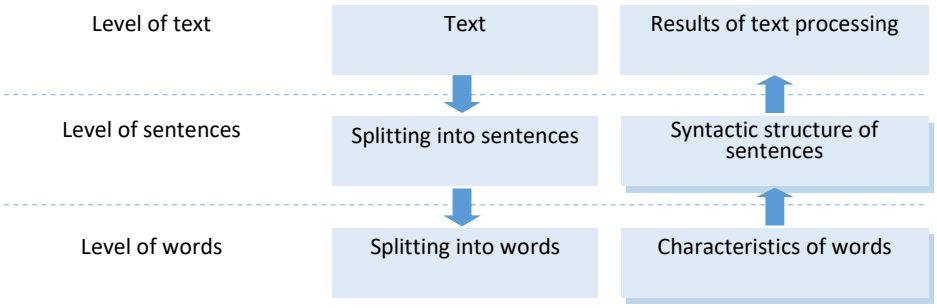

Figure 1. General scheme of text processing

Almost simultaneously with the advent of computers, appeared the first computer programs to calculate the quantitative characteristics of texts. Currently, there is a large number of software products for various natural languages, including free programs. However, to view the statistics of the text, that is, to calculate the number of characters, words, sentences and other quantitative indicators, you can do without additional programs. Modern text editors have this ability, that is, to calculate the number of characters, words, sentences, paragraphs and other quantitative characteristics of the text. For example, the Word text editor, after checking the spelling, provides statistics on the readability of the text. Of course, everything here, first of all, is suitable for texts in English; for example, the number of syllables in the text is calculated based on the rules of the English language.

In this research, were developed two software for calculating the statistics of Azerbaijani texts and assessing their complexity level. One of them is the application program "**Mətn analizi**" ("Text Analysis"), and the other is the website "**www.oxunabilir.az**".

2. The concept of "readability" and readability formulas

To indicate the text complexity level and the degree of its perception is used the term "*readability*" [4]. A question may arise: how is it determined which characteristics of a text affect its complexity?

The principle of identifying the complexity factors is simple: several texts are taken – one texts are simple, the other are complex. These texts are compared to individual indicators, such as logical structure, topic, length of sentences, and other parameters. If the characteristic value changes from easy texts to hard ones, this characteristic is one of the complexity factors of the text. For example, comparison of simple and hard texts shows that the hard texts contain more unknown words and long sentences. Hence, the familiarity of words and the length of sentences depend on the complexity of the text. [5]

Initial research on the assessment of complexity of texts began in the United States, and the first readability formulas also appeared there. Hundreds of such formulas have been developed for English texts, but only a small number of them have gained popularity. [6]

The most pupular readability formula is ***Flesch reading-ease formula***. In this formula, two variables are used as factors that affect "ease of reading": the average length of sentences in words and the average length of words in syllables:

$$\boldsymbol{K_{en}} = 206.835 – (1.015 \times \boldsymbol{S}) – (84.6 \times \boldsymbol{W}), \quad (1)$$

where $\boldsymbol{K_{en}}$ – reading ease score, $\boldsymbol{S}$ – average sentence length in words (the number of words divided by the number of sentences), $\boldsymbol{W}$ – average word length in syllables (the number of syllables divided by the number of words).

For the easy application of the formula, offers the following procedure: 100 words are taken from any text; the average length of sentences in words and the average length of words in syllables are calculated. The value of the readability may vary within 100 (very easy text) and 0 (very difficult text). (Table 1)

Table 1. Flesch's reading-ease scores

Reading ease score (K)	**Description**
0 – 29	very difficult
30 – 49	difficult
50 – 59	fairly difficult
60 – 69	standard
70 – 79	fairly easy
80 – 89	easy
90 –100	very easy

Another widely used readability formula – ***Flesch-Kincaid grade-level formula*** is an improved version of the Flesch reading-ease formula. The value received from the application of this formula, in fact, indicates not only the readability of the text, but also the U.S. grade level it is intended for. This allows teachers, parents, librarians and other professionals who work with texts to judge the readability level of various books and texts. The formula is as follows:

$$\boldsymbol{GL_{en}} = (0.39 \times \boldsymbol{S}) + (11.8 \times \boldsymbol{W}) – 15.59 \quad (2)$$

The result ($\boldsymbol{GL_{en}}$) of applying the formula indicates the appropriate grade level. For example, the value 8.2 means that the text will be understood as an 8th grade student (usually 12-14 years). As you can see from the formula, the smaller the average length of sentences and the average length of words in a text, the lower the value of grade level, respectively, the text can be perceived by younger readers.

It should be borne in mind that both formulas are obtained experimentally for texts in English. To apply these formulas to the texts in the Azerbaijani language, the coefficients of the variables $\boldsymbol{S}$ and $\boldsymbol{W}$ must be adjusted, since the Azerbaijani language, unlike the inflectional English language, is agglutinative according to the morphological structure. In other words, the average sentence length in Azerbaijani is less than in English, and the average length of words, on the contrary, is more.

Table 2. Comparison of the quantitative characteristics of the identical literary samples in English and Azerbaijani (the fictions marked with the * are used separate fragments, not the full text)

	Names of fictions	Number of sentences	Number of words	Number of syllables	Average sentence length in words, ASL	Average number of syllables per word, ASW	ASL_{en} / ASL_{az}	ASW_{en} / ASW_{az}
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	Ernest Hemingway. ***Nobody Ever Dies*** (1939)	540	5264	7063	**9.75**	**1.34**	0.67	1.84
	Ernest Heminquey. **Heç kim heç vaxt ölmür** (translated by: Kamran Nazirli)	594	3865	9489	**6.51**	**2.46**		
2	*Gabriel Garcia Marquez.* **Monologue of Isabel Watching It Rain in Macondo** (1955)	178	2902	3529	**16.30**	**1.22**	0.72	2.04
	Qabriel Qarsia Markes. **İsabel Makondada yağışa baxır** (translated by: Natig Safarov)	157	1847	4602	**11.76**	**2.49**		
3*	Mark Twain. **The Adventures of Tom Sawyer** (1876)	142	1946	2371	**13.70**	**1.22**	0.72	1.83
	Mark Tven. **Tom Soyerin macəraları** (translated by: Shafiga Aghayeva)	158	1549	3460	**9.80**	**2.23**		
4*	John Galsworthy. **Beyond** (1917)	352	7119	9001	**20.22**	**1.26**	0.76	2.02
	Con Qolsuorsi. **Ölümdən güclü** (translated by: Aslan Guliyev)	366	5566	14199	**15.27**	**2.54**		
5	Herbert Wells. **The Crystal Egg** (1897)	301	6878	9324	**22.85**	**1.36**	0.62	1.90
	Herbert Uels. **Büllur yumurta**	346	4889	12603	**14.13**	**2.58**		
6	Jack London. **Grit of Women** (1900)	362	5675	7428	**15.68**	**1.31**	0.69	1.82
	Cek London. **Qadın cəsarəti** (translated by: Sevda Abuzarli)	400	4344	10337	**10.86**	**2.38**		
7	John Steinbeck. **The Chrysanthemums** (1937)	448	4220	5742	**9.42**	**1.36**	0.88	1.87
	Con Steynbek. **Xrizantemlər** (translated by: Ramiz Abbasli)	369	3077	7743	**8.34**	**2.52**		
8	William Somerset Maugham. **Louise** (1969)	149	2118	2437	**14.21**	**1.15**	0.77	2.03
	Uilyam Somerset Moyem. **Luiza** (translated by: Kamran Nazirli)	196	2158	5040	**11.01**	**2.34**		
9*	Agatha Christie. **Murder on the Orient Express** (1934)	334	4737	6047	**14.18**	**1.28**	0.63	1.99
	Aqata Kristi. **Şərq ekspessində qətl** (translated by: Parviz Jabrayil)	425	3795	9683	**8.93**	**2.55**		
10*	Oscar Wilde. **The Picture of Dorian Gray** (1890)	363	4950	6097	**13.63**	**1.23**	0.76	1.93
	Oskar Uayld. **Dorian Qreyin portreti** (translated by: Kamran Nazirli)	422	4369	10414	**10.35**	**2.38**		
	T o t a l							
						average	**0.72**	**1.93**
						variance	0.0059	0.0076
						min	0.62	1.82
						max	0.88	2.04

To determine the relationship between the average length of sentences and the average length of a word in both languages, the following methodology was used (*for details see*: [5]). Firstly, were calculated the statistical indicators of various identical texts in Azerbaijani and English, and then the ratio of the values of certain parameters. Then are determined the average values of these parameters and are derived the numbers, which show how much the coefficients in the formulas must be corrected. To ensure that the texts in both languages have a high level of content and style, some of the well-known literary samples in English and their translations into the Azerbaijani language were used (Table 2).

However, since fictions and their translations largely depend on the style of the writer and translator, various academic texts and their translations into English, taken from the portal *azerbaijan.az* and the official website of the President of the Republic of Azerbaijan (*www.president.az*), were investigated similarly way. In order to make the study more comprehensive, also taken into account the statistical indicators of separate sentences in English and their translations into the Azerbaijani language.

The study showed that the sentences in English in comparison with the sentences in the Azerbaijani language are 0.77 times longer on average, and the words in syllables are 1.91 times shorter. [7] Thus, the coefficient of the average sentence length (***S***) in formula (1) was adjusted 0.77 times, and the average word length in syllables (***W***) – 1.91 times. As a result, ***Flesch reading-ease formula for the Azerbaijani text*** was as follows:

$$\boldsymbol{K_{az}} = 206.835 - (1.318 \times \boldsymbol{S}) - (44.3 \times \boldsymbol{W}), \quad (3)$$

where: $\boldsymbol{K_{az}}$ – reading ease score, ***S*** – average sentence length in words, ***W*** – average word length in syllables.

Similarly, the Flesh-Kinside formula is adapted for the Azerbaijani language texts. Thus, ***Flesch-Kincaid grade-level formula for the Azerbaijani text*** was as follows:

$$\boldsymbol{GL_{az}} = (0.51 \times \boldsymbol{S}) + (6.18 \times \boldsymbol{W}) - 15.59, \quad (4)$$

where: $\boldsymbol{GL_{az}}$ – grade level score, ***S*** – average sentence length in words, ***W*** – average word length in syllables.

3. The "Mətn analizi" ("Text Analysis") application

The application program "Text Analysis" is designed to automate the process of calculating statistics and evaluating the readability of texts in the Azerbaijani language. At the same time, this program also calculates the frequency of words in the text and allows to sort the frequency list in alphabetical or frequency order. The program is written in the Delphi programming environment and runs under the Windows operating system with text documents such as .doc or .docx.

The program window (Figure 2) conditionally consists of three main sections. The text for which the statistics will be calculated and the readability will be evaluated is entered into the edit box found on the left side of the window. Fields for various statistical indicators and readability values are placed in the middle of the window. The list box on the right hand side is for individual words of the text and their frequencies.

The main objectives of the program are:

- to determine the quantitative indicators and the level of readability of the entire text in the Azerbaijani language or its selected fragment;
- to save the calculated statistical indicators in a separate file;
- to select the parts (paragraphs) of the text that the readability values do not match the given value;
- to define the frequencies of separate words in the text.

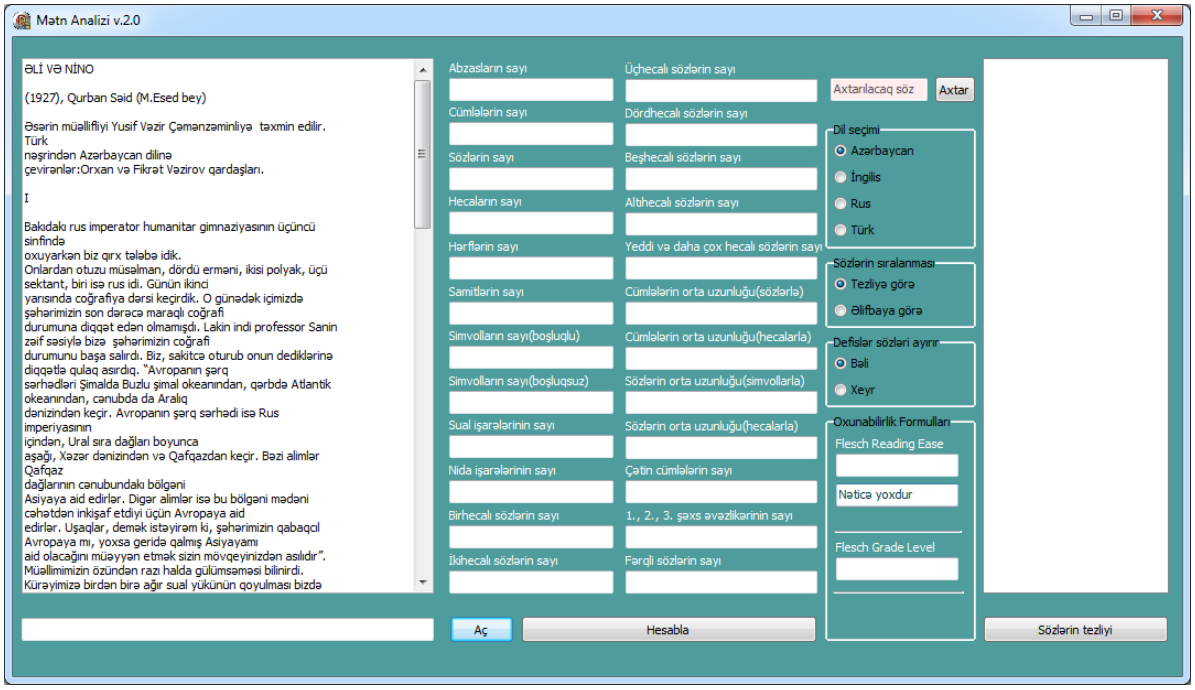

Figure 2. Screenshots of "Mətn analizi" ("Text Analysis") application

The program allows to calculate the following quantitative indicators required to evaluate complexity of texts:

- number of paragraphs
- number of words
- number of letters
- number of characters (with spaces)
- number of question marks
- number of monosyllabic words
- number of 3-syllable words
- number of 5-syllable words
- number of 7 or more syllable words
- number of first, second, and third-person pronouns
- number of sentences
- number of syllables
- number of consonants
- number of characters (no spaces)
- number of exclamation marks
- number of 2-syllable words
- number of 4-syllable words
- number of 6-syllable words
- number of different words

Other indicators that are calculated in the app are:

- average sentence length in words – the number of words divided by the number of sentences;
- average sentence length in syllables – the number of syllables divided by the number of sentences;

- average word length in symbols – the number of symbols (no spaces) divided by the number of words;
- average word length in syllables – the number of syllables divided by the number of words;
- percentage of polysyllabic (containing three or more syllables) words.

It's easy to calculate some of the listed parameters (e.g. number of letters, consonants, symbols, question marks, exclamation marks): to do this, you can simply use the standard functions of the programming language. There is also no difficulty in calculating parameters such as the number of paragraphs, words, syllables.

Number of paragraphs. In electronic texts, the symbol ¶ is used to indicate the end of a paragraph. This symbol in the Windows operating system is entered using the combination of the <Ctrl> key and the <-> key on the numeric keypad. Thus, with the statistical analysis of texts, the number of paragraphs can be calculated on the basis of these symbols.

Number of words. The sequence of characters (letters, digits, special characters) that does not contain spaces and punctuation marks (except hyphen) is considered as a *word***.** (e.g. oxunabilirlik, tərcümeyi-hal, BMT, 1962-ci).

Usually words are separated by a space character. However, between the two words there may be other symbols:

- punctuation marks (period, question mark, exclamation point, colon, semicolon, ellipsis, comma, dash, parenthesis, quotation marks);
- symbols of four arithmetic operations (+, –, ×, /);
- special characters (e.g. &, ∧, ∨, →, etc.).

The program should take into account one nuance: a dash (–), should not be treated as a word, even if there are spaces on both sides.

```
#include <bits/stdc++.h>
using namespace std;

int main()
{
  string s;
  getline(cin, s);
  int ans = 0;
  for(int i = 1; i <= s.length(); i++)
  {
    if (s[i] == ' ' && s[i-1] != ' ' || s[i] == '\n' &&
       s[i-1] != '\n' || s[i] == '\t' && s[0]!='\t' ||
       s[i] == '\0' && s[i-1] != '\0')
      ans++;
  }
  cout << ans << endl;
  return 0;
}
```

Figure 3. C++ program code that determines the number of words in the text

Number of syllables. The number of syllables in Azerbaijani is determined by the number of vowels. So, in order to find the number of syllables in the texts in Azerbaijani language, it is enough to just count the number of vowels (a, ı, o, u, e, ə, i, ö, ü).

However, there are some parameters that their calculation is not as simple as it might seem.
Number of sentences. Although at first glance it seems simple, building a module for splitting text into sentences is not an easy task. Let's take a closer look at this issue. ***Sentence*** is a unit, grammatically formed on the basis of the laws of each language and expressing a complete thought. Usually the sentence is defined as follows: "A combination of a few words or a single word that express a complete thought is called sentence". Although this is a simple explanation of the sentence, there are three important features for the sentence:

1) The sentence must express a complete thought;
2) The sentence consists of a combination of several words;
3) The sentence can consist of a single word.

Depending on purpose and intonation sentences are divided into four types: *declarative sentence, interrogative sentence, exclamatory sentence, imperative sentence*.

In the writing, to separate and distinguish sentences or their parts is used *punctuation*. Punctuation was created much later than the alphabet. Despite the diversity of linguistic systems and alphabets, punctuation is common to most peoples. Although there were significant differences between punctuation marks in different languages at the time of theirs creation, they are now almost the same. To Azerbaijani language punctuation marks were derived from the Russian language. True, in the manuscript of the Book of Dede Korkut (in his Dresden copy) there is a sign between the sentences. This sign looks like the letter *o*, but its inner was completely filled. According to some experts, in addition to the sentence separator, this sign also had a different function in the manuscript. [8]

The punctuation marks used in the Azerbaijani writing are: *period, question mark, exclamation point, colon, semicolon, ellipsis, comma, dash, parenthesis, quotation marks*. Some linguists include a paragraph in punctuation.

In the writing, the sentence usually starts with a capital letter and ends with terminal symbols, such as period, exclamation point, or question mark. However, it should be borne in mind that the sentence may not start with a capital letter, and the period sign may not indicate the end of the sentence. At the end of the sentence, there may be other symbols.

The most commonly used punctuation mark, to indicate the end of a sentence, that is, to separate a sentence from someone else, is the period sign.

The period sign (**.**) is first put forward at the end of the declarative sentences and separates them from other sentences. As you know, the sentence may be simple or complex; complete or incomplete. The above statement applies to all of these. The period sign also serves to separate nominal sentences. At the end of the imperative sentences, which are not mentioned with high excitement, is putting the period sign.

The question mark (?) is place at the end of the interrogative sentence, which is used to get the answer. The interrogative sentences may be simple or complex; complete or incomplete; for example:

Necə yəni? (*How so?*)

Nə? (*What?*)
Eləmi? (*Really?*)
Sən gedirsən? (*Are you going*?)
Adınız və soyadınız? (*Your name and surname?*)
Cavab ver görək: Səni kim öyrədib? (*Answer me: Who taught you?*)
Sən bilmək istəyirsənmi, o nə vaxt qayıdacaq? (*Would you like to know when he/she will return*?)

Is first, ***the exclamation point*** (!) puts in the end of the exclamatory sentences; for example:

E y v a z. Mən düşmənəm sizin insan əti yeyən ikibaşlı qartalınıza! Mən düşmənəm sizin süngü və pulemyot üstündə duran hökmranlığınıza!

E y v a z. I am the enemy of your double-headed eagle devouring people! I am the enemy of your domination, holding on bayonets and machine guns! (J.Jabbarli)

At the end of the imperative sentences, which are called high emotion, also is placed the exclamation point; for example: *Sağa dön!* (*Turn right!*) *Yerinizdən danışmayın!* (*Do not speak from a place!*)

As noted, in the end most sentences are put a period, an exclamation or question mark. If to exclude the appearance of these symbols in the middle of the sentence, then the number of sentences in the text can be calculated using the following code (Figure 4).

```
#include <iostream>
using namespace std;

bool check(char c) {
   switch(c) {
      case '.':
      case '!':
      case '?':
         return true;
      default:
         return false;
   }
}

int main() {
   int say = 0;
   string str;
   getline(cin, str);
   for(int i = 1; i < str.length(); i++) {
      if(check(str.at(i))) {
         if(!check(str.at(i - 1))) {
            say++;
         }
      }
   }
   cout << say << endl;
   return 0;
}
```

Figure 4. The code in C++, which defines the number of sentences in the text according to the sentence ending marks such as period, question mark and exclamation point.

However, these evidences (that is, accepting the characters ".", "?", "!" as the end of the sentence) are not enough that the computer program defined the sequence of characters as a sentence. As the above examples show, these three symbols can be found in the middle of the sentence. On the other hand, other symbols may appear at the end of the sentence. Let's find out more about these special cases to ensure that such cases are considered by the program.

In addition to splitting sentences, ***the period*** has other tasks. In the dialogue of the plays, after the name of the characters, the period is put. As well as, after a short header at the beginning of the line, a period is inserted (the text starts with the same line, after the period).

The period is also used as a abbreviation:

a) Name, father's name is shortened, i.e. their first letters are written, then the period is put; for example: *C.Cabbarlı*, *S.Vurğun*, *M.İbrahimov*, etc.

Therefore, such expressions should not increase the number of sentences, but should be taken simply as a word. In the application "Mətn analizi" one more case is taken into account: two consecutive text elements are checked, and if the first element consists of one character, and the next element is a period, then these two elements are accepted as initials or as an abbreviation.

b) After the fragments taken from the works of individual authors, on the basis of the adopted table of abbreviations, only the initial letters of the authors' names are indicated; for example: instead of (*Cəfər Cabbarlı*) – (*C.C.*), instead of (*Səməd Vurğun*) – (*S.V.*), etc.

c) The academic ranks, such as *professor*, *dosent*, *akademik* can be reduced, after which also puts a dot; for example: *prof.*, *dos.*, *akad.*

d) With the abbreviation of such words as "*və bu kimi*", "*və sair*", "*məsələn*" (*et cetera*, *for example*) the period is used: *və b.k.*, *və s.*, *məs.* (*etc.*, *e.g.*)

e) The period is used as a thousands separator after every three digits in a number, counting from right to left; for example: *123.523*. In English-speaking countries, a comma is used for this purpose, and the dot plays the role of a decimal separator, i.e., separating the integer and fractional part.

f) Sometimes a dot is used as a separating symbol within a sentence; for example, as a separator in Internet resource addresses, a date and time record; for example: *10.11.2017*, *www.ict.az.*

g) Sometimes a dot is inserted after the numbering of list items or algorithm steps; for example: *1. Bakı, 2. Gəncə, 3. Sumqayıt*. [9]

The exclamation points and ***question marks*** can also be used in the middle of the sentence. Often these signs indicate the end of the fragment given in parentheses or in quotation marks:

- *Getdikcə xəyalımda yaranan bu aləm mənə real həyatdan, təsadüf elədiyim insanlardan (yalnız anamdan başqa!) daha artıq xoş gəlirdi.*
  This world, which gradually developed into my dreams, was more pleasant for me than the real life, the people I met (except my mother!). (I.Afandiyev)
- *Gördü bir qarı bir keçəlin əlindən tutub "ay qul alan!" deyib dad edir.*
  He sees an old woman taking a bald man's hand and shouting "hey, who wants to buy a slave!".

Sometimes the sign that indicates the end of the sentence is given in the form of a combination of a question mark and an exclamation mark:

- *Pədərsüxtə, kəvakibin afətindən məni qorxudarsan və əlacını gizlərsən?!*
  How dare you, the dog son, threaten me with a misfortune, hiding in the stars, without telling me about the means against her?! (M.F.Akhundzade)

Sometimes the end of the sentence is indicated in the form of an ***ellipsis*** (...).The ellipsis, on the one hand, is close to the comma, on the other hand, to the period. Usually, the ellipsis are given as three

dots. In the quotations, where the speech of another person is interrupted (at the beginning, middle or end of the quoted text), the ellipsis is placed. The ellipsis in the middle or end of the sentence indicates that the thought is not finished. Thought may not be completed for various reasons: "... or the author considers it unnecessary to continue it, or not to speak it out of secrecy, or there are such words that it is unacceptable for society, or it is forgotten, or the speach is interrupted by someone else, or the speaker is agitated and he needs to talk with pauses". [9]

- *Bacım uşaqları dayılarına qənim kəsiləcəklər. Gəlini gərdəkdə, küçüyü ... Bəs o dünyada necə?*
  My sister's children will hate their uncle. The bride in bed, the puppy in ... But how in the other world? (F.Karimzade)
- *O, hara isə əfsanəvi bir aləmə gedir və o aləmdə onu, kim bilir, nələr gözləyir ...*
  He goes some sort of legendary world, and in this world, who knows, what's in store for him ... (I.Afandiyev)
- *– Mirzə Qələndər, mənə də yox də ... Mən gimnaziya qurtarmışam ... Özüm də Allaha inanıram, amma siz mollaların alimibiəməlliklərinizi də bilirəm axı.*
  – Mirza Galandar, to someone, but not me ... I graduated from gimnasium, from university, I'm literate ... I believe in God, but I also know all your stuff ... (I.Afandiyev)
- *Qaldı sünnü olmağa ... Bayram heylə şeylərə fikir verən deyil.*
  As for being Sunni ... Bayram does not pay attention to such things. (I.Afandiyev)

In the application "Mətn analizi", must also take into account one nuance: in electronic texts, in addition to the sequence of three dots, the ellipsis can also be in the form of one symbol (UNICODE code is U+2026).

Thus, the main punctuation marks that indicate the end of the sentence are: a period, an exclamation point, a question mark and an ellipsis. There are two more punctuation marks, which can sometimes be regarded as symbols that indicate the end of the sentence. This is a semicolon and a colon.

A ***semicolon*** (;) is a punctuation mark, which occupies the middle position between a period and a comma, and requires a pause less than a period, and greater than a comma. When lists the homogeneous terms of a simple sentence, or when there is a comma in one of the components of a complex sentence, then between the components that have a longer pause, puts a semicolon; for example:

- *Əmir İnanc Qətibənin Nizamiyə hüsn-rəğbət göstərməsinə yol verməklə bərabər, onu Hüsaməddinə də vəd edirdi; Dilşadı Bağdada göndərməyə hazırladığı halda, Fəxrəddinin də başını aldadır və Dilşadın Fəxrəddinlə olan tanışlığına maneə törətmirdi.*
  Amir Inanj encouraged Gatiba's fascination to Nizami and at the same time promised her to Husameddin; he was going to send Dilshad to Baghdad the Caliph and at the same time he was deceiving Fakhreddin, not interfering with his meetings with his beloved. (M.S.Ordubadi)
- *O, gözəl olduğu qədər də macəraçı və iftiraçı idi; o, bir gün də olsa, böhtan və iftira toxumadan yaşaya bilməzdi.*
  She was as adventurous and slanderer as she was beautiful; she could not survive a day without slander and calumny. (M.S.Ordubadi)
- *Sən o şairi çox da həqir hesab etmə; o, gəncdir, fəqət bizim kimi gənclərdən deyil.*
  Do not think that the poet is too poor; he is young, but he is not like us. (M.S.Ordubadi)

In the first of this samples a semicolon was separate the homogeneous terms of a simple sentence, and in the next ones – the components of a complex sentence. [9]

The ***colon*** (:) is used for clarification purposes. There are three main uses of the colon: between two main clauses in cases where the second clause explains or follows from the first, to introduce a list and before a quotation, and sometimes before direct speech; for example:

- *O həm gözəl idi, həm də gözəlliyindən bir silah kimi istifadə edirdi: əlini şairin əlindən çəkmək istəmirdi, bütün naz, qəmzə, utanmaq, qısqanmaq, rəng verib rəng almaq, hətta göz yaşları axıtmağı... hamısını təcrübədən keçirirdi.*
  She was beautiful, and used her beauty as a weapon: the hand still remained in the hand of the poet, she did not tear it away, her face expressed both tenderness, and sadness, and coquetry, and shame, and jealousy. (M.S.Ordubadi)
- *Lakin təxir etmək də yaramazdı: qız uzaqlaşır və axşam qaranlağının içərisində itirdi*.
  But it was impossible to postpone: the girl withdrew, dissolving in the evening darkness. (M.S.Ordubadi)
- *Edama toplaşmış camaatdan səs çıxmırdı: nə qoca dindi, nə cavan, nə qız, nə gəlin.*
  The people did not make a sound: the old and the young, and the girls and young women – all seemed suddenly numb. (Y.Samadoghlu)
- *O öz qəlbində deyirdi: "Ərəb, türk və rum qanından yaradılan bu möcüzə gənc şairin qəlbini əsir edə bilər".*
  He said in his heart: "This miracle of Arab, Turkic and Greek blood can captivate the heart of a young poet." (M.S.Ordubadi)

How to use the "Mətn analizi" application

1. Start the "Mətn analizi" application. The program window will open.
2. Click the Aç (Open) button at the bottom of the window. The Open dialog box opens.
3. Navigate to the folder where the text file you want to evaluate is located and select it. Then click the Open button. The document is opened in the text box.
4. If necessary, make appropriate changes to the document.
5. Click on the button Hesabla (Calculate). The statistical value of the text should be calculated and displayed in the corresponding fields. At the same time, in the Oxunabilirlik düsturları (Readability formulas) section, the readability indicators of the text will be displayed using both Flesch reading-ease formula (3) and Flesch-Kincaid grade-level formula (4).
6. Click the Sözlərin tezliyi (Word frequency) button under the list box. The frequencies of all words in the text will be calculated and the frequencies will be reflected next to the words in the list box.
7. The list is sorted by the frequency of the words. Arrange the list in alphabetical order to see if there is any word in the text. To do this, select the Əlifbaya görə (Alphabetically) option in the Sözlərin sıralanması (Sort words) section.
8. To quickly find the exact location of a certain word in the text, enter this word in the text box to the left of the Axtar (Find) button and click the button. The word found will be highlighted in the text.
9. Clicking the Axtar (Find) button again and again, you can switch to the other positions where the search word is found.

Below is a screenshot of the main window of the program, after counting the statistics and evaluating the level of readability of the text (Figure 5).

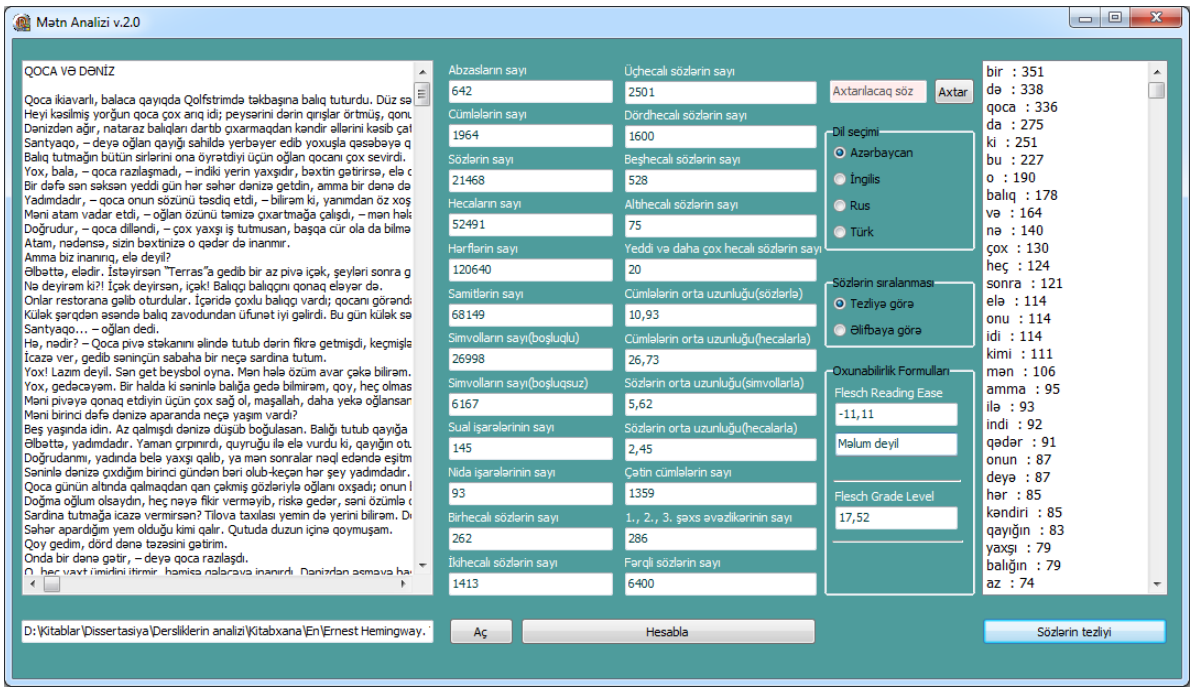

Figure 5. A screenshot of the main window of the "Mətn analizi" application, after counting the statistics and evaluating the level of readability of the text

The "Mətn analizi" application also has very important functionality for researchers. This is saving the results of calculations for all indicators in the database. The following table shows some records of the main database table (Table 3).

4. The web-site www.oxunabilir.az

The main goal of creating this Internet resource is to present the capabilities of the "Mətn analizi" application to a wide audience. As in the "Mətn analizi" application, this resource calculates the statistical indicators of the text entered from the keyboard or from an existing .doc file.

Table 3. A sample to the main table of the database in which the results of the "Mətn analizi" application are stored

№	*Date*	*Title*	K_{az}	GL_{az}	*Number of paragraphs*	*Number of sentences*	*...*
1	03.07.2018	Ernest Heminquey. **Heç kim heç vaxt ölmür**	89	3	221	594	...
2	03.07.2018	Qabriel Qarsia Markes. **İsabel Makondada yağışa baxır**	81	6	79	157	...
3	03.07.2018	Herbert Uels. **Büllur yumurta**	74	8	74	346	...

Then, based on parameters such as the average sentence length in words and the average word length in syllables, the readability of the text is evaluated by two formulas: Flesch reading-ease formula (3) and Flesch-Kincaid grade-level formula (4). (Figure 6)

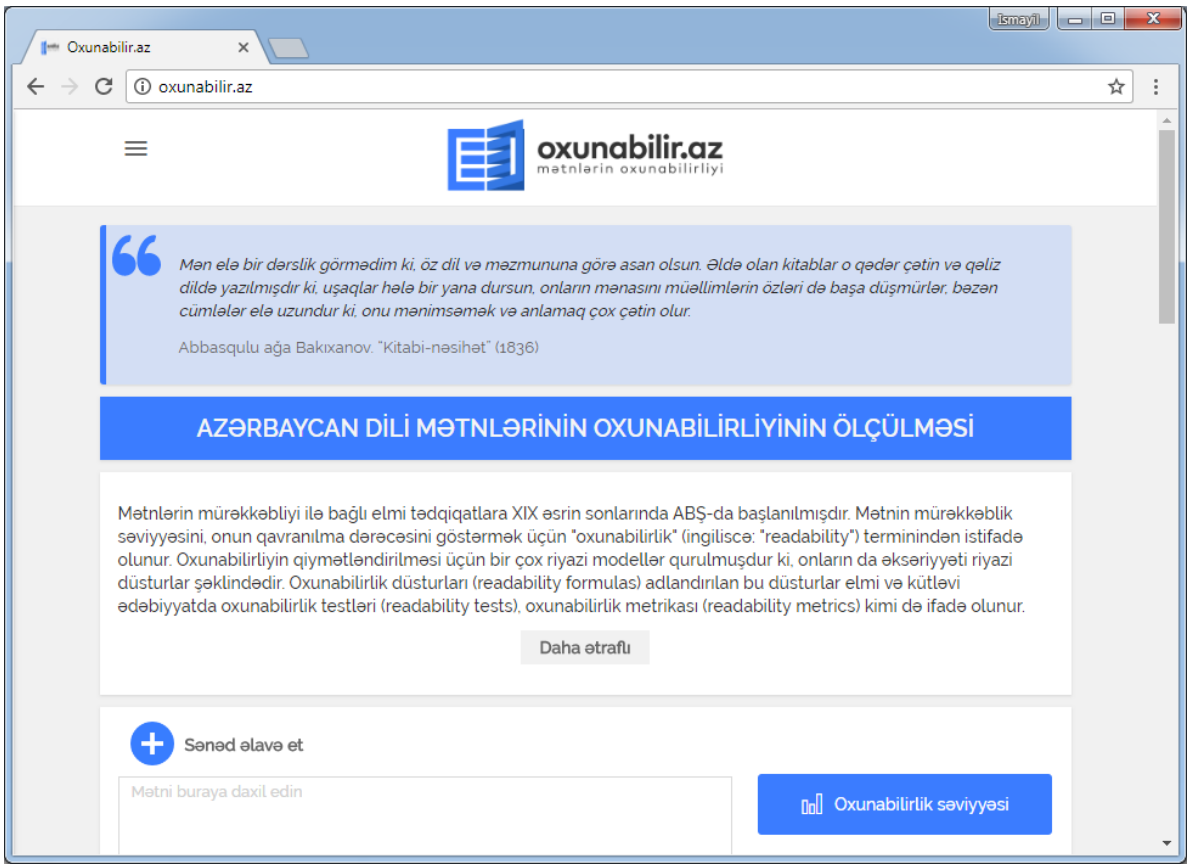

Figure 6. Home page of the web-site *oxunabilir.az*

The procedure for using the site is as follows: the sample text is entered in the text box "Mətni buraya daxil edin" ("Enter text here") from the keyboard. Of course, typing large texts from the keyboard directly into the text box is not very convenient. Therefore, it is also possible to work with existing text files (.doc format). To do this, click the link Sənəd əlavə et (Add document) and in the opening Open dialog box, find and open the desired file. You can make any changes to the text that appears in this way in the text box. Then click the Oxunabilirlik səviyyəsi (Readability level) button on the right side of the text box.

The results will be displayed below the Oxunabilirlik səviyyəsi (Readability level) button. Statistical values of the text will be reflected in the Mətnin statistikası (Text Statistics) section under the text field. (Figure 7)

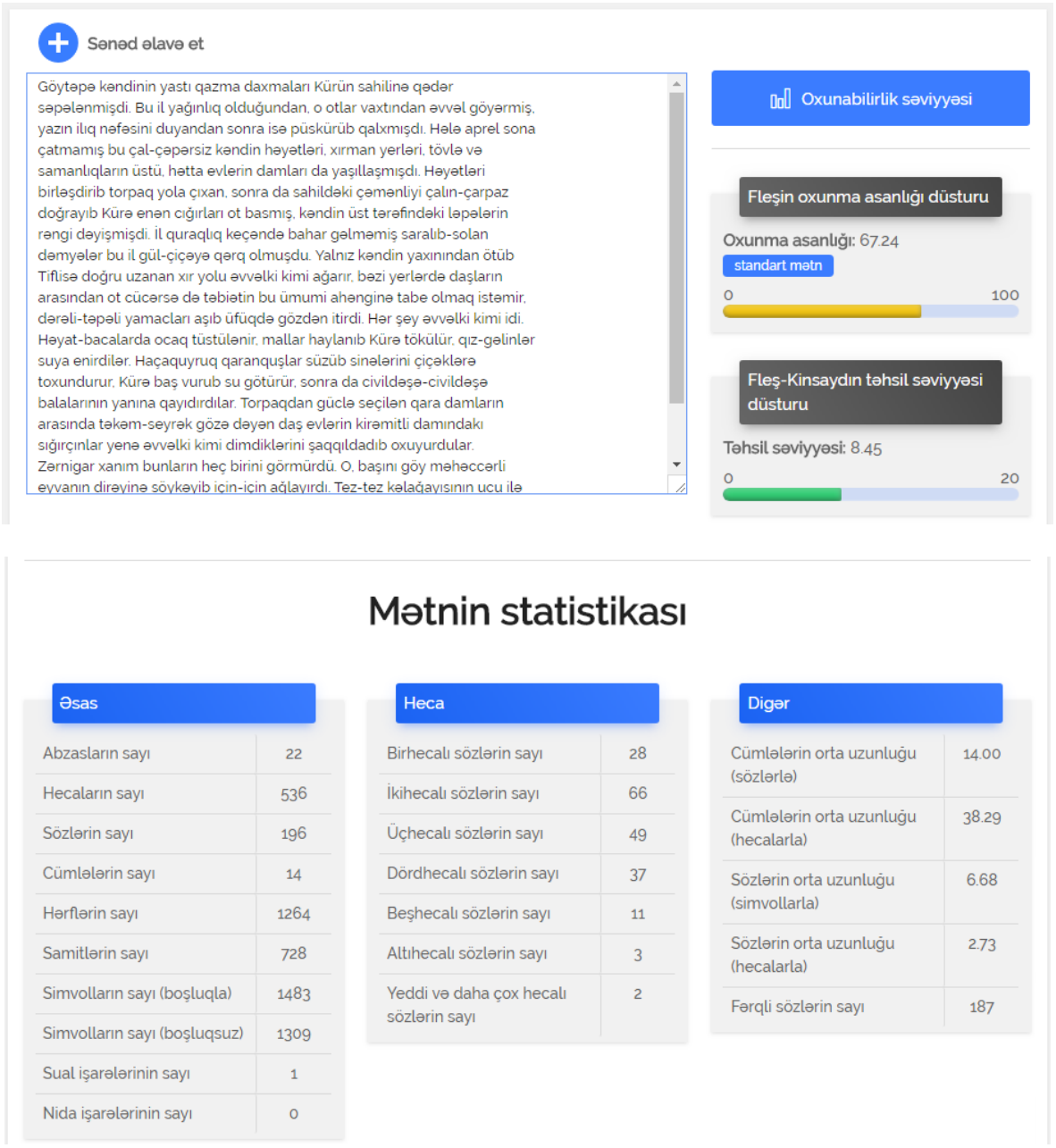

Figure 7. The readability score and statistics of the selected text

5. Conclusion

A software was developed to automate the process of calculating the quantitative characteristics of texts in the Azerbaijani language. The developed software also evaluates the readability of texts on the basis of the Flesch reading-ease and Flesch-Kincaid grade-level formulas, modified for the estimation of texts in the Azerbaijani language.

The rules used in the calculation of many parameters are suitable for most natural languages. However, this can not be said about some parameters: for example, the number of syllables in the Azerbaijani language is determined by the number of vowels, but it is known that this rule does not apply to English.

The Turkish and Azerbaijani languages belonging to the same language group are very close to each other in terms of their lexical, morphological and syntactic aspects. Therefore, the rules and methods given in this article are also relevant for the Turkish language.

References

[1]. "Новое в зарубежной лингвистике. Выпуск XXIV, Компьютерная лингвистика", Прогресс, МОСКВА, 1989.
[2]. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С., "Автоматическая обработка текстов на естественном языке и анализ данных", НИУ ВШЭ, МОСКВА, 2017.
[3]. Селезнев К., "Обработка текстов на естественном языке", Открытые системы. СУБД, МОСКВА, 2003, №12.
[4]. Sadıqov İ., "Mətnlərin murəkkəbliyi və onun qiymətləndirilməsi yolları (II yazı)", Kurikulum, BAKI, 2013, №3, ss. 11–29.
[5]. Sadıqov İ., Mətnlərin murəkkəbliyi və onun qiymətləndirilməsi yolları (I yazı)", Kurikulum, BAKI, 2013, №2, ss. 30–42.
[6]. DuBay W.H., "The Classic Readability Studies", Impact Information, COSTA MESA, 2006.
[7]. Sadıqov İ., "Azərbaycan dili mətnlərinin mürəkkəbliyinin qiymətləndirilməsi üçün modifikasiya olunmuş Fleş düsturu", İnformasiya texnologiyaları problemləri, BAKI, 2018, №1, ss. 46–58.
[8]. Abdullayev Ə., Seyidov Y., Həsənov A., "Müasir Azərbaycan dili. Sintaksis", Şərq-Qərb, BAKI, 2007.
[9]. Kazımov Q.S., "Müasir Azərbaycan dili. Sintaksis", Təhsil, BAKI, 2007.