

REPUBLIC OF AZERBAIJAN

On the rights of the manuscript

ABSTRACT

of the dissertation for the degree of Doctor of Philosophy

DEVELOPMENT OF METHODS AND ALGORITHMS FOR EVALUATION OF READABILITY OF TEXTS IN AZERBAIJANI LANGUAGE ON THE BASIS OF STATISTICAL ANALYSIS

Speciality: 3338.01 – “System analysis, control and
information processing”
(information technologies)

Field of science: Technical sciences

Applicant: **Ismayil Jalal oglu Sadigov**

Baku – 2022

The work was performed at the Institute of Information Technologies of the Azerbaijan National Academy of Sciences (ANAS).

Scientific supervisor: Full member of ANAS, Doctor of technical sciences, Prof. **Rasim Mahammad oglu Alguliyev**

Official opponents: Doctor of Technical sciences, Prof.

Nadir Bafadin oglu Agayev

Doctor of Technical sciences, Prof.

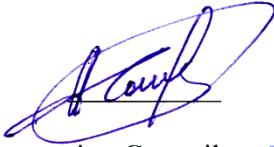
Ramin Rza oglu Rzayev

PhD in Technical sciences

Narmin Eldar gizi Rzayeva

Dissertation Council ED 1.35 of Supreme Attestation Commission under the President of the Republic of Azerbaijan operating at the Institute of Information Technologies of the ANAS.

Chairman of the Dissertation Council:



Full member of ANAS,
Doctor of technical sciences, Prof.
Rasim Mahammad oglu Alguliyev

Dissertation Council
scientific secretary:



Doctor of Philosophy in Technical science,
Assoc. Prof.
Farhad Firudin oglu Yusifov

Chairman of the scientific seminar:



Doctor of Technical science
Mutallim Mirzaahmed oglu Mutallimov



GENERAL CHARACTERISTICS OF WORK

Relevance of work. The expansion of book printing in Europe at first created the idea that printed books would become a means of mass learning, and that this would lead to the rapid enlightenment of the people. However, such an optimistic end has not come, and today, even in the most developed countries, education is not at the desired level. The success rate of graduates is low, and the percentage of pupils and students expelled from secondary and higher education is very high. As early as the 17th century, the founder of pedagogical science, Comenius, saw one of the reasons for this situation in education as the lack of textbooks¹.

Undoubtedly, the quality of education depends on the quality of textbooks, teaching aids and other teaching materials used in the teaching process. Thus, the main content of education is given in textbooks, and it is the textbooks that form the skills and habits of students. The quality of the textbook depends on how the text is perceived, because both the content of the subject and the methodological apparatus are presented in the form of text in the textbook.

Although the study of the complexity of texts began in the United States in the late 19th century², in the first half of that century this issue attracted the attention of the prominent Azerbaijani educator Abbasgulu aga Bakikhanov. In the preface to his 1836 book, “The Book of Advice” (“Kitabi-nəsihət”), he complained about the inconsistency of the books on education and training, written in difficult language: *“I have not seen a textbook that is easy in its language and content. The books available are written in such a difficult and complex language that not only children but also teachers themselves do not understand their meaning; sometimes sentences are*

¹ Подласый И.П. Педагогика. Москва: Юрайт-Издат, 2009.

² DuBay W.H. The Principles of Readability. Costa Mesa, California: Impact Information, 2004.

so long that it is very difficult to master and understand.”³ As can be seen from this example, A. Bakikhanov focused on two factors that specifically complicate the texts - the abundance of complex words and the length of sentences.

According to researchers, during the Soviet era, “... *amateurish continued in the creation of textbooks for all levels of education, and each author invented a textbook for himself, the quality of which has been estimated very approximately to this day*”⁴. This view of textbooks in Russian can be equally applied to Azerbaijani textbooks, as the vast majority of textbooks in Azerbaijan during the Soviet era were translated from Russian. The annual collections "Problems of the school textbook" (20 books) published in Moscow in 1974-1991 and devoted to this issue did not change the situation.⁵

After gaining independence in 1991, Azerbaijan began to prepare its own textbooks, primarily in the humanities and gradually in other disciplines. In the 1992-1993 academic year alone, 84 new textbooks were prepared and published. In 2005, a new stage in the field of work with textbooks began in Azerbaijan. Thus, the document “Textbook policy in the general education system”, prepared in the same year, sets out the basic principles of textbook policy in the Republic of Azerbaijan, the main requirements for the content of textbooks, the language of textbooks, electronic textbooks and other issues. One of the main innovations is that new textbooks are not evaluated formally, on the basis of subjective considerations, but on the basis of approved specific criteria. Starting from 2008, textbooks and teaching materials developed on the basis of new subject programs (curricula) and evaluated according to new criteria are a step forward in solving this problem. However, in our opinion, this is not enough to completely resolve the problem, since among the criteria for evaluating textbooks

³ Abdullayev A.S. Azərbaycan dilinin tədrisi tarixindən. Bakı: Maarif, 1966.

⁴ Беспалько В.П. Учебник. Теория создания и применения. Москва: НИИ школьных технологий, 2006.

⁵ Проблемы школьного учебника: XX век. Итоги Сборник (под ред. Зуева Д.Д.). Москва: Просвещение, 2004.

there is no statistical analysis of texts and a mathematical assessment of their complexity.

It is true that there are mathematical models for assessing the complexity of any text, especially educational texts, taking into account the age characteristics of schoolchildren. However, these models, on the one hand, are mainly intended for texts in English, and, on the other hand, the use of some of these models is inappropriate without appropriate software.⁶ Therefore, the adaptation of existing mathematical models to texts in the Azerbaijani language and the development of appropriate software for these models is one of the most important issues.

This research is the first in Azerbaijan on readability. It is true that the issues of statistical analysis of texts have been the subject of research by some Azerbaijani scholars. However, none of these studies touched on the complexity of texts, their measurement (evaluation) and methods of reducing complexity.

The popularity of readability formulas is growing. Once available only in English, these formulas are now available in a number of languages. With the help of these formulas, the level of complexity of textbooks is assessed, and information on the readability of fiction helps librarians to give the right advice to their readers in choosing the right books. Many journals use these formulas to evaluate the articles presented to them.

The aim of the work. The main goal of the dissertation is to develop mathematical models (readability formulas) and automation tools (corresponding software) for assessing the complexity of texts in the Azerbaijani language. At the same time, on the basis of mathematical and information models, an objective assessment is made of the correspondence of the texts of textbooks used in general education schools in Azerbaijan to the age of students. In accordance with this goal, the following scientific questions are posed in the dissertation:

⁶ DuBay W.H. *The Classic Readability Studies*. Costa Mesa, California: Impact Information, 2006.

- Study of international experience in assessing the complexity of texts by various methods, mainly mathematical and statistical methods;
- Development of mathematical models for assessing the level of complexity of texts and substantiation of the need to modify existing assessment formulas for English texts, taking into account the specifics of the Azerbaijani language;
- Development of methods for modifying Flesch reading-ease formula and Flesch-Kincaid grade-level formula, one of the most popular readability formulas for English texts, for Azerbaijani language texts; modifying Flesch reading-ease formula and Flesch-Kincaid grade-level formula for Azerbaijani texts using these methods;
- Development of a special methodology for the preparation of lists of 1000 and 3000 words, which form the main vocabulary of the Azerbaijani language and are understood by the majority of primary school students, and the preparation of the above-mentioned lists for the Azerbaijani language on the basis of this methodology; modification of New Dale-Chall readability formula and Spache readability formula for Azerbaijani texts based on the percentage of unfamiliar words in the text (ie not on the list of 3000 and 1000, respectively);
- Development of special algorithms for automating the process of calculating statistical indicators of texts in the Azerbaijani language (number of sentences, number of words, number of syllables, etc.), as well as developing software based on these algorithms to automate the process of calculating statistical indicators of texts;
- Non-subjective assessment of the readability of textbooks in the Azerbaijani language and the degree of compliance of texts in these textbooks with the age level of students using the developed software; including the ability to track how the level of complexity of texts in individual subjects changes when moving from one class to another;

- To create an opportunity for textbook authors, teachers preparing methodological materials, experts engaged in the examination of teaching aids to objectively assess the level of complexity of texts on a number of parameters, as well as to quickly examine and improve recently popular e-learning resources.

Research methods. Regression analysis, mathematical statistics, computer linguistics, NLP-technologies, bibliometry, pedagogical methods were used in solving the set problems.

The main provisions of the defense:

1. Usefulness of readability formulas for Azerbaijani texts (Flesch reading-ease formula, Flesch-Kincaid grade-level formula, New Dale-Chall readability formula and Spache readability formula) based on English-compliant formulas as a result of comparing the statistical indicators of parallel texts in English and Azerbaijani.
2. The effectiveness of algorithms and software tools (both desktop and web applications) developed to automate the process of obtaining estimates of a number of statistical parameters and the degree of complexity of texts in the Azerbaijani language.
3. The results of the analysis of textbooks based on the developed methodology and the correspondence of textbooks to the age level of students, as well as recommendations and prospects for improving the readability of texts.

The scientific novelty of the dissertation. As part of the dissertation, the following new scientific **results** were obtained:

- A methodology has been developed to modify two of the most popular readability formulas for English texts (Flesch reading-ease and Flesch-Kincaid grade-level formulas);
- On the basis of a special methodology Flesch reading-ease formula and Flesch-Kincaid grade-level formula were modified for texts in Azerbaijani developed;

- For the first time, according to a special methodology, lists of 1000 and 3000 words were developed, which make up the basic vocabulary of the Azerbaijani language and are understandable to most primary school students;
- Based on the percentage of unfamiliar words in the text (not on the list of 3000 and 1000, respectively), New Dale-Chall readability formula and Spache readability formula have been modified for Azerbaijani texts;
- Special algorithms have been developed to automate the process of calculating statistical indicators of texts in the Azerbaijani language;
- For the first time, with the help of special software, the readability of a series of textbooks in the Azerbaijani language and the adequacy of the texts in these textbooks to the age level of students were evaluated on the basis of statistical indicators.

Practical significance of the work. The practical value of the dissertation is that:

- Any person will be able to determine the statistical indicators of any text in the Azerbaijani language, as well as the level of readability of the text, using the Internet resource *www.oxunabilir.az*, prepared in this dissertation and open for use;
- Authors of textbooks and teaching aids in the Azerbaijani language, as well as other books, will be able to assess the level of complexity of the texts prepared by them for the target audience with the help of this Internet resource;
- Recommendations for improving the readability of textbooks will help the authors to prepare appropriate texts;
- The mentioned Internet resource will help the members of the expert groups to automatically detect non-compliant, ie very long words and sentences during the evaluation of textbooks;
- Translators from English will have the opportunity to compare the readability of the texts they translate into Azerbaijani with the readability of the original English texts.

Approbation of the work. The process of approbation of scientific-theoretical and practical results of the researches was carried out through speeches and discussions at various scientific conferences, seminars:

- “Fasiləsiz pedaqoji təhsildə elektron təlim texnologiyalarının tətbiqi” respublika elmi-metodik konfransı (Bakı, Azərbaycan Dövlət Pedaqoji Universiteti, 18–19 iyun 2010);
- “İnformasiya sistemləri və texnologiyalar: nailiyyətlər və perspektivlər” beynəlxalq elmi konfransı (Sumqayıt Dövlət Universiteti, 15–16 noyabr 2018);
- “International Congress of Engineering and Natural Sciences Studies”, Ankara/Turkey, 07-09 May 2021;
- Scientific seminars of the Institute of Information Technologies of ANAS (2011, 2017).

At the same time, the scientific-theoretical and practical results of the research were applied in the process of preparation and evaluation of textbooks and teaching aids in various institutions that prepare teaching resources.

Scientific publications. 11 scientific works on the topic of the dissertation were published. 8 of them were published in peer-reviewed scientific journals, 2 reports in the materials of the International scientific conference, 1 report in the materials of the republican conference and 1 "Express-information".

The structure and volume of the work. The dissertation consists of an introduction, 4 chapters, a conclusion, a list of 120 references and one appendix, 18 tables and 13 figures.

BRIEF OVERVIEW OF THE WORK

In the introduction substantiates the relevance of the dissertation, the purpose of the study and the issues to be resolved. The scientific novelty and practical significance of the results obtained are shown.

The first chapter is devoted to the study of quality indicators of texts and methods of their evaluation.

It is known that people, especially students, get most of their knowledge from books. According to some researchers, the role of textbooks in the formation of knowledge in students is greater than the explanations of teachers. From a psychological point of view, reading is the reception and processing of graphically encoded information in a text. According to research, one of the main factors hindering the development of reading skills in students is the complexity of textbooks. Students do not understand overly complex texts and become accustomed to memorizing them.

Although often used interchangeably in ordinary speech, *complexity* and *difficulty* are, in fact, different concepts. The *complexity of the text* does not depend on the person reading it, it is an objective feature of the text.⁷ Its complexity can be changed by changing the value of the components of the text, ie the length of the sentences, the ratio of abstract words and other parameters.

The complexity of the text makes it difficult to understand. In general, difficulties in the teaching process are natural, because students develop by overcoming difficulties. The *difficulty of the text* depends not only on its complexity, but also on the level of preparation of the reader. The same text can be easy for a prepared reader and difficult for an unprepared reader.⁸

The *difficulty of a text* is determined experimentally by the results of its comprehension. The *complexity of the text* is determined by analyzing the text (for example, the percentage of unfamiliar words,

⁷ Лурия А.Р. Основные проблемы нейролингвистики. Москва: Изд-во МГУ, 1975.

⁸ Микк Я.А. Оптимизация сложности учебного текста. Москва: Просвещение, 1981.

the length of sentences, the complexity of the logical structure and other components).

A number of methods have been developed to analyze the quality of the language of texts and it has been proved that it is possible to reduce the complexity of texts and make them more understandable. In other words, you can adjust the text to the required level by defining and changing the parameters that complicate the text.

The existing methods for determining the level of complexity of texts can be divided into two major categories: *psychological methods* and *statistical methods*. Currently, there are many methods that belong to both the first and second categories. Since this dissertation is entirely devoted to statistical analysis of texts, this chapter provides a brief overview of the available methods for assessing the complexity of the text by psychological methods.

Several experimental methods have been developed in psychology to study the quality of texts, especially teaching texts: questioning, complementary methods, expert assessment, generalization of content, planning of the text, letter-by-letter finding, intonation reading, speaking in one's own words, reading speed and so on.

In psychology, several experimental methods have been developed for studying the quality of texts, especially educational texts: expert assessment, generalization of content, drawing up a text plan, letter-by-letter finding of a text, reading by intonation, reading speed, and so on.

The second chapter looks at the history of statistical assessment of the complexity of texts, explains the essence of the concept of "readability", provides information on the widely used formulas for English, as well as formulas developed for other languages. At the end of the chapter, the shortcomings of readability formulas are noted.

Scientific research on the complexity of texts began in the United States in the late 19th century.

The term "*readability*" is used in English to indicate the level of complexity of the text, its degree of comprehension. The word is used in Russian as "*читабельность*" and in Turkish as "*okunabilirlik*" in the "English-Azerbaijani dictionary" as "*asan oxuna bilmə*", "*oxunuşu asan olma*" ("not easy to read"). Of course, none of these translations

can be used as a term. Since the term "*oxuna bilir*" more accurately characterizes texts with the intended features, we will use the term "readability" in our language as "*oxunabilirlik*".

William Gray and Bernice Leary identified more than 200 variables that affect readability and grouped them into the following four categories: *content*, *style*, *design*, and *structure*.⁹ They found that content, format, and structure could not be measured statistically (although many researchers have since attempted to measure content statistically). Although these three groups of factors are not overlooked, W. Gray and B. Leary focused on 80 style variables, 64 of which can be reliably calculated.¹⁰

The authors used correlation coefficients to determine the most important of these 64 variables that affect the complexity of the text, in other words, the best readability. As a result, 17 factors that most affected the readability of the text were identified (Table 1).

Table 1.
Variables related to reading difficulty

№	The name of the parameter
1	Average sentence length in words
2	Percentage of easy words
3	Number of words not known to 90% of sixth-grade students
4	Number of "easy" words
5	Number of different "hard" words
6	Minimum syllabic sentence length
7	Number of explicit sentences
8	Number of first, second, and third-person pronouns
9	Maximum syllabic sentence length
10	Average sentence length in syllables
11	Percentage of monosyllables
12	Number of sentences per paragraph

⁹ Gray W.S., Leary B. What makes a book readable. Chicago: Chicago University Press, 1935.

¹⁰ DuBay W.H. The Principles of Readability. Costa Mesa, California: Impact Information, 2004.

13	Percentage of different words not known to 90% of sixth-grade students
14	Number of simple sentences
15	Percentage of different words
16	Percentage of polysyllables
17	Number of prepositional phrases

In the 1920s, educators in the United States discovered a way to use word difficulty and sentence length to predict the level of complexity of a text. By 1980, there were more than 200 readability formulas, and thousands of published studies confirmed that these formulas were theoretically and statistically very powerful.

Readability formulas are simply mathematical formulas obtained by regression analysis. This procedure finds an equation that expresses the relationship between two variables. Thus, one of these variables indicates the extent of difficulty that people face when reading a given text, and the other indicates the extent of the linguistic characteristics of the text.¹¹

In addition to the term *readability formulas*, scientific and popular literature also uses the terms *readability tests* and *readability metrics*.

Although not all readability formulas are popular, the following widely used formulas should be noted:

- Flesch Reading Ease Formula;
- Flesch-Kincaid Grade Level Formula;
- Fry Readability Formula;
- Gunning Fog Index;
- Dale-Chall Readability Formula;
- SMOG Readability Formula;
- Spache Readability Formula;
- Powers-Sumner-Kearl Readability Formula;
- FORCAST Readability Formula;
- Automated Readability Index.

¹¹ McLaughlin G.H. SMOG grading – a new readability formula, *Journal of reading*, 1969, 22: pp.639–646.

It should be noted that most of the proposed formulas are based on the linear regression model, and the variables of this model are the statistical parameters of the text:

$$f(x, b) = b_0 + \sum_{i=1}^k b_i x_i \quad (1)$$

where b_i – regression parameters (coefficients); x_i – regressors (model complexity factors); k is the number of factors in the model.

Readability formulas based on the linear regression model typically used 2–3 parameters related to vocabulary and syntax. In these formulas, the main application of which is the evaluation of teaching texts, the regression coefficients are chosen so that the results show the level of education or age of the reader in order to understand the proposed text.

Much of the chapter is devoted to the description of the readability formulas mentioned above. At the same time, a summary of research in the field of readability for the Turkish language is given. The paragraph is also devoted to readability formulas for Russian, German, Swedish, Chinese, Korean, Estonian and Ukrainian languages.

The most popular of the readability formulas is *the Flesch reading ease formula*. This formula uses two variables as factors that affect "ease of reading" – the average length of sentences in the text and the average number of syllables in words:

$$K_{en} = 206.835 - (1.015 \times S) - (84.6 \times W), \quad (2)$$

where K_{en} – reading ease score, S – average sentence length in words (the number of words divided by the number of sentences), W – average word length in syllables (the number of syllables divided by the number of words).¹²

¹² Flesch R. Marks of a readable style. Columbia University contributions to education, 1943, no. 897. New York: Bureau of Publications, Teachers College, Columbia University.

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. Reading ease score can range from 100 to 0 (Table 2).

Table 2.
Flesch’s reading-ease scores

Reading ease score (K_{en})	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, *the Flesch-Kincaid grade level formula*, is an improved version of the Flash readability 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:

$$GL_{en} = (0.39 \times S) + (11.8 \times W) - 15.59 \quad (3)$$

The result (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.

One of the most popular and reliable readability formulas was proposed in 1948 by the American educator Edgar Dale and

psychologist and writer Jeanne Chall.¹³ *The Dale-Chall readability formula* is designed to overcome some of the shortcomings of the Flash readability formula. The average sentence length and percentage of “hard words” were used in this formula. By "hard words" we mean words that are not included in the list of 3,000 simple words compiled by the authors, 80% of which are known to the reader at the 4th grade level. The main idea is that the reader is more receptive to and understands familiar words, which means that the use of more frequently used words improves the quality of comprehension of the text.

In 1995, E. Dale and J. Chall revised the 3,000-word list and published a new version of their formula.¹⁴ *The new Dale-Chall readability formula* takes into account the advantages of its predecessors and more accurately assesses the readability of the text. This formula is as follows:

$$\text{Raw Score} = 0.1579 \times PDW + 0.496 \times ASL, \tag{4}$$

where *PDW* – percentage of difficult words (words not on the Dale-Chall word list); *ASL* – average sentence length in words.

If the *PDW* is greater than 5%, 3.6365 is added to the initial score to get the final score, ie *Adjusted Score* = *Raw Score* + 3.6365, otherwise *Adjusted Score* = *Raw Score*.

Table 3.
Table of correspondence to the level of education of the value
obtained from the Dale-Chall formula

Adjusted Score	Grade Level
4.9 and below	Grade 4 and below

¹³ Dale E., Chall J.S. A formula for predicting readability. Educational research bulletin. 1948, Jan. 21 and Feb. 17, 27: 1-20, 37–54.

¹⁴ Chall J.S., Dale E. Readability revisited, the new Dale-Chall readability formula. Cambridge, MA: Brookline Books, 1995.

5.0 to 5.9	Grades 5-6
6.0 to 6.9	Grades 7-8
7.0 to 7.9	Grades 9-10
8.0 to 8.9	Grades 11-12
9.0 to 9.9	Grades 13-15 (college
10 and above	Grades 16 and above (college graduate)

The Dale-Chall formula is one of the most reliable readability formulas and is widely used in scientific research.

Proposed in 1953 and called *the Spache readability formula*, the formula is designed to assess the readability of texts for primary education (grades 1–4).¹⁵ As with the Dale-Chall formula, this formula is based on the average sentence length in the text and the percentage of "difficult" words. However, George Spache offered his own list of about 1,000 words that elementary school students knew.

$$\text{Index Spache} = 0.141 \times \text{ASL} + 0.086 \times \text{PDW} + 0.839, \quad (5)$$

where **PDW** – percentage of difficult words (words not on the Spache word list); **ASL** – average sentence length in words.

The list of words was determined as a result of an experiment with 3rd graders in US schools. Thus, the word that 80% of students mentioned as they knew it was included in the *list of familiar words*.

Although readability formulas for measuring the complexity of texts are very useful and objective, they also have shortcomings. The shortcomings of these formulas have been discussed and debated by a number of researchers. Some of the major shortcomings of readability formulas identified are:

- *They cannot measure conceptual complexity*: no formula takes into account the content of the document being evaluated. For example, according to the results of some formulas, the

¹⁵ Spache G. A new readability formula for primary-grade reading materials // Elementary school journal, 1953, 53: pp.410–413.

readability of Einstein's theory of relativity is at the level of 5th grade.¹⁶

- *They cannot verify that the expression is incomprehensible:* if the words in the text are moved and the text is encrypted, the readability score remains the same.
- *The results of the readability formulas for the same text are inconsistent:* for example, for a text rated by the Flesch-Kincaid formula at 16.7, the Gunning index may be 22.6 and the SMOG value may be 18.5. The reason for this difference is that different formulas use different variables and different evaluation criteria.
- *They believe that all readers are the same:* readability formulas do not take into account the characteristics of readers. That is, they do not take into account the fact that readers have different goals, their level of maturity and skills.¹⁷

The third chapter is devoted to the modification of four popular readability formulas for English based on a specific methodology: Flesch reading-ease, Flesch-Kincaid grade-level, new Dale-Chall readability and Spache readability formulas.

It should be noted that all these formulas were obtained experimentally for texts in English. To apply these formulas to Azerbaijani texts, it is necessary to adjust the coefficients of variables *S* (*ASL*) and *W*, because unlike English, which is an inflectional language type, Azerbaijani is an agglutinative language in terms of morphological type. The average sentence length in this language is shorter than in English, while the average word length, on the contrary, is longer. To determine the percentage of difficult words (*PDW*), it is necessary to prepare similar lists for the Azerbaijani language.

Such a methodology was used to determine the ratio between the average sentence length and the average word length in both languages. First, the statistical indicators of different texts in the

¹⁶ U.S. State Department. A Plain English Handbook: How to create clear SEC disclosure documents, 1998.

¹⁷ Redish J. Readability Formulas Have Even More Limitations Than Klare Discusses // ACM Journal of Computer Documentation, 2000, 24(3), pp.132–140.

Azerbaijani language and the English originals of those texts and the ratio of the values of the defined parameters in both languages were calculated. Then, the average values of these parameters were calculated and the ratio to which the corresponding coefficients in the Flash formula should be corrected was determined. Some popular examples of English literature and their translations into Azerbaijani were used to ensure that the texts in both languages were level in terms of content and style (Table 4).

Table 4.

Comparison of the quantitative characteristics of the identical literary samples in English and Azerbaijani

	Names of fictions	Number of sentences	Number of words	Number of syllables	A S L	A S W	ASL _{en} / ASL _{az}	ASW _{en} / ASW _{az}
1	2	3	4	5	6	7	8	9
1	Ernest Hemingway. Nobody Ever Dies	540	5264	7063	9.75	1.34	0.67	1.84
	Ernest Heminquey. Heç kim heç vaxt ölmür	594	3865	9489	6.51	2.46		
2	Gabriel García Marquez. Monologue of Isabel Watching It Rain in Macondo	178	2902	3529	16.30	1.22	0.72	2.04
	Qabriel Qarsia Markes. İsabel Makondada yağışa baxır	157	1847	4602	11.76	2.49		
3*	Mark Twain. The Adventures of Tom Sawyer	142	1946	2371	13.70	1.22	0.72	1.83
	Mark Tven. Tom Soyerin macəraları	158	1549	3460	9.80	2.23		
4*	John Galsworthy. Beyond	352	7119	9001	20.22	1.26	0.76	2.02

	Con Qolsuorsi. Ölümdən güclü	366	5566	14199	15.27	2.54			
5	Herbert Wells. The Crystal Egg	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	362	5675	7428	15.68	1.31	0.69	1.82	
	Cek London. Qadın cəsarəti	400	4344	10337	10.86	2.38			
7	John Steinbeck. The Chrysanthemums	448	4220	5742	9.42	1.36	0.88	1.87	
	Con Steynbek. Xrizantemlər	369	3077	7743	8.34	2.52			
8	William Somerset Maugham. Louise	149	2118	2437	14.21	1.15	0.77	2.03	
	Uilyam Somerset Moyem. Luiza	196	2158	5040	11.01	2.34			
9*	Agatha Christie. Murder on the Orient Express	334	4737	6047	14.18	1.28	0.63	1.99	
	Aqata Kristi. Şərq ekspessində qətl	425	3795	9683	8.93	2.55			
10 *	Oscar Wilde. The Picture of Dorian Gray	363	4950	6097	13.63	1.23	0.76	1.93	
	Oskar Uayld. Dorian Qreyin portreti	422	4369	10414	10.35	2.38			
	T o t a l (en)	3169	45809	59039					
	T o t a l (az)	3433	35459	87570					
							average	0.72	1.93
							variance	0.0059	0.0076
							min	0.62	1.82
							max	0.88	2.04

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.¹⁸ Thus, the coefficient of the average sentence length (*S*) in formula (2) 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:

$$K_{az} = 206.835 - (1.318 \times S) - (44.3 \times W). \quad (6)$$

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:¹⁹

$$GL_{az} = (0.51 \times S) + (6.18 \times W) - 15.59 \quad (7)$$

The next modified formulas for Azerbaijani texts are *the new Dale-Chall readability formula* (4) and *the Spache readability formula* (5).

There is no problem editing the coefficient of the average sentence length (*ASL*) in formulas (4) and (5): English sentences are on average 0.77 times longer than sentences in Azerbaijani, so in the both new Dale-Chall readability formula and Spache readability formula, the mean sentence length coefficient (*ASL*) should be corrected 0.77 times. After making this correction, *new Dale-Chall readability formula for Azerbaijani texts* will look like this:

¹⁸ Sadıqov İ.C. Azərbaycan dili mətnlərinin mürəkkəbliyinin qiymətləndirilməsi üçün modifikasiya olunmuş Fleş düsturu // "İnformasiya texnologiyaları problemləri" jurnalı, 2018, №1, s.46–58.

¹⁹ Sadıgov I.J. Mathematical and information models for evaluating readability of texts in Azerbaijani language // *El-Cezeri Journal of Science and Engineering*, V.5 – N.3, 2018, pp. 888–903.

$$\begin{cases} 0.1579 \times \mathbf{PDW} + 0.644 \times \mathbf{ASL} + 3.6365, & \mathbf{PDW} > 5\% \\ 0.1579 \times \mathbf{PDW} + 0.644 \times \mathbf{ASL}, & \mathbf{PDW} \leq 5\% \end{cases} \quad (8)$$

Table 3 shows the grade level to which the Dale-Chall formula corresponds.

After adjusting the Spache readability formula in the same way, *the Spache readability formula for Azerbaijani texts* will be as follows:²⁰

$$\mathbf{GL} = 0.183 \times \mathbf{ASL} + 0.086 \times \mathbf{PDW} + 0.839 \quad (9)$$

The main problem is to calculate the value of the **PDW** parameter, ie the percentage of compound words, to apply formulas (8) and (9) to Azerbaijani texts. Because for this, it is important to have analogues for the Azerbaijani language of the 3,000-word list prepared for English by E. Dale and J. Chall, as well as the list of about 1000 words known to primary school students proposed by J. Spache. However, as there are no similar lists for the Azerbaijani language so far, such lists have been prepared as part of this research.

For the first time in this dissertation, lists of the most frequently used 1000 and 3000 words in the Azerbaijani language were prepared. Such a methodology was used for this purpose. First, the words in the "Explanatory Dictionary of the Azerbaijani language" ("Azərbaycan dilinin izahlı lüğəti") were considered one by one, and the words that are intuitively considered to be known by everyone in our language were included in the list.²¹ In order to make a decision on the inclusion in the list of words that may raise questions, the frequency of their use in the "Frequency Dictionary of the Azerbaijani Language"

²⁰ Sadıgov İ.C. Azərbaycan türkçesindeki metinlər için kelime sıklıklarına dayalı okunabilirlik formülleri // "Uluslararası Mühendislik ve Doğa Bilimleri Çalışmaları Kongresi", Ankara, 07-09 may 2021

²¹ Azərbaycan dilinin izahlı lüğəti. Dörd cildə, Bakı: Şərq-Qərb, 2006.

("Azərbaycan dilinin tezlik lüğəti") was considered.²² Then the texts in the textbook "Azerbaijani language", as well as in textbooks on other subjects currently used in primary classes, were processed by the software and words were added that were used in them, but were not included in the list. The list of more than 3,000 words (3333 words) received was distributed to people of different age groups and they were asked to mark the words they found difficult. The most frequently mentioned words were reviewed by experts, the list was amended and 3,000 words were retained. The list of 3,000 words was once again evaluated by experts, and based on it, a list of 1,000 words that are not difficult for an elementary school student was prepared.

The fourth chapter is devoted to the automation of the procedure for assessing the complexity of texts in the Azerbaijani language.

Almost simultaneously with computers, the first computer programs that calculated the quantitative characteristics of texts also appeared. Currently, there are many such software products for different natural languages, including free (free) programs. In order to assess the level of complexity of the texts in the Azerbaijani language, two software programs have been developed within the framework of this dissertation. One of them is "*Mətn analizi*" ("*Text Analysis*") application, and the other is the website *www.oxunabilir.az*.²³

The program "*Mətn analizi*" is designed to calculate the statistics of texts in the Azerbaijani language. The program also calculates the frequency of occurrence of words in the text, and allows you to sort this frequency list by both alphabetical and frequency values.

The main purpose of creating the website *www.oxunabilir.az* is to present the opportunities provided by the "*Mətn analizi*" program to a wide range of users. Like the "*Mətn analizi*" program, this resource calculates the statistics of text entered from a keyboard or an existing *.doc* file. Then, based on parameters such as average sentence length

²² Mahmudov M., Fətullayev Ə. və b. Azərbaycan dilinin tezlik lüğəti, I cild. Bakı: Elm, 2010.

²³ Sadıqov İ.C. Azərbaycan dili mətnlərinin mürəkkəbliyinin qiymətləndirilməsi üçün modifikasiya olunmuş Fleş düsturu // "İnformasiya texnologiyaları problemləri" jurnalı, 2018, №1, s.46–58.

and average word length, it evaluates the readability of the text according to the Flesch reading-ease formula (6) and Flesch-Kincaid grade-level formula (7).

As noted, the first studies to assess the complexity of texts began in the United States, where the first readability formulas also appeared. It is natural that the first applications of these formulas were in the United States. The first to benefit from the application of these formulas were school teachers and librarians. Thus, the selection of textbooks for the teaching process, which assessed the level of readability, greatly facilitated the work of teachers. At the same time, the results of other books (especially fiction) involved in the readability assessment were sent to libraries so that librarians could advise readers on choosing the right book. In addition, for many years now, publishers in the United States have used these formulas to measure the level of complexity of their manuscripts and return extremely complex manuscripts.

The formulas developed in the former Soviet Union were also applied primarily to teaching texts.²⁴

Textbooks were also the first applications of readability formulas developed for the Turkish language.²⁵

The first field of application of readability formulas developed for texts in the Azerbaijani language, of course, were textbooks.²⁶ Thus, using formulas (6) and (7), the level of readability of individual texts in some textbooks used in the primary grades of secondary schools in the 2017-2018 academic year was first assessed (Table 5).

²⁴ Микк Я.А. Оптимизация сложности учебного текста. Москва: Просвещение, 1981.

²⁵ Zorbaz K.Z. Türkçe Ders Kitaplarındaki Masalların Kelime-Cümle Uzunlukları ve Okunabilirlik Düzeyleri Üzerine Bir Değerlendirme // "Eğitimde Kuram ve Uygulama", 2007, Sayı: 3 (1), s.87-101.

²⁶ Алгулиев Р.М., Садыгов И.Дж. Оценивание удобочитаемости учебников на азербайджанском языке // "ScienceRise", 2018, №11(52), s.50–57

Table 5.
Indexes of readability of texts in some textbooks of primary classes
of secondary schools.

The name of the topic	Words / Sentences	Syllables / Words	Flesch reading-ease	Flesch-Kincaid grade-level
"Azərbaycan dili", 3-cü sinif. Fitnə	7.13	2.34	93.80	3
"Azərbaycan dili", 3-cü sinif. Onluq say sistemi	10.00	2.26	93.35	3
"Azərbaycan dili", 4-cü sinif. Hikmət xəzinəsi	12.11	2.39	85.20	5
"Azərbaycan dili", 4-cü sinif. İlanlar niyə rəqs edir	10.14	2.62	77.32	6
"Həyat bilgisi", 3-cü sinif. Göbələklər	8.50	2.97	64.03	7
"Həyat bilgisi", 3-cü sinif. Təhlükəli təbiət hadisələri	10.00	2.76	71.39	7
"Həyat bilgisi", 4-cü sinif. Coğrafi təbəqə	12.07	2.77	68.08	8
"Həyat bilgisi", 4-cü sinif. Uşaq hüquqları konvensiyası	11.59	3.01	58.14	9
"İnformatika", 3-cü sinif. Budaqlanma	8.09	2.69	76.96	5
"İnformatika", 3-cü sinif. Qovluq	8.00	2.73	75.50	5
"İnformatika", 4-cü sinif. Elektron poçt və İnternet	8.23	2.75	74.27	5
"İnformatika", 4-cü sinif. Qrafik redaktorun alətləri	10.71	2.53	80.68	5

"Texnologiya", 3-cü sinif. Naxışkəsmə. Qayçının tarixindən	9.50	2.92	64.91	7
"Texnologiya", 3-cü sinif. Texnoloji maşınlar. Yerüstü və yeraltı nəqliyyat vasitələri	10.84	2.85	66.10	8
"Texnologiya", 4-cü sinif. Emal texnologiyaları və vasitələri	12.82	3.06	54.21	10
"Texnologiya", 4-cü sinif. İstehsalat müəssisələri və istehsal	16.11	3.04	50.87	11
"Musiqi", 4-cü sinif. Azərbaycan xalq musiqisi janrları	13.17	2.80	65.55	8
"Musiqi", 4-cü sinif. Azərbaycan bəstəkarlarının yaradıcılığında ərəb xalq musiqisi	19.50	2.75	59.31	11

At the same time, this dissertation examines how the readability of textbooks changes from lower grades to higher grades.

Statistical analysis has shown that some textbooks contain sufficiently complex texts that do not correspond to the age level of students. This complexity is related to both the average sentence length and the average word length.²⁷

At the end of the chapter, a number of recommendations were made to improve the readability of textbooks.

²⁷ Sadıqov İ.C. Azərbaycan dilindəki mətnlər üçün oxunabilirlik düsturları əsasında dərsliklərin mürəkkəbliyinin qiymətləndirilməsi // "Kurikulum" jurnalı, 2018, №3, s.12–24.

RESULTS

The scientific problems raised in the process of research on the topic of the dissertation were solved by the author and the following results were obtained:

1. The need to establish mathematical models to assess the level of complexity of the texts and to modify the existing assessment formulas for English texts, taking into account the specifics of the Azerbaijani language, is justified;
2. A technique has been developed for modifying the Flesch reading-ease formula and Flesch-Kincaid grade-level formula, one of the most popular formulas for the readability of English texts, for texts in the Azerbaijani language;
3. Based on the developed special technique, the Flesch reading-ease formula and Flesch-Kincaid grade-level formula for Azerbaijani texts were modified;
4. For the first time, a special methodology was prepared and applied for compiling lists of 1000 and 3000 words that make up the basic vocabulary of the Azerbaijani language and are understandable to most primary school students;
5. Based on the percentage of unfamiliar words in the text (not on the list of 3000 and 1000, respectively), new Dale-Chall readability and Spache readability formula have been modified for Azerbaijani texts;
6. Special algorithms have been developed to automate the process of calculating the statistical indicators (number of sentences, number of words, number of syllables, etc.) of texts in the Azerbaijani language;
7. Relevant software tools have been developed on the basis of the developed algorithms in order to automate the process of calculating the statistical indicators of the texts in the Azerbaijani language, as well as the application of the modified readability formulas;
8. For the first time, the readability of a number of textbooks in the Azerbaijani language and the correspondence of the texts of these textbooks to the age level of students were assessed

without subjective considerations, with the help of special software;

9. For the first time, mathematical models and relevant software developed for texts in the Azerbaijani language made it possible to observe how the level of complexity of texts in individual subjects changed from class to class;
10. Scientific-based tools have been created for textbook authors in the Azerbaijani language, teachers preparing methodological materials, experts engaged in the examination of teaching aids to objectively assess and improve the level of complexity of texts on a number of parameters in a short time.

The following scientific works on Dissertation materials have been published:

1. Əliquliyev R.M., Sadıqov İ.C. Tədris mətnlərinin keyfiyyəti və onların psixoloji metodlarla qiymətləndirilməsi // “Fasiləsiz pedaqoji təhsildə elektron təlim texnologiyalarının tətbiqi” respublika elmi-metodik konfransının materialları, Bakı, 18–19 iyun 2010, s.58–61.
2. Sadıqov İ.C. Müasir dövrdə tədris mətnlərinin qiymətləndirilməsi zərurəti və problemləri haqqında // “İnformasiya cəmiyyəti problemləri” jurnalı. 2011, №1, s.97–102.
3. Sadıqov İ.C. Mətnlərin mürəkkəbliyi və onun qiymətləndirilməsi yolları // Ekspres-informasiya. İnformasiya cəmiyyəti seriyası. Bakı: İnformasiya Texnologiyaları, 2012, 72 s.
4. Sadıqov İ.C. Mətnlərin mürəkkəbliyi və onun qiymətləndirilməsi yolları (I) // "Kurikulum" jurnalı, 2013, №2, s.30–42.
5. Sadıqov İ.C. Mətnlərin mürəkkəbliyi və onun qiymətləndirilməsi yolları (II) // "Kurikulum" jurnalı, 2013, №3, s.11–29.
6. Sadıqov İ.C. Azərbaycan dilindəki mətnlər üçün işlənmiş riyazi və informasiya modelləri əsasında dərsləklərin oxunabilirliyinin qiymətləndirilməsi // “İnformasiya sistemləri və texnologiyalar: nailiyyətlər və perspektivlər” Beynəlxalq elmi konfransın materialları, Sumqayıt, 15-16 noyabr 2018, s.415–417.
7. Sadıqov İ.C. Azərbaycan dili mətnlərinin mürəkkəbliyinin qiymətləndirilməsi üçün modifikasiya olunmuş Fleş düsturu // "İnformasiya texnologiyaları problemləri" jurnalı, 2018, №1, s.46–58.
8. Sadıqov İ.C. Azərbaycan dilindəki mətnlər üçün oxunabilirlik düsturları əsasında dərsləklərin mürəkkəbliyinin qiymətləndirilməsi // "Kurikulum" jurnalı, 2018, №3, s.12–24.
9. Sadigov I.J. Mathematical and information models for evaluating readability of texts in Azerbaijani language // El-

Cezeri Journal of Science and Engineering, V.5 – N.3, 2018, pp. 888–903.

10. Алгулиев Р.М., Садыгов И.Дж. Оценивание удобочитаемости учебников на азербайджанском языке // "ScienceRise", 2018, №11(52), s.50–57.
11. Алгулиев Р.М., Садыгов И.Дж. Трудности вычисления статистических показателей текстов на азербайджанском языке и алгоритмы их решения // Телекоммуникации, 2019, №3, s.42–51.
12. Sadıgov İ.C. Azerbaycan türkçesindeki metinler için kelime sıklıklarına dayalı okunabilirlik formülleri // “Uluslararası Mühendislik ve Doğa Bilimleri Çalışmaları Kongresi”, Ankara, 07-09 may 2021, s.14.

The personal role of applicant in works published with co-authors:

[1, 10, 11] – the statement of the problem belongs to RM Alguliyev, the development of methods and experiments were carried out by another co-author.

The defense will be held on **29 April** at **15⁰⁰** at the meeting of the Dissertation council ED 1.35 of Supreme Attestation Commission under the President of the Republic of Azerbaijan operating at the Institute of Information Technology of the Azerbaijan National Academy of Sciences.

Address: AZ1141, Azerbaijan Republic, Baku, B.Vahabzade str., 9A

Dissertation is accessible at the library of the Institute of Information Technology of the ANAS.

Electronic versions of dissertation and its abstract are available on the official website of the Institute of Information Technology of the ANAS (*ict.az*).

Abstract was sent to the required addresses on **15 Mart 2022**.

Signed for print: 14.03.2022
Paper format: 60 × 80^{1/16}
Volume: 35052 characters
Number of hard copies: 20