TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Silva Sammelsaar 153018IASM

VIABILITY OF PROGNOSING THE SUCCESS OF ACADEMIC WRITING THROUGH TEXT READABILITY ANALYSIS

Master's thesis

Supervisor: Tarmo Robal
PhD

Tallinn 2018

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Silva Sammelsaar 153018IASM

AKADEEMILISE KIRJUTISE EDUKUSE PROGNOOSIMISE VÕIMALIKKUSEST TEKSTI LOETAVUSE ANALÜÜSI BAASIL

magistritöö

Juhendaja: Tarmo Robal
PhD

Tallinn 2018

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Silva Sammelsaar

06.01.2019

Abstract

The present thesis investigates readability of academic writing and possible associations to the human evaluation of the writing. Human evaluation is a manually assigned grade that indicates the success of the writing. The readability grade of the writing is computed by applying the calculation methods of five readability algorithms. Readability grade level indicates, how many years of education is needed to easily understand a certain text. The obtained manual and automatically computed scores are then analysed to determine the strength of the correlation between the grades. The aim of the thesis is to understand, whether the manual grade of the writing could be affected by linguistic elements of the text – number of words, sentences, characters, syllables, the use of complex words. Conducting the readability analysis answered the research question, whether the writings with higher manually assigned grades are easier to read.

The outcome of the readability analysis of academic writings proved that there is no strong significant correlation between manual and automatically computed grades. The analysis identified only a moderate correlation between one of the readability methods and the success of the writing. The readability analysis also showed that the readability level of writings with lowest grade also tend to be lower. Though not determining any firm correlation between the grades, the academic writings with highest human assigned score tend to be less intelligible and more difficult to read. As moderate correlation is not sufficient to indicate the potential success of the writing, the results of the thesis suggest that the success of the academic writing cannot be predicted through text readability analysis.

This thesis is written in English and is 45 pages long, including 6 chapters, 4 figures and 13 tables.

Annotatsioon

Akadeemilise kirjutise edukuse prognoosimise võimalikkusest teksti loetavuse analüüsi baasil

Käesolev töö uurib akadeemilise kirjutise loetavust, viies selleks läbi teksti loetavuse analüüsi. Töö eesmärgiks on selgitada välja võimalikud seosed kirjutise loetavuse ja talle retsensendi poolt määratud hinde vahel. Töö loetavuse hindamiseks rakendatakse viite loetavuse algoritmi, mis tuvastavad, milline on sisendteksti keerukuse tase. Teksti loetavus näitab, millisele tasemele peaks vastama lugeja omandatud haridus, selleks et antud tekst oleks kergelt arusaadav ja mõistetav. Akadeemiliste kirjutiste loetavuse hindamine ja teksti keerukuse määramine viiakse läbi selleks arendatud eksperimentide keskkonnas. Loetavuse eksperimentide käigus leitakse vastus küsimusele, milline on seos tööle antud hinde ja tema loetavuse vahel ning kas teksti loetavuse analüüsi abil on võimalik prognoosida töö edukust. Antud töö uurib, kas akadeemilise kirjutise edukus on mõjutatud teksti semantilistest omadustest – lausete koguarv, lausete pikkus, silpide, sõnade ja tähemärkide arv, keeruliste sõnade esinemissagedus. Samuti selgitakse korrelatsiooni arvutuste tulemuste põhjal välja, milline töös kasutatavatest loetavuse algoritmidest annab parimaid tulemusi analüüsi läbiviimiseks ja kas kõrgema hinde saanud tööl on ka parem loetavus.

Akadeemiliste kirjutiste teksti loetavuse analüüsi tulemuste põhjal ei selgunud märkimisväärset seost töö edukuse ja tema loetavuse vahel. Analüüsi tulemusel selgus, et akadeemilise kirjutise ja tema teksti loetavuse vahel esineb tagasihoidlik korrelatsioon. Kuigi teksti analüüs näitas, et madalaima hinde saanud tööl on ka madalam ehk parem loetavus, ei ole loetavuse analüüsi põhjal saadud tagasihoidlik korrelatsioon piisavalt tugev, et prognoosida töö edukust.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 45 leheküljel, 6 peatükki, 4 joonist, 13 tabelit.

List of abbreviations and terms

AES	Automated Essay Scoring
APTA	American Physical Therapy Association
ARI	Automated Readability Index
AS	Analytical score
AVG_CP	Average contribution and presentation
AVG_OCP	Average originality, contribution and presentation
AVG_R	Average readability
BRI	Bormuth's Readability Index
CLI	Coleman-Liau Index
DC	Dale-Chall Score
FINAL_G	Final grade
FK	Flesch-Kincaid Grade Level
FRE	Flesch Reading Ease
GF	Gunning Fog
GUI	Graphical User Interface
LW	Linsear Write
PC	Presentation and contribution
PS	Presentation score
REV_G	Reviewer's grade
SMOG	Simple Measure of Gobbledygook
TC	Theoretical contribution
US	United States

Table of contents

List of figures

List of tables

1 Introduction

There is a lot of various textual material that surrounds us every day. Whether it is a book, article, content of a website, blog, official document, educational material, questionnaire or advertisement – its content has to be easily comprehendible for the reader. To evaluate the reading difficulty of some certain textual content, readability measurement can be used. Readability is a computable metric, which purpose is to evaluate the level of reading difficulty of a given text by measuring its syntactic properties [1] [2]. Though there are multiple different approaches to assess the readability of a text, there are other features that impact, how easily understandable a writing is for the reader. These features include the background, motivation and knowledge of the reader, also the skill and writing style of the author. By measuring the syntactic properties of the text (word length, use of difficult words, sentence length), readability metrics can be used to indicate whether the writing is suitable for the respective audience.

There are hundreds of different readability formulas designed in the past [2] and some of them are used more frequently to evaluate readability – The Gunning Fog Index [3], the Flesch-Kincaid Grade Level [4], the Dale-Chall formula [2], FORCAST Formula [5] and many more. Many of the mostly used readability metrics are linear models with a few simple parameters based on words and sentences. These parameters include for example the average number of syllables per word, average number of words per sentence, total number of sentences, characters per word [6].

Current thesis tries to prognose the success of academic writings through evaluating their level of reading difficulty by using different readability approaches. To determine the relations between readability and success of the writing, the academic score of the writing is used in the analysis. During the analysis process, this thesis tries to answer the following research questions:

1. What is the correlation between readability level and manually assigned score of academic writing? Do works with higher score have better readability?

2. Which of the readability assessment methods gives the best results for analysing the correlation?
3. Which approach of readability evaluation performs better in terms of indicating potentially successful and high-scoring written work – the formulas based on syllable count or on character count?
4. Is there a significant difference in automatically computed readability values for academic writings by research scientists versus master level students?

To answer the aforementioned four questions, readability metrics for each of the academic writing will be computed using five different methods – Flesch-Kincaid Grade Level, SMOG (Simple Measure of Gobbledygook) Grade Level, Dale-Chall Score, Automated Readability Index and Coleman-Liau Index. The readability methods used in this thesis were selected to get variety of results based on distinct approaches and to select the best performing method in the context of academic writings. The selection criteria of chosen readability methods is discussed in more detail in Section 3.2.

The mentioned readability methods have been implemented in a readability experiment environment developed specifically for the use of this thesis. The automated tool calculates the readability metrics of the given file and outputs the reading difficulty as a US (United States) grade level. After the experiments have been executed, the analysis to search for relations between the readability results and academic scores can be carried out.

The following chapter of the thesis describes the research and studies done in readability field that are relevant in terms of the analysis. Next, Chapter 3 explains the essence of readability in more detailed manner by giving an overview of different readability algorithms used in the computerised readability calculation tool. Chapter 4 is describing the experiment environment and preliminary work that needs to be done before the readability tests can be executed. The following Chapter 5 presents readability experiments and analysis of the result. The final results of the analysis and answers for the research questions are explained in the Conclusions.

2 Related works

Though readability formulas cannot measure prior knowledge and interest level of the reader, its' algorithms can be applied in many different domains to measure the complexity of text features. As readability can indicate whether a text is possibly intelligible to a certain target group, the use of readability metrics to evaluate the complexity of a textual content, is widespread. Due to the ease of use of readability formulas and the availability of multiple automated readability computing tools, the amount of studies conducted in the field is outstanding. Much research can be found in numerous domains – for instance medicine, where the focus is on comprehension of health-related materials by patients [7]. Also, quite a lot of studies have been conducted in education area – for example evaluation of readability of academic writings, scientific articles, educational literature [8] [9] [10]. The studies concerning educational literature are especially important, while educational institutes might use readability metrics when choosing appropriate reading material for age groups. In addition, remarkable amount of readability studies have explored the understandability of textual content of different blogs and websites [11].

A study conducted in the field of education [9] estimated the level of readability of more than 700 000 scientific articles published between 1881 and 2015, gathered from 123 scientific journals. The clarity of the articles were assessed with using Dale-Chall and Flesch Reading Ease method. Dale-Chall is also used in the readability analysis in this thesis. The scientific articles analysed in the study included writings from the field of general, biomedical and life sciences. The researchers examined how the readability of an article's abstract relates to the year of its publication. The study showed a strong increasing trend of average yearly Dale-Chall and decreasing tendency in average yearly Flesch Reading Ease scores. The average number of syllables (counted by Flesch Reading Ease) and the percentage of difficult words (component of Dale-Chall formula) showed significant increase over the years. The results of the study proved that the readability of scientific articles is steadily decreasing over time and that could impact the overall accessibility and reproducibility of research findings [9].

Another study that investigated the readability level of patient educational materials on the American Physical Therapy Association (APTA) consumer website [11], analysed 14 educational brochures and obtained them a computed readability score using Flesch-Kincaid Grade Level, Flesch Reading Ease, Fry Readability and SMOG readability formulas. The formulas included in this study and also used in current thesis, are Flesch-Kincaid and SMOG. The aim of the study was to determine whether the readability level of materials published on the APTA website is too high for patients to comprehend. According to the conducted readability analysis, Flesch-Kincaid and Flesch-Reading Ease determined that over 90% of the brochures were suitable for readers with more than 6 years of education. SMOG and Fry Readability formulas also indicated that the reading difficulty of the educational materials published on the website was significantly greater than 6^{th} grade level. The results of the study suggested that majority of the patient education information available on the analysed website of health organization, obtained a level of reading difficulty that were too high for the average consumer to comprehend [11]. This study provides a valuable example where readability methods and the evaluation of reading difficulty of certain texts can be used for the benefit of the reader – the obtained computed readability level indicates the suitability of the content of the website for the targeted audience and can therefore be edited to make it more clear and intelligible for the reader.

Though much research has been done in readability area, there are not many studies to be found where the numeric value of readability is further investigated and compared to some other method of evaluating a piece of writing – e.g. scores assigned manually by human evaluator. Analysis that is related to readability and its correlation to human ratings, is often included in the studies that are assessing the validity of different Automated Essay Scoring (AES) technologies (e-rater [12], IntelliMetric [13] [14]). The purpose of AES is to grade essays automatically and thus developing models to reduce the involvement of human raters [15]. However, the impact of AES tools on replacing the human raters in evaluating the quality of essays, has not been sufficiently investigated. Critics of the AES suggest, that this kind of grading system might send wrong message to students, that overall meaning and nature of the writing is not important, since the audience of the writer is replaced by a machine. Instead, students might start to focus on writing essays that are formatted to match the highest-score algorithm. There are also studies conducted that claim the opposite and demonstrate a strong correlation between

AES and human scoring [14]. Considering the conflicting results in AES related research, the debate over the validity of automated essay scoring is still ongoing.

IntelliMetric[1], being one of the AES tools, analyses more than 400 semantic-, syntactic- and discourse-level features to form a sense of meaning of a given text and to provide a holistic score [16]. Readability under the scoring tool's domain of "Structure" is one of the features being analysed. The validity of IntelliMetric has been analysed in a study that compares automated essay scoring and human scoring [13]. The study is validating the automated essay score by comparing group of mean scores assigned by human raters and by an AES tool IntelliMetric. The test writings were collected from over 100 college students, from whom majority were of Hispanic origin. All of the participants' native language was English. The purpose of this was to represent a different population from the one whose essays were used to train the scoring model of IntelliMetric and therefore to further investigate how well the AES can be applied to scoring. The aim of the study was to find out whether the group mean score assigned by IntelliMetric differs significantly from the group mean score given by human raters on the standardized writing test. As a result of the study, the group mean scores comparison expressed relevant differences between the mean scores assigned by IntelliMetric and those given by human raters. It appeared that the AES tool tends to assign higher scores than do human raters. The descriptive statistics obtained in the study also showed that IntelliMetric appointed much higher passing rate and much lower failing rate [13].

A study carried out by M. Azizi and M. Nemati [8] compares the score calculated from readability formulas and score given by human raters. The purpose of the study was to answer the question to what extent the readability indices of text written by learners of English as a foreign language could substitute the scores given to the same texts by human users. To perform the study, they gathered writing samples from group of participants who were learning English as a foreign language. The writings were scored by two experienced raters who gave scores in the range from 1 to 17. The same samples were then analysed in terms of readability indices through the use of six different readability formulas – Flesch-Kincaid, Flesch's Reading Ease, Gunning Fog's index, SMOG index, Fry's Graph and Dale-Chall index. These include three of them that are also used in this

[1] http://www.vantagelearning.com/products/intellimetric/intellimetric-how-it-works/

thesis (Flesch-Kincaid index, SMOG index, Dale-Chall index). For formula calculations they used computerised readability assessment tools available on the Internet. The obtained readability indices and scores given by raters were then analysed to find correlation coefficient. As a result of the study, the six readability formulas and scores given by human raters to the participants' writing samples, appeared to have almost no relationship.

Another research is evaluating the performance of IntelliMetric automated scoring system to scores of human raters [14]. For the purpose of the study, 770 essays were used in the analysis. Those included 270 essays used for training IntelliMetric and 500 for validation of scores. Besides the 500 validation essays, another 13 essays were fabricated, in order to test the ability of scoring software to detect common cheating techniques. Results from the evaluation tests proved that IntelliMetric AES replicates the scores given by human raters and very few of them needed to be adjudicated to human ratings [14].

A study conducted by Nigam [17], explores the automated essay scoring for non-native English speakers by evaluating readability as one of the six features measured in nearly 900 essays. The essays were collected from undergraduate students, whose native language was not English. The length of the essays was between 150 to 400 words and they were manually scored by two raters on scale of 1 – 10. For measuring the readability score, the Flesch-Kincaid grade level algorithm was used. The results of the test showed that the readability metrics alone had a strong correlation with human scoring.

The validity of some of the commonly used readability algorithms have been analysed by Pooneh [18] in his study to explore the correlations between readers' evaluation of text readability and four readability formulas. The study concentrates on the validity of Flesch Reading Ease Score, Gunning's Fog Index, Flesch-Kincaid Grade Level and SMOG Index. The readability algorithms analysed in the research and that are also used in this thesis, are Flesch-Kincaid and SMOG. The study was conducted with 118 participants who were learning English as a foreign language. To carry out the study, 5 reading passages were randomly selected from a textbook which was being taught to the participants in their level of reading course. Then the readability scores of the 5 passages were calculated. The passages were accompanied by 10 comprehension questions, including 5 true or false and 5 multiple choice questions to determine the difficulty level of text from participants' responses. The results of the study suggested that there is no

significant correlation between human evaluation of text and scores obtained with readability formulas.

Considering the various researches conducted on the validity of readability and its comparison to scores assigned by human raters, it can be said that the results of the studies are conflicting. There are studies that find the readability formulas providing credible results and suggest that there is strong correlation with human raters' scores. On the opposite side, there are studies that did not find the correlation between automated scoring tools and manually assigned scores significant. In addition, there is one study [18] that investigated two of the formulas that are also being analysed in the current thesis – Flesch-Kincaid, SMOG – and found no remarkable correlations between evaluation of text by readers and by using the aforementioned readability formulas.

Therefore, based on the diverse research results, further investigation and analysis of the use of readability formulas and their involvement in automated essay scoring tools, is needed.

3 Readability

Dale and Chall [2] defined readability as "the sum total of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at an optimal speed or find it interesting" [2]. In natural language, the readability of a certain text depends on its visual presentation (font size, line height, line length) and on the content (the complexity of its vocabulary and syntax) [19]. This thesis focuses on readability based on the content of the text only by measuring specific text characteristics. There are numerous algorithms to calculate readability but their concepts do not differ much from each other – the typically used inputs in readability formulas are the number of words, number of sentences, character or syllable count and in some cases word frequency or list of familiar words. Readability score is a calculated index which indicates approximately the level of obtained education a reader needs to be able to read and understand a piece of text easily [1]. The aim of this thesis is to analyse and find relations between automatically computed readability scores and manually given scores to academic texts written in English by students and researchers.

3.1 Comparison of readability methods

There exist multiple formulas for calculating the readability of a given text content. These include for instance FORCAST Formula [5], Flesch Reading Ease [2], Flesch Kincaid Grade Level (FK) [4], Automated Readability Index (ARI) [20], Dale-Chall score (DC) [2], The Coleman-Liau Index (CLI) [21], The Bormuth Readability Index [22], Gunning Fog Scale Level [3], McLaughlin's SMOG Grade [23], Fry Graph Readability Formula [24], the Powers-Sumner-Kearl Readability Formula [5], Linsear Write Readability Formula [5], Spache Readability Formula [5]. The aforementioned readability formulas are amongst the most used methods for assessing readability level of a text but not all of the formulas are used and implemented in the readability calculator software developed for the present thesis. The readability algorithms in the scope of the thesis were selected for their ease of usability, difference in input parameters (some formulas use syllable count and sentence length while others use characters per word count or previously

defined list of familiar words), widespread usage and suitability to evaluate the readability of the academic writings. The Powers-Sumner-Kearl and Spache readability formulas are aimed to assess the readability of texts suitable for primary age children and therefore are not considered ideal to evaluate academic writings [5]. Fry Graph readability formula is using a specific graph-based scale to determine the reading difficulty of a text [24]. To follow the selection criteria of ease of usability and implementation, also the suitability for evaluating the readability of academic writings, Spache, Powers-Sumner-Kearl and Fry Graph formulas are excluded from further comparison of readability algorithms.

One of the selection criteria for the variety of methods involved in the readability evaluation, was to include commonly used equations that differ from input parameters – an algorithm that relies on character count, another one that counts syllables to measure word length, one with polysyllabic words count and an algorithm that uses a special list of familiar words to find the percentage of complex words in the text. The comparison of variables used in each of the previously mentioned equations (except for Spache, Fry Graph and Powers-Sumner-Kearl) is shown in Table 1,

Table 1. Comparison of readability algorithms.

Formula	***c***	***sy***	***psy***	***w***	***s***	***list***	***p***
BRI	+	-	-	+	+	+	4
FORCAST	-	+	-	-	-	-	1
FK	-	+	-	+	+	-	3
FRE	-	+	-	+	+	-	3
GF	-	-	+	+	+	-	3
CLI	+	-	-	+	+	-	3
DC	-	-	-	+	+	+	3
ARI	+	-	-	+	+	-	3
SMOG	-	-	+	-	+	-	2
LW	-	-	+	+	+	-	3

where *c* stands for characters, *sy* for syllables, *psy* for polysyllables, *w* for words, *s* for sentences, *list* for a special list of words and *p* shows the total number of input parameters per formula. The usage of the aforementioned variables by the readability methods in Table 1 is marked as either "+" (is used) or "-" (not used). The most frequently used

variable in the formulas is the count of sentences, following with the count of words. Only Forcast formula is not using the number of sentences in the equation and relies merely on the number of syllables. In the formulas (except Forcast), the sentence length measure is calculated from words per sentence. Word length is computed either from syllables per word or characters per word. As seen from Table 1, CLI, BRI and ARI algorithms are using character count while the algorithms FK, FRE and Forcast are using syllables per word. Dale-Chall and Bormuth algorithms are unique in the formula list for using a special list of predefined familiar words. While Bormuth and Dale-Chall are both based on the same Dale-Chall list of commonly used words, Bormuth's approach is to count the number of familiar words and Dale-Chall uses the percentage of difficult words in text to calculate readability. SMOG, Gunning Fog and Linsear Write formulas compute readability grade level by counting polysyllabic, i.e. difficult words. A word is considered polysyllabic, if it contains three or more syllables. While the majority of readability formulas were initially developed for the use of manual calculations, there is one method included in the comparison of readability algorithms, which was created to enable automated computing of readability – the Automated Readability Index (ARI) [20].

In addition to the several input variables shown in Table 1, the readability algorithms use constant values in the equations. These constants have been derived during the initial development and testing of the formulas by their authors and the main objective of these values is to adjust the output of the calculation to appropriate grade level [2].

The purpose of the comparison of input parameters (Table 1) was to determine a list of various readability algorithms that rely on different variables – character, syllable and polysyllable count and a specific list of familiar words. Also considering other criteria like the total number of parameters, widespread usage and ease of usability, the selection of readability methods based on similar input parameters is explained as follows:

- character count – ARI, CLI and BRI. Bormuth Readability Index is using four input parameters to compute the level of readability. ARI and Coleman-Liau Index both use characters per word and words per sentence to assess the readability grade level of the given piece of text. The difference of ARI, BRI and CLI lies in the constants and the initial purpose of the algorithm. As ARI is the only method in the list that was initially created for automated readability

calculations and Bormuth is in addition using a list of familiar words, the methods included in the readability analysis based on character count are ARI and CLI.

- syllable count – FK, FRE, Forcast. Forcast formula is using only the number of syllables to determine the readability of a text. Flesch-Kincaid and Flesch Reading Ease rely exactly on the same input parameters. As Flesch-Kincaid is an improved version of Flesch-Reading Ease, the method included in the analysis based on syllable count, is Flesch-Kincaid.
- polysyllable count – there are three methods that count polysyllables to determine the reading ease of a text – SMOG, Gunning Fog and Linsear Write. The selection of a method based on polysyllable count is done by taking into account the overall usage of the method. Based on several previous research and studies [8] [11] [18] conducted in the area of readability (few of them also discussed in Section 2) and information from readability calculator tools [5], the most commonly used formula in these three appears to be SMOG. Therefore, the method that relies on polysyllable count to compute readability and will be implemented in the readability software in current thesis, is SMOG.
- list of words – Bormuth and Dale-Chall methods are both using the same predefined list of familiar words, the total number of sentences and total number of words to calculate readability. In addition, Bormuth is also counting characters in a text. As there are already two methods included in the list of selected formulas that are based on character count (CLI, ARI) and also that Bormuth relies on the same list of words as Dale-Chall, the method selected to proceed with the readability analysis, is Dale-Chall.

The usage of various readability algorithms helps to determine, which of the formulas produces the best results for the correlation analysis. Also, the correlation analysis can refer, whether the formulas using syllable count and character count variables have any differences in associations to manually assigned scores. The aforementioned selection criterion resulted in 5 readability methods that are used in the readability assessment of academic writings – Flesch-Kincaid, Dale-Chall, Automated Readability Index, Coleman-Liau and SMOG.

The more detailed overview of the aforementioned readability assessment algorithms used in the thesis, is described in Section 3.2.

3.2 Overview of selected readability methods

Flesch-Kincaid Grade Level is an improved version of Flesch Reading Ease readability method created by Flesch in collaboration with Kincaid [2] [4]. It uses core measures as word length and sentence length. The formula to compute the value of readability with Flesch-Kincaid method is given by Equation 1:

$$FK = 0.39 \cdot \frac{w}{s} + 11.8 \cdot \frac{sy}{w} - 15.59 \quad (1)$$

where *w* is the number of words, *s* the number of sentences and *sy* the number of syllables in the analysed text. *FK* is the score of Flesch-Kincaid Grade Level formula, measuring the complexity of the text in grade levels. The core measures words per sentence and syllables per word cannot be converted and are not directly comparable. This score does not have an upper limit though the lowest score could be in theory -3.40. As already mentioned, the constant values in Equation 1 (and the following readability equations) have been acquired during the formation of the formula and are used to adjust the computed score to grade level. The Flesch-Kincaid method relies more on sentence length rather than word length [2] [5].

The Coleman-Liau Index is a readability test created by Meri Coleman and T. L. Liau [13]. The main difference between Coleman-Liau Index and some other formulas (e.g. Flesch formulas) lies in the measure of words. The Coleman-Liau counts characters in a word instead syllables per word. Due to this reason Coleman-Liau is said to be more accurately calculated by computer programs. The formula to calculate Coleman-Liau Index is given below [5] [21]:

$$CLI = 0.0588 \cdot \frac{c}{w} \cdot 100 - 0.296 \cdot \frac{w}{s} \cdot 100 - 15.8 \quad (2)$$

where *c* is the number of characters, *w* the number of words and *s* the number of sentences in the input text. Variable *CLI* presents the output of the formula, measuring the text complexity in grade levels.

The Automated Readability Index (ARI) is a readability test developed in 1967 to evaluate the understandability of a text [20]. Similarly, to other readability tests, it produces an approximate representation of the US grade level needed to easily understand the text. The formula to calculate ARI is given by Equation 3:

$$ARI = 4.71 \cdot \left(\frac{c}{w}\right) + 0.5 \cdot \left(\frac{w}{s}\right) - 21.43 \quad (3)$$

where *c* is the number of characters, *w* the number of words and *s* the number of sentences in the analysed text. *ARI* is a variable to present the complexity of the text in grade levels. Like Coleman-Liau formula, the ARI also relies on word difficulty (characters per word) and sentence difficulty (words per sentence), instead the usual syllables per word readability method. Characters in a word are faster to calculate, as the number of characters is more accurately counted by computer programs, rather than syllables. ARI is the only method amongst the ones used in this thesis, that was developed and adjusted to be calculated by computer programs. More precisely, it was designed for real-time monitoring of readability on electric typewriters [5] [20].

SMOG is an acronym for Simple Measure of Gobbledygook and was developed in 1969 by G. Harry McLaughlin [23]. It was meant to substitute for the Gunning Fog Index, due to its better accuracy and being more easily calculated. The SMOG Index, also known as SMOG Grade, estimates the years of education needed to understand fully the text assessed. SMOG Index can be calculated as described below [5]:

$$SMOG = 1.0430 \cdot \sqrt{\left(psy \cdot \left(\frac{30}{s}\right)\right)} + 3.1291 \quad (4)$$

where *psy* is the number of polysyllabic words (words with three or more syllables) and *s* is the number of sentences in the text. Variable *SMOG* is the output of Equation 4 that indicates the difficulty of the text in grade levels. The SMOG Index formula should be applied to texts with at least 30 sentences, as the formula was normed on 30-sentence samples. Otherwise the results are statistically invalid [23].

Dale-Chall Score was created by Edgar Dale and Jeanne S. Chall in 1948 [2]. The formula outputs a raw score which is then mapped to final score to get the grade-equivalent level. The Dale-Chall Formula uses a count of "hard" words. Hard words are words that do not appear on a specific list of words designed by E. Dale and J. Chall. The original list consisted of 763 common words familiar to most 4th-grade students. The list was renewed and expanded to common 3000 words in 1995 [25]. Due to the use of the common words list, the Dale-Chall formula is considered to be more accurate. Dale-Chall formula uses the percentage of difficult words and average sentence length in words for calculating the raw score of readability. Formula is given by Equation 5:

$$DC = 0.1579 \cdot \left(\frac{dcw}{w} \cdot 100\right) + 0.0496 \cdot \left(\frac{w}{s}\right) \quad (5)$$

where *dcw* is the number of difficult words (words, that do not occur in the Dale-Chall list of commonly used words), *w* the total number of words and *s* the number of sentences in the input text. Variable *DC* is presenting the raw score of Dale-Chall readability calculation. If the percentage of difficult words is more than 5%, then 3.6365 should be added to the raw score, to correct the result at the higher grades [2]. Otherwise the raw score is equal to the adjusted score [5].

All of the readability calculation algorithms in Table 1, except Dale-Chall, present the result as grade levels. As DuBay [2] stated in his study about principles of readability, the number of years of obtained education does not necessarily correlate to one's reading level. A person who seeks for special domains of knowledge might develop higher reading skills in that specific domain, compared to their general reading level. Also, university students and graduates may prefer more difficult readings in their own speciality and materials that are appealing to them, though their reading level for general classroom material might be lower [2]. Therefore, readability grade level of a text can only refer the number of years of education required to easily understand a text.

To obtain the computed grade level for each of the formulas, the raw score of the result is rounded to the closest integer - for example, if the output of a formula is 7.4, then the text is considered to be suitable for 7th grade student (or a reader with 7 years of education). A raw score of 7.6 indicates that the text should be intelligible for an 8th grade student. The higher the grade level score, the less intelligible the text is for the reader. However, the Dale-Call formula computes the raw score of readability which is not tied to a grade scale. To be able to compare the results of the Dale-Call formula readability evaluation this score needs to be converted to an equivalent grade level. For this, Dale and Chall included a chart to convert the raw score to appropriate grade level [2]. The conversion of raw scores is presented below in Table 2.

Grade levels above 12 indicate that the text under the assessment is suitable for readers with college education and/or higher professional qualification [23]. As the input data collected for the readability analysis consist of academic writings that have been written by university students and researches, the expected grade level for academic texts should be over 12 - that is, they are intelligible for readers with higher education.

Table 2. Dale-Chall raw score conversion [2].

Dale-Chall score	**Grade level**
<= 4.9	<= 4th
5.0-5.9	5th-6th
6.0-6.9	7th-8th
7.0-7.9	9th-10th
8.0-8.9	11th-12th
9.0-9.9	13th-15th/college
10 and above	16th/college graduate

4 Methodology to conduct the analysis

In order to carry out the analysis and determine the relations between the scores assigned manually by human and scores obtained as a result of readability tests, a bulk of academic writings needs to be evaluated and assigned a readability score. The computing of readability grade level for each of the academic writing is done automatically by using the readability application developed for the purpose of the current thesis. The readability scores are then analysed and compared with academic scores assigned by the reviewers of the academic writings.

4.1 Data for the analysis

The input data for the readability analysis is gathered from 87 academic papers written in English (non-native speakers) by computer science students and researchers. The writings are divided into 2 groups – 33 essays and 54 articles. Essays have been written by first year master students, are a little shorter than articles and consist of 21 to 96 sentences and contain 6 581 words on average. Scientific articles, which length is 11 pages and 20 063 words on average, are written by researchers in computer science field. The scientific articles have been graded through reviewing process by 2 to 4 reviewers in the score range of 1 to 5, where 5 is the highest value. For academic essays, the highest value of the final grade is 10. All of the writings are assessed individually by each of the reviewer and finally given an average score computed from individual scores. The more detailed overview of source data by input variables is described in Table 3. It contains the total number of academic writings (total), in addition the average (avg), minimum (min) and maximum (max) values of all core parameters necessary for computing the readability grade levels. The descriptive data in Table 3 is partitioned by the length of academic writing and is presented separately for essays and articles. The parameter *c* in Table 3 stands for characters, *sy* for syllables, *w* for words and *s* for sentences. Words that consist of three or more syllables are denoted as *dw* (difficult words), *wps* stands for words per sentence, *cpw* for characters per word and *spw* for syllables per word variable. This means, *wps* is measuring sentence length in words, *cpw* and *spw* word length in characters or syllables respectively.

Table 3. Overview of source data variables.

Parameter	**Essay**			**Article**		
	min	**max**	**avg**	**min**	**max**	**avg**
c	541	1 942	1 293	1 096	6 197	3 850
sy	164	574	348	380	1 966	1 189
w	2 837	10 453	6 581	6 044	33 682	20 062
s	21	96	55	40	326	179
dw	960	3 529	2 209	2 038	11 603	6 907
dcw	117	445	243	238	1 578	845
wps	13	41	25	16	37	22
cpw	4.6	5.5	5	4.7	5.9	5
spw	14.8	28.4	19	14.7	21.3	17
total	33			54		

Though the difference in length of the writings was known beforehand and the variation of variables was expected, the source data appears to be the most similar in *cpw*, *wps* and *spw*. It could imply that the reading difficulty of the academic writings might not vary significantly for methods that measure word and sentence length as main input parameters – these are Flesch-Kincaid, Coleman-Liau and ARI.

The collection of academic writings is processed one by one by the readability application (discussed in Section 4.2) and assigned 5 readability grade level scores according to the measurement methods. To get the most accurate readability scores as possible, the needless data in terms of readability calculations has been removed from the writings. In particular, figures and tables would increase the number of words and characters in the text, but not affecting substantially the number of sentences. As all of the 5 readability formulas use total number of sentences as an input parameter, the invalid ratios would express in incorrect results — the data in tables and figures typically does not contain punctuation marks and that would keep the number of sentences unchanged, while adding characters, syllables and words to the writing. Also, the reference list of the used literature has been removed from the academic writings. The information in the reference list holds no value in scope of readability of the original work and could cause errors when calculating for instance Dale-Chall index. Dale-Chall uses the list of well-known words

in English, while some of the references contain non-English literature. That would cause the increase of Dale-Chall score and affect readability analysis' accuracy.

After all the academic writings have been processed, the readability scores are rounded up to 2 decimal places, in order to keep the accuracy for the correlation calculations. In case of Dale-Chall formula, if the grade level is for example between 11-12, it is presented as 11.5. After the preparatory work has been done, the manually assigned grades and their readability scores can be analysed. The simplified diagram of process flow of the conducted document readability analysis is described in Figure 1.

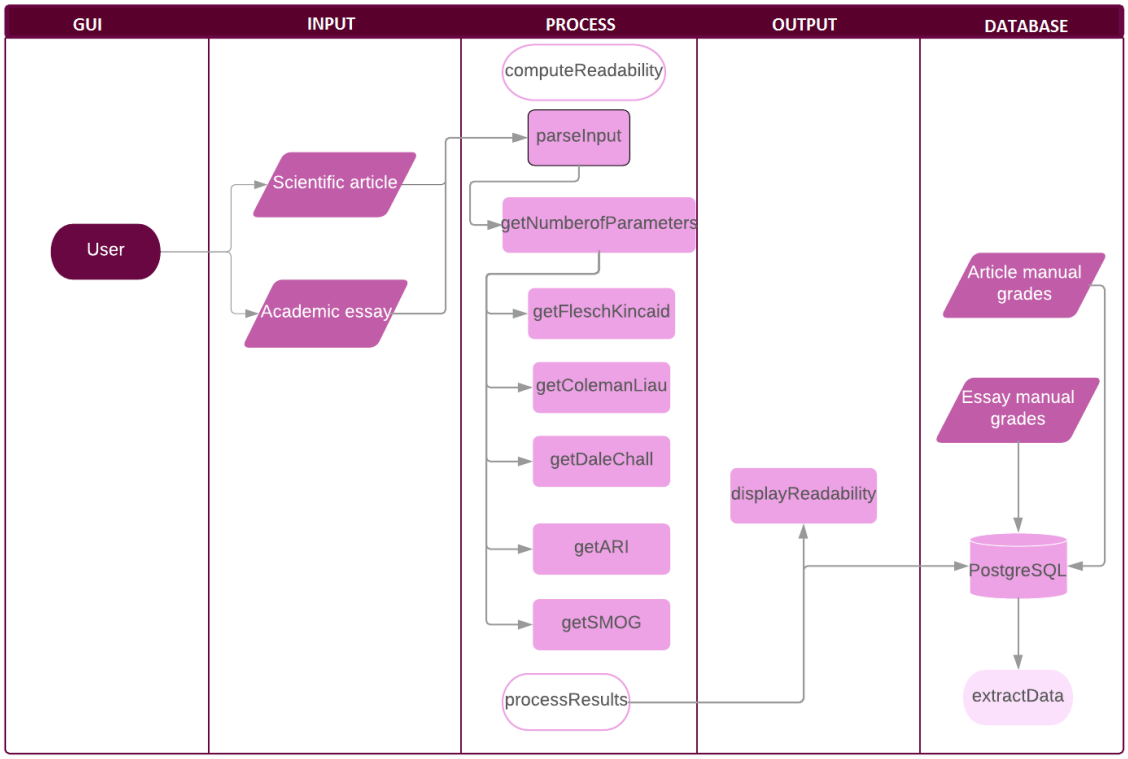

Figure 1. Process flow of readability analysis.

After the academic essay or scientific article has been uploaded into the experiment environment by a user, the input file is parsed and obtains a total number of input parameters needed to calculate a readability score. Readability score is calculated for all of the five formulas and after getting the results, the statistical data and readability grades are outputted in the GUI (Graphical User Interface). Files for the analysis need to be uploaded one by one. To provide data for further and more thorough analysis, the readability calculation results and parameter statistics are stored into PostgreSQL database. The reviewers' evaluations of the essays and articles are inserted separately into the database. The obtained data of readability grades and manual scores of the academic

writings are then joined and extracted from the database, to further carry out the analysis. The readability analysis includes comparing the manual evaluations with readability grade levels and finding the correlation coefficient between the different type of grades of academic writings.

4.2 Readability application

The readability calculation tool (also named as Experiment environment in the thesis) is a Java desktop application that computes the readability scores of given input text. The readability application has been developed by the Author and is meant to support the readability analysis by implementing the formulas of the five readability algorithms (Flesch-Kincaid, ARI, SMOG, Dale-Chall, Coleman-Liau). To compute the Dale-Chall raw score, the Dale-Chall list of 3000 familiar words was implemented [25]. The graphical user interface of the experiment environment (Figure 2) is divided into two main sections – first section displays the grade levels of the readability assessment of the input document. Dale-Chall raw score conversion to the corresponding grade level is done during the calculation. The second section shows the overall statistics of the input parameters used in the calculations of readability – total number of words, sentences, syllables, characters, average number of words per sentence. As additional information, it also provides the average readability grade level, number of complex words (words with three or more syllables) and average number of characters per word of the text under the assessment. As shown in Figure 2, the GUI of the readability tool is simple and does not need detailed explanation. More information about the usage of Experiment environment can be found in Appendix 1.

To store the experiment results and carry out further analysis to answer the set research questions, PostgreSQL database with a schema consisting of four data tables was established. The tables of the schema hold the results of computed readability scores and manually assigned scores for each of the academic writing. The description of the tables is explained below in more details. The database diagram of relevant tables is shown in Figure 3.

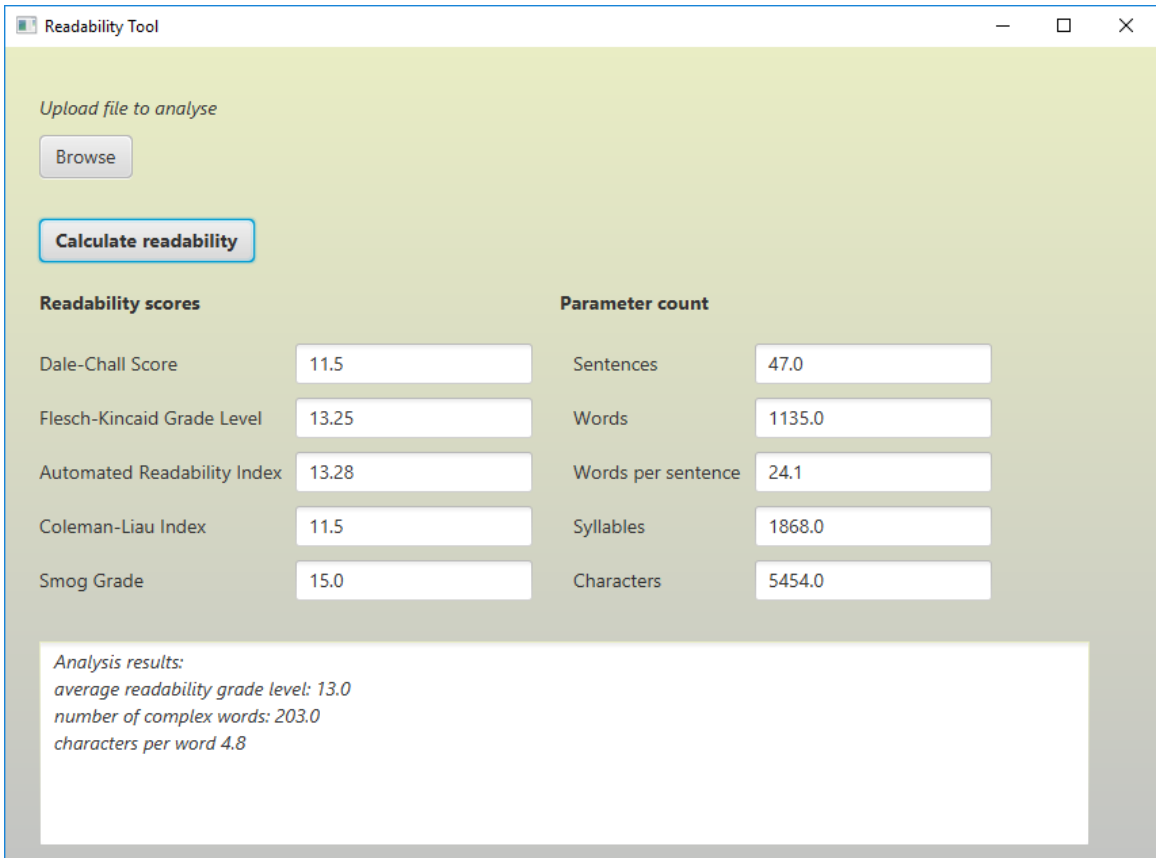

Figure 2. GUI of readability tool.

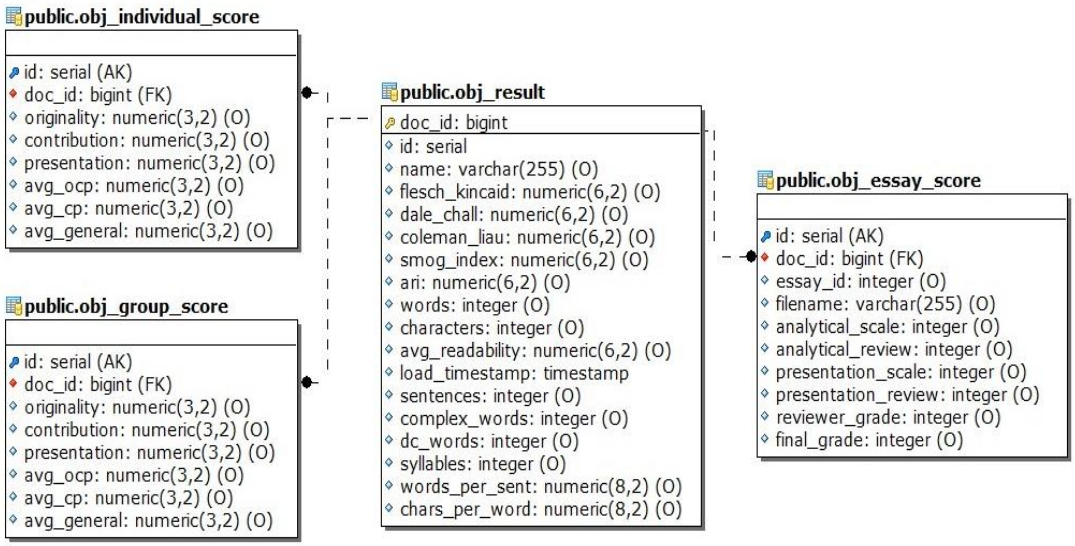

Figure 3. PostgresSQL database diagram.

The schema of established PostgreSQL database has the following tables:

- obj_result is the main table that holds the records of the automatically computed readability results and the overall parameters of the input text obtained in the experiment environment,
- obj_individual_score – reviewers' individually assigned scores of the scientific articles,
- obj_essay_score – individually assigned scores by the reviewers for the essays,
- obj_group_score – stores the manually assigned group grades of the articles.

To be certain that the results of the experiment environment are valid and can be used in the analysis, the grade level results were validated against other readability analyser applications available on the Web. The validation is done by using two computerised readability calculation tools – Readability Calculator[1] and Readability Analyser[2]. Readability Calculator offers the computing of readability for each of the five methods used in the current study, while Readability Analyser computes the results for Flesch-Kincaid, Dale-Chall and SMOG Grade. To provide more reliable results in the validation of the experiment environment, both of the readability applications are used. To confirm the results from the experiment environment, the grade levels are compared to the results from the aforementioned two readability tools for the same input text and formulas. It is important to note, that the computerised scoring may vary amongst different readability applications for the same formula and same textual content. Computerised readability calculator's validity depends on the implementation of the formulas and on the input text – for instance, some tools may require preparing the text before the calculation. Before uploading a text to a certain readability calculator, removing unnecessary elements in terms of readability (graphs, tables, figures) and correcting improper punctuation may improve the accuracy of the calculation results. As there is not any standard documentation to be found on the implementation of the readability algorithms for the automated calculations, the verification of the scores from the experiment environment is not entirely reliable and can only indicate, whether the results are credible.

To thoroughly validate the results, 5 academic essays and 5 articles were randomly selected and analysed with Readability Analyser and the Readability Calculator tools.

[1] http://www.readabilityformulas.com/free-readability-calculators.php

[2] https://datayze.com/readability-analyzer.php

The results were compared to the results obtained from the Experiment environment established for the thesis, to determine any wider deviations between grade levels. The average (AVG) and standard deviation (STD) values of the results across the readability tools are presented in Table 4.

Table 4. Readability scores comparison for the validation of the Experiment environment.

Formula	**Experiment environment**		**Tool: Readability Analyser**		**Tool: Readability Calculator**	
	AVG	**STD**	**AVG**	**STD**	**AVG**	**STD**
Flesch-Kincaid	12.72	1.59	10.93	1.68	11.04	1.41
Coleman-Liau	12.76	1.99	n/a	n/a	11.30	2.05
Dale-Chall	12.35	1.78	13.65	2.48	12.15	2.55
ARI	13.50	2.15	n/a	n/a	10.55	1.95
SMOG	13.43	1.31	13.04	1.34	10.97	1.42

The readability verification test reveals that the average grade levels across the readability calculators vary from 11 to 14. Grade levels for ARI, which counts characters per word, seem to be the most inconsistent – the average readability level is in the range of 11 to 14. For Flesch-Kincaid, Coleman-Liau and SMOG, the readability grades are in the scale of 11 to 13. Dale-Chall scores are between 12-14. Standard deviation for SMOG grade, which counts polysyllables in the text, is the smallest for results from each of the tools and indicates, that SMOG is the most consistent in assigning grades in each of the tools. Flesch-Kincaid is computed from the ratios syllables per word and words per sentence and has standard deviation values from 1.41 to 1.68 across the tools, referring that Flesch-Kincaid method is also quite consistent within a tool. Dale-Chall grades are the most dispersed in Readability Analyser and Readability Calculator, while for Experiment environment the standard deviation of Dale-Chall does not show significant dispersion. Furthermore, what stands out in the comparison of average values, is that the grade levels somewhat differ amongst the validation tools – for instance, SMOG has an average value of 10.97 in Readability Calculator and 13.04 in Readability Analyser. Similarly, for Dale-Chall there is a slight discrepancy between grade levels from validation tools.

To further validate the Experiment environment, the comparison of the minimum and maximum values of the results for the same 5 essays and 5 articles, are introduced in Table 5.

Table 5. Readability min and max scores comparison for the validation of the Experiment environment.

Formula	**Experiment environment**		**Tool: Readability Analyser**		**Tool: Readability Calculator**	
	MIN	**MAX**	**MIN**	**MAX**	**MIN**	**MAX**
Flesch-Kincaid	8.88	14.24	6.39	12.64	8.40	12.80
Coleman-Liau	9.06	15.51	n/a	n/a	8.00	14.00
Dale-Chall	9.50	14.00	9.50	16.00	7.50	16.00
ARI	9.29	15.18	n/a	n/a	8.40	13.20
SMOG	8.25	15.42	9.76	14.28	7.70	12.80

The minimum values for Flesch-Kincaid vary from 6.39 to 8.88, making it the most dispersed across the three readability tools. The maximum value varies the most for SMOG formula – from 12.80 to 15.42. The most equal grade levels across the tools were obtained with Dale-Chall method – minimum values vary from 7.5 to 9.5, while maximum values are in the range of 14 to 16.

The comparison of the minimum, maximum, average and standard deviation values gained from the three automated readability calculation tools revealed, that the average grade levels vary the most for ARI algorithm, minimum and maximum scores vary the most for Flesch-Kincaid. The average values from Table 4 showed, that the most stable in evaluating readability level, is SMOG formula, having relatively modest standard deviation – varying from 1.31 to 1.42. The lowest standard deviations values are amongst the scores obtained from the Experiment environment, making it the least dispersed across the tools.

As already mentioned, the specification for automated readability computing is uncertain and the differences in readability scores computed by different readability calculation tools is affected from the way a certain tool is counting the linguistic elements of the text. Similar conclusion was stated by a study which investigated the consistency of well-known readability formulas by estimating the readability level of design standards [26]. The way the text elements – words, sentences, syllables, numbers and abbreviations are counted and how punctuation, tables and figures are treated, is vague and varies between equations and tools, having an effect on the final calculated score.

The Experiment environment validation analysis confirms that the results obtained from the readability application used in the present thesis, are not significantly different from other readability analysers and therefore can be used to execute readability experiments and proceed with analysis.

5 Readability analysis of academic writings

To carry out the readability analysis of academic texts and investigate its correlations to manually assigned scores, the input data is divided into two sets that are being evaluated separately – academic essays and scientific articles. The purpose of partitioning the academic writings is to get more detailed information about possible correlations between the success of the writing and its obtained scores with manual and computerised methods. The manually assigned scores for both of the data sets include the individual grades assigned by each of the reviewer and the group or final grade, which is an overall score computed from several individual score components.

Each of the manual score component's value assigned by reviewers is compared against the automatically obtained readability scores. In case of scientific articles, the comparison is conducted for the individually assigned scores and for the group scores separately. In addition, the components of manual evaluation are compared against the average value over the computerised readability assessment results (AVG_R). To determine the correlation between human and computerised scores, the correlation coefficient is calculated with Spearman's Rank-Order Correlation formula [27]. Spearman's correlation coefficient measures the monotonic association between two ranked variables. The Spearman's coefficient is calculated for the individual scores, group scores and automatically computed readability grades of the academic writings using the formula given by Equation 6 [27]:

$$\rho = \frac{cov(rg_x - rg_y)}{\sigma_{rg_x}\sigma_{rg_y}} \tag{6}$$

where ρ the denotes the coefficient, $cov(rg_x - rg_y)$ the covariance and σ_{rg_x}, σ_{rg_y} the standard deviations of the ranked variables. The coefficient can have values from -1 to +1, where +1 indicates perfect positive association and -1 a perfect negative association of ranks. The closer the coefficient is to 0, the weaker is the correlation between the variables. The statistical significance of the correlation depends on probability value (p-value). If probability value is less than 0.05, the correlation is statistically significant [28]. In the following analysis, the significance of the correlation is merely marked as *s*

(statistically significant) or *ns* (statistically not significant). The strength of the correlation can be described using the following interpretation of Spearman's coefficient absolute value [29]:

- 0.00– 0.19 – very weak
- 0.20– 0.39 – weak
- 0.40– 0.59 – moderate
- 0.60 – 0.79 – strong
- 0.80 – 1.00 – very strong

After finding the correlation coefficient for all of the relevant components, it is possible to determine, what is the nature of correlation between the manual evaluation of academic writing as a whole and computerised readability scores.

5.1 Readability analysis of scientific articles

To analyse the readability of scientific articles, 54 academic articles were assigned grade levels during the execution of readability experiments. The writings have been assigned a manual grade in scale of 1 to 5, where 5 is the highest possible score. Group scores for each of the scientific article consists of the average results of all of the reviewers' evaluations for each of the manual assessment component. The overall grade assigned by the reviewers consists of the following elements:

1. theoretical contribution (TC) - expresses the content and analytical nature of the text
2. presentation and readability (PR) – human evaluation of the readability of the work
3. originality – level of innovation
4. AVG_OCP - average score over originality, contribution and presentation
5. AVG_CP - average score over contribution and presentation

The assessment results by components that are compared to automatically assigned readability, are TC, PR, AVG_OCP and AVG_CP. The manually assigned scores for each of the academic writing are linked to the readability results - Flesch-Kincaid (FK), Coleman-Liau (CL), Dale-Chall (DC), ARI, SMOG - that are obtained for the same input text. The sample extract of result data over all of the score types (manual and computerised) used in the analysis of scientific articles, is shown in Figure 4.

id	tc	pr	avg_ocp	avg_cp	fk	cl	dc	ari	smog	avg_r
4	2	1.7	1.9	1.8	14.2	13.81	14	15.34	14.33	14.34
6	3	4.3	3.7	3.7	11.63	12.63	11.5	13.8	11.29	12.17
7	3.3	5	4.1	4.2	14.24	15.51	14	15.42	14.58	14.75
8	3.7	4.3	3.8	4	13.2	13.08	14	15.11	12.34	13.55
9	3.2	3.8	3.5	3.5	13.69	12.63	14	15.13	13.48	13.79
10	2.5	2	2.3	2.3	13.47	13.75	14	15.12	13.07	13.88
11	1.7	3.7	2.7	2.7	13.46	14.02	16	14.77	13.18	14.29
12	2.3	2.3	2.2	2.3	14.82	12.63	14	15.34	15.29	14.42
14	3	3.3	3.4	3.2	10.95	12.15	11.5	13.41	10.06	11.61
15	2	2	2	2	13.2	14.58	14	14.82	12.64	13.85
16	3	2.3	2.7	2.7	13.42	12.32	14	15.16	12.93	13.57
17	2	2	2.2	2	12.87	16.45	14	14.63	13.09	14.21
18	4	4.5	4.2	4.3	14.72	13.99	14	16.12	14.74	14.71

Figure 4. Extract of scientific articles' results from SQL Manager.

The basic features of all of the received manual score components are shown in Table 6. These include the minimum and maximum values, standard deviation (STD), the average of individual scores and readability formulas (AVG) and the value that occurs most often (MODE).

Table 6. Statistical overview of manual grades of scientific articles.

	AVG	**STD**	**MODE**	**MIN**	**MAX**
TC	2.61	0.95	3.00	1.00	5.00
PR	3.15	1.10	4.00	1.00	5.00
AVG_OCP	3.08	0.76	3.50	1.25	4.75
AVG_CP	2.88	0.87	2.00	1.00	4.50

The descriptive statistics shows, that the average manually assigned scores AVG_OCP = 3.08 and AVG_CP = 2.88 on the scale of 1-5. Amongst the human assigned scores, the presentation and readability obtained the largest standard deviation value – 1.10, while the most consistent seems to be AVG_OCP. The minimum and maximum values for human assigned scores are in the range of 1 to 5. The most frequently manually assigned score for AVG_OCP = 3.50 and for AVG_CP = 2. The articles obtained the highest average grades for presentation and readability, while the grades for theoretical contribution appear to be the lowest – 2.61 on average.

The statistical features of the obtained computed readability scores for scientific articles are presented in Table 7.

Table 7. Statistical overview of readability scores of scientific articles.

	AVG	STD	MODE	MIN	MAX
FK	14.25	2.06	13.20	10.95	20.29
CLI	14.01	1.51	12.63	10.99	17.78
DC	14.03	1.40	14.00	11.50	16.00
ARI	14.24	2.49	12.34	10.06	21.68
SMOG	15.66	1.53	15.34	13.16	19.78

Amongst the computed readability scores, the average grades are between 14.01 and 15.66. Therefore, the average readability of scientific articles meets the expectations and is over 12, with reading difficulty suitable for readers with higher education. The minimum and maximum values for computed readability are in the range of 10.06 to 21.68, where the highest and the lowest score is obtained with ARI algorithm. In grade levels, the text with readability score over 14, is considered as "extremely difficult to read". Though the standard deviation is the most dispersed for ARI scores, the relatively low average and mode values indicate that there are not many writings with ARI grade level over 20. The higher ARI grade could be inherent to writings that consist of many long words. The most consistent in assessing readability in the automated readability results occurs to be the Dale-Chall method – the average grade level is 14.03 and the value that occurs most often, is also 14.

The following Table 8 contains data of Spearman's correlation coefficient calculations between the human assigned scores and computed readability. The correlation coefficients are calculated using Equation 6 and the obtained coefficients help to determine the strength of the association between the ranked manually assigned and computed readability score pairs. Table 8 holds data for individual scores comparison and group scores comparison. In addition to the correlation coefficient, statistical significance of each of the result is labelled with *(s)* or *(ns)*. Since the probability value (p-value) is impacted by the number of input cases, the equal value for individual and group scores' coefficient does not imply, that the statistical significance is also the same.

Table 8. Article correlation coefficient of individual and group scores.

	FK	**CLI**	**DC**	**ARI**	**SMOG**	**AVG_R**
Individual scores						
TC	**0.21 *(s)***	-0.02 *(ns)*	0.067 *(ns)*	0.15 *(ns)*	**0.24 (s)**	0.15 *(ns)*
PR	0.089 *(ns)*	0.015 *(ns)*	0.067 *(ns)*	0.037 *(ns)*	0.10 *(ns)*	0.072 *(ns)*
AVG_OCP	0.18 *(s)*	-0.003 *(ns)*	0.053 *(ns)*	0.13 *(ns)*	**0.21 *(s)***	0.12 *(ns)*
AVG_CP	0.16 *(s)*	0.004 *(ns)*	0.073 *(ns)*	0.10 *(ns)*	0.19 *(s)*	0.12 *(ns)*
Group scores						
TC	**0.33 *(s)***	-0.027 *(ns)*	0.061 *(ns)*	0.26 *(ns)*	**0.36 *(s)***	0.23 *(ns)*
PR	0.036 *(ns)*	-0.028 *(ns)*	0.06 *(ns)*	-0.025 *(ns)*	0.056 *(ns)*	0.028 *(ns)*
AVG_OCP	0.22 *(ns)*	-0.021 *(ns)*	0.11 *(ns)*	0.15 *(ns)*	0.24 *(ns)*	0.16 *(ns)*
AVG_CP	0.18 *(ns)*	-0.029 *(ns)*	0.10 *(ns)*	0.11 *(ns)*	0.21 *(ns)*	0.13 *(ns)*

According to the data in Table 8 and using the previously introduced scale [29] to describe the strength of Spearman's correlation coefficient absolute values, there occurs to be a weak positive correlation between individual theoretical contribution and Flesch-Kincaid grade level and SMOG. The same tendency stands out in group scores comparison, where the correlation between TC, FK and SMOG is somewhat stronger. For all the remaining relations, the correlation is either statistically insignificant or close to zero. This means that there appears to be almost no association between the obtained manual grades and most of the computed readability grades of the scientific articles. It also indicates that three of the selected readability algorithms (SMOG and FK not included) with different input parameters behave in a similar way when evaluating the reading difficulty of the scientific article. What also stands out in Table 8 is that none of the manually evaluated grade components have a significant association with average readability (AVG_R). It could refer that the computed readability level should not be treated as an average value over several methods, and the scores of readability algorithms give stronger correlations when handled separately. This kind of matter could be caused by the computed average readability value being less accurate compared to the accuracy of each of the computed readability method separately and that could affect the correlation to manual grades.

The weakest associations are between manually assigned scores and CLI. Like ARI, the Coleman-Liau index relies on characters per word and words per sentence parameters, though having distinct constant values. ARI and Coleman-Liau both demonstrate no significant correlation between any of the manual score elements. Flesch-Kincaid and SMOG are based on syllable (or polysyllable) count and both of them obtained a weak positive correlation between theoretical contribution of the scientific article. As there is no relevant difference in associations with the assessment components in individual and group scores coefficients, it seems that evaluating the correlation coefficient separately adds no significant value.

To further explore the associations between the success of the academic work and readability, the average scores assigned by reviewers and average readability results by grade groups is presented in Table 9. Analysing the average results by grade groups helps to determine, whether the academic works with higher scores have a better readability.

Table 9. Article average scores by grade groups.

Grade	**AVG_OCP**	**AVG_CP**	**AVG_R**	**FK**	**CLI**	**DC**	**ARI**	**SMOG**
<=2	1.86	1.85	13.58	13.06	13.98	13.17	13.00	14.72
>2 and <=3	2.53	2.55	14.48	14.48	13.99	14.15	14.43	15.65
>3 and <=4	3.41	3.42	14.50	14.42	13.83	14.14	14.31	15.80
>4 and <=5	4.10	4.23	15.38	15.32	15.36	14.00	15.52	16.69

The average readability level increases noticeably along with higher academic score. The difference between the readability level for the lowest and highest manually assigned grade group is 2.52, obtained with ARI algorithm. ARI method also stands out with having the lowest average readability grade level - 13. The highest readability level is obtained with SMOG formula – 16.69. Therefore, it is fair to say that though the average reading difficulty of the academic paper increases considerably, the complexity of the writings in all grade groups is already significantly high and is suitable to readers with higher education (readability grade >= 13). Comparing the results of readability formulas to the group with lowest grade (<=2) and with the highest grade (>4 and <=5) indicates that the lowest grades also have lower readability grade level. Though the same

association applies for all of the readability methods, the increase in readability scores over the grade groups is the most noticeable for ARI and Flesch-Kincaid methods, 2.52 and 2.26 accordingly.

In the readability analysis of scientific articles, 54 writings were processed by the readability application tool and assigned a computed readability score. The computed readability grades were compared against the scores assigned manually by the reviewers of the academic writings to determine the nature of associations. The average value of obtained readability scores varied from 14-16 in grade levels and are therefore intelligible for readers with higher education. The Spearman's correlation coefficient calculation results showed that there is only a weak correlation between some of the readability formulas and manual score components. More precisely, a weak positive correlation exists between theoretical contribution and two of the readability algorithms - Flesch-Kincaid and SMOG. The correlation analysis revealed that the readability method developed initially for the use of computerised readability calculations (ARI), does not perform better compared to readability methods developed for manual calculations in terms of prognosing the success of the scientific article. The part of the analysis, where the manual scores by grade groups were compared against readability scores, showed that the writings with the lowest manually assigned score, also tend to have a lower readability grade level.

5.2 Readability analysis of academic essays

The readability analysis of academic essays is carried out similarly to scientific articles – 33 academic essays were evaluated in the experiment environment and assigned 5 automatically computed readability scores. The readability scores of the input text were linked with human assigned evaluations for the same text. The human evaluation is a score assigned by the reviewers (teaching assistants and students) and consists of the following grade elements:

1. analytical effort score (AS) – a mathematical grade, based on scores in scale of 0 to 8
2. presentation score (PS) – a mathematical grade, that represents the style, clearness and readability of the essay, based on scores in scale of 0 to 2
3. REV_G – reviewer's (subjective) evaluation of the essay overall, highest possible value is 10

4. FINAL_G – final grade of the essay, combined from all of the individual scoring elements assigned by the reviewers, maximum value of the final score is 10.

Each of the essay scoring component is compared against the obtained computerised readability grades. The readability analysis includes finding the correlations between all of the individual manually assigned and readability grades with Spearman's correlation coefficient, computed with Equation 6. Then, the average values of each of the academic essay score components is compared in grade groups, to determine possible associations between received manual grade and reading difficulty of the text.

Considering the results of scientific articles' readability analysis, the expected outcome of the readability analysis of academic essays is that readability methods based on same or similar input parameters presumably obtain similar correlation results. In particular, Flesch-Kincaid and SMOG are expected to have a weak correlation with few of the manual score elements. In addition, the presumable computed readability level of academic essays should be lower in lowest human assigned grade group and highest in the highest manual grade group. The average computed readability grade level of academic essays is expected to be over 12.

The descriptive statistics of each of the obtained manual score elements of the academic essays is presented in Table 10.

Table 10. Statistical overview of manual grades of academic essays.

	AVG	**STD**	**MODE**	**MIN**	**MAX**
AS	6.20	1.75	8.00	2.00	8.00
PS	1.39	0.69	2.00	0	2.00
REV_G	7.70	2.07	9.00	2.00	10.00
FINAL_G	8.18	1.59	8.00	4.00	10.00

In manual scores, the most deviated is the reviewer's score – 2.07, while the lowest standard deviation value was obtained by the score for presentation. The average final grade = 8.18, while the minimum and maximum final grades are in the range of 4 to 10. Reviewers assigned most often a grade "9" for the essays, though the average of the reviewer's grade is a bit lower – 7.70. The minimum and maximum values assigned by reviewers are 2 and 10 respectively.

Overview of statistical features of the gained computed readability scores is presented in Table 11. Two of the essays contained less than 30 sentences and are not included in any of the calculations that are related to SMOG formula.

Table 11. Statistical overview of readability scores of academic essays.

	AVG	**STD**	**MODE**	**MIN**	**MAX**
FK	13.98	2.50	11.56	10.10	20.20
CLI	13.16	1.55	14.80	10.02	15.90
DC	12.68	1.86	14.00	9.50	16.00
ARI	14.67	1.55	12.04	9.87	22.92
SMOG	13.70	1.75	13.80	12.5	18.82

In obtained readability scores, the highest standard deviation value belongs to Flesch-Kincaid method. The least dispersed readability scores are for methods Coleman-Liau and ARI, indicating that they are the most consistent in assigning readability grades. As Coleman-Liau and ARI use character count in equations to compute readability, it confirms the expectations that methods with similar inputs perform alike when assessing readability. The biggest gap between minimum and maximum values in readability formulas, occurs in the scores of ARI – the scores for ARI vary from 9.87 to 22.92. But as the mode for ARI is 12.04 and the average 14.67, it refers that there are not many works with readability grade over 20. The same pattern with the scores of ARI stood out in the descriptive statistics of the scientific articles, where few of the writings scored readability level over 20. The average values of computed readability scores vary from 12.68-14.67, indicating that similarly to scientific articles, the academic essays are the most suitable for readers with higher education.

The results of the calculations of Spearman's correlation coefficient between reviewers' manually assigned scores and automatically computed readability grades of the academic essays is shown in Table 12. The table holds data for the correlations between the components of manually assigned scores and each of the readability method, including the overall average computed readability.

Table 12. Essay correlation coefficient of individual scores.

	FK	**CLI**	**DC**	**ARI**	**SMOG**	**AVG_R**
Reviewers' scores						
AS	0.032 *(ns)*	0.032 *(ns)*	0.15 *(ns)*	0.020 *(ns)*	**0.30 *(s)***	**0.22 *(s)***
PS	-0.010 *(ns)*	0.082 *(ns)*	0.07 *(ns)*	-0.006 *(ns)*	0.13 *(ns)*	0.12 *(ns)*
REV_G	0.11 *(ns)*	0.12 *(ns)*	0.18 *(ns)*	0.10 *(ns)*	**0.36 *(s)***	**0.29 *(s)***
FINAL_G	0.17 *(ns)*	0.14 *(ns)*	**0.22 *(s)***	0.17 *(ns)*	**0.49 *(s)***	**0.39 *(s)***

The Spearman's correlation calculation results show, that the strength of the correlation between manual scores and computerised scores of the academic essays, can be divided into three subsets:

1. moderate correlation – here is the correlation analysis pair, that demonstrates a moderate significant correlation - SMOG and FINAL_G with correlation coefficient of 0.49. The association between SMOG formula and the final score of the essay, is the strongest amongst all of the essay correlation coefficients. SMOG's coefficient is the lowest for presentation scale score - 0.13.
2. weak correlation – pairs, that have a weak significant association with manual grade elements – DC and FINAL_G, SMOG and REV_G, AS. Also, there is weak correlation between AVG_R and AS, REV_G, FINAL_G. The association between average readability and final grade, is the strongest in this group – 0.39. In addition, again SMOG formula stands out with having a considerable correlation with reviewer's points – 0.36.
3. very weak or no correlation – correlation analysis pairs, that are close to zero or have no significant association. Here belong all the remaining manual scoring elements and their associations to readability scores that are represented in Table 12. Out of the 5 readability methods, only SMOG has a weak or moderate correlation to some of the manual score components. Flesch-Kincaid, Coleman-Liau, Dale-Chall and ARI have no significant association to any of the manual score elements.

The correlation analysis of academic essays reveals that there is no noticeable association between four of the readability methods and manual scoring. The only readability formula having a weak or moderate association with manual scoring, is SMOG. This indicates,

that counting the polysyllabic words in a text gives the best results when predicting the success of an academic essay. Moreover, the overall average of automatically computed readability demonstrates a weak correlation with final score of the essay. That is conflicting with the results obtained from the analysis of scientific articles, where the average readability had no significant association to manual score components. Since amongst the reviewers of the essays were also the students who wrote the essays and evaluated each other's writings, the manual grades for the essays might be affected more by the reading ease of the text. It means, that students evaluated the essays according to their different background knowledge and motivation and therefore the obtained final grade has stronger association with overall readability level of the essay.

To further explore the level of reading difficulty of academic essays, the average values of manual score components and each of the readability formulas, is presented in Table 13, divided by grade groups. The minimum final score of the essays was 4 and none of the essays obtained a manual score 5, therefore the data in Table 13 does not contain information about grade groups 0 to 3 and 5, the calculations start with grade group "4".

Table 13. Essay average scores by grade groups.

Grade	AS	PS	REV_G	AVG_R	FK	CLI	DC	ARI	SMOG
4	3.90	0.64	4.10	12.25	14.06	13.00	12.86	14.36	6.97
6	4.25	0.88	5.50	12.63	15.30	12.73	11.75	16.43	6.90
7	5.39	1.44	7.00	12.72	12.50	12.64	11.28	13.19	13.97
8	6.50	1.28	7.75	14.46	14.53	13.58	13.38	15.19	15.65
9	7.26	1.65	9.00	12.53	12.19	12.37	11.83	12.56	13.70
10	7.04	1.73	9.15	15.24	15.70	13.75	13.71	16.77	16.29

The average readability scores by grade groups are not increasing evenly along with the final grade of the essay. For essays with grade 4, the average readability is 12.25. For essays that obtained the highest final score (10), the average readability level is also significantly higher – 15.24. The same tendency applies to all readability formulas – having a lower readability score in grade group 4 and the highest level in grade group 10. The largest deviation of average scores, has the SMOG formula – varying from 6.90 to 16.26. The readability scores in general are behaving quite volatile along the increase of the final grade – for instance, Flesch-Kincaid has a readability score of 12.72 in grade

group 7 but obtained 14.53 in grade group 8. Again, in grade group 9, the score of Flesch-Kincaid is lower - 12.19. The inconsistency could indicate that in grade group "9" there are couple of works that obtained relatively low computed readability scores. As there were overall 33 academic essays to conduct the readability analysis, the significantly lower or higher readability grades of couple of writings in certain smaller subset (e.g. group of essays divided by manual grades) could express in noticeable change of average readability scores in the whole group of works. Though this kind of fluctuation is happening amongst all of the 5 readability formulas, the average score analysis in Table 13 confirms, that the essays with a lower manually assigned score tend to also have a lower readability grade. The average results in Table 13 align with the outcome of the readability analysis by grade groups of the scientific articles – the writings that obtained the lowest reviewers' grades tend to be easier to read than the ones with the highest grades.

5.3 Results of readability analysis

By executing the automated readability experiments and conducting the analysis between human assigned evaluation scores for the work as a whole and calculated readability scores, current thesis tried to find answers to the following research questions.

Question #1: What is the correlation between readability level and manually assigned score of academic writing? Do works with higher score have better readability?

The Spearman's correlation coefficient comparison of scientific articles in Table 8 showed, that there is no moderate or strong association between human assigned scores and automatically computed readability. On the opposite side, the correlation analysis of academic essays in Table 12 indicated that there is a moderate positive correlation between SMOG formula and the final score of the essay – meaning that when the manual score of the essay increases, so does the SMOG grade. In addition, the average value of readability grades also had a positive weak, nearly moderate (0.39) association with the final manual score of the essay. The outcome of the correlation calculation results refer that the correlation between readability level and manually assigned score of academic writing, could depend on the length of the writing. In addition, as the reviewers of scientific articles and academic essays were with somewhat different background, knowledge and motivation, the manual score of the academic essays could be more

dependent on the reading level and interest of the reviewer. Nevertheless, even the maximum obtained strength of association – moderate correlation – is not sufficient to use only automated readability scoring to evaluate academic writings. The obtained moderate correlation between manual scores and computed readability scores of academic essays refers that automated readability evaluation could be used as one of the components of manual grading as a whole.

"Better readability" can be interpreted as having a lower readability grade. Lower readability grade level means that the text contains less polysyllabic words and long sentences, less difficult and long words. Lower readability score indicates that the text is easier to read than the one with higher score. Therefore, to have a better readability, the writing should have a lower readability score. The analysis of academic writings by grade groups between average automated readability and human assigned scores revealed that as the academic score rises, so does the readability score, i.e. the difficulty of reading. The analysis for both, scientific articles and academic essays, showed that writings with lowest grade also have a low readability grade level. The increase in average readability was significant enough in both data sets – essays and articles – to indicate, that the works with higher score, are more difficult to read and they do not have better readability.

Question #2: Which of the readability assessment methods gives the best results for analysing the correlation?

Though ARI formula was initially developed for the use of automated readability calculations, the correlation analysis results proved that ARI does not perform better than the algorithms aimed for manual readability calculations in terms of indicating the potential success (higher manually assigned grade) of the writing. The readability methods that had some kind of significant association in the analysis in general, were Flesch-Kincaid, Dale-Chall and SMOG. Flesch-Kincaid had a weak correlation in the analysis of the articles, while obtained no significant coefficient in the analysis of the essays. Dale-Chall was represented with significant correlation only once, in Table 12 – analysis of the essays. SMOG method had weak positive correlation between manually assigned score (theoretical contribution and AVG_OCP) and readability level of scientific articles. In addition, SMOG had the strongest associations with analytical (AS), reviewer's (REV_G) and final score (FINAL_G) of the essays. SMOG algorithm obtained the highest correlation coefficient in the readability analysis overall – 0.49,

which indicates a moderate correlation between manual evaluations and automatically computed readability scores. Therefore, SMOG algorithm performed the best and gave valuable results in terms of correlation analysis.

Question #3: Which approach of readability evaluation performs better in terms of indicating potentially successful and high-scoring written work – the formulas based on syllable count or on character count?

As already mentioned, the readability methods that had at least weak or moderate correlation in the analysis, were Flesch-Kincaid, Dale-Chall and SMOG. Coleman-Liau and ARI, which both rely on character count in the equations, demonstrated a very weak or nonsignificant correlation. It adds credibility to the results of the readability calculations and analysis that two of the methods with similar input parameters, behave the same way when prognosing the success of the academic writing. The same applies to Flesch-Kincaid and SMOG, as both rely on syllable count as one of the parameters. SMOG, more precisely, is taking into account the total number of complex words by counting polysyllables in the text. Dale-Chall is also counting the complex words in the text, but in its own approach, by using a special list of familiar words. As Dale-Chall obtained a significant weak correlation only once, it can be said, that the formulas based on syllable count perform better in terms of indicating potentially successful and high-scoring written academic work. Specifically, SMOG method performed the best in terms of the analysis.

Question #4: Is there a significant difference in automatically computed readability values for academic writings by research scientists versus master level students?

The overall readability level of a writing is in addition to several linguistic elements also impacted by the author of the text – more precisely, the previous experience, obtained education, skills and knowledge, moreover the writing style of the author influence the readability of the writing. As the academic essays were written by first year master students and scientific articles by researchers in the field of computer science, the expected readability level for both of the writings was above 12-13 in grade levels. The statistical overview of obtained readability scores showed that the average grade level of readability methods varied from 14 to 16 in case of scientific articles and from 13 to 15 for the academic essays. Therefore, the outcome of readability experiments was as

expected. There is a slight variation in readability levels of the writings – academic essays obtained one grade lower readability level than scientific articles. The one grade gap in readability levels of the writings could be because of the various background of authors of the works – more precisely, the researchers and master students presumably have different experience and knowledge in the domain of computer science. The minor variation in readability levels of academic essays and scientific articles refer that academic essays are with slightly lower level of reading difficulty and therefore are more intelligible for the reader. The overall average reading difficulty of the academic writings as a whole still obtained relatively high maximum value – 16 in grade levels. This indicates that the academic writings are quite complex in the structure and vocabulary, and are therefore the most suitable for the audience with sufficient knowledge and/or education in the domain of computer science.

Overall, although some of the readability formulas demonstrated a weak or moderate correlation in readability analysis of the academic writings, the results of the analysis are not sufficient to firmly indicate the potential success of the writing by evaluating the readability level of the text. One of the reasons for insufficient strength of the correlation between manually and automatically assigned grades of the academic writings could be that the writings were reviewed by the targeted audience – as the average readability level of all of the writings varied between 13 and 16, the complexity of the writings was suitable for readers with higher education and therefore the human evaluation of the work was not considerably affected by the reading difficulty of the academic writing.

6 Conclusions

The purpose of this thesis was to investigate whether it is possible to prognose the success of academic writings through text readability analysis. The thesis also tried to answer the question, which of the used readability formulas performed best in evaluating the academic writings and whether the performance of readability algorithm is dependent on the input parameters. Text readability analysis included the semantic analysis of 87 academic writings. More precisely, the writings consisted of shorter academic essays and longer scientific articles, which were written by university students and researchers in computer science field, whose native language is not English. The readability analysis was conducted first by computing readability grade levels for each of the writings by using five well-known readability algorithms – Flesch-Kincaid, Coleman-Liau, Dale-Chall, ARI and SMOG. The calculations of the readability algorithms were done in a specific readability experiment environment, developed by the Author for the usage of this thesis. The obtained readability grade levels were then compared against manually assigned scores for the same writing. Comparison was done by calculating Spearman's correlation coefficient between manual grade and readability grade. The coefficient analysis helped to investigate and determine possible associations between the scores and was done separately for shorter writings (essays) and for articles, that were more capacious. In addition, the manually assigned evaluations and obtained readability grade levels were analysed in grade groups to investigate the conformity of reading difficulty with different grade groups.

The readability analysis of essays and articles proved, that there is a weak or moderate positive correlation between SMOG formula and (few elements of) manual grading. The association between manual scores and computed scores of the remaining four readability formulas were not significant or had a value of close to zero. Therefore, SMOG formula also performed the best and produced valuable results in terms of readability analysis. ARI was expected to perform somewhat better in the analysis, as it is developed for the purpose of automated readability calculations. Then again, ARI and CLI, which rely on character count in computing readability, both obtained insignificant correlation with manual scores – this confirms the presumptions that readability methods with similar

input parameters behave the same in evaluating readability level. However, as none of the readability algorithms demonstrated a strong correlation with human assigned grades, the results of the analysis do not support the viability of prognosing the academic success with readability algorithms. Moderate correlation does not give sufficient evidence to use merely the automated readability evaluation to prognose the potential success of academic writing. The overall average computed readability for the writings varied between 13 to 16, which met the expectations that the academic writings are the most intelligible for readers with higher education. There was a slight difference in the readability levels of academic essays and scientific articles – the average readability for academic essays varied between 13 and 15, while for scientific articles the average computed readability grades were in the range of 14 to 16. The small variation of average readability levels of academic writings could be due to the different authors of the essays and articles – these were students and researchers respectively.

On the opposite side, the analysis conducted in grade groups showed that there are some dependences between grade groups and readability levels. More specifically, the readability methods obtained a lower score in lowest grade group and the opposite – the academic writings with the highest manually assigned grade also had a higher readability level. There appeared to be some fluctuations in the average readability scores of academic essays – more precisely, the readability scores dropped suddenly in grade group "9". The reason behind this could be that couple of writings in manually assigned grade group "9" obtained considerably low computed readability scores and that had an effect on the average grades for the whole group. But, as the overall reading difficulty of the writings was lower in lowest grade group and higher for the highest grade, it implies that the writings with highest scores are more complex in linguistic elements – longer sentences, more complex and polysyllabic words. Therefore, as works which performed the best in manual evaluation have higher readability, the obtained manual score does not mean that the writings have better readability. The readability analysis by grade groups gave evidence to the opposite conclusion – writings with low(est) manual grades are more intelligible for the reader.

This thesis supports the outcome of some of the previously conducted studies on readability and human assigned grades, which suggested that there is no strong significant correlation between readability and manually assigned grades [8] [18]. On the other hand, the results of conducted readability analysis are in conflict with a study that found strong

correlation between readability metrics and human assigned scores [17]. The study estimated the readability of the text by using Flesch-Kincaid Grade Level method (which is also implemented and used in the readability analysis of present thesis) and found that readability alone had a strong correlation with manual scoring [17]. In addition, the research done on the validity of AES tool IntelliMetric, which evaluates readability of the text as one of the components of automatically assigned grade, proved that the automated scoring tool replicates the scores of human raters [14].

The results of the conducted readability analysis of this thesis does not recommend using automated readability evaluation as the only method to assign a grade to academic writings. In addition to linguistic elements of the text, reading difficulty also depends on the motivation, background and interest of the reader. Therefore, readability methods and computed grades could merely assist and be used as a part in the process of manual grading.

The discrepancies in the results of the previous studies and this thesis, also the conflicting outcome of readability analysis in grade groups and the correlation analysis overall, suggest that further investigation in terms of readability and manual grading might be needed. In further research in the area of readability, SMOG formula is recommended as one of the readability assessment approaches to be used in future studies. The analysis conducted in grade groups suggests that readability of the writings in different levels of manually assigned grades should be investigated more thoroughly – for instance, enlarging the amount of writings to be analysed and more reviewers with various background could provide more detailed insight into the nature of the correlation between manually assigned and computed readability grades.

References

[1] "What is readability and how can it help you?," [Online]. Available: https://readable.io/blog/what-is-readability/. [Accessed 15 October 2018].

[2] W. H. DuBay, "Principles of Readability," Impact Information, Costa Mesa, 2004.

[3] R. Gunning, The Technique of Clear Writing, Madison: McGraw-Hill, 1952.

[4] J. P. Kincaid, R. P. Fishburne, R. L. Rogers and B. S. Chissom, "Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel," *CNTECHTRA Research Branch Report,* pp. 8-75, 1975.

[5] "Readability Formulas," [Online]. Available: https://www.readabilityformulas.com/. [Accessed 5 December 2018].

[6] E. Fry, "Readability," in *Reading Hall of Fame,* 2006.

[7] A. Vivekanantham, J. Protheroe, S. Muller and S. Hider, "Evaluating on-line health information for patients with polymyalgia rheumatica: a descriptive study," *BMC Musculoskelet Disord,* vol. 18, 2017.

[8] M. Nemati and M. Azizi, "Readability index of essays as an alternative to the scoring procedure in L2 academic writing," *Journal of Paramedical Sciences,* vol. 4, 2013.

[9] P. Plavén-Sigray, G. Matheson, B. Schiffler and W. Thompson, "The readability of scientific texts is decreasing over time," *eLife,* vol. 6, no. e27725, 2017.

[10] Y. Ma, E. Fosler-Lussier and R. Lofthus, "Ranking-based readability assessment for early primary children's literature," in *North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Montreal, 2012.

[11] N. Falconer, E. Reicherter, B. Billek-Swahney and S. Chesbro, "An Analysis of the Readability of Educational Materials on the Consumer Webpage of a Health Professional Organization: Considerations for Practice," *The Internet Journal of Allied Health Sciences and Practice,* vol. 9, no. 3, 2011.

[12] D. E. Powers, J. Burstein, M. Chodorow, M. E. Fowles and K. Kukich, "Stumping E-Rater: Challenging the Validity of Automated Essay Scoring," ETS Research Report, Princeton, 2001.

[13] J. Wang and M. Brown, "Automated Essay Scoring Versus Human Scoring: A Comparative Study," *The Journal of Technology, Learning, and Assessment,* vol. 6, no. 2, 2007.

[14] L. Rudner, V. Garcia and C. Welch, "An Evaluation of the IntelliMetric Essay Scoring System," *The Journal of Tecnology, Learning and Assessment,* vol. 4, no. 4, 2006.

[15] E. Amorim, M. Cancado and A. Veloso, "Automated Essay Scoring in the Presence of Biased Ratings," in *North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* New Orleans, 2018.

[16] “IntelliMetric: How it works,” [Online]. Available: http://www.vantagelearning.com/products/intellimetric/intellimetric-how-it-works/. [Accessed 28 October 2018].

[17] A. Nigam, “Exploring Automated Essay Scoring for Nonnative English Speakers,” New Delhi, 2017.

[18] H. Pooneh, “The Validity of Some Popular Readability Formulas,” *Mediterranean Journal of Social Sciences,* vol. 3, no. 2, 2012.

[19] L. Feng, M. Jansche, M. Huenerfauth and N. Elhadad, “A Comparison of Features for Automatic Readability Assessment,” in *23rd International Conference on Computational Linguistics*, Beijing, 2010.

[20] E. A. Smith and R. J. Senter, “Automated Readability Index,” *Wright-Patterson Air Force Base,* no. AMRL-TR-6620, 1967.

[21] M. Coleman and T. L. Liau, “A computer readability formula designed for machine scoring,” *Journal of Applied Psychology,* vol. 60, no. 2, pp. 283-284, 1975.

[22] J. R. Bormuth, “Readability: A new approach,” *Reading Research Quarterly,* vol. 1, no. 3, pp. 79-132, 1966.

[23] G. H. McLaughlin, “SMOG Grading - a New Readability Formula,” *Journal of Reading,* vol. 22, pp. 639-646, 1969.

[24] E. B. Fry, “A readability formula that saves time,” *Journal of Reading,* vol. 11, pp. 513-516, 1968.

[25] “Count Wordsorth. Dale-Chall Easy Word List,” [Online]. Available: http://countwordsworth.com/download/DaleChallEasyWordList.txt. [Accessed 4 November 2018].

[26] S. Zhou, H. Jeong and P. A. Green, “How Consistent Are the Best-Known Readability Equations in Estimating the Readability of Design Standards?,” *IEEE Transactions on Professional Communication,* vol. 60, no. 1, pp. 97-111, 2017.

[27] “Laerd Statistics: Spearman's Rank-Order Correlation,” [Online]. Available: https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php. [Accessed 10 November 2018].

[28] “SPSS Tutorials,” [Online]. Available: https://www.spss-tutorials.com/statistical-significance/. [Accessed 10 November 2018].

[29] “Explorable: Statistical correlation,” [Online]. Available: https://explorable.com/statistical-correlation . [Accessed 11 November 2018].

[30] K. A. Danielson, “Readability Formulas: A Necessary Evil?,” *Reading Horizons,* vol. 27, no. 3, 1987.

Appendix 1 – Experiment environment

The readability application tool developed by the Author for the use of this thesis included a Java desktop application that implemented the selected readability algorithms. The source code of the application has been uploaded to the GitLab[1] of Department of Computer Systems.

Besides the source code, the executable file of Readability Tool - *ReadabilityTool.jar* - is also available in the aforementioned GitLab directory. When running the executable file, the readability application starts and allows the user to estimate the readability of a text by uploading a file into the environment. The computed readability scores are then outputted in the GUI as shown in Figure 5.

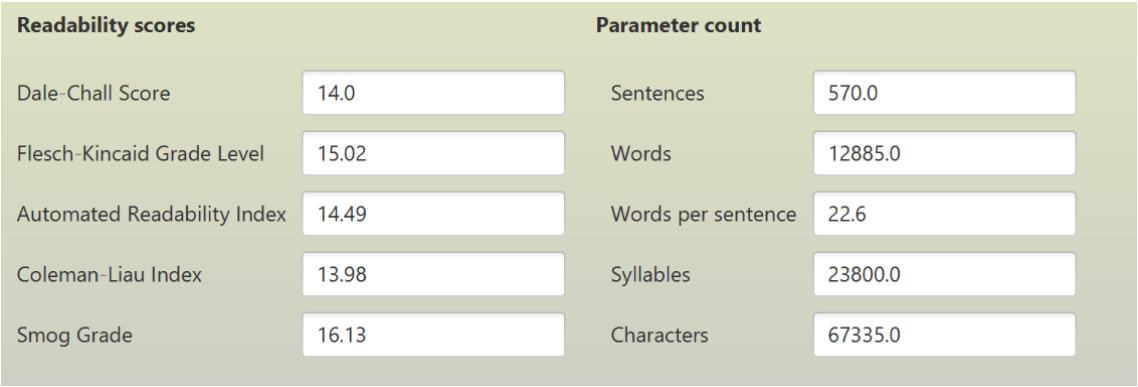

Figure 5. Computed readability results of sample text in Experiment environment.

Before uploading a document to estimate its readability, it is strongly recommended to prepare the text if it contains certain elements like figures, tables, graphs. Removing unnecessary content in terms of readability improves the accuracy of calculations.

As the Experiment environment was initially developed to support the conducted readability analysis in this thesis, the application needs further improvements to enable wider usage of the readability tool.

[1] https://gitlab.pld.ttu.ee/silva.s/ReadabilityTool