# Leesbaarheid: An Empirical Exploration of the Applicability of Dutch Traditional Readability Formulas to Texts Targeting Children

**Joris Voogt**[1]

**Supervisor: Maria Soledad Pera**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Joris Voogt
Final project course: CSE3000 Research Project
Thesis committee: Maria Soledad Pera, Examiner: Pradeep Murukannaiah

## Abstract

Readability plays a vital role in the transfer of knowledge or information. This is especially true for children who are more than anyone gaining new knowledge every day, yet their reading abilities are still in active development. Although readability assessment of children's literature is an active topic of research in many different languages, research into the assessment capabilities of well-known Dutch traditional readability formulas on texts aimed at Dutch children remains in its infancy. In this paper, we explore the performance of four well-known Dutch traditional readability formulas when applied to different types of texts targeting Dutch children. More specifically, we examine the applicability of the Flesch-Douma, Leesindex A, CLIB and CILT formulas on excerpts of school materials, books, and media for children aged 6-14 years old. Outcomes from this empirical exploration reveal that the readability formulas do not perform similarly across reading materials. As a result, choosing a readability formula based on both the type of reading material and a child's age can improve the readability estimation, which in turn can help connect children to understandable resources.

## 1 Introduction

Readability, being able to read and understand a text, is important for gaining new knowledge or understanding information. Hence, the intended audience of a text should have no issues with the difficulty of said text. The process which estimates a text's complexity is referred to as readability assessment and serves two main purposes [22]. Its first purpose is to signify if a text requires readability improvement such that it better fits its target audience. For example, it allows teachers to assess if their created school materials match the readability comprehension of their students [21]. Its second purpose, the opposite of its first, is to connect an audience to texts that they can understand. For instance, the age labels on books in libraries give an indication of the age a child should be such that they are able to read and understand them.

Besides the traditional physical texts like books and newspapers, there are also plenty of digital reading materials such as subtitles for movies and television and web pages on the internet. Likewise, these can also benefit from readability assessment. Namely, using readability assessment to help investigate if government websites and online documents are understandable by the average adult [16]. Recent research has shown that children specifically can have trouble understanding web resources they obtain from popular search engines like Google and Bing [5; 2], and so another example is that readability assessment can help filter these web resources to only contain those resources that children can comprehend. As a result, automated readability assessment tools have been and still are being researched to improve the transfer of knowledge and information from a text to an audience [34; 1].

Research in automatic readability assessment in the previous century yielded the so-called readability formulas [24]. These are now referred to as traditional readability formulas and calculate a score based on simple features inferred from a text. For example, in the English language, the Flesch Reading Ease formula [17] takes the average sentence length in words and the average word length in syllables of a text to produce a difficulty score. In 1975, this formula was recalculated to produce a U.S. grade level instead and this variation is known as the Flesch-Kincaid formula [20], one of the most popular traditional readability formulas. These formulas have been criticized for making use of the so-called "shallow" features of a text [7; 22] which impacts their estimations of text readability [3]. Hence, more recent research focuses on the use of machine models in readability assessment [25; 13]. However, traditional readability formulas are significantly easier to implement due to their simplistic nature when compared to state-of-the-art machine models and do not suffer from opacity when it comes to explaining which textual features affect its readability score [1]. Moreover, the authors in [1] show the potential of traditional readability formulas when estimating the readability of English texts targeting children. Naturally, the question arises of how traditional readability formulas of other languages fare when estimating the readability of children's texts.

As there are many different languages, each with their own peculiarities, to control the scope of this work, we focus on a singular language. In recent years, the reading proficiency of Dutch children has gone down substantially [18] and so, readability assessment could help them connect to understandable texts. As a result, this work explores readability assessment of texts targeting children using traditional readability formulas in the **Dutch** language.

There are four well-known traditional readability formulas in Dutch: the Flesch-Douma [14], the Leesindex A [6], the Cito leesbaarheidsindex voor het basisonderwijs (CLIB) [34] and the Cito leesindex technisch lezen (CILT) [33]. Of these four, the Leesindex A has been used in the readability estimation of texts targeting children in the past [37]. Nowadays, the CLIB and CILT formulas created by the Cito[1] organization are used instead to not only assess the readability of children's texts, but also test the readability comprehension of children in Dutch primary schools [37; 34; 33]. All four formulas have already been tested on adult texts and the results showed that both the CLIB and CILT formulas are not as effective in estimating the difficulty of adult texts when compared to the older and more general Flesch-Douma and Leesindex A formulas [41; 43]. However, as the CLIB and CILT formulas were designed specifically for children's readability estimation, we can assume that they will function more effectively when assessing texts targeting children. Since the CLIB and CILT formulas have not been tested on a large corpus of Dutch children's texts [41], and neither have the Flesch-Douma and Leesindex A formulas to our knowledge, we explore the readability estimations of these four for-

[1](Centraal Instituut voor Toets Ontwikkeling) The Dutch Central Institute for Test Development which creates tests and exams used in education [9].

mulas on Dutch children's texts. As children not only consume physical texts like books but also a lot of media [28], we compare the readability estimations of the four formulas on different types of texts targeting children.

With this work, we aim to answer the following research questions:

RQ1 **How do Dutch traditional readability formulas fare when estimating the readability of Dutch texts targeting children?** We analyze the effectiveness of the Flesch-Douma, the Leesindex A and the CLIB and CILT Dutch traditional readability formulas in predicting the readability of children's texts by comparing and contrasting their performance across grade levels and text types. We posit that the CLIB and CILT formulas should give a better readability estimation as opposed to the Flesch-Douma and Leesindex A because the CLIB and CILT formulas have a stronger empirical basis and their purpose is to assess both the difficulty of children's literature and the reading abilities of children [23; 33]. Nowadays, reading materials include not only physical texts but also media. This leads us to another question.

RQ2 **Is there a significant difference between readability estimation of school materials, books and media texts using the Dutch traditional readability formulas?** Using the results from the first research question, we take a closer look at the readability estimations of the four formulas on different types of texts, in particular school materials, books and media. Similarly to the first research question, we compare and contrast performance across grade levels and text types.

The analysis of our results show that none of the readability formulas estimate readability consistently across different text types and grades. This implies that the choice of readability formula should depend on the type of reading material and the age of the target audience. As this is an empirical exploration into Dutch readability of texts targeting children, our results and conclusions provide a foundation for other research to build upon. Therefore, we provide our code and shareable data in the following repository: https://github.com/JorisVoogt/RP-Dutch-Readability.

The rest of this paper is organized as follows: In the next section, we describe the methodology that was used. Section 3 states the found results and includes a discussion of these results. In Section 4, we reflect critically on the ethical aspects of this research. Finally, in Section 5, the conclusions are summarized and we discuss limitations and future work.

## 2 Method

In this section, we describe the readability formulas, data set, and metrics of the proposed empirical exploration and end with an explanation of our experimental setup. Any code used in this research can be found on: https://github.com/JorisVoogt/RP-Dutch-Readability.

### 2.1 Readability formulas

The Dutch traditional readability formulas of which we explored their readability assessment capabilities on texts aimed at children are the following:

- Flesch-Douma (FD) [14] from 1960, based upon the Flesch Reading Ease [17].
- Leesindex A (LiA) [6] from 1963, also based upon the Flesch Reading Ease and was used to calculate the AVI levels, a level stating the reading proficiency of a child in primary education, in the Netherlands until 2008 [37].
- CLIB: Cito readability index for primary education [34] from 1994 which focuses on if a child understands what they are reading.
- CILT: Cito readability index technical reading [33] from 1997 which focuses on the reading proficiency of a child and is the successor of the Leesindex A for calculating the AVI levels from 2008 onwards [37]. To obtain the CILT-value, the value which can be mapped to a school grade, the CILT score must be subtracted from 150 [32].

Each readability formula is shown in Table 1 where a positive slope signifies that a higher score equals a more difficult text and a lower score an easier text. The negative slope works in the opposite way where a lower score equals a more difficult text and a higher score an easier text. In Table 2, each variable in the readability formulas is explained in greater detail. Both Tables 1 and 2 are adaptations of tables constructed by Oosten et al. [41].

In the remainder of this section, we describe how we calculated each variable in Table 2 and we explain our choice of score-to-grade mappings for each readability formula.

**Calculation of formula variables**

In order to calculate most of the variables used in the readability formulas, we made use of Python's Textstat library [4]. However, the amount of unique words in a text is not supported by Textstat and so we made use of a Python set to store and count these.

As stated in Table 2, the *freq77* word lists, Dutch frequency word list with a cumulative frequency of 77%, are adaptations of a frequency list by Schrooten & Vermeer[2] [31]. Our choice for this frequency list was based upon the fact that it is the largest and newest corpus for texts aimed at Dutch children after the BasiLex-corpus [36]. We opted to use two word lists, one which contains Dutch stop words as part of its cumulative frequency, whereas the other does not. In the latter case, stop words were removed and counted using the NLTK Python package [27] before the *freq77* was calculated. Both word lists contain lemmatized words, also referred to as the dictionary form, and therefore the spaCy Python package [15] was used to lemmatize the texts before calculating the *freq77*. Furthermore, the frequency list by Schrooten & Vermeer also differentiates words that are identical in writing but have different meanings. However, this was not something that we took into account in this research.

The syllable count of a text was obtained by checking if a word is contained in a CELEX [29] data set[3] of close to

[2]The list can be downloaded from https://annevermeer.github.io/woordwerken.html

[3]This data set can be obtained from the following GitHub repository: https://github.com/KBNLresearch/scansion-generator

| Name | Acronym | Formula | Slope |
|---|---|---|---|
| Flesch-Douma [14] | FD | $206.84 - 0.93 \times (w/sen) - 77 \times (syl/w)$ | $-$ |
| Leesindex A [6] | LiA | $195 - 2 \times (w/sen) - 66.67 \times (syl/w)$ | $-$ |
| CLIB: Cito leesbaarheidsindex voor het basisonderwijs [34] | CLIB | $46 + 0.474 \times freq77 - 6.603 \times (let/w)$ $-0.364 \times ttr + 1.425 \times psw$ | $+$ |
| CILT: Cito leesindex technisch lezen [33] | CILT | $114.49 + 0.28 \times freq77 - 12.33 \times (let/w)$ | $+$ |

Table 1: The four well-known Dutch traditional readability formulas used in this paper. A positive slope signifies that a higher score equals a more difficult text, whereas a negative slope signifies the opposite. Table 2 contains an explanation of the variables in the formulas. This table is based upon a similar one first introduced by Oosten et al. [41].

| Variable | Description |
|---|---|
| (w/sen) | The average sentence length in words. |
| (syl/w) | The average word length in syllables. |
| (let/w) | The average word length in letters. |
| psw | The percentage of sentences per word. |
| ttr | Type/token ratio, the percentage of unique words in a text. |
| freq77 | The percentage of words which are included in a Dutch frequency word list with a cumulative frequency of 77%. For this work, the frequency lists considered are adaptations of the one from Schrooten & Vermeer [31] (see project repository). |

Table 2: The variables used in the readability formulas in Table 1. Percentages are between 0 and 100. It is based upon a similar table by Oosten et al. [41].

| Score | | | | Grade |
|---|---|---|---|---|
| FD | LiA | CLIB | CILT | |
| >100 | >100 | <8 | <59 | 1 |
| >100 | ≥89 | 8-20 | 59-63 | 2 |
| >100 | 79-100 | 21-35 | 64-67 | 3 |
| 90-100 | 74-88 | 36-48 | 68-71 | 4 |
| 80-90 | 69-83 | 49-61 | 72-74 | 5 |
| 70-80 | 69-73 | 62-74 | ≥75 | 6 |
| 60-70 | 69-73 | 75-87 | ≥75 | 7 |
| 60-70 | 50-73 | ≥ 88 | ≥75 | 8 |
| 45-60 | 50-68 | - | - | 9 |
| 45-60 | 35-68 | - | - | 10 |
| 30-45 | 20-55 | - | - | 11 |
| 30-45 | 20-40 | - | - | 12 |
| <30 | <40 | - | - | Col./Uni. |

Table 3: Used score-to-grade mappings for the readability formulas.

400,000 Dutch words which includes their syllables. An initial accuracy check showed that about 52% of the unique words in our preprocessed BasiLex data (see Table 5) was known and therefore correctly counted. Splitting words to deal with compound words increased the accuracy to 79%. The remaining 21% mostly consists of human and location names, numbers, and misspelled words which are not contained in the data set and were therefore handled by Textstat's syllable counter. A short list of words with their syllable counts including common words in children's texts showed that this counter is 96% accurate, whereas our own dictionary is 100% accurate on this same word list. For the full code implementation and the short word list, we refer to our repository.

**Score-to-grade mappings**

As the goal of this research is to compare the given grade level of a text to its estimated one, the score a readability formula produces had to be translated to a grade, i.e., the ground truth provided in the data set. In Table 3, the approximate score-to-grade mapping for each readability formula is given. We explain how we obtained the score-to-grade mappings as follows.

In the case of the Flesch-Douma [14] formula, the education system mentioned in its mapping has changed since then and grades were therefore estimated using the Mammoetwet [8]. This law was the end of the Dutch education system used during the creation of the Flesch-Douma formula and forms the basis of the current one.

Leesindex A [6] uses book/text genres as a difficulty measure as opposed to grades/ages for texts aimed at humans past primary education. The idea behind this is that certain book/text genres are seen as more complex. Of course, this is more subjective which is also shown in the readability levels for the ages of 12-18 in the Netherlands [11; 12]. As a result, we chose to take the averages of those readability levels for ages 12-15 [11] and 15-18 [12] and use these to approximate the grades. As a score between 80 and 100 signifies primary education as a whole, further investigation resulted in a more refined score-to-grade mapping for different grades in primary education [30]. Finally, when a score can be mapped to multiple grades, all these grades were taken into account.

The CLIB and CILT formulas [34; 33] already have existing mappings of scores to end of school years [35; 42], however, these are limited to primary education. As our data set in Table 5 contains entries up to and including grade eight, a second source of CLIB score-to-grade mappings was used to add the first two years of secondary school (grades seven and eight) [26], whereas, for the CILT, no additional mappings could be found. As such, a score equal to or higher than 75 was used for grades six through eight.

## 2.2 Data set

In order to analyze the Dutch traditional readability formulas on texts aimed at children, we required a data set that not only contains a large enough volume of texts but also is an-

| | School | Books | Media | All types |
|---|---|---|---|---|
| Grade 1 | 1,067 | 6,441 | - | 7,508 |
| Grade 2 | 9,285 | 19,056 | - | 28,341 |
| Grade 3 | 39,329 | 2,542 | 14,983 | 56,854 |
| Grade 4 | 52,376 | 2,991 | 17,082 | 72,449 |
| Grade 5 | 51,306 | 23,858 | 2,153 | 77,317 |
| Grade 6 | 46,917 | 1,646 | 608 | 49,171 |
| Grade 7 | - | 10,356 | - | 10,356 |
| Grade 8 | - | 457 | - | 457 |
| All grades | 200,280 | 67,347 | 34,826 | **302,453** |

Table 4: Statistics of the BasiLex-corpus after the first round of preprocessing.

| | School | Books | Media | All types |
|---|---|---|---|---|
| Grade 1 | 4 | 178 | - | 182 |
| Grade 2 | 132 | 5,921 | - | 6,053 |
| Grade 3 | 1,523 | 1,329 | 583 | 3,435 |
| Grade 4 | 3,031 | 986 | 7,734 | 11,751 |
| Grade 5 | 4,337 | 9,111 | 140 | 13,588 |
| Grade 6 | 3,771 | 1,060 | 11 | 4,842 |
| Grade 7 | - | 5,207 | - | 5,207 |
| Grade 8 | - | 335 | - | 335 |
| All grades | 12,798 | 24,127 | 8,468 | **45,393** |

Table 5: Statistics of the BasiLex-corpus after the second round of preprocessing. This resulting data set was used to explore the applicability of the Dutch traditional readability formulas.

notated with a grade or age to validate the use of the readability formulas. The BasiLex-corpus[4] [36] satisfies both these properties while also being the newest and largest corpus of Dutch texts aimed at children [36] and was, therefore, our data set of choice. Furthermore, it contains multiple types of children's texts: School-related texts, books, and media. Therefore, providing a broader test scenario and the opportunity for comparing the readability estimations of different types of texts.

An entry in the BasiLex-corpus is an annotated piece of text in the Format for Linguistic Annotation (FoLiA) [39; 40]. For the purpose of this research, any annotation was of no interest, and therefore the BasiLex-corpus was first preprocessed. From an entry, we extracted the actual text, the type of the text and the Dutch primary school grade which reflects at which age a child should be able to understand the text. Table 4 shows the number of entries for each type and grade after the first round of preprocessing. It should be noted that of the total of 305,994 entries in the BasiLex-corpus, 443 entries could not be opened using the Python library FoLiA [38] and another 241 entries do not contain grades rendering these of no use to this research. Furthermore, anything below grade one (i.e., preschool and kindergarten) have been excluded from this research. The reason behind this is that children in the Netherlands do not learn to read and write until grade one of primary education. Therefore, texts in the BasiLex-corpus for those earlier ages are either supposed to be read by adults to children or are picture books with simple words [36]. As a result, these texts are either not directly aimed at children or are simply too short for a meaningful readability assessment.

The second round of preprocessing consisted of selecting entries with enough "volume" (words and sentences) for our proposed research. The challenge here is that grade one entries are quite short due to children first learning to read in this grade. After examining the entries in grade one, and using the fact that known implementations of some of the readability formulas require at least 75 words [32], the choice was made to keep all entries consisting of at least 75 words and five sentences. The exact amount of entries can be found in Table 5. Of note, the number of entries in grade one school and grade six media are considerably lower when compared to other grade and type combinations. This was taken into account when we evaluated the results in Section 3.

[4]Due to copyright reasons, the corpus cannot be directly shared. However, for research purposes, a download can be obtained through http://hdl.handle.net/10032/tm-a2-n4.

## 2.3 Metrics

In our exploration, we consider a number of well-known metrics for evaluation purposes. As the predicted grade can be a range of grades, depending on the metrics, the grade closest to the expected grade is the one used as ground truth. We define:
$n =$ number of samples in our data set,
$y_i =$ the predicted grade by a given readability formula of the ith sample in the data set, and
$x_i =$ the ground truth of the ith sample, i.e., the grade assigned to the ith sample in the data set.

**Accuracy**
In order to examine the accuracy of the predictions generated by each formula, we used Equation 1. A higher accuracy score equals a better readability estimation.

$$Accuracy = \frac{|MatchedSamples|}{n} \tag{1}$$

where $|MatchedSamples|$ refers to the number of samples in $n$ for which the predicted readability estimation matches the ground truth.

**MAE: Mean absolute error**
To examine the average error rate of the readability formulas, we used the MAE as seen in Equation 2. A lower MAE means a better readability estimation as it signifies lower average error rate.

$$MAE = \frac{\Sigma_{i=1}^{n}|y_i - x_i|}{n} \tag{2}$$

**RMSE: Root mean square error**
We used the RMSE as observed in Equation 3 to check if any readability estimations were considerably missing their target grade. Like with the MAE, a lower RMSE means a better readability estimation.

$$RMSE = \sqrt{\frac{\Sigma_{i=1}^{n}(y_i - x_i)^2}{n}} \tag{3}$$

### 2.4 Experimental setup

Using the data set described in Table 5, we predict the grade for each of the samples in our data set using the formulas introduced in Section 2.1. Based on the predicted grades and the expected grades, i.e., the ground truth, we compare and contrast performance using the metrics introduced in Section 2.3 across the formulas, as well as juxtaposed across grades and text types. In order to demonstrate the significance of the comparisons reported in Section 3, we performed the one-way ANOVA and the Tukey post-hoc test with a $p < 0.05$. We found that the average absolute error rate or MAE has a statistically-significant difference according to the readability formulas ($p < 2e\text{-}16$) and showed significant pairwise differences between all the readability formulas with a $p < 0.005$ with the exception of the CLIB_stop[5] and CILT pair which does not have a significant pairwise difference. We also found a statistically-significant difference on the MAE per readability formula with respect to the text types ($p < 2e\text{-}16$) where most of the text type pairs showed a significant difference for each of the six readability formulas ($p < 0.05$). The exceptions include no significant difference between school materials and books for both the Leesindex A and CLIB_stop.

## 3 Results and Discussion

Here, we present and discuss our results.

### 3.1 Results

We first examine the overall accuracy, MAE and RMSE across all readability formulas on the entire preprocessed data set. These results can be found in Table 6. The Leesindex A shows the highest overall accuracy, however, it also has the highest overall MAE and RMSE. The accuracy could be explained by the fact that the score-to-grade mapping for the Leesindex A as seen in Table 3 contains many overlapping score ranges. In our case, that means as long as the expected grade, i.e., the ground truth, appears in the list of predicted grades, it is counted as a correct prediction for the accuracy metric. One of the reasons for the higher MAE and RMSE could be attributed to the very small score range for grades six and seven and the surrounding score ranges mapping to grades which are a few grades higher/lower. Investigation into the average scores show that for grade six the average score is 67. This means the prediction on average misses the expected grade by two grades, yet the score is off by only two points.

The CLIB, on the other hand, has the lowest overall MAE and RMSE while its accuracy of 0.19 is just below the average of 0.21 across all six readability formulas. Given its low MAE and RMSE that indicate that its predictions are on average within 1.5 grades from the expected grade, we posit that the CLIB formula gives the closest estimation of expected grades of texts targeting children. Nevertheless, its readability estimation may differ across grades and text types.

In order to inspect the consistency of the estimations made by the readability formulas across grades and text types, we examine the absolute error rates across grades, text types and readability formulas as seen in Figure 1. We also look at the overall metric values for each text type displayed in Table 7. In the remainder of this section, when we mention error rates, we are actually referring to the absolute error rates.

One immediate difference we notice is that the newer CLIB and CILT formulas have smaller outliers overall when compared to the older Flesch-Douma and Leesindex A formulas. However, the score-to-grade mappings in Table 3 show that the CLIB and CILT formulas are limited to grades one through eight while the Flesch-Douma and Leesindex A formulas can be mapped to grades one through twelve and College/University (which is internally represented as grade 13). Consequently, the error rate limit of the CLIB and CILT is lower when compared to the other two formulas. Another observation we can make is that for school materials and books the CLIB formulas have a particularly low error rate on grade four while getting an increasingly bigger error rate on grades further away from grade four. If we take a look at the complete data we gathered on the CLIB formulas in Appendix A, Tables 10 and 11, we notice a similar trend where grade four has the highest accuracy and the lowest MAE and RMSE. Grades further away get progressively higher MAE and RMSE while below grade three and above grade five the accuracy is zero.

When we examine school materials, the Leesindex A has a very consistent low error rate on grades one through four, yet its overall MAE of 1.44 for school materials is higher when compared to Flesch-Douma's MAE of 1.24. An explanation is that when we look back at our data distribution in Table 5, most of the school materials have an expected grade of four, five, or six. Particularly on grade five, the Leesindex A shows a wide variation in error rates resulting in that higher MAE. As observed, the CLIB formulas have low error rates on grades three and four but show relatively high error rates on the other grades, especially grades one and two when compared to the other formulas. Both CILT formulas show a consistent low error rate across all the grades. The CILT_stop variant has an overall lower MAE when compared to CILT which we can attribute to the data distribution as the CILT_stop variant has lower error rates on the higher grades.

For books, the CILT formulas have consistent error rates on grades one through five, yet their overall MAE rates of 1.91 for CILT and 1.75 for CILT_stop are higher than the MAE rates of 1.59 for the Flesch-Douma and 1.49 for Leesindex A. The data distribution in Table 5 shows that most books have an expected grade of two, five, or seven. Particularly on grades five and seven, the MAE of both the Flesch-Douma and Leesindex A are smaller compared to the MAE of the CILT formulas, therefore explaining why the overall MAE rates of the CILT formulas are higher than the other two formulas. Overall, the Flesch-Douma and Leesindex A have lower error rates on the higher grades compared to both the CLIB and CILT formulas. As said before, the CLIB formulas only have low error rates around grade four. In the case for books, both grades three and four show low error rates (and grade five for the CLIB variant), and grades further away from these get progressively higher error rates.

Finally, examining the media error rates, we see that the

[5] _stop signifies that the *freq77* was calculated using a frequency word list containing stop words.

| Metric | Flesch-Douma | Leesindex A | CLIB | CLIB_stop | CILT | CILT_stop |
|---|---|---|---|---|---|---|
| Accuracy | 0.20 | 0.32 | 0.19 | 0.13 | 0.20 | 0.21 |
| MAE | 1.64 | 1.77 | 1.44 | 1.59 | 1.61 | 1.52 |
| RMSE | 2.07 | 2.47 | 1.81 | 1.94 | 2.04 | 1.92 |

Table 6: The overall accuracy, MAE and RMSE on each readability formula. In the case of the CLIB and CILT formulas, **_stop** signifies that the *freq77* was calculated using a frequency word list containing stop words.

CLIB formulas have low error rates across grades three through five. Our data distribution for media entries in Table 5 shows that almost all of them have an expected grade of four, and the entries are limited to grades three through six. As we have established, the CLIB formulas predict particularly well on grade four and its surrounding grades, therefore it is not surprising that it performs so well on media entries. The Flesch-Douma and Leesindex A error rates are also consistent with the exception of grade four. As grade four contains most of the media entries, it explains the high MAE rates of 2.37 for the Flesch-Douma and 3.08 for the Leesindex A compared to the low MAE rates of 0.50 and 0.77(stop variant) for the CLIB formulas. The CILT formulas show a trend where higher grades have higher error rates. Because they have low error rates on grade four, their overall MAE of 1.21 and 1.45(stop variant) is quite lower than Flesch-Douma's MAE of 2.37 even though Flesch-Douma's error rates are more consistent.

### 3.2 Discussion

Recall that we aimed to answer:

RQ1 **How do Dutch traditional readability formulas fare when estimating the readability of Dutch texts targeting children?** There is no Dutch traditional readability formula which predicts text readability with a consistent error rate across all the grades and text types. However, the CILT formulas do have the most consistent error rates on both school materials and books for grades one through six. Of the two CILT formulas, the CILT variant has lower error rates on grades one through four, whereas the CILT_stop variant has the lower error rates on grades five through eight. Leesindex A has also got similar error rates on both school materials and books compared to the CILT formulas and even shows lower error rates on grades seven and eight for books. We posit that since the error rates are similar across both school materials and books, Leesindex A would perform well on school materials for grades seven and eight. For media entries on the other hand, the CLIB formulas, in particular the CLIB variant, show very consistent low error rates, with the exception of grade six. Nevertheless, considering how the CLIB formulas perform on both school materials and books, it is very likely that on media entries with different expected grades than we tested here, the error rates increase. As the older Flesch-Douma and Leesindex A formulas have consistent error rates on media entries with the exception of grade four, we looked at the difference in the entries between grade four and the other grades. Grade four almost entirely consists of online news articles while the other grades contain only subtitles of TV shows. Therefore, we posit that both the Flesch-Douma and Leesindex A perform well when estimating the readability of subtitles of children's TV shows.

RQ2 **Is there a significant difference between readability estimation of school materials, books and media texts using the Dutch traditional readability formulas?** Observing Figure 1 suggests a significant difference between media entries on the one hand and school materials and books on the other. Using significance testing, we found that there are statistically-significant differences on the MAE between each of these text types for most of the readability formula as explained in Section 2.4. There is no significant difference on the MAE between school materials and books for both the Leesindex A and CLIB_stop.

When we compare our results to how the English traditional readability formulas fare on children's texts [1], we see that overall, the Dutch traditional readability formulas have smaller error rates, in particular at lower grades. The best English traditional readability formulas, Spache, Spache-Sven and Spache-Allen, have a similar trend compared to the Dutch CLIB formulas. In both the English and Dutch language, these formulas are relatively more reliable on media entries.

It was shown that the CLIB and CILT formulas are not as effective in readability estimation of adult texts when compared to the Flesch-Douma and Leesindex A [41; 43]. Our results do not contradict those findings as the CLIB and CILT formulas show considerably higher MAE and RMSE on grades seven and eight when compared to the Flesch-Douma and Leesindex A. This is also not much of a surprise as the CLIB and CILT formulas were created using texts aimed at primary school children [34; 33]. Their use outside of primary education is therefore inherently limited.

## 4 Responsible Research

Here, we emphasize several aspects of responsible research with regards to our work.

**Data sets**

The three data sets used in this research are the BasiLex-corpus [36], a CELEX [29] data set and a frequency word list [31]. The BasiLex-corpus, although not open-source, has both a research paper and a manual explaining its contents in great detail. Furthermore, it is available for free with a non-commercial license at the Instituut voor de Nederlandse Taal[6]. We obtained the corpus through this institute and used

[6]The institute for the Dutch language which collects and describes the spoken and written Dutch language [19].

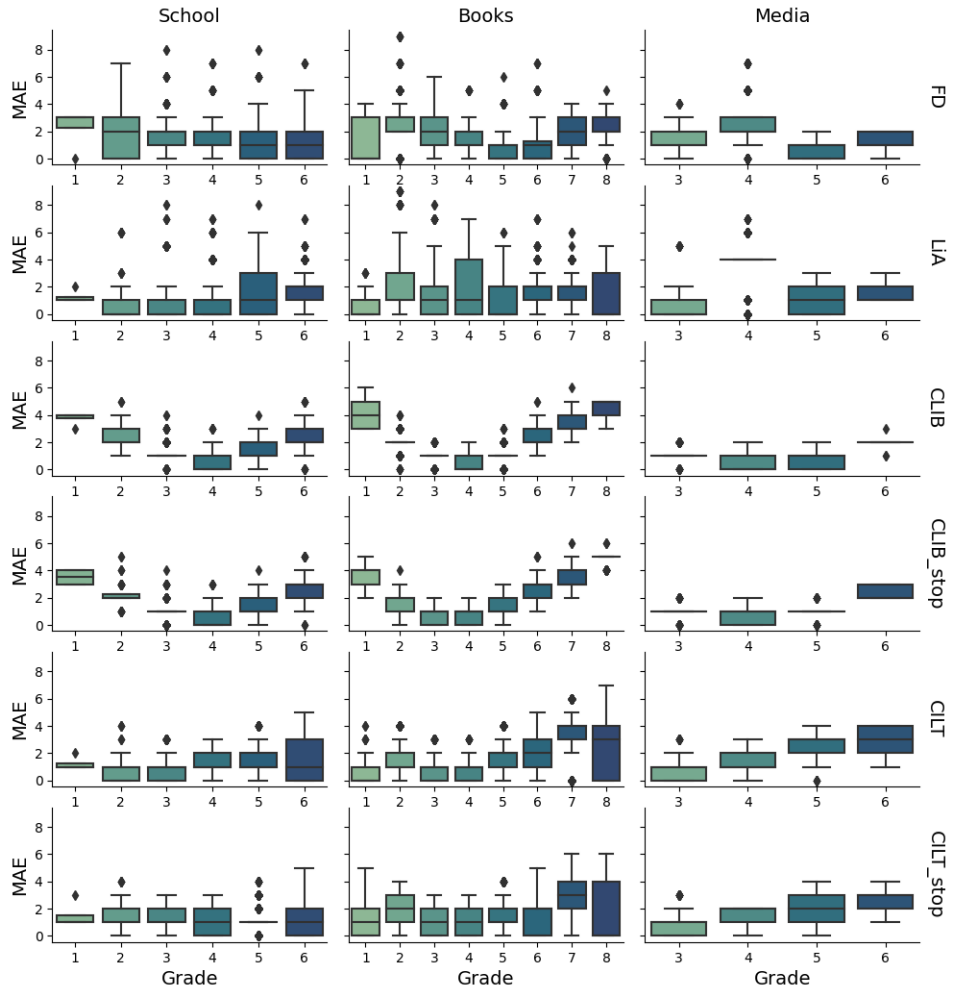


Figure 1: MAE across grades and types of texts for each readability formula. In the case of the CLIB and CILT formulas, **_stop** signifies that the *freq77* was calculated using a frequency word list containing stop words.

| Metric | School | Books | Media | All types |
|---|---|---|---|---|
| Accuracy | 0.25 | 0.23 | 0.02 | 0.20 |
| MAE | 1.24 | 1.59 | 2.37 | 1.64 |
| RMSE | 1.66 | 2.04 | 2.64 | 2.07 |

(a) Flesch-Douma

| Metric | School | Books | Media | All types |
|---|---|---|---|---|
| Accuracy | 0.39 | 0.35 | 0.14 | 0.32 |
| MAE | 1.44 | 1.49 | 3.08 | 1.77 |
| RMSE | 2.08 | 2.19 | 3.55 | 2.47 |

(b) Leesindex A

| Metric | School | Books | Media | All types |
|---|---|---|---|---|
| Accuracy | 0.21 | 0.06 | 0.52 | 0.19 |
| MAE | 1.31 | 1.84 | 0.50 | 1.44 |
| RMSE | 1.62 | 2.13 | 0.74 | 1.81 |

(c) CLIB

| Metric | School | Books | Media | All types |
|---|---|---|---|---|
| Accuracy | 0.18 | 0.05 | 0.26 | 0.13 |
| MAE | 1.48 | 1.94 | 0.77 | 1.59 |
| RMSE | 1.79 | 2.25 | 0.91 | 1.94 |

(d) CLIB_stop

| Metric | School | Books | Media | All types |
|---|---|---|---|---|
| Accuracy | 0.26 | 0.16 | 0.23 | 0.20 |
| MAE | 1.29 | 1.91 | 1.21 | 1.61 |
| RMSE | 1.67 | 2.37 | 1.46 | 2.04 |

(e) CILT

| Metric | School | Books | Media | All types |
|---|---|---|---|---|
| Accuracy | 0.29 | 0.18 | 0.15 | 0.21 |
| MAE | 1.11 | 1.75 | 1.45 | 1.52 |
| RMSE | 1.48 | 2.19 | 1.64 | 1.92 |

(f) CILT_stop

Table 7: The accuracy, MAE and RMSE across text types and readability formulas. In the case of the CLIB and CILT formulas, **_stop** signifies that the *freq77* was calculated using a frequency word list containing stop words.

the manual to guide us in how to use the data set. To aid in accessibility, in Section 2.2, we supplied a direct link to the website where the corpus can be obtained. In that same section, we explain in detail how we preprocessed the corpus including tables showing how many entries are left in our used data set. We also believe that the choice made to remove entries with fewer than 75 words is justified as it is based upon other research [32].

The authenticity of the CELEX data set could be regarded as questionable. Due to time constraints, we could not obtain a membership to the Linguistic Data Consortium [29] which provides the official CELEX data set on CD-ROM. However, a CELEX data file was found on a non-published GitHub repository linking to the web version of CELEX. We compared it to the official CELEX manual and were able to discern a great resemblance in terminology. We would have preferred to get the data set through official means, and therefore do not share it in our repository but rather share the link to the repository it was found in. We omitted to explain that words without their syllable counts in the data set are not considered as this is clear from our use case and the code shared in our repository.

The frequency word list is the result of research [31] and can be obtained for free. We share the link to the list, however, we did not upload it to our repository in case of copyright issues. Our adaptations of this list that we use in our research have been uploaded to our repository.

**Results**

As we are exploring the performance of readability formulas, we are not omitting any data in the results as this would defeat the purpose. We carefully considered what each result meant by comparing it to the other results and looking at the data distribution. In our case, average results could be skewed because of data not being distributed equally across the different types that we were comparing, and so, we keep pointing that out in Section 3. Our conclusion is based upon all our results, and in our case that means we have multiple conclusions as the results showed that different readability formulas are more appropriate on different types of texts and grades. However, as our conclusion should not be a repeat of the discussion section, the focus of the conclusion is upon the main results and takeaways.

**Reproducibility**

We believe that with the detailed method section and our full code being shared on a GitHub repository, anyone can reproduce our results. The data sets are not directly shared, but we do provide links in order to obtain the data sets. Furthermore, a data set containing the intermediate results, this includes the expected grades, text types, and the scores of each readability formula, is made available in our GitHub repository. Therefore, anyone without the data sets can still obtain the same results, and is also able to adjust the score-to-grade mappings for research purposes.

## 5 Conclusions and Future Work

In this paper, we explored how four well-known Dutch traditional readability formulas fared on texts aimed at children. Analysis of our results suggests that the CILT formulas, particularly the CILT variant without stop words in its frequency list, are more appropriate to estimate readability of both school materials and books for grades one through six than the other formulas. On media entries, the best fit are the CLIB formulas although we posit that on grades below grade three and above grade five, this is not the case. We also investigated the difference between readability estimations across different text types and found that overall, these estimations are indeed significantly different from one another, and therefore care should be taken which readability formula is used on which text type. An implication which follows is that depending on the reading material and the age group of the target audience, specific readability formulas are more suited to estimate the difficulty of the reading material.

This work can serve as a base or stepping stone in the research of discovery and retrieval of online resources for Dutch children. Furthermore, it can aid in the creation of algorithms and applications that rely on readability estimation. Particularly, for simplistic applications or ones that require the user to understand why a text matches a certain difficulty level, readability formulas can be more favorable when compared to the complex and opaque state-of-the-art machine models [1].

As we limited ourselves to the BasiLex-corpus, the media entries we examined were limited to one type of news website. An improvement on this paper would be to research how the Dutch traditional readability formulas fare on a wide range of websites aimed at children. We were also limited by the syllable counter with its accuracy of roughly 80% and an improvement on this counter could have an effect on the readability estimations of both the Flesch-Douma and Leesindex A formulas. Finally, the used frequency word lists for both the CLIB and CILT formulas are based upon a dated frequency word list [31]. These lists could be improved by basing them on newer corpora and extracting frequent words from online resources targeting children.

We recommend research into the CLIB formula because in [32], it became clear from examples that its score should be transformed like with the CILT (see Section 2.1) to obtain the final score which can be mapped to a grade. We posit that this score transformation will positively affect the grade estimations and should give a more consistent error rate across grades. Since we were not able to obtain this "complete" CLIB formula in time for this research, we opted to use the CLIB as stated in its research paper [34] and in other research [41; 43; 10]. Another recommendation is research into the score-to-grade mappings of the Flesch-Douma and Leesindex A formulas. Particularly, making new mappings based upon observations of their scores on a wide range of texts on different age groups. Their original mappings are difficult to translate as Flesch-Douma's mapping is based upon an obsolete education system and Leesindex A's mapping uses book/text genres which we argue are not an objective way of establishing readability. Different people enjoy different genres, and therefore find those easier to read and understand. Furthermore, within genres, there can be differences in text difficulty.

## References


[1] Garrett Allen, Ashlee Milton, Katherine Landau Wright, Jerry Alan Fails, Casey Kennington, and Maria Soledad Pera. Supercalifragilisticexpialidocious: Why using the "right" readability formula in children's web search matters. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, editors, *Advances in Information Retrieval*, pages 3–18, Cham, 2022. Springer International Publishing.

[2] Oghenemaro Anuyah, Ashlee Milton, Michael Green, and Maria Soledad Pera. An empirical analysis of search engines' response to web search queries associated with the classroom setting. *Aslib Journal of Information Management*, 72(1):88–111, 2020.

[3] Alan Bailin and Ann Grafstein. The linguistic assumptions underlying readability formulae: A critique. *Language & communication*, 21(3):285–301, 2001.

[4] Shivam Bansal and Chaitanya Aggarwal. textstat 0.7.3 [Python library]. https://pypi.org/project/textstat/, 2022. Accessed: 2023-05-26.

[5] Dania Bilal and Li-Min Huang. Readability and word complexity of serps snippets and web pages on children's search queries: Google vs bing. *Aslib Journal of Information Management*, 71(2):241–259, 2019.

[6] R. H. M. Brouwer. Onderzoek naar de leesmoeilijkheden van nederlands proza. *Pedagogische Studiën*, 40:454–464, 1963.

[7] Peter Burger and Jaap de Jong. Zin en onzin van leesbaarheidsformules. *Onze Taal*, 65:73–77, 1996.

[8] CBS, Centraal Bureau voor de Statistiek. Mammoetwet. https://www.cbs.nl/nl-nl/nieuws/2018/40/50-jaar-mammoetwet-bijna-iedereen-gaat-nu-naar-school/mammoetwet. Accessed: 2023-06-01.

[9] Cito, Centraal Instituut voor Toets Ontwikkeling. Cito website. https://www.cito.nl/. Accessed: 2023-04-29.

[10] Daelemans, Walter and De Clercq, Orphée and Hoste, Veronique. STYLENE : an environment for stylometry and readability research for Dutch. In Odijk, Jan and Van Hessen, Arjan, editor, *CLARIN in the Low Countries*, pages 195–210. Ubiquity Press, 2017.

[11] de Bibliotheek, jeugd. De leesniveaus 12-15 jaar. https://www.jeugdbibliotheek.nl/12-18-jaar/lezen-voor-de-lijst/12-15-jaar/de-niveaus.html. Accessed: 2023-06-01.

[12] de Bibliotheek, jeugd. De leesniveaus 15-18 jaar. https://www.jeugdbibliotheek.nl/12-18-jaar/lezen-voor-de-lijst/15-18-jaar/de-niveaus.html. Accessed: 2023-06-01.

[13] O. De Clercq and Veronique Hoste. All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *COMPUTATIONAL LINGUISTICS*, 42(3):457–490, 2016.

[14] W.H. Douma. *De leesbaarheid van landbouwbladen : een onderzoek naar en een toepassing van leesbaarheidsformules*. Number no. 17 in Bulletin / Afdeling sociologie en sociografie van de Landbouwhogeschool Wageningen. 1960.

[15] Explosion. spacy 3.5.3 [Python library]. https://pypi.org/project/spacy/, 2023. Accessed: 2023-06-01.

[16] Catherine Ferguson, Margaret Merga, and Stephen Winn. Communications in the time of a pandemic: the readability of documents for public consumption. *Australian and New Zealand Journal of Public Health*, 45(2):116–121, 2021.

[17] Rudolph Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233, 6 1948.

[18] Floor Bevaart and Mathilde Jansen. Lezen is verplichte kost! https://www.nemokennislink.nl/publicaties/lezen-is-verplichte-kost/, January 2020. Accessed: 2023-06-16.

[19] ivdnt. /instituut voor de Nederlandse taal/. https://ivdnt.org/. Accessed: 2023-06-22.

[20] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch, 1975.

[21] George R Klare. Assessing readability. *Reading research quarterly*, pages 62–102, 1974.

[22] Rogier Kraf and Henk Pander Maat. Leesbaarheidsonderzoek: oude problemen, nieuwe kansen. *Tijdschrift voor taalbeheersing*, 31(2):97–123, 2009.

[23] Camilla Lindholm and Ulla Vanhatalo. *Handbook of easy languages in Europe*. Frank & Timme, 2021.

[24] Ion Madrazo Azpiazu and Maria Soledad Pera. Is cross-lingual readability assessment possible? *Journal of the Association for Information Science and Technology*, 71(6):644–656, 2020.

[25] Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179, 2021.

[26] Nieuwsbegrip. Wat is de CLIB-score van de teksten? https://www.nieuwsbegrip.nl/page/660/wat-is-de-clib-score-van-de-teksten. Accessed: 2023-06-01.

[27] NLTK Team. nltk 3.8.1 [Python library]. https://pypi.org/project/nltk/, 2023. Accessed: 2023-06-03.

[28] Ofcom. Children and Parents: Media Use and Attitudes. https://www.ofcom.org.uk/__data/assets/pdf_file/0027/255852/childrens-media-use-and-attitudes-report-2023.pdf, March 2023. Accessed: 2023-06-17.

[29] R. Piepenbrock R. H. Baayen and L. Gulikers. The celex lexical database. https://catalog.ldc.upenn.edu/LDC96L14, 1995. Accessed: 2023-06-01.

[30] Rovict. Informatie m.b.t. de Rovict-toetsen In ESIS webbased. https://www.rovict.nl/downloads/ROVICT-toetsen_in_ESIS_-_januari_2016_v.pdf, 2016. Accessed: 2023-06-06.

[31] Walter Schrooten and Anne Vermeer. *Woorden in het basisonderwijs. 15 000 woorden aangeboden aan leerlingen*. Tilburg University Press, 1994.

[32] Rosemary Slagmolen. Dit boek is lekker makkelijk! Een inventarisatie van de problemen met het nieuwe AVI-systeem. Master's thesis, Faculteit Geesteswetenschappen, Universiteit Utrecht, Drift 21, 3512 BR Utrecht, June 2008.

[33] G. Staphorsius and N.D. Verhelst. Indexering van de leestechniek. *Pedagogische studiën*, 74(3):154–164, 1997.

[34] Gerrit Staphorsius. *Leesbaarheid en leesvaardigheid, De ontwikkeling van een domeingericht meetinstrument*. Cito, Arnhem, 1994.

[35] Gerrit Staphorsius, Ronald Krom, Frans Kleintjes, and Norman Verhelst. Verantwoording Verslag van het kalibratie-, validerings- en normeringsonderzoek. https://docplayer.nl/26567361-Verantwoording-verslag-van-het-kalibratie-validerings-en-normeringsonderzoek.html, 2004. Accessed: 2023-06-01.

[36] Agnes Tellings, Micha Hulsbosch, Anne Vermeer, and Antal van den Bosch. Basilex: an 11.5 million words corpus of dutch texts written for children. *Computational Linguistics in the Netherlands Journal*, 4:191–208, 2014.

[37] Christine Stam van Gent. Nieuwe manier om lezen te meten. Vertrouwde AVI-systeem is rijp voor een opvolger. *Reformatorisch Dagblad*, page 11, 26-03-2007. Accessed: 2023-05-26 on https://www.digibron.nl/viewer/collectie/Digibron/id/tag:RD.nl,20070326:newsml_f558dc3f7029d9763606a403bcfd8ba3.

[38] Maarten van Gompel. FoLiA 2.5.8 [Python library]. https://pypi.org/project/FoLiA/, 2022. Accessed: 2023-05-24.

[39] Maarten van Gompel and Martin Reynaert. Folia: A practical xml format for linguistic annotation – a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3:63–81, Dec. 2013.

[40] Maarten van Gompel (proycon). FoLiA, Format for Linguistic Annotation. http://proycon.github.io/folia/. Accessed: 2023-05-24.

[41] Philip van Oosten, Dries Tanghe, and Véronique Hoste. Towards an improved methodology for automated readability prediction. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 5 2010. European Language Resources Association (ELRA).

[42] Alma van Til, Frans Kamphuis, Jos Keuning, Martine Gijsel, and Anja de Wijs. Wetenschappelijke verantwoording LVS-toetsen AVI. https://www.cito.nl/-/media/files/kennisbank/cito-bv/109--cito_lvs-avi_gr-3-tm-halverwege-gr-8-wet-verantwoording.pdf?la=nl-nl, 2018. Accessed: 2023-06-01.

[43] Vincent Vandeghinste and Bram Bulté. Linguistic proxies of readability: Comparing easy-to-read and regular newspaper dutch. *Computational Linguistics in the Netherlands Journal*, 9:81–100, 2019.

# A Results across formulas, metrics, grades, and text types

| | School | Books | Media | All types |
|---|---|---|---|---|
| Grade 1 | 0.25 | 0.68 | - | 0.67 |
| Grade 2 | 0.27 | 0.04 | - | 0.04 |
| Grade 3 | 0.06 | 0.02 | 0.08 | 0.05 |
| Grade 4 | 0.25 | 0.17 | 0.01 | 0.09 |
| Grade 5 | 0.31 | 0.45 | 0.30 | 0.40 |
| Grade 6 | 0.28 | 0.30 | 0.09 | 0.28 |
| Grade 7 | - | 0.09 | - | 0.09 |
| Grade 8 | - | 0.17 | - | 0.17 |
| All grades | 0.25 | 0.23 | 0.02 | **0.20** |

(a) Accuracy

| | School | Books | Media | All types |
|---|---|---|---|---|
| Grade 1 | 2.25 | 1.02 | - | 1.05 |
| Grade 2 | 1.90 | 2.83 | - | 2.81 |
| Grade 3 | 1.80 | 2.11 | 1.51 | 1.87 |
| Grade 4 | 1.26 | 1.36 | 2.46 | 2.06 |
| Grade 5 | 1.10 | 0.64 | 0.83 | 0.79 |
| Grade 6 | 1.13 | 1.21 | 1.27 | 1.15 |
| Grade 7 | - | 1.84 | - | 1.84 |
| Grade 8 | - | 2.12 | - | 2.12 |
| All grades | 1.24 | 1.59 | 2.37 | **1.64** |

(b) MAE

| | School | Books | Media | All types |
|---|---|---|---|---|
| Grade 1 | 2.60 | 1.82 | - | 1.84 |
| Grade 2 | 2.35 | 3.03 | - | 3.02 |
| Grade 3 | 2.07 | 2.32 | 1.72 | 2.12 |
| Grade 4 | 1.66 | 1.66 | 2.72 | 2.41 |
| Grade 5 | 1.56 | 0.92 | 1.04 | 1.16 |
| Grade 6 | 1.54 | 1.74 | 1.41 | 1.59 |
| Grade 7 | - | 2.08 | - | 2.08 |
| Grade 8 | - | 2.42 | - | 2.42 |
| All grades | 1.66 | 2.04 | 2.64 | **2.07** |

(c) RMSE

Table 8: Metrics across grades and types of texts on predictions by the Flesch-Douma.

| | School | Books | Media | All types |
|---|---|---|---|---|
| Grade 1 | 0.00 | 0.58 | - | 0.57 |
| Grade 2 | 0.52 | 0.11 | - | 0.12 |
| Grade 3 | 0.57 | 0.36 | 0.74 | 0.52 |
| Grade 4 | 0.51 | 0.44 | 0.09 | 0.23 |
| Grade 5 | 0.44 | 0.60 | 0.36 | 0.55 |
| Grade 6 | 0.15 | 0.16 | 0.00 | 0.15 |
| Grade 7 | - | 0.17 | - | 0.17 |
| Grade 8 | - | 0.67 | - | 0.67 |
| All grades | 0.39 | 0.35 | 0.14 | **0.32** |

(a) Accuracy

| | School | Books | Media | All types |
|---|---|---|---|---|
| Grade 1 | 1.25 | 0.68 | - | 0.69 |
| Grade 2 | 0.98 | 2.19 | - | 2.16 |
| Grade 3 | 1.02 | 1.84 | 0.50 | 1.25 |
| Grade 4 | 1.22 | 1.48 | 3.31 | 2.62 |
| Grade 5 | 1.56 | 0.89 | 0.94 | 1.11 |
| Grade 6 | 1.67 | 1.91 | 1.55 | 1.72 |
| Grade 7 | - | 1.64 | - | 1.64 |
| Grade 8 | - | 1.01 | - | 1.01 |
| All grades | 1.44 | 1.49 | 3.08 | **1.77** |

(b) MAE

| | School | Books | Media | All types |
|---|---|---|---|---|
| Grade 1 | 1.32 | 1.12 | - | 1.13 |
| Grade 2 | 1.84 | 2.93 | - | 2.91 |
| Grade 3 | 1.92 | 2.76 | 1.21 | 2.20 |
| Grade 4 | 2.07 | 2.30 | 3.69 | 3.24 |
| Grade 5 | 2.22 | 1.54 | 1.26 | 1.79 |
| Grade 6 | 1.98 | 2.37 | 1.68 | 2.07 |
| Grade 7 | - | 1.98 | - | 1.98 |
| Grade 8 | - | 1.79 | - | 1.79 |
| All grades | 2.08 | 2.19 | 3.55 | **2.47** |

(c) RMSE

Table 9: Metrics across grades and types of texts on predictions by the Leesindex A.

| | School | Books | Media | All types |
|---|---|---|---|---|
| Grade 1 | 0.00 | 0.00 | - | 0.00 |
| Grade 2 | 0.00 | 0.00 | - | 0.00 |
| Grade 3 | 0.08 | 0.15 | 0.07 | 0.11 |
| Grade 4 | 0.68 | 0.69 | 0.56 | 0.60 |
| Grade 5 | 0.12 | 0.05 | 0.29 | 0.08 |
| Grade 6 | 0.00 | 0.00 | 0.00 | 0.00 |
| Grade 7 | - | 0.00 | - | 0.00 |
| Grade 8 | - | 0.00 | - | 0.00 |
| All grades | 0.21 | 0.06 | 0.52 | **0.19** |

(a) Accuracy

| | School | Books | Media | All types |
|---|---|---|---|---|
| Grade 1 | 3.75 | 4.08 | - | 4.08 |
| Grade 2 | 2.47 | 1.91 | - | 1.93 |
| Grade 3 | 1.18 | 0.91 | 1.18 | 1.08 |
| Grade 4 | 0.34 | 0.32 | 0.45 | 0.41 |
| Grade 5 | 1.17 | 1.10 | 0.76 | 1.12 |
| Grade 6 | 2.27 | 2.40 | 1.91 | 2.30 |
| Grade 7 | - | 3.22 | - | 3.22 |
| Grade 8 | - | 4.57 | - | 4.57 |
| All grades | 1.31 | 1.84 | 0.50 | **1.44** |

(b) MAE

| | School | Books | Media | All types |
|---|---|---|---|---|
| Grade 1 | 3.77 | 4.17 | - | 4.16 |
| Grade 2 | 2.57 | 1.97 | - | 1.99 |
| Grade 3 | 1.32 | 1.02 | 1.29 | 1.21 |
| Grade 4 | 0.61 | 0.57 | 0.67 | 0.65 |
| Grade 5 | 1.33 | 1.18 | 0.93 | 1.23 |
| Grade 6 | 2.36 | 2.47 | 1.98 | 2.39 |
| Grade 7 | - | 3.28 | - | 3.28 |
| Grade 8 | - | 4.60 | - | 4.60 |
| All grades | 1.62 | 2.13 | 0.74 | **1.81** |

(c) RMSE

Table 10: Metrics across grades and types of texts on predictions by the CLIB with a frequency word list not containing stop words.

| | School | Books | Media | All types |
|---|---|---|---|---|
| Grade 1 | 0.00 | 0.00 | - | 0.00 |
| Grade 2 | 0.00 | 0.01 | - | 0.01 |
| Grade 3 | 0.24 | 0.38 | 0.16 | 0.28 |
| Grade 4 | 0.58 | 0.47 | 0.27 | 0.37 |
| Grade 5 | 0.05 | 0.02 | 0.18 | 0.03 |
| Grade 6 | 0.00 | 0.00 | 0.00 | 0.00 |
| Grade 7 | - | 0.00 | - | 0.00 |
| Grade 8 | - | 0.00 | - | 0.00 |
| All grades | 0.18 | 0.05 | 0.26 | **0.13** |

(a) Accuracy

| | School | Books | Media | All types |
|---|---|---|---|---|
| Grade 1 | 3.50 | 3.81 | - | 3.80 |
| Grade 2 | 2.18 | 1.65 | - | 1.66 |
| Grade 3 | 0.89 | 0.63 | 0.94 | 0.80 |
| Grade 4 | 0.44 | 0.54 | 0.75 | 0.65 |
| Grade 5 | 1.46 | 1.37 | 0.94 | 1.39 |
| Grade 6 | 2.55 | 2.68 | 2.36 | 2.58 |
| Grade 7 | - | 3.48 | - | 3.48 |
| Grade 8 | - | 4.79 | - | 4.79 |
| All grades | 1.48 | 1.94 | 0.77 | **1.59** |

(b) MAE

| | School | Books | Media | All types |
|---|---|---|---|---|
| Grade 1 | 3.54 | 3.89 | - | 3.89 |
| Grade 2 | 2.31 | 1.73 | - | 1.75 |
| Grade 3 | 1.09 | 0.81 | 1.07 | 0.99 |
| Grade 4 | 0.69 | 0.76 | 0.89 | 0.83 |
| Grade 5 | 1.59 | 1.46 | 1.09 | 1.50 |
| Grade 6 | 2.63 | 2.75 | 2.41 | 2.65 |
| Grade 7 | - | 3.53 | - | 3.53 |
| Grade 8 | - | 4.81 | - | 4.81 |
| All grades | 1.79 | 2.25 | 0.91 | **1.94** |

(c) RMSE

Table 11: Metrics across grades and types of texts on predictions by the CLIB with a frequency word list containing stop words (_stop).

| | School | Books | Media | All types |
|---|---|---|---|---|
| Grade 1 | 0.00 | 0.54 | - | 0.53 |
| Grade 2 | 0.29 | 0.19 | - | 0.19 |
| Grade 3 | 0.31 | 0.34 | 0.26 | 0.31 |
| Grade 4 | 0.25 | 0.31 | 0.23 | 0.24 |
| Grade 5 | 0.16 | 0.10 | 0.04 | 0.12 |
| Grade 6 | 0.35 | 0.23 | 0.00 | 0.32 |
| Grade 7 | - | 0.12 | - | 0.12 |
| Grade 8 | - | 0.36 | - | 0.36 |
| All grades | 0.26 | 0.16 | 0.23 | **0.20** |

(a) Accuracy

| | School | Books | Media | All types |
|---|---|---|---|---|
| Grade 1 | 1.25 | 0.69 | - | 0.70 |
| Grade 2 | 0.95 | 1.46 | - | 1.45 |
| Grade 3 | 1.01 | 0.91 | 0.95 | 0.96 |
| Grade 4 | 1.14 | 0.96 | 1.20 | 1.17 |
| Grade 5 | 1.28 | 1.68 | 2.44 | 1.56 |
| Grade 6 | 1.53 | 1.95 | 2.91 | 1.62 |
| Grade 7 | - | 3.28 | - | 3.28 |
| Grade 8 | - | 2.54 | - | 2.54 |
| All grades | 1.29 | 1.91 | 1.21 | **1.61** |

(b) MAE

| | School | Books | Media | All types |
|---|---|---|---|---|
| Grade 1 | 1.32 | 1.13 | - | 1.13 |
| Grade 2 | 1.28 | 1.80 | - | 1.79 |
| Grade 3 | 1.34 | 1.25 | 1.20 | 1.28 |
| Grade 4 | 1.41 | 1.23 | 1.44 | 1.42 |
| Grade 5 | 1.57 | 1.93 | 2.65 | 1.83 |
| Grade 6 | 2.07 | 2.38 | 3.07 | 2.15 |
| Grade 7 | - | 3.64 | - | 3.64 |
| Grade 8 | - | 3.25 | - | 3.25 |
| All grades | 1.67 | 2.37 | 1.46 | **2.04** |

(c) RMSE

Table 12: Metrics across grades and types of texts on predictions by the CILT with a frequency word list not containing stop words.

| | School | Books | Media | All types |
|---|---|---|---|---|
| Grade 1 | 0.00 | 0.44 | - | 0.43 |
| Grade 2 | 0.22 | 0.08 | - | 0.08 |
| Grade 3 | 0.24 | 0.25 | 0.30 | 0.25 |
| Grade 4 | 0.28 | 0.31 | 0.14 | 0.19 |
| Grade 5 | 0.18 | 0.18 | 0.06 | 0.18 |
| Grade 6 | 0.47 | 0.32 | 0.00 | 0.43 |
| Grade 7 | - | 0.21 | - | 0.21 |
| Grade 8 | - | 0.50 | - | 0.50 |
| All grades | 0.29 | 0.18 | 0.15 | **0.21** |

(a) Accuracy

| | School | Books | Media | All types |
|---|---|---|---|---|
| Grade 1 | 1.50 | 1.02 | - | 1.03 |
| Grade 2 | 1.27 | 1.96 | - | 1.95 |
| Grade 3 | 1.27 | 1.21 | 0.92 | 1.19 |
| Grade 4 | 1.09 | 1.01 | 1.48 | 1.34 |
| Grade 5 | 1.07 | 1.27 | 2.13 | 1.22 |
| Grade 6 | 1.11 | 1.48 | 2.55 | 1.20 |
| Grade 7 | - | 2.71 | - | 2.71 |
| Grade 8 | - | 1.85 | - | 1.85 |
| All grades | 1.11 | 1.75 | 1.45 | **1.52** |

(b) MAE

| | School | Books | Media | All types |
|---|---|---|---|---|
| Grade 1 | 1.73 | 1.50 | - | 1.51 |
| Grade 2 | 1.60 | 2.25 | - | 2.24 |
| Grade 3 | 1.61 | 1.55 | 1.21 | 1.53 |
| Grade 4 | 1.36 | 1.29 | 1.65 | 1.55 |
| Grade 5 | 1.30 | 1.55 | 2.37 | 1.48 |
| Grade 6 | 1.68 | 1.94 | 2.73 | 1.74 |
| Grade 7 | - | 3.18 | - | 3.18 |
| Grade 8 | - | 2.68 | - | 2.68 |
| All grades | 1.48 | 2.19 | 1.64 | **1.92** |

(c) RMSE

Table 13: Metrics across grades and types of texts on predictions by the CILT with a frequency word list containing stop words (_stop).