Universiteit Leiden

Validating LiNT readability formula on various text categories

Anneroos van Diermen (s3907112)

Supervisors:
Wessel Kraaij & Natalia Amat Lefort

Bachelor Thesis– Data Science and Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

July 16, 2026

Abstract

Measuring the accessibility of written text is an important task in the domains of education and formal communication. Readability describes the ease with which a written text can be understood. Well-known metrics include the Flesch reading ease formula and CEFR, but reliable and valid measurement is not an easy task. Many formula-based metrics use too simplistic predictors, such as sentence length, which aren't proven to causally relate to readability. CEFR and other similar metrics are originally meant to examine reader skill, not text difficulty, and tools to determine such scores can give widely different scores. For this reason, the University of Utrecht created the LiNT tool. This tool uses a readability formula with the textual features word frequency, content words per clause, concrete words, and maximum syntactic dependency length.

This thesis examines the external validity of the LiNT tool by testing whether the LiNT algorithmic predictions match real-world human reading difficulty for text genres beyond those on which the LiNT tool was created.

To measure this real-world reading difficulty, we performed a cloze study with secondary school students with 15 texts across 3 different text genres, namely Fiction (specifically novels), Terms & Conditions, and texts from the experiments used in the original LiNT study.

These cloze tests were "semantically graded", so we also ran an experiment to test the subjectivity with inter-rater agreement metrics, such as Cohen's Kappa and Krippendorff's Alpha. To compute these scores, we had 15 independent raters score a total of 45 cloze gaps, 3 per text. The findings of the cloze scores compared with the LiNT analysis show that LiNT successfully ranks relative difficulty between text genres, but differs significantly from actually obtained cloze scores per text. There was a non-significant medium correlation between the two scores. LiNT proved more accurate for the formal, legal texts of the Terms & Conditions than for the more creative texts in the other two categories. Inter-rater reliability metrics were below standard thresholds, which shows that semantic grading of cloze texts is a difficult task with significant subjective elements.

These results show that LiNT can be a valuable tool for comparisons, but does not give a perfect estimation of reading difficulty. Furthermore, measuring human reading comprehension remains difficult from the aspect of validity as well as reliability. It requires a significant time investment of subjects and human raters, which is a notable obstacle in the creation and testing of readability metrics.

Acknowledgements

This thesis is written for the Leiden Institute of Advanced Computer Science at Leiden University as part of the bachelor's degree in Data Science and Artificial Intelligence. I would like to thank the following people for helping me create this thesis:

Wessel Kraaij and Natalia Amat Lefort, for supervising this project, providing feedback and guidance.

Karen Jansze and Frederique van Es, for their help in running the experiment during their classes and collecting the data.

Antal van de Bosch, for creating a LiNT account for us to access the LiNT tool.

Henk Pander Maat, for providing me access to the original texts used in their study and additional information.

My family, for supporting me during the process and helping transfer physically written data to a digital format for analysis.

And all the other people who helped to bring this project to fruition, by participating in the experiment, responding to the survey, or providing tools for analysis.

Contents

1 Introduction

Readability plays an important role in the transfer of knowledge and information. Texts that are too difficult for the intended readers will not be able to effectively communicate information to them. As such, the extent to which a text is readable is fundamentally a concern of accessibility. This is often discussed when a text is important to the general public, such as government communication [Corsius, 2023].

Due to the importance of readability, there have been many approaches to quantify, measure, or predict it. This has led to the creation of various tools and formulae that assess some features of the text, and compute an expected readability for the text.

Especially with the digitization of information, reading skill, behaviour, and attention have changed drastically. A report from the OECD [OECD, 2021] discusses the implications of this shift to accessing information primarily through digital platforms. It specifically mentioned how the skills necessary to comprehend information have shifted to distinguish more effectively between fact and opinion, and how reading through a digital environment affects various reading skills. This report shows both positive and negative effects of digital access, such as the positive correlation between access to a computer at home and performance for distinguishing between fact and opinion, and less awareness of their own performance, which means a significantly larger difference between their performance and their own estimation of their performance, on assessments based on reading from screens. This has a number of implications for knowledge and learning, which makes it much more important for information to be accessible to as many people as possible, without excluding groups such as non-native speakers, people with learning disabilities such as dyslexia, and people with lower literacy levels.

This means that it is necessary to explore how effective measures of readability are at addressing such problems. Which metrics adequately predict readability, and do the changes that these metrics inspire actually reduce the difficulty?

1.1 The situation

Several metrics are commonly used to determine the readability of a text. Some examples are the Flesch reading-ease formula [Flesch, 1948] and its adaptations of the Flesch-Kincaid scale [Kincaid, 1975] and Flesch-Douma scale [Douma, 1960], but also the common European language learning standard of CEFR [Counsil of Europe, 2001]. Formulaic approaches like the Flesch formulae often use a few specific textual features to determine difficulty, such as the number of words per sentence or the number of letters or syllables per word.

However, these methods have a number of flaws in their application. As such, the University of Utrecht set out to create a readability tool that would fix such issues: LiNT [Pander Maat, 2023]. LiNT is also based on a formula of textual features, though the features it uses are more in-depth. This is possible due to developments in text analysis, specifically the tool TScan. The features used are word frequency, content words per clause, concrete nouns and maximum syntactic dependency length.

However, since this is such a new method, it is still uncertain how generalizable it is [Pander Maat, 2023]. Their tool was created based on school books and educational texts. This is a fairly limited selection. As such, this research will examine the generalizability through examining its performance on different text categories it has not yet been tested or trained on, which leads to the following

research question:

1.2 Research question

How accurate is the LiNT tool in predicting the readability of untested text categories?

1.3 Thesis overview

In this thesis, our objective is to test the external validity of the LiNT tool. This refers to the ability of a measure outside of the specific circumstances on which it was based or trained. This is different from internal validity, which examines the internal procedure for its ability to isolate causal factors, and reliability, which examines the consistency of outcomes.

This thesis will be structured as follows: This chapter contains the introduction; In section 2 we will discuss general background about readability and the relevant history of readability research; In section 3 we will focus in on LiNT: its origins, development and applications; In section 4 we will discuss the details of the methods used in this paper, including text collection, analysis with LiNT, creation of the cloze tests, a practical experiment, and testing the inter-rater agreement; In section 5 we will show the results of the experiments we did; In section 6 we will analyze our findings and attempt to draw conclusions from the results; In section 7 we will discuss potential limitations of this study; In section 8 we will propose further avenues of research.

2 Background

The term readability describes how easy or difficult it is to read and properly understand written work. A more readable text is easier for a reader to understand. Readability is determined by both the content of the text, and the structure and presentation. These two factors are called conceptual difficulty and stylistic difficulty, respectively. Furthermore, there are two main purposes for measuring readability: readability prediction and readability improvement. Readability prediction merely wants to assess the difficulty of a text, possibly for selection or categorization purposes, but makes no changes to the text. Readability improvement wishes to assess specifically how the structure of the text can be changed to make it more readable. [Kleijn, 2018]

2.1 Comprehension measurement methods

In order to test how easily a text can be understood, we first need to be able to test how much someone has understood a text. There are many ways to measure reader comprehension, so here we discuss several of these methods.

2.1.1 Self-assessment

The simplest way to measure reading comprehension is through self-reported understanding. This merely requires respondents to describe their experienced difficulty. However, it has been found that there is very little correlation between self-reported comprehension and achieved scores on other comprehension measures, such as comprehension questions. Specifically, it was noted that people often considerably overestimated their comprehension of a text, to the point that it was dubbed the "illusion of knowing" [Glenberg, 1982]. Because of this discrepancy, self-assessment is generally not considered to be a good measure of reading comprehension.

2.1.2 Comprehension questions

The second, and generally most common, way to measure reader comprehension is with specific comprehension questions. This can refer to any kind of questions asked after a person has read the text that attempt to extract the understanding of the content of the text. In particular, this is often the method used in many standardized tests. This method is considered to be a fairly reliable way of testing reader comprehension, but there are a number of problems in implementation. [Kleijn, 2018]

One practical issue is the fact that the questions must be tailored to the specific text being tested. Because of this, it takes a lot of time and effort to create the questions for each text. For multiple-choice questions, the answer options need to be carefully crafted so that the other options are close enough to be reasonable, while distinct enough that there is still only one correct answer. However, for open question, responses need to be carefully analyzed for intent, and can include a significant degree of subjectivity in scoring those answers. Especially on experiments that intend to examine a large number of texts, the amount of time and resources to set up the experiment grows exponentially.

Another issue is the variety of possible questions and question types. Over this variety of questions, not all implementations are of high quality, and different types of questions can require different levels of abstraction. The precise content and phrasing of a comprehension question has a

significant effect on the difficulty, which means that the results of comprehension questions are determined not only by the readability of the examined text, but also by the details of the questions themselves. [Luo, 2022]

Such reading comprehension questions are questionable in independent comprehension assessment, but if they are used as a dependent variable in the creation of a readability formula, they can significantly propagate biases in the models [Luo, 2022]. As such, it isn't an optimal method for our experiment.

2.1.3 Recall and think-aloud procedures

To circumvent the influence of comprehension questions on the difficulty, test makers can make use of recall or think-aloud procedures. In a free-recall procedure, the respondent is asked to read the text and afterwards describes everything they remember about the text. This shows both what information was remembered and how this is structured in memory. In a think-aloud procedure, the respondent is asked to speak their thought aloud while reading a text. This method places less focus on the recall of information, and more on the processes that lead to understanding. Both of these methods contain the same major flaw that scoring such tests is very labour-intensive and subjective. [Kleijn, 2018]

2.1.4 Ordering, sorting and mental model tasks

Another category of comprehension measures takes the form of various ordering, sorting, and mental modelling tasks. These tasks are grouped together due to their focus on testing understanding of relations in a text, such as causal or temporal relations between events. An understanding of such relations is essential for a coherent situational model, which is the deepest level of understanding in the construction-integration model of text comprehension. [van Silfhout, 2014]

Practically, these tasks will ask the respondent to order or sort events, facts or other parts of content through the lens of some coherence relation. For example, a commonly used method is a time-ordering task, which asks respondents to put events in a text in the temporal order in which they happened.

However, some issues with this comprehension measure is that multiple categorizations may be correct, and some categorizations may be partially correct. Furthermore, such methods cannot be applied to any text, since for example a time ordering task would require multiple events to happen in a specific order within the text, so it isn't widely applicable. [Kleijn, 2018]

2.1.5 Cloze test

The term cloze test "encompasses any procedure that omits portions of a text or discourse and asks readers or listeners to resupply the missing elements" [John W. Oller, 1994]. In the case of a text, this usually means that words are removed from the text and replaced by a blank. There are many variations in the specifics of how a cloze test is set up. These specifics can be explained according to the following 11 characteristics: [Kleijn, 2019]

1. Deletion strategy
2. Deletion ratio

3. Deletion distance
4. Number of gaps (per 300 words)
5. Number of starting points
6. Excluded text segments
7. Excluded words
8. Pre-cloze or post-cloze testing
9. Open or closed answer format
10. Marking of deletion
11. Scoring

The first of these characteristics is perhaps the most important. Deletion strategy describes the strategy for choosing words to be deleted. There are three deletion strategies: Mechanical, Rational and Random. Mechanical deletion is the most commonly used method. With this strategy, every nth word is removed. With rational deletion, the words are chosen to be removed based on some constraints on features of the word, such as grammatical class. For example, one could remove only verbs. Random deletion includes any method where some randomized method is used to select the words to be removed. [John W. Oller, 1994]

The number of starting points is mostly relevant for any cloze test that uses a (partly) mechanical deletion strategy. When making a mechanical deletion, you can get different cloze texts based on which word you start counting from, which is your starting point. If you decide to remove every 5th word for example, there are 5 different cloze texts you can make based on your starting point. These different cloze texts include all different removed words and can greatly impact the results, so many experiments with cloze tests will test multiple cloze versions of the same text. This is what the number of starting points refers to. [Kleijn, 2019]

The answer formatting of a cloze test is also quite relevant. In an open answer format, the examinee has to produce an answer themselves, while a closed answer format gives some options for the examinee to choose from. Furthermore, a blank in a text is usually marked as a single fixed length, but the blank could also be marked to show the length of the removed word, which could give an examinee additional support. [Kleijn, 2019]

Deletion ratio and number of gaps are very closely related, since they both describe the amount of words that are to be removed from the text. Deletion distance refers to the amount of words between two removed words, as well as whether this is fixed or varied. Excluded text segments and words describe which parts of the text or types of words are excluded and will never be removed, such as the title or proper names. Finally, pre- or post cloze testing refers to whether the examinee was shown the full text prior to the cloze test.

2.2 Readability metrics

All of these comprehension measurement methods have in common that they require respondents. This takes a lot of time, effort, and resources. This makes it pretty unhelpful for some of the most common use cases of readability. For this reason, there have been several attempts to derive some metric that can automatically calculate the readability of a text.

2.2.1 Flesch

One notable metric here is the Flesch reading ease formula [Flesch, 1948]. This formula takes the amount of words per sentence and the amount of syllables per word as inputs. It returns a grade on a scale of 0 to 100, where 0 is the most difficult and 100 is the easiest, and also categorized these scores into reading grades. These describe which grade of school a person would need to have completed before being able to understand the text. The Flesch reading ease formula was later adapted into the Flesch-Kincaid scale [Kincaid, 1975] for English, and the Flesch-Douma scale [Douma, 1960] for the Dutch language. Both of these metrics still only use the same two variables. This is often considered to be such a simplification that it fails to account for the variety of factors that determine the readability of a text.

2.2.2 CEFR

Another common way to scale readability can be found internationally in CEFR (Common European Framework of Reference for Languages), which classifies texts into A1,A2,B1,B2,C1 and C2. While this measure is easily available through online CEFR rating tools, CEFR was not created for the purpose of readability. CEFR was created to measure a person's proficiency in a non-native language. As such, it provides general concepts for rating the language skills of a person. [Counsil of Europe, 2001]

Furthermore, CEFR is much less formally designed. There is not some clear formula that inputs some variables and returns a difficulty score. There are just general requirements and guidelines for judging language proficiency. Because of this vague definition, there is a significant level of inconsistency between online CEFR rating tools. [Boersma, 2013]

3 LiNT - Leesbaarheidsinstrument voor Nederlandse Teksten

Given the central role that the LiNT tool plays in this research, we dedicate a separate section to LiNT specifically. We discuss the origin of the formula used by this tool, how this formula was developed further into the LiNT tool, and how LiNT has already been used.

3.1 Suzanne Kleijn [Kleijn, 2018]

The formula used by the LiNT tool originated from a dissertation by Suzanne Kleijn, published in 2018. The dissertation is named "Clozing in on readability : How linguistic features affect and predict text comprehension and on-line processing." [Kleijn, 2018]. In this dissertation, she examines the effects of various linguistic features on text comprehension. This was specifically done as part of the "LIN-project" to provide the basis of empirical data for what was then called the "LIN-tool". The "LIN", which stands for "LeesbaarheidsIndex voor het Nederlands" ("Dutch Readability Index"), was later changed to LiNT, which stands for "Leesbaarheidsinstrument voor Nederlandse Teksten" ("Readability instrument for Dutch Texts").

3.1.1 Critique of prior readability research

In the first part of her thesis, she briefly explains some common points of critique in prior readability research. This serves as a preface to the explanation of her own approach, which attempts to subvert or overcome the issues. A central concept in this section is the distinction between readability prediction and readability improvement. Readability prediction merely assesses the difficulty of a text, usually for categorization or sorting purposes, while readability improvement actually attempts to improve the readability of a text. Specifically, a common condition for text improvement is to adapt the stylistic features of the text, while maintaining the content. Noticeably, readability prediction uses all available features of a text, while readability improvement only uses stylistic features, since it builds on the assumption that the content of the text is unchangeable. The problem that results from this distinction is that many predictors used in readability formulae have been designed for readability prediction, which does not ensure that any improvements based on these formulae would actually increase the readability of the text. [Kleijn, 2018]

Another point of criticism she discusses is the lack of causal relevance of commonly used predictors. She specifically mentions the two predictors used by the Flesch formulae: sentence length and word length. She describes these parameters as "shallow predictors", which are either not causally related to readability, or simply correlate to some other characteristic which does predict difficulty. She mentions that the correlation between sentence length and readability is likely not a direct causal relation, but that sentence length simply correlates to sentence structure, where longer sentences may often have more complicated structures, and that that is the actual causal relationship.

Furthermore, she discusses how research often neglects to include reader characteristics and how statistical measures tend to artificially inflate the explained observed variance. Both of these issues tend to result from limitations of time and effort. Testing the readability of a text on the target audience is necessary to properly account for reader characteristics, but this is an extensive and difficult task, so a lot of research will take expert judgments or graded corpora as proxy. Simplifying

the features of a text to a single metric or value leads to missing a lot of necessary context, but actually analyzing a text in every smaller part would greatly increase the amount of work and difficulty of the analysis. [Kleijn, 2018]

3.1.2 Approach

In the next section, she discusses the approach she will take in her experiments, specifically referring to how she will avoid the pitfalls discussed in the previous section. The analysis of the effect of linguistic features begins with reliably determining the linguistic features of a text. For this purpose, she uses a tool called T-Scan. This is an analysis tool that automatically extracts more than 400 linguistic features from texts.

There are four ways in which she distinguishes her approach from prior methods. The first of these is that she does not limit herself to one aspect of readability. A lot of research focuses exclusively on reader comprehension, but she will also measure processing ease. This is done because a text is only considered readable if it can be understood with a "reasonable amount of effort" [Kleijn, 2018]. In order to measure processing ease, she uses eye movement registration.

The second difference is a focus not only of what makes a text difficult, but specifically on what makes a text difficult for which reader. For this purpose, she includes both low-proficiency and high-proficiency readers in her experimentation. As such, standardized tests are included to approximate their reading ability, as well as having each respondent do multiple texts to better understand the effect of reader characteristics.

The third difference is in the examination of the causality of linguistic features. Although causality is not by definition necessary for readability prediction, the purpose of the LIN-project is to develop an understandable tool to assess and adapt texts, so she will conduct a controlled experiment to measure causality. In this experiment, they used two versions of each text, one with altered linguistic features. There were three linguistic features they altered, with each text being assigned to one of the three features. The three features examined in this way were lexical complexity, which refers to the difficulty of words, syntactic complexity, which refers to the difficulty of syntax and grammar structures, and coherence marking, which refers to textual devices to convey logical structure and connections between sentences and clauses. Through only changing the structure of the text, but not the conveyed content, the structural difficulty is changed, while the conceptual difficulty was maintained. If there is a significant difference in comprehension between the two versions of a text, this would imply a causal relationship between the altered linguistic feature and readability. [Kleijn, 2018]

Finally, the results are not averaged over texts and readers. The various sources of data, both in the extraction from T-Scan and the two validation measures give localized data, which they analyze using multilevel analysis techniques.

3.1.3 Method

In the selection of which texts to include in the experiment, they also used a stratified random sampling method. To divide the text into groups, they used a three-dimensional matrix based on tertile scores of linguistic features and selecting one text from each cell (plus three more randomly selected texts to round to 30 texts). The three linguistic features used for this step are referential coherence, concreteness, and personal style [Kleijn, 2018]. Referential coherence refers to words

and clauses that refer to a specific entity discussed earlier or later in the text. Concreteness refers to how tangible or conceptual the words used are. By using this selection method, they can analyze the correlative effect of those three linguistic features, on top of the other three features already analyzed, which allows her to cleanly analyze six different linguistic features.

Characteristics	Standard cloze	HyTeC-cloze
Deletion strategy	Mechanical	Mechanical-Rational
Deletion ratio	20%	10%
Deletion distance	Fixed: every 5th word	Varied: at least 1 word in between
Number of gaps (per 300 words)	60	30
Number of starting points	3 to 5	2
Excluded text segments	Title and first sentence	Title and first sentence
Excluded words	Proper names and numbers	Locally predictable words and guess words
Pre-cloze or Post-cloze testing	Pre-cloze	Pre-cloze
Open or closed answer format	Open	Open
Marking of deletion	Fixed length marking	Fixed length marking
Scoring	Exact + spelling errors	Semantic + spelling errors

Table 1: comparison of standard and HyTeC cloze procedure (adapted from [Kleijn, 2018] p.23)

For their experiment, they developed their own cloze method: HyTeC. HyTeC stands for Hybrid Text Comprehension. This method was created to address some common critiques of cloze tests and to better suit the purpose of measuring text comprehension specifically.

There are several important differences between a standard cloze as described in section 2 and the hybrid method they applied. The most important feature they changed is in the deletion strategy. Rather than choosing one of the three methods, they created a hybrid version of rational and mechanical selection. By merging these strategies, they take advantage of the strengths of both methods, while mitigating some of their individual weak points. Rational deletion allows for the experimenter to limit the types of words that are removed, which can make it a better measure for comprehension, and also allows for direct comparison between different experimental versions of the same text, but a completely rational deletion is sensitive to biases from the experimenter. To combat this bias sensitivity, the final selection is done with mechanical deletion, so that the test will be a more accurate representation of the full text. [Kleijn, 2018]

In practice, this means that they first extract candidates for deletion based on some rational method, in this case they selected specifically content words, while excluding names, numbers and field-specific terms, and then select which of the candidates to remove mechanically. To do this, they calculate the ratio of candidates to all words, select 10% of the total amount of words to select one in every n. For example, if the candidates are 40% of the text, every 4th word from candidates is removed. [Kleijn, 2019]

Another notable difference is the amount of words removed. In a standard cloze, 20% of all words are removed from a text, while the HyTeC method only removes roughly 10%. This is done because the rational selection gives much fewer options for removal, and in their experimentation with different versions of the same text, they also couldn't remove any of their own changes, since this would limit the experimental comparison between the two versions. Furthermore, if too many words are removed, there can be a floor effect with little to no observable variance.

Finally, there is a major difference in how the cloze tests are scored. In a standard cloze test, only the exact answer is given points (usually also accepting minor differences such as spelling mistakes), but to measure the understanding of a reader, that is quite strict. As such, they scored the responses semantically instead. This means that synonyms or other worlds that are similar in meaning to the precise answer should also be counted as correct. To limit judgement bias, each answer was graded by two independent judges, with disagreements settled by a third judge. They state that "All judges received a short training to familiarize them with the scoring procedure.".[Kleijn, 2018]

3.1.4 Results

They collected responses from 2749 students across 120 texts. These students were enrolled in various levels of secondary education, namely pre-vocational education ("vmbo-bb", "vmbo-kb" and "vmbo-gt"), general secondary education ("havo"), and pre-university education ("vwo"). On this data, they performed linear regression with the cloze score as dependent variable, and the linguistic features of the text as the independent variables. In the end, there were 5 linguistic features that proved to be successful in unilevel regression: word frequency, content words per clause, adjectival past participles, proportion of concrete nouns, and maximum syntactic dependency length (SDL). In the multilevel regression analysis, it was the same 5 features that proved to be the best predictors.

In their analysis, they also found that reader characteristics such as education level, grade and reading ability accounted for 34% of the variance in their data. The model they formed based on the 4 best predictors explained an additional 23% of the variance in the data, compared to Flesch-Douma and CLIB, which both only managed to explain 19% of the variance. [Kleijn, 2018]

Those four best predictors that formed the basis of the LiNT formula are:

1. Word frequency. This describes how commonly used the words in the text are. It has a positive relation with the cloze scores, which implies that a text with words that are used more frequently generally results in a higher cloze score.
2. Content words per clause. This predictor refers to the average amount of content words in a clause. It is negatively related to cloze scores, so a text with more content words per clause generally leads to a lower cloze score.
3. Concrete nouns. This describes the amount of nouns that describe something concrete, such as observable objects, as opposed to something abstract, such as concepts or theories. It is

positively related to cloze scores, so a text with more concrete nouns tends to lead to higher cloze scores.

4. Maximum syntactic dependency length (SDL). This refers to the longest distance in words between two words that are grammatically related. It has a negative relation to cloze scores.

All of these predictors together form the following LiNT formula:

$$\begin{aligned} ClozeScore = 3.204 + 15.845 * WordFrequency + 13.096 * ConcreteNouns \\ -1.331 * SDL - 3,829 * ContentWords \end{aligned} \tag{1}$$

3.2 Development of the tool

After the data was collected and analyzed, it was further developed into a usable tool. Two important steps in this process are validating the formula, and furthering the interpretability of the formula. Afterwards, we will discuss the resulting tool. [Pander Maat, 2023]

3.2.1 Validation

To examine the generalizability of the LiNT formula outside of the training data, it would be best to test on new independent data, but due to the lack of this data, they explored further with the data they already had (specifically from Suzanne Kleijn, 2018 [Kleijn, 2018]). This was done in two parts: examining the two text genres in the experiment separately, and cross-validation.

There are two 'genres' in the dataset, namely schoolbooks and educational texts. For each genre, they developed a new formula using the LiNT predictors. This showed that SDL is a weaker predictor for the schoolbooks, while concrete nouns are a weaker predictor for educational texts. Furthermore, the established LiNT formula performs better for schoolbooks than educational texts (81% explained variance for schoolbooks as opposed to 63% for educational texts). [Pander Maat, 2023]

For cross-validation, they performed a k-fold cross-validation with 4 folds. The data was randomly split into a 90 text training set and 30 text training set 4 separate times, such that every text was in the test set exactly once. Then, a formula was created with the four LiNT variables on the training set, and evaluated on the test set. Each of the four validation rounds produced a robust formula, with the minimum R of a test set being 0.795. The performance of the four validation formulae were compared with the performance of the LiNT formula on the same subsets of the data. This showed that the variations were equal or very slightly better on their own training set than LiNT (differences range between 0 and 0.004 for R), while being slightly worse on their smaller test set than LiNT (differences range between 0.001 and 0.035 for R). Overall, the LiNT formula had less variation in its performance over the various subsets of data. [Pander Maat, 2023]

They also discuss the causality of the four predictors. Since adaptations to a text are not reliably effective to reduce difficulty without a causal relationship between the changed variable and readability, they examine whether these four variables have a previously proven causal relationship with readability. For word frequency and SDL, several different studies have found causal effects [Pander Maat, 2022] [Norman, 1992]. The ratio of concrete nouns has some correlational and some experimental evidence, through an experiment that is less reliable due to comparing different texts instead of different experimental versions of the same text, which does not account for content difficulty or other factors [Sadoski, 2000] and through the more indirect evidence of adding examples

increasing readability [Beishuizen, 2003], since adding examples usually increases the concreteness score of a text. Content words per clause only has correlational evidence, though it is often used as a proxy for proposition density. [Pander Maat, 2023]

Finally, they scored adaptations of the same text that were simplified with the LiNT tool. These adaptations keep the content difficulty the same, while only editing the structure of a text, and are intentionally made to be a more easily understandable version of the text. This analysis show that the various simpler versions of texts get significantly lower LiNT scores, and huge differences in the three of the four LiNT predictors, excluding word frequency. [Pander Maat, 2023]

3.2.2 Interpretability

They develop the interpretability of the LiNT formula for the tool in three particular ways.

The first way is through reworking the results of the LiNT formula, which predict the expected cloze score, to a LiNT score, which describes the difficulty of the text. Since the resulting variable is inverse (cloze scores are higher with readability, while difficulty is lower), they calculate the LiNT score as $LiNTscore = 1 - ClozeScore$.

The second way is through splitting up the 0-100 score range into 4 separate difficulty levels. These levels are based on how many Dutch readers could realistically get a cloze score of higher than 50%. This changes the LiNT score from a number without context to an easily understood measure of difficulty. The first difficulty level is taken so that only the portion of Dutch people that is functionally illiterate are unable to pass the 50% cloze score requirement. They estimated this population to be roughly 15%, based on some prior estimations from studies and the Court of Audit (in Dutch: "Algemene Rekenkamer) that ranged from 12%-14% and taken a slightly higher value to account for some omissions in prior data (such as citizens aged 65+) and future predictions (such as a higher percentage of students being at risk to become part of the illiterate population). So, the LiNT score where 15% of people fall off is the boundary. The second level is taken to be double that, so that it reaches until 30% of people fail. The third and fourth levels are then chosen to each increase the percentage of 'dropouts' by 25%, so that they go until 55% and 80% of people respectively. These levels are then called "Eenvoudig" (LiNT score 0-33), "Acceptabel" (LiNT score 34-45), "Moeilijk" (LiNT score 46-60), and "Ontoegankelijk" (LiNT score 61-100), or "Simple", "Acceptable", "Difficult" and "Inaccessible" in english. [Pander Maat, 2023]

The third way in which the interpretation of LiNT scores is made more intuitive is through comparison of common text genres on the LiNT scale. These genres span a range of informal and formal texts aimed at adults. Specifically, the genres tested were in order of average LiNT score: travel blogs, tabloids, vmbo-book, news articles, medical advice, opinion pieces, vwo-book, election programs, court rulings, and statutes. The easiest genres fall just below the level 1/2 border on average, and the most difficult genres fall just above the level 3/4 border on average. [Pander Maat, 2023]

3.2.3 LiNT tool

Finally, after these modifications, the LiNT tool was integrated into a website. It is not publicly available, but it can be freely used by anyone who works at a government organization after contact with the University of Utrecht. On this website, a user can make a project with up to 100 texts to analyze at once.

The website uses the natural language processing tool T-Scan to analyze the text. This returns many linguistic text features. The main feature shown is the LiNT score, which shows the score and difficulty level based on the LiNT formula. Aside from that, the texts can be analyzed in detail with specific features which can be marked in the text, the texts can be compared generally to a wide variety of text genres in a graph, and the values for several linguistic features can be shown in a bar graph, with the largest and smallest of the text genres for comparison. A screenshot of the website is shown in Figure 1.

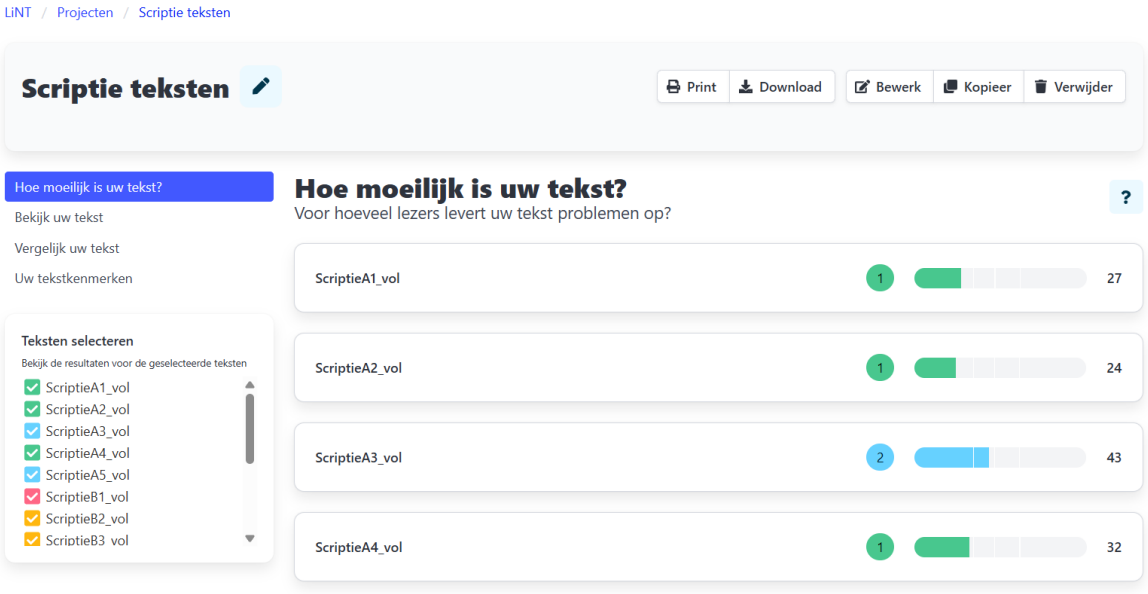

Figure 1: Screenshot of the LiNT website, the navigation for the four pages of results is shown on the left side

4 Method

In order to test the accuracy of the LiNT formula, we will compare the difficulty scores that LiNT assigns to a text, to the actual difficulty. To get a measure for the actual difficulty, we will replicate the experiment done by Suzanne Kleijn, so that the results will be comparable.

4.1 Text collection and modification

We will be testing on two text categories that LiNT has not been developed or tested on. Different text categories have significantly different writing styles and conventions. We believe that it would be valuable to examine whether LiNT remains accurate given these differences, or whether these might require adaptations to the formula, or an entirely new method. This would help establish how generalizable LiNT is.

The two text categories we chose to test with are fiction and Terms & Conditions.

We chose to test fiction due to its variation and focus in writing styles, the prevalence of descriptive language and wide range of difficulty in word choice. As such, the way a fictional text reads will be very different from many of the text categories that LiNT has examined. Due to the incredible variety of genres, intended readers and difficulty in fiction, we specifically decided to focus on books that could be read by secondary students for their literature list. This guarantees that all the books are written in Dutch, specifically not translated, and are assumed to be readable to the secondary school students we will be testing.

We chose to test Terms & Conditions due to the conflicting nature of being a formal and legal document on the one hand, and yet intended to be read by an average person who likely has little to no previous knowledge of the subject matter or of legal document on the other. The prevalence of Terms & Conditions and their legal validity should imply that the average person should be able to understand them, yet an article in NRC[Özdemir, 2025] found that the reading level of these texts were often way too difficult.

Aside from the two new text categories, we will also include a selection of texts that were included in the original experiment as a control group. These texts originally came from school books and other educational texts. We will call these texts "LiNT texts" for simplicity.

For each text category, we first selected 10 texts. For fiction, we collected the texts from the public library website, which contained example PDFs for some e-books. For Terms & Conditions we collected the texts from public websites. For the original texts used for LiNT, we contacted the authors of the study designing LiNT, who provided us with the original texts. Each category got a simple label:

- Category A - Fiction
- Category B - Terms & Conditions
- Category C - LiNT texts

Given the constraints on time and resources for the practical experiment, we further narrowed the selection down to 5 texts for each category, which led to 15 texts in total. This selection was done by assigning each text a number, and selecting with a random number generator.

Finally, all texts were shortened, so that the length of each text was between 300 and 400 words. For the original texts from LiNT, this was not necessary since they were already at this length.

Where possible, the texts were taken from the beginning, so no necessary prior context would be missed, and cut of at a split in a paragraph, so no necessary later context would be missed.

4.2 LiNT analysis

Each text was examined with the online LiNT tool. The title of the text was excluded from the analysis. This tool gives a wide variety of information about the text, which is divided into four categories:

1. "Hoe moeilijk is uw tekst?" ("How difficult is your text?"). This page shows the LiNT score calculated for each text, and classifies them according to the four difficulty levels scaled by LiNT.
2. "Bekijk uw tekst" ("Examine your text"). This page show you the content of the text, with the option to mark several linguistic features: word frequency, concrete versus abstract nouns, "tangconstructies" (which is a specific grammatical construction, where different parts of a grammatical element are split to surround another grammatical element, forming a metaphorical pair of pliers "tang" around it), sentences with many clauses, sentences with many descriptors, and references to people. For word frequency, it shows words with a low frequency, words with an extremely low frequency, and words that were skipped in this analysis. Notably, words can be skipped in this analysis if they believe that it is incorrectly judged as uncommon based on a mismatch with the corpus. The corpus they use for word frequency contains subtitles for english television programs and was made in 2010. As such, it misrepresents some common household words, words that are much more common in Dutch culture, such as "cycling", and words that gained a lot of popularity after 2010, such as technology terms like "app".
3. "Vergelijk uw tekst" ("Compare your text"). This page shows all of the texts plotted on a graph with word frequency and amount of descriptors per clause as axes. This graph also plots several text genres for comparison.
4. "Uw tekstkenmerken" ("Your text characteristics"). This page shows bar graphs for a number of different text characteristics. Every graph contains the text genre from the previous section with the lowest score for that feature, and the text genre with the highest, for comparison. The characteristics are split into three categories: words, which contains word frequency, abstract nouns, and new words which haven't been used in the 50 words preceding it; sentences, which contains the length of clauses, tang constructions, descriptors, enumerations, amount of clauses per sentence, and sentence length; and personal, which contains references to people, pronouns, directly addressing the reader, and active verbs.

4.3 Cloze text creation

4.3.1 Cloze procedure

For the practical experiment, we first need to create cloze versions of each text. For our cloze procedure, we attempted to match the HyTeC-cloze used by LiNT in many ways, though there are some differences. The overview of the cloze characteristics is given in table 2.

The biggest difference between the HyTeC-cloze and our procedure is in the deletion strategy. They used a mechanical-rational deletion strategy, where the first round of selection is done rationally, and those candidates are selected mechanically in the second round. In our procedure, we matched the first part by selecting the candidates with the same rational procedure, but then we selected randomly from these candidates rather than mechanically. The reason for this is that we didn't have the resources and sample size to test multiple cloze versions for each text, so we determined that a random selection would give every candidate the same chance of being selected, which makes it free from text and experimenter biases. A drawback to this method is that there is no control over the distribution of cloze gaps. This difference also leads to the possibility that gaps can be in direct succession, which might give respondents less immediate context to fill those gaps. This is why the deletion distance for the HyTeC cloze has at least 1 word in between, while ours does not.

Characteristics	Standard cloze	HyTeC-cloze (LiNT)	Our cloze
Deletion strategy	Mechanical	Mechanical-Rational	Random-Rational
Deletion ratio	20%	10%	10%
Deletion distance	Fixed: every 5th word	Varied: at least 1 word in between	Varied
Number of gaps (per 300 words)	60	30	30
Number of starting points	3 to 5	2	1
Excluded text segments	Title and first sentence	Title and first sentence	Title
Excluded words	Proper names and numbers	Locally predictable words and guess words	Locally predictable words and guess words
Pre-cloze or Post-cloze testing	Pre-cloze	Pre-cloze	Pre-cloze
Open or closed answer format	Open	Open	Open
Marking of deletion	Fixed length marking	Fixed length marking	Fixed length marking
Scoring	Exact + spelling errors	Semantic + spelling errors	Semantic + spelling errors

Table 2: comparison of standard and HyTeC cloze [Kleijn, 2019] procedure with our cloze procedure

Another more minor difference is in the choice not to exclude the first sentence from the possibility of having words removed. Both the standard cloze and the HyTeC-cloze do this, with the purpose of allowing relevant context to be guaranteed to stay. We decided not to exclude this, because the length of the first sentence varied wildly, with the longest containing more than 10% of

all words in the text. To exclude such a sentence from selection, the ratio of gaps in the rest of the text would go from 1:10 to 1:9, which is an increase from 10% to 16,7%. Such a difference in ratio would clearly influence the difficulty of the text.

For the scoring method, we maintained the semantic dichotomous (0 = false, 1 = true) scoring based on the global appropriateness criterion, which means that the answer has to fit the entire text, not just the direct sentence, just like the HyTeC-cloze. However, in Kleijn's study each answer was scored by two independent judges from a pool of 16 judges total, with a third judge to be the tie-breaker if there was disagreement. We did not have the time nor resources for such a system, so scoring had to be done by a single individual. In order to examine how reliable this scoring is, we did have a number of people rate a sample of the words to examine how much different people would agree.

4.3.2 Cloze algorithm

To do this, we created an algorithm in python. There were a number of requirements for the cloze texts, since it had to match the specifications of the LiNT experiment. In their experiment, they specifically describe that they only removed content word, but excluded names, numbers and field-specific terms, since these would reasonably never be guessed regardless of how understandable the text is.

In order to select only content words, we have to be able to parse the grammatical type of the words. For this, we used a natural language processing library called Spacy. Using this, we parsed the texts, and selected all nouns, verbs, adjectives and adverbs as possible words to remove, Then, we randomly selected roughly 10% of the total words in the text (so between 30-40) from that list. Each word was removed and replaced by a blank of fixed length.

4.4 Cloze study

In order to compare the LiNT scores to some measure of actual readability, we replicated the experiment done by Suzanne Kleijn, which led to the creation of LiNT [Kleijn, 2018]. By using the same experiment to measure readability, it ensures that any differences between scores are not due to differences in measurement, but reflect real differences between expected and actual scores.

For this experiment, we contacted some teachers at a secondary school to run the cloze experiment in their class. The teachers received a page with instructions on how to run the experiment. The tests were given on paper. Every student received two random cloze texts, with both texts belonging to different categories. The students were asked for their name, grade and education type. The name was used to identify which texts were filled in by the same person, and in the data was replaced by a unique number for each student, so that any identifying information was removed.

4.4.1 Participants

In order to replicate the experiment as closely as possible, we also tried to get respondents from the same demographic. They did their experiment with secondary school students in year 2,3 or 4, from a large variety of education types. As such, we attempted to run our experiment with participants from the same categories. We were partially successful in this. Our experiment had 110 participants, the distribution of which is shown in Table 3. Notably, most of our respondents ended up being from the education type Vwo and school year 3 or 4.

School year Education type	Havo	Vwo	Total
3	0	45	45
4	13	42	55
5	0	10	10
Total	13	97	110

Table 3: Participating students per education type and school year

Every student made 2 texts, which leads to a total of 220 responses. The average amount of responses per text is 14.67 (SD 1.66), with all texts having between 10 and 17 responses.

4.5 Inter-rater agreement on semantic scoring

4.5.1 Agreement form

Due to the subjectivity present in semantic scoring, we did another experiment to measure how much people would agree in the grading. We randomly selected 3 words from every text for a total of 45 words. Then we created a questionnaire where respondents got the sentence, the correct word, and, in some cases, additional context, and then had to choose from all the answers given by students which they would grade as correct.

This method does introduce a difference in grading, since the respondents of this questionnaire could not grade according to the global appropriateness criterion without having the full text, so they only graded for local appropriateness, which means the answer only has to fit in the context of the immediate sentence. This is suboptimal, but it would be unreasonable to ask respondents to read multiple texts in their entirety. This means that the alternative would be to examine one or maybe two texts, but we decided that this would not be representative of scoring all the different texts, with their variations in categories and difficulty.

The questionnaire was answered by 15 individuals, with a wide variety of features. Age ranged from 22 to 79, there were 8 women and 7 men, people had various educational backgrounds, and worked in various different sectors.

The results from this questionnaire will be analyzed with several different reliability metrics, which will be explained below.

4.5.2 Percentage agreement

Percentage agreement is by far the simplest of the reliability measures. This merely takes the amount of times that the two raters agreed (so if both evaluate an item to be in the same category) and divides this by the total amount of items. This measure only works when comparing two raters, not more, since it does not account for possibilities of multiple but not all people agreeing. As such, when we implemented this, we calculated the percentage agreement between every permutation of raters. This means that we calculated the agreement for all combinations of the 16 raters (15 respondents plus our own grading), and took the average score of all $\frac{16*15}{2} = 120$ possible pairs. The division by two eliminates doubles (for example, the combination 1,2 is identical to 2,1). We tested whether including the doubles had any effect, but they did not change any of the metrics in

any way.

However, due to its simplicity, many researchers argue that percentage agreement is not a suitable reflection of reliability. This is because of three reasons. The first reason is that a percentage only has one meaningful anchor for reliability (namely 100%), which means that any score different than 100% is uninterpretable for as far as data reliability is concerned. The second reason is that a high percentage agreement becomes more difficult to achieve as the number of categories increases. The third reason is that it doesn't account for chance agreement. There is always a chance that agreement is not because of true agreement, but rather due to chance. As such, a comparison between measured agreement and total disagreement is less useful when examining reliability than a comparison between measured agreement and the theoretical agreement that would occur if every rater picked randomly. [Krippendorff, 2004]

4.5.3 Cohen's kappa

Cohen's kappa accounts for chance agreement in order to achieve a more accurate view of how significant inter-rater agreement is. As such, Cohen's kappa is calculated as follows:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{2}$$

Where P_o is the observed agreement, which is the same as the percentage agreement, and P_e is the expected agreement, also called the chance agreement. The difficulty is that chance agreement relies on the distribution of rater decisions, which is not generally known [Krippendorff, 2004]. To solve this, methods attempt to create an estimate for the chance agreement by using the distribution of the available rater sample. Different methods will often have different ways to estimate the expected agreement, but in the case of Cohen's kappa, it is calculated as follows:

$$P_e = \frac{1}{N^2} * \sum_k n_{k1} n_{k2} \tag{3}$$

Where N is the total number of observations, and n_{ki} the amount of times rater i predicted category k. [McHugh, 2012]

The issue with this calculation of the chance agreement is that the marginal may or may not actually estimate the amount of chance agreement [McHugh, 2012]. Their method of computing the chance agreement is similar to the χ^2 method, but this is dedicated to testing association, not agreement. As such, this method is not independent of the rater preferences. When voters have different preferences, the kappa will change even if the estimated proportions remain the same. [Krippendorff, 2004]

The interpretation of kappa is as follows: [McHugh, 2012]

- kappa = 0-0.20, level of agreement = none
- kappa = 0.21-0.39, level of agreement = minimal
- kappa = 0.40-0.59, level of agreement = weak
- kappa = 0.60-0.79, level of agreement = moderate
- kappa = 0.80-0.90, level of agreement = strong
- kappa > 0.90, level of agreement = almost perfect

This metric is, much like the percentage agreement, designed to compare the judgement of two raters, not more. For this metric we once again took the average of all combinations.

4.5.4 Krippendorff's alpha

Unlike the previous metrics, Krippendorff's alpha is much more generic and can compare multiple raters and over several different data types. As such, we could calculate the alpha without averaging every combination of raters, which was done with the Fast Krippendorff library[Castro, 2017]. The function is defined as follows:

$$\alpha_\pi = \frac{4n\sum n_{ii} - \sum(n_{i.} + n_{.i})^2}{4n^2 - \sum(n_{i.} + n_{.i})^2} \tag{4}$$

Where $\sum n_{ii}$ is the sum over the diagonals and the term $\sum(n_{i.} + n_{.i})^2$ is the sum of squares of the joint marginal entries by two observers. [Krippendorff, 1970]

5 Results

5.1 LiNT analysis

For the results of the LiNT analysis, we will discuss the main results for each of the four sections displayed by the LiNT tool.

5.1.1 LiNT difficulty

The most important results, which will be used for the comparison, are naturally the LiNT scores and difficulty levels. These are meant to provide a clear overview of the difficulty of each text. The results are shown in table 4.

Text	LiNT score	LiNT level	Text	LiNT score	LiNT level	Text	LiNT score	LiNT level
A1	27	1	B1	62	4	C1	43	2
A2	24	1	B2	52	3	C2	53	3
A3	43	2	B3	56	3	C3	37	2
A4	32	1	B4	61	4	C4	45	2
A5	45	2	B5	50	3	C5	50	3

Table 4: Results of LiNT analysis for each text with the LiNT score (0-100) and difficulty level (1-4)

A notable feature of these scores is the difference in difficulty between the three categories. Category A (Fiction) has scores ranging from 24-45, category B (Terms & Conditions) has scores ranging from 50-62, and category C (LiNT texts) falls between with scores ranging from 37-53.

5.1.2 Text examination

The second section of results displayed by the LiNT tool shows us various features marked within the text. We will not discuss all of these, but there are some things to note.

The first notable exception is that the text B2 appears to have some difficulty properly computing, since most features do not mark anything, or barely anything in the text, while "tangconstructies" marks the entire text. This is likely because this text is constructed as a bullet point list. As such, the grammar is very unlike most regular texts. This seems to be only an issue of visualization, since the text does have reasonable values for all text characteristics shown in the last section. Since the visualization would imply extreme scores with a number of characteristics that would be completely 0, but the actual values for these characteristics are not such extremes, the LiNT score is almost certainly unaffected by this visualization error.

The second note is that the texts all had words excluded from examination of word frequency (possibly except for B2, since we have no insight there). The amount of excluded words ranged from 2 to 20, with most falling between the 10-15 range. Prior to the analysis, we had concerns that the exclusions might disproportionally affect texts in category B, due to the mention that words that had become significantly more common after 2010, such as "app", would be excluded, because the

Terms & Conditions are written for and consistently mention apps and websites. However, these texts did not have significantly more exemptions.

Thirdly, very few texts had any sentences marked with a lot of clause, but text B3 had 3 long sentences marked. These three sentences contained 185 words together, in a text of 340 words total. This means that these three sentences contained over half of the words in the text.

5.1.3 Text comparison

The third page draws a graph to compare the inputted texts with each other and with some established text genres. This graph is shown in Figure 3. Note that in this and other figures taken from the LiNT tools, the texts are labeled with "Scriptie[text id]_vol", and they are coloured based on the LiNT difficulty level. The legend to show the difficulty level associated with each colour is shown in Figure 2.

Figure 2: Legend colouration LiNT difficulty

Figure 3: Comparison graph for all texts and established text genres by LiNT. The x-axis is unknown words, the y-axis is descriptors per sentence.

Some notably features of this graph are in how the categories are placed. Category B consistently contains the most uncommon words, most even falling higher than scientific papers, though there is a wide spread in how uncommon. Category C has very similar scores for word frequency, while it also has a wide variance of descriptors per sentence.

5.1.4 Text characteristics

Finally, the text characteristics contain a lot of interesting implications between text categories, most commonly separating category B from the other two. The specific text characteristics used for the LiNT score are shown in Table 5. This page on the website attempts to increase clarity by having all variables in a section having the same relation to difficulty. In the non-personal variables, they plot each variable where a higher score is considered more difficulty, so the concrete nouns variable is shown as its inverse abstract nouns. Furthermore, they replaced word frequency with another variable called unknown words, and content words per clause with just words per clause. Unknown words is based on word frequency, since it is based on the same frequency dictionary, though this variable is higher when a text includes more uncommon words, as opposed to word frequency. This is not simply an inverse, like abstract versus concrete words, so we included this variable and the words per clause as is.

Genre	Unknown words	Concrete nouns	SDL	Words per clause
Fiction (A)	1,50 (0,23)	0,70 (0,09)	6,18 (2,71)	7,82 (1,25)
Terms & Conditions (B)	2,12 (0,17)	0,18 (0,10)	6,40 (1,71)	9,98 (1,20)
LiNT texts (C)	1,64 (0,04)	0,53 (0,17)	6,60 (1,71)	9,40 (0,96)

Table 5: Mean and standard deviation scores for LiNT predictors per category. Unknown words stands in for word frequency, though word frequency is higher when there are less unknown words. Words per clause stands in for content words per clause.

For word frequency, or more specifically the unknown words, it was already fairly clear in the previous section, but the texts can nearly perfectly be sorted into word frequency based on the category, with A having the most common, then C, and finally B with the most uncommon. The only exception is A5, which has a considerably higher score and falls between B4 and B5.

For abstract versus concrete words, category B has significantly more abstract words than the other two, which are fairly mixed. Text C2 stands out with 75,4% abstract words, which is more than 25% higher than the second-highest score for category C.

The amount of "new words" in the text is consistently less for category B. "New words" is defined by the percentage of content words that were not present in the 50 words preceding it. As such, it refers to the amount of repetition within a text, claiming that texts that often repeat the same key terms are often easier, while texts that discuss a lot of different concepts or refer to those concepts with many different words are often more difficult for readers. This means that texts in category B repeat words more often.

Finally, texts in category B directly address the reader significantly more than other text categories, despite category A containing more references to people. This is logical, since Terms & Conditions are a direct agreement between the reader and the company, so it naturally addresses that reader more directly.

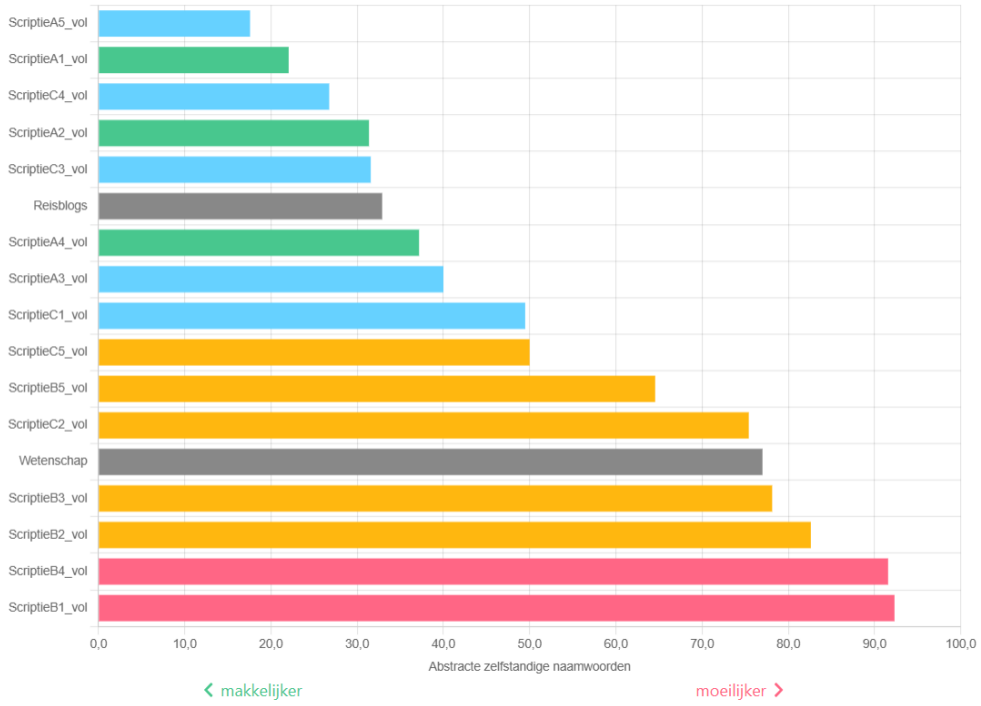

Figure 4: LiNT abstract nouns per text

5.2 Cloze study

For the practical experiment, we received 220 responses over 15 texts. These responses were on paper, so they first had to be transferred to a digital medium for analysis. Then, for each gap in each text, the various answers given by students were examined and scored semantically. The scoring is dichotomous, so an incorrect answer receives zero points and a correct answer receives 1.

28 students did not finish their texts. This is 25,5% of all 110 students. For nearly all of these students, one text was finished, and another was not. In total, 30 text scores were considered unfinished, which is 13,6% of all 220 texts. In some cases (9 students, 8,2%), students missed parts of the text on the back of the paper. We considered that being unable to finish a text could be partially due to difficulty of the text, but since the unfinished tests are spread out over the texts and categories fairly equally, we decided it was more likely a lack of time or reader specific feature. In these cases, we modified their score by lowering their total achievable score. In most cases, this was set to the last gap they filled out. Two of the students that appeared unable to finish showed a strategy of leaving many gaps open and only answering those they were sure about. In that case, we decided that taking the last filled in gap and total achievable score did not represent how many gaps they actually would have been able to fill in. For these two students, we took two thirds of their final filled in gap as an estimate. The amount to take as an estimate had to be larger than the amount of gaps they filled in, but less than the final gap they filled in. The amount of gaps filled in was for both around 15%, and the last gap filled in was for both a little above 75% of the way in. Taking two thirds of the last filled in gap then results in the estimate being around half of the total points, which we concluded was likely a reasonable expectation of what they could have filled in, due to the fact that only 3 non-finishing students filled in less than 50% of the test.

The mean and median cloze scores for each text are shown in Table 6. A box plot showing the distribution of the cloze scores per text is shown in Figure 5.

Text	Cloze score mean (sd)	Median	Text	Cloze score mean (sd)	Median	Text	Cloze score mean (sd)	Median
A1	0.35 (0.12)	0.35	B1	0.31 (0.10)	0.27	C1	0.33 (0.11)	0.34
A2	0.56 (0.10)	0.56	B2	0.21 (0.08)	0.21	C2	0.31 (0.13)	0.31
A3	0.28 (0.08)	0.26	B3	0.47 (0.17)	0.42	C3	0.43 (0.16)	0.46
A4	0.62 (0.15)	0.67	B4	0.29 (0.13)	0.27	C4	0.67 (0.07)	0.66
A5	0.43 (0.16)	0.40	B5	0.41 (0.08)	0.41	C5	0.39 (0.07)	0.39

Table 6: Mean cloze scores, standard deviation and median scores for each text

The average score for a text in category A is 0.45 (SD 0.18), in category B is 0.33 (SD 0.14), and in category C is 0.42 (SD 0.17).

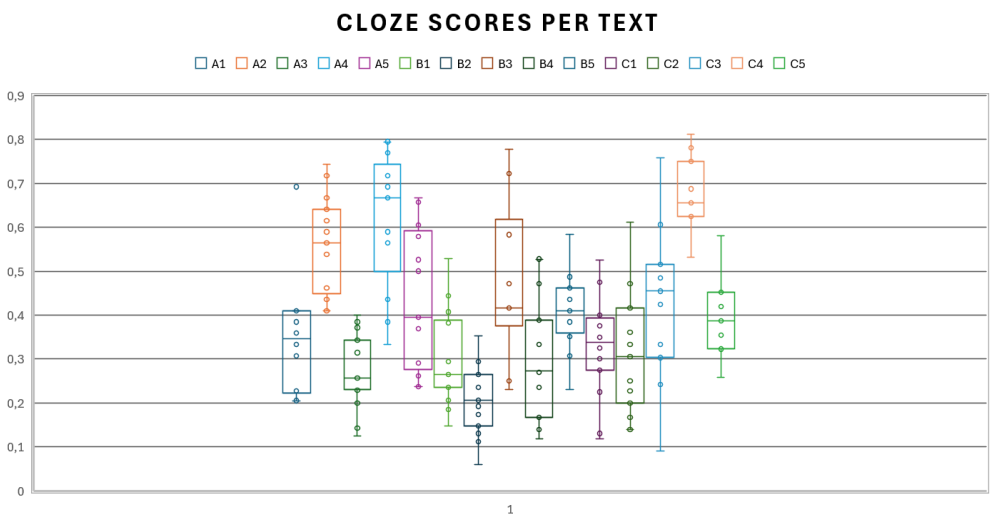

Figure 5: Box plot with the cloze scores per text

There are several notable outliers in Figure 5:

1. For text A1, the highest score is 0.69, which is far higher than the second-highest score of 0.41. This student received 0.48 for C1, which is the second-highest score for that text.
2. For text A4, the lowest score is 0.33, which is a little lower than the second-lowest score of 0.38, but much lower than the median of 0.67. This student received 0.25 for C2, which is somewhat below average.
3. For text B3, the highest score is 0.78, which is a little higher than the second-highest score of 0.72, but much higher than the median of 0.42. This student received 0.47 for C2, which was the second-highest score for that text.

4. For text B3, there were two much lower scores than others: 0.23 and 0.25. The student with 0.23 had the second-highest score of 0.67 on A2. It should be noted that this student is the only one on this list of outliers who did not finish one of their texts, and that the score of 0.23 is after taking their last filled in gap as total. The other had 0.32 for C5, which was the shared second-lowest score of that text.
5. For text C3, the highest score is 0.76, which is far higher than the second-highest score of 0.61. This student received a score of 0.79 for A4, which shares a spot as the highest score for that text. Having the (shared) highest score for both texts, it can be assumed that at least part of the height of this score is due to the student.
6. For text C3, the lowest score is 0.09, which is far lower than the second-lowest score of 0.24. In this text, this student had 3 correct answers. This student received a score of 0.39 for B4, which is above average.

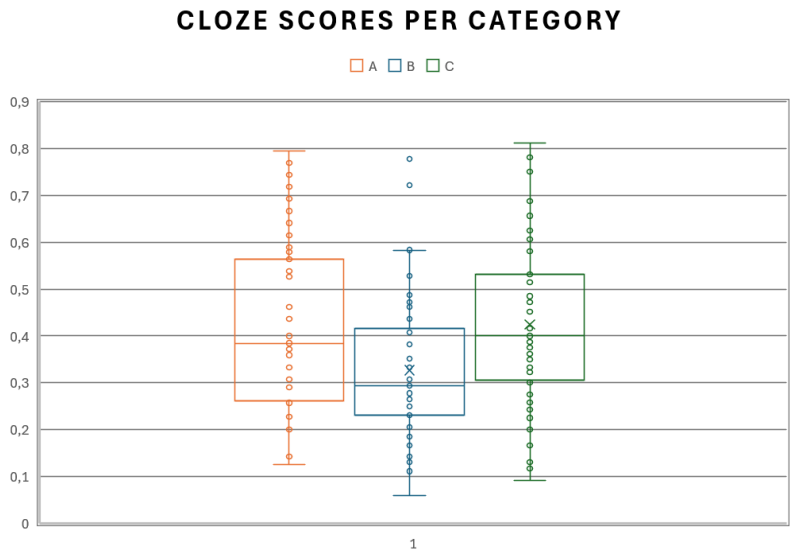

Figure 6: Box plot with the cloze scores per category

5.3 Inter-rater agreement on semantic scoring

5.3.1 Form results

We now present the experimental data for our inter-rater agreement study for the semantic scoring step. The agreement form was filled in by 15 raters. There are some observations from how many people chose specific results, as well as from questions asked during and after the taking of the questionnaire.

The most notable is the effect of grammatical correctness on grading. A number of respondents asked whether and to what extent they should include the grammatical correctness of an answer in the sentence in their consideration. In the responses, these were some of the most controversial answers, often nearing an even split of correct to incorrect. Some examples of this: for one answer, the same answer was given in both the present and past tense. The correct answer for that question

was in the present tense, and the present tense answer received nearly three quarter correct ratings, while the same verb in past tense received slightly less than half. Another case is when the correct word was "zonden" ("sins"), and options included that, "zondes" (which is an alternative correct spelling of sins), and "zonde" (which is the singular form). Notably, only half of people graded "zondes" as correct, and only one third graded "zonde" as correct, despite the identical meaning.

Another observation is that there is quite a wide variety on leniency between different raters. The questionnaire had 45 questions, with a total of 299 possible answers that were rated. The most 'strict' rater only graded 54 of those as correct (note that 27 answers were the given correct answer), while the most 'lenient' rater graded 138 answers as correct. The average amount of answers graded correct is 83,7. For this selection, our grading counted 76 answers as correct.

Finally, one question included a gap in a proverb. This question was the only one where the answers were unanimous, only the exact word in the proverb is graded as correct by all raters, and no raters graded any other answer as correct.

5.3.2 Agreement metrics

The results of the agreement form were measured with several different inter-rater agreement metrics. Both percentage agreement and Cohen's kappa only work for a pair of two raters, so these were run over all combinations of the 16 raters (15 from the form, plus our own ratings), and the average of all combinations was taken. Some raters graded the given correct answer as incorrect, so for each measure, we have a result for the original responses, a result for after all the answers for the correct answer have been graded as correct, and a result where all given correct answers are excluded. These results are shown in Figure 7. The correct answers were graded as correct in 91,9% of cases for all raters and gaps.

There was quite a variation in the agreement measures between different people. The highest scoring combination of raters had a percentage agreement of 0.973 and a kappa of 0.933. These two raters gave the same answer to 291 of the 299 total answer possibilities. The lowest scoring combination had a percentage agreement of 0.719 and a kappa of 0.195. These two raters gave the same answer to 218 of the 299 total answer possibilities.

Metric	Original score	Correct answers modified	Correct answers excluded
Percentage agreement	0.793	0.806	0.788
Cohen's Kappa	0.487	0.531	0.367
Krippendorff's Alpha	0.486	0.527	0.371

Table 7: Results of the agreement score metrics. For Percentage agreement and Cohen's Kappa this is the average value of comparisons between each permutation of two of the sixteen raters.

5.4 Comparison

The LiNT score is the inverse of the expected cloze score, so the comparisons are done on 1 - cloze score. The difference is shown in Table 8. The average difference for texts in category A is 0.213, in category B is 0.111, and in category C is 0.120.

Text	Difference	Text	Difference	Text	Difference
A1	0.382	B1	0.074	C1	0.245
A2	0.199	B2	0.271	C2	0.160
A3	0.291	B3	-0.030	C3	0.204
A4	0.065	B4	0.098	C4	-0.123
A5	0.118	B5	0.091	C5	0.106

Table 8: Difference between inverse of cloze scores and LiNT score per text ((1-cloze score) - LiNT score)

Notably, for most texts, the actual cloze score was lower than LiNT predicted, which means that the text was more difficult. The only negative values, where the cloze score was higher than LiNT predicted, are for texts B3 and C4. The difference for B3 is rather small, so it is not much easier than predicted, but students scored over 0.12 higher on C4 than predicted. It was also the text with the highest score according to the mean and the second-highest score according to the median, as shown in Table 6. This text is about fire safety, so it is possible that students considered this text easier due to the familiarity or concreteness of the subject. However, these characteristics of the subject should be represented by the word frequency and noun concreteness predictors in the LiNT formula, so this would not properly explain the difference between the two scores.

In order to test whether these differences are statistically significant, we will do a one-sample t-test. The t-test can only be applied when the data is normally distributed, so we examined the normality with an online Shapiro-Wilk test calculator. For all texts but one (specifically A1), there was no significant evidence to reject the null hypothesis of normality. A one-sample t-test compares a distribution of observations with a hypothetical expected mean. In this case, we will take the LiNT score as the expected mean, since the LiNT score should be identical to the mean if it was a perfect predictor. The test statistic t is calculated as follows:

$$t = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{5}$$

Where $\overline{x}$ is the sample mean (in this case, 1 - cloze score mean), μ is the expected mean (in this case, LiNT score), σ is the standard error, and n is the sample size. This test statistic is then compared to a critical t value based on the α and the degrees of freedom (=n-1). If the actual t is larger than the critical value, then we reject the null hypothesis, and the difference is statistically significant. The p-values are shown in Table 9.

Text	p-value	Text	p-value	Text	p-value
A1	<0.0001	B1	0.0109	C1	<0.0001
A2	<0.0001	B2	<0.0001	C2	0.0003
A3	<0.0001	B3	0.5863	C3	0.0003
A4	0.1376	B4	0.0095	C4	<0.0001
A5	0.0191	B5	0.0006	C5	<0.0001

Table 9: P values for one-sample t-test for each text

Aside from direct analysis per text, we will also test the correlation between the obtained inverse cloze scores and LiNT scores. We do this with a Pearson correlation test. This test calculates a linear regression on the data. For this, we take X as the inverse cloze score and Y as the LiNT score. This test relies on several assumptions:

1. Continuous variables. This method only works for continuous variables (interval or ratio).
2. Linearity. This method specifically assesses the linear correlation between X and Y, if the relation is non-linear, the results would not accurately reflect the size of the correlation.
3. Normality. This method works on the assumption that the data is distributed normally. This assumption was checked with the Shapiro-Wilk test, which gave a p of 0.5201 (the normality assumption can't be rejected).

For interpretation, Cohen's guidelines [Cohen, 1988] state the following:

Correlation value (r)	Level
$\|r\| < 0.1$	Very small
$0.1 \leq \|r\| < 0.3$	Small
$0.3 \leq \|r\| < 0.5$	Medium
$0.5 \leq \|r\|$	Large

Table 10: Interpretation of correlation value

Results of the Pearson correlation indicated that there is a non significant medium positive relationship between X and Y, ($r(13) = 0.46$, $p = 0.085$). The best fitted line this calculation finds is:

$$Y = 0.401X + 0,214 \tag{6}$$

This line and the data points are shown in Figure 7.

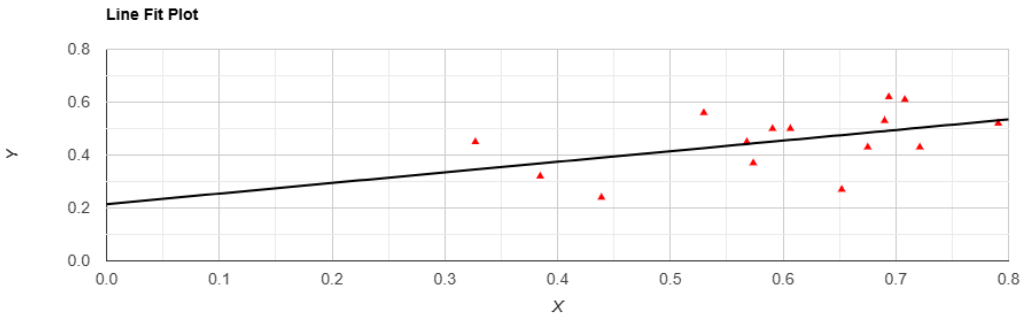

Figure 7: Pearson correlation. X is the inverse cloze score, and Y is the LiNT score.

There are some points that have a large difference between the actual data point and the value expected by the regression. The largest difference is for the point (0.652, 0.27) and the expected value of (0.652, 0.476), which has a difference of over 0.2. This is text A1, which also had the largest absolute difference, as shown in Table 6. This text is from the book "Joe Speedboot", which

is known for having simple words, but a lot of metaphors and other figurative language, most of which are humorous and ironic. This prevalence of short but varied descriptors might make the gap words more difficult for students to guess, which could inflate the difficulty of the cloze test. If this text is removed, the Pearson correlation shows a significant large positive relationship ($r(12) = .574$, $p = .032$).

6 Discussion and Conclusions

The purpose of this thesis was to examine the generalizability of the LiNT tool to assess the readability level of text categories beyond those for which LiNT was designed. This was important to validate due to the limited variety in the original study and the fact that there was a significant difference found even in the two genres they tested. In the process we created an algorithm to create cloze texts, performed a cloze study with secondary school students, and performed an experiment to determine the subjectivity of semantic scoring.

6.1 Inter-rater agreement on semantic scoring

The results of the agreement contained a lot of variation, both per rater and per word. The amount of agreement is very dependent on the people rating, as shown by the fact that a pair of raters could get a kappa of 0.973, which is nearly perfect agreement, while another had only 0.195, which is considered no level of agreement. The mean obtained scores for the agreement metrics are below the threshold for reliability, being interpreted as a minimal to weak level of agreement, but they are noticeably higher than the expected score of 0 if the agreement were only due to chance. Even the two raters with the highest rate of disagreement between them agreed on more than 70% of all cases. This implies that there is a significant effect of subjectivity in the grading, though agreement is measurably more than chance.

The results of the inter-rater agreement experiment also showed evidence of certain patterns of thought, or possible internal rules for the semantic grading. Perhaps these implicit rules could form a basis for more explicit instructions for the graders. With these clearer instructions, differences between rater judgements could likely be lessened significantly.

6.2 LiNT

There is a large and statistically significant difference between the obtained cloze scores and the LiNT score for almost all of the texts. This would either imply that our version of the experiment had some fundamental difference in difficulty, or that the LiNT score does not accurately predict the expected cloze scores of these texts. Almost all of the texts received lower cloze scores than expected, but C4 got a significantly higher cloze score than expected. This contradicts the hypothesis that our cloze texts were simply systematically more difficult than those done by Kleijn, though it is only a single data point, so this does not disprove the possibility.

The correlation between the inverse cloze score and the LiNT score was non-significant and medium positive. The fact that the correlation is not significant might be due to the size of the effect of outliers, such as shown by the large difference when excluding the largest outlier of text A1. Perhaps with a larger sample size, the results could be more definitive and significant. The medium positive relationship shows that for each 1 point increase, the expected value of the LiNT score also increases by a little over 0,4.

Aside from the direct difference per text, there is also a more global overview per category. The LiNT scores per category show that the categories ordered from least to most difficult are A, C, B. This same pattern is shown in the cloze results, where category A received the highest average score, followed by category C, then B. Furthermore, considering the relations between different texts, the LiNT scores and cloze score agree on which text is more difficult in 63,8% of cases. As

such, LiNT does seem to maintain many relationships between texts, so that it can predict which texts would provide more or less difficulty for texts.

Finally, the LiNT score is closer to the inverse cloze score for category B, than for the other two. This might suggest that LiNT is better at predicting more formal texts, like the legal formality of Terms & Conditions. Otherwise, it could be possible that LiNT is more accurate for more difficult texts, since category B was the most difficult by both LiNT and cloze scores. The differences between LiNT and category A were less consistent over the multiple texts, but it had some of the largest absolute differences. This variance could be explained by the wide variation within the fiction texts of category A.

6.3 Final conclusions

Considering these findings, we conclude that LiNT is a tool that properly predicts and identifies several areas of difficulty, but is far from a perfect predictor of difficulty. The tool had a limited experimental pool and its scores are based on cloze scores, which retains some unreliability and variance from the human interpretation in the grading, as shown by the effect of subjectivity in the semantic scoring.

The tool maintains many useful relations and provides significant insight into textual difficulty and factors that influence these. The differences between accuracy per category are fairly inconclusive, largely due to the wide variance within each category. It seems that there is some effect of genre on the prediction accuracy, due to the different distributions of accuracy over the different categories, though what features specifically affect this is unclear.

6.4 Implications

This study has shown some of the limitations of the LiNT tool. While it successfully preserves relative difficulty, the absolute scores differ significantly. This suggests that the formula for the LiNT tool relies on a number of genre-specific features, and the scores may miss some nuances in other, less formal, genres. Therefore, LiNT scores can be valuable as a relative indicator of difficulty, but should not be interpreted as absolute measure of readability, especially when considering diverse text genres.

Secondly, the results of the inter-rater agreement experiment highlight the effect of human subjectivity in semantic grading. This implies that any metric that attempts to use semantically scored human "ground truth" will inherit some of this unreliability, and future research should be careful to mitigate this variance as much as possible, such as through more explicit instructions or through a large enough sample of raters to average over.

Finally, while there are several discrepancies, the medium positive correlation and consistent ordering of categories suggests that LiNT captures the underlying direction of difficulty correctly, and mostly falters in the magnitude of readability. This implies that the LiNT tool has a proper basis for assessing readability, which could likely be significantly improved by applying genre-specific factors or expanding the training data to include more different text categories.

7 Limitations

There are a number of factors that might influence the reliability of our conclusions. One practical limitation is the amount of texts we could examine per category. Especially for a genre as wide as fiction, 5 texts can hardly represent all the variety in the genre.

Secondly, our respondents were almost all in the education type of vwo, which doesn't allow us to examine the texts over the general public, or the differences between various education types. Not only this, but we lacked younger respondents in school year 2, which were present in the experiment done by Suzanne Kleijn. This difference of respondents could introduce further differences.

Thirdly, the specifics of our cloze procedure were close but not identical to the cloze procedure utilized with LiNT. In theory, we might assume this would create generally more or less difficult cloze tests, so the results could simply be modified to account for this. In reality however, some texts had very accurate LiNT scores, and C4 was significantly easier than the LiNT score predicted. This seems unusual if the cloze tests were simply more difficult than those done by LiNT.

Finally, another limitation is present in the subjectivity of the semantic scoring. This means that the results are significantly less reliable, when a significant amount of variance is introduced by the person grading the test. We examined the effect of this with the agreement form, which showed the effects clearly. Depending on who is grading, the amount of answers graded correct could be up to 2.5 times more (based on the lowest and highest rater amounts), which would give a 2.5 times higher score on average. There are some factors that might help to reduce this variation, such as explicit instructions for some of the implicit patterns we found, such as the effect of grammatical incorrectness, or otherwise to provide the full context of the text rather than only the sentence. For the second option, we chose to prioritize getting opinions from more people on more answer possibilities, rather than getting more accurate answers by providing the full context of each text, since time was limited and reading and grading a full text would be very time-intensive. The extent to which this further context would affect rater judgements is rather unclear.

8 Further Research

There are several avenues of further research possible. The focus could be put on the genres, by doing further experiments with more texts for a genre, or to explore different genres. This could provide more certain evidence for the effect of genres on prediction accuracy. It could also reduce the effect of outliers on the correlation analysis, which might help further clarify the actual relationship between the LiNT scores and the cloze scores.

Another interesting possibility would be to create a new regression line for each of the categories with the four LiNT variables, just like they did for their own genres, which would show the effect of each variable for different kinds of texts.

Another possibility is to explore the effects of different cloze procedures on the outcome of the cloze test. Which features impact difficulty the most, and what procedure would be a most accurate representation of readability?

Finally, the semantic scoring could also be examined further. What is the effect of different practical implementations, and how subjective is a subjective scoring procedure? Can differences in rates judgements be mitigated by providing further context to the word gaps or by providing detailed instructions which clarify whether certain factors should influence grading?

References

[Beishuizen, 2003] Beishuizen, Jos ; Asscher, J. . P. F. . E.-M. M. (2003). Presence and place of main ideas and examples in study texts. *British journal of educational psychology*, 73 (3):291–316.

[Boersma, 2013] Boersma, C. J. . N. (2013). Meten is weten? over de waarde van de leesbaarheidsvoorspellingen van drie geautomatiseerde nederlandse meetinstrumenten. *Tijdschrift voor Taalbeheersing*, 35, Issue 1:47 – 62.

[Castro, 2017] Castro, S. (2017). Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. https://github.com/pln-fing-udelar/fast-krippendorff.

[Cohen, 1988] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (2nd ed.)*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

[Corsius, 2023] Corsius, Mischa ; van der Pool, E. . S.-T. W. (2023). Begrijp jij het? mixed-method monitor van overheidsteksten. *Tijdschrift voor Taalbeheersing*, 45, Issue 1:40 – 65.

[Counsil of Europe, 2001] Counsil of Europe, C. (2001). Common european framework of reference for language learning, teaching and assessment.

[Douma, 1960] Douma, I. W. H. (1960). de leesbaarheid van landbouwbladen. *Afdeling sociologie en sociogragie van de lanbouwhogeschool Wageningen*.

[Flesch, 1948] Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, 32 (3):221–233.

[Glenberg, 1982] Glenberg, Arthur M. ; Wilkinson, A. C. . E.-W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory cognition*, 10 (6):597–602.

[John W. Oller, 1994] John W. Oller, J. . J. J. (1994). *Cloze and coherence*. Lewisburg, PA: Bucknell University Press.

[Kincaid, 1975] Kincaid, J P ; Fishburne, J. R. P. . R. R. L. . C. B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

[Kleijn, 2019] Kleijn, S ; Pander Maat, H. . S. T. (2019). Cloze testing for comprehension assessment: The hytec-cloze. *Language Testing*, 36 (4):553–572.

[Kleijn, 2018] Kleijn, S. (2018). Clozing in on readability : How linguistic features affect and predict text comprehension and on-line processing. *Universiteit Utrecht, LOT dissertation series ; 493.*

[Krippendorff, 1970] Krippendorff, K. (1970). Bivariate agreement coefficients for reliability of data. *Sociological methodology*, 2:139–150.

[Krippendorff, 2004] Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30 (3):411–433.

[Luo, 2022] Luo, Dehong ; Gong, J. . L. Y. (2022). The effect of reading comprehension questions on linguistic predictors of readability. *Educational studies*, 48 (5):659–675.

[McHugh, 2012] McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22 (3):276–282.

[Norman, 1992] Norman, S. ; Kemper, S. . K. D. (1992). Adults' reading comprehension: Effects of syntactic complexity and working memory. *Journal of gerontology*, 47 (4):258–265.

[OECD, 2021] OECD (2021). 21st-century readers developing literacy skills in a digital world.

[Pander Maat, 2022] Pander Maat, Henk ; Gravekamp, J. (2022). Kan een tekst te simpel zijn? hoe lager en hoger opgeleiden oordelen over eenvoudige taal. *Tijdschrift voor taalbeheersing*, 44(2):62–90.

[Pander Maat, 2023] Pander Maat, Henk; Kleijn, S. . F. S. (2023). Lint: een leesbaarheidsformule en een leesbaarheidsinstrument. *Tijdschrift voor Taalbeheersing*, 45, Issue 1:2 – 39.

[Sadoski, 2000] Sadoski, Mark ; Goetz, E. T. . R. M. (2000). Engaging texts: Effects of concreteness on comprehensibility, interest, and recall in four text types. *Journal of educational psychology*, 92 (1):85–95.

[van Silfhout, 2014] van Silfhout, Gerdineke ; Evers-Vermeul, J. . S. T. J. (2014). Establishing coherence in schoolbook texts: how connectives and layout affect students' text comprehension. *Dutch journal of applied linguistics*, 3 (1):1–29.

[Özdemir, 2025] Özdemir, M. (11 october 2025). Alle kleine lettertjes lezen? dat zou voor je meest gebruikte apps acht werkdagen kosten. *NRC*.

A Cloze algorithm

This section contains the code used to generate the cloze tests. The text to be edited was called `Test_file.txt` and the code was run separately for each text. The resulting cloze test and removed words are printed and then manually copied to a separate file.

```
import spacy
import random

if __name__ == "__main__":
    #First load in Spacy in Dutch
    nlp = spacy.load("nl_core_news_md")

    #Read the full text
    text = read_text_file('Test_file.txt')

    #Read the word frequency dictionary
    frequency_dict = read_frequency_file("ANW.lemma.txt")

    #Rational selection of possible removals
    doc, content_words = get_possible_content_words(text)

    #Random selection from those possible words
    selected_content_words, selected_contend_words_indexes =
       select_content_words(content_words, frequency_dict)

    #Generate text with the selected words removed and replaced with a
        blank space
    clozentext = generate_clozentext(doc,
       selected_contend_words_indexes)

    #Sort the removed words and return the list
    sorted_content_words = sort_content_words(selected_content_words,
       selected_contend_words_indexes)

    print(clozentext)
    print(sorted_content_words)


def read_text_file(fname):
    #Open, read and store the full text as a variable
    f =  open(fname, 'r', encoding="utf-8")
    text = f.read()
    f.close()
    return text
```

```
def read_frequency_file(fname):
    #Create empty dictionary to input the data from the frequency file
    frequency_dict = {}
    total_observations = 0
    f = open(fname, 'r')
    while True:
        #Read every line one by one (since each line is a single entry
           ), until there are no more lines
        line = f.readline()
        if not line:
            break

        #Split the line into the first half before the tab, which is
           the word, and the other half, which is the number of times
           that word was observed
        modified_line = line.strip('\n').split('\t')
        total_observations += int(modified_line[1])
        frequency_dict[modified_line[0]] = int(modified_line[1])

    #Originally we wanted to store the frequency as a relative rather
       than absolute value, as follows
    #for i in frequency_dict:
        #frequency_dict[i] = frequency_dict[i]/total_observations

    #A lot of the detail in the frequency dictionary becomes
       unnecessary, since we only exclude words that are not in the
       frequency file at all

    return frequency_dict


def get_possible_content_words(text):
    #Process the text with the Spacy model
    doc = nlp(text)

    #Filter for content words: nouns (NOUN), verbs (VERB), and
       adjectives (ADJ)
    content_words = [token.text for token in doc if token.pos_ in ["
       NOUN", "VERB", "ADJ"]]

    return doc, content_words


def select_content_words(content_words, frequency_dict):
```

```
    #Empty lists for storing the selected words and their indexes
    content_words_indexes = []
    content_words_words = []

    #Loop over every word to check whether it's in the frequency
       dictionary
    for index, token in enumerate(doc):
        if token.text in content_words:
            if token.lemma_ in frequency_dict:
                content_words_indexes.append(index)
                content_words_words.append(token.text)

    #Calculate the amount of words to be removed, which is 10% of the
       full text
    #Text_length was inputted manually, since len(doc) included non-
       word characters like punctuation
    amount = int(Text_length / 10)

    #Select that amount of words randomly from the indexes
    random_content_words_indexes = random.sample(content_words_indexes
       , k=amount)
    random_content_words_words = []

    #Use indexes to get the chosen words
    for i in random_content_words_indexes:
        index = content_words_indexes.index(i)
        word = content_words_words[index]
        random_content_words_words.append(word)

    return random_content_words_words, random_content_words_indexes


def generate_clozentext(doc, content_words_indexes):
    # Write the text, but replace the selected words with __________
    clozenstring = ""
    for index, token in enumerate(doc):
        if index in content_words_indexes:
            rep_string = "_" * 10
            clozenstring += rep_string + "-"
        else:
            clozenstring += token.text_with_ws
    return clozenstring


def sort_content_words(content_words, content_words_indexes):
```

```
    #Create a sorted list of the removed words (random selection can
       shuffle them) for grading and oversight
    sorted_content_words = []
    sorted_content_words_indexes = sorted(content_words_indexes)
    for item in sorted_content_words_indexes:
        index = content_words_indexes.index(item)
        word = content_words[index]
        sorted_content_words.append(word)
    return sorted_content_words
```

B Experimental results

This section contains the full results of our experiments

B.1 Cloze study

The results of the cloze study are shown in the table below. A grey cell shows that the student did not fill in this text. An asterisk is placed after a score that had a modification.

Student	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	C1	C2	C3	C4	C5
1					0.37										0.35
2				0.67		0.38									
3					0.29*						0.13*				
4									0.28			0.20*			
5			0.13*										0.52		
6		0.67						0.23*							
7							0.13*					0.14			
8	0.21*													0.53	
9							0.06						0.24		
10									0.17				0.33		
11			0.26					0.42							
12		0.44									0.12*				
13				0.38						0.23					
14		0.64											0.49		
15						0.44*						0.42			
16			0.20			0.26									
17		0.62					0.26								
18	0.33														0.39
19						0.21							0.45		
20	0.31														0.45
21			0.26										0.48		
22		0.72						0.47							
23									0.53					0.81	
24										0.31	0.35				
25					0.66			0.72							
26			0.34								0.53				
27			0.37				0.35								
28				0.77					0.53						
29			0.34											0.66	
30				0.59					0.33						
31										0.41	0.40				
32				0.56								0.31			
33		0.56								0.35*					
34					0.61		0.24								

Student	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	C1	C2	C3	C4	C5
35				0.67								0.61			
36						0.41*								0.75	
37					0.67*				0.24*						
38								0.58							0.45
39								0.42			0.33				
40				0.72										0.69	
41							0.29						0.30		
42					0.50*										0.39
43	0.41									0.58*					
44				0.69		0.26									
45									0.39				0.45		
46								0.78				0.47			
47						0.53								0.69	
48			0.38*							0.44					
49							0.29							0.78	
50									0.47					0.63	
51					0.53						0.30				
52				0.79									0.76		
53		0.74										0.33			
54	0.69										0.48				
55					0.24							0.42			
56	0.38							0.42							
57	0.41					0.39*									
58		0.42							0.14						
59								0.25							0.32
60						0.19*									0.45
61										0.36	0.28				
62	0.33											0.36			
63							0.26								0.39
64		0.59								0.49					
65			0.24*										0.30		
66					0.39				0.24*						
67				0.79			0.19*								
68	0.21										0.38				
69								0.42					0.52		
70							0.15					0.14			
71										0.46					0.39
72	0.36					0.26									
73						0.24					0.40				
74									0.38				0.10		
75			0.26	0.44											
76						0.29					0.38				

Student	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	C1	C2	C3	C4	C5
77	0.38														0.32
78			0.14						0.12*						
79		0.64												0.63	
80										0.38				0.63	
81		0.41													0.32
82		0.54								0.38					
83				0.33								0.25			
84			0.34											0.63	
85		0.41				0.21									
86									0.17					0.66	
87						0.15					0.28				
88					0.58				0.27*						
89					0.29*									0.75	
90							0.11*					0.23*			
91					0.24								0.42		
92			0.31										0.61		
93							0.21					0.31			
94		0.46					0.17*								
95					0.26*									0.63	
96			0.23				0.21								
97			0.26									0.31			
98									0.28						0.58
99										0.41					0.42
100						0.38					0.23				
101			0.4								0.38				
102	0.41									0.44					
103				0.59					0.14*						
104										0.49				0.66	
105										0.41		0.17			
106		0.54				0.26									
107		0.56				0.24									
108		0.59									0.28				
109	0.23*														0.42
110	0.21														0.26

B.2 Inter-rater agreement

The results of the inter-rater agreement experiment is shown in the tables below. Table 12 shows how many answers each rater (with rate 0 referring to our judgement) graded as correct versus incorrect from the total 299 possibilities. Table 13 shows the percentage agreement between pairs of raters and Table 14 shows the kappa for pairs of raters. Both of these use the unmodified complete data, including judgements where the given correct answer was marked as incorrect.

Person	Graded correct	Graded incorrect
Rater 0	76	223
Rater 1	60	239
Rater 2	138	161
Rater 3	83	216
Rater 4	112	187
Rater 5	83	216
Rater 6	54	245
Rater 7	73	226
Rater 8	54	245
Rater 9	112	187
Rater 10	80	219
Rater 11	107	192
Rater 12	111	188
Rater 13	58	241
Rater 14	51	248
Rater 15	80	219

Table 12: The amount of possible words graded correct by each rater of the total 299.

Raters	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	1	0.90	0.77	0.84	0.84	0.84	0.86	0.81	0.77	0.72	0.79	0.74	0.81	0.86	0.86	0.79
1	0.90	1	0.72	0.85	0.81	0.85	0.90	0.82	0.80	0.73	0.80	0.76	0.80	0.87	0.89	0.79
2	0.77	0.72	1	0.79	0.83	0.78	0.71	0.72	0.65	0.77	0.71	0.76	0.81	0.72	0.71	0.71
3	0.84	0.85	0.79	1	0.82	0.97	0.86	0.80	0.74	0.78	0.78	0.83	0.82	0.85	0.87	0.82
4	0.84	0.81	0.83	0.82	1	0.82	0.79	0.78	0.70	0.78	0.77	0.80	0.88	0.81	0.79	0.79
5	0.84	0.85	0.78	0.97	0.82	1	0.85	0.81	0.76	0.80	0.78	0.83	0.83	0.85	0.87	0.82
6	0.86	0.90	0.71	0.86	0.79	0.85	1	0.82	0.77	0.73	0.77	0.79	0.77	0.91	0.91	0.81
7	0.81	0.82	0.72	0.80	0.78	0.81	0.82	1	0.73	0.72	0.79	0.73	0.78	0.84	0.83	0.80
8	0.77	0.80	0.65	0.74	0.70	0.76	0.77	0.73	1	0.68	0.75	0.68	0.70	0.79	0.79	0.72
9	0.72	0.73	0.77	0.78	0.78	0.80	0.73	0.72	0.68	1	0.75	0.80	0.78	0.73	0.74	0.77
10	0.79	0.80	0.71	0.78	0.77	0.78	0.77	0.79	0.75	0.75	1	0.74	0.76	0.80	0.80	0.77
11	0.74	0.76	0.76	0.83	0.80	0.83	0.79	0.73	0.68	0.80	0.74	1	0.79	0.78	0.77	0.78
12	0.81	0.80	0.81	0.82	0.88	0.83	0.77	0.78	0.70	0.78	0.76	0.79	1	0.78	0.79	0.81
13	0.86	0.87	0.72	0.85	0.81	0.85	0.91	0.84	0.79	0.73	0.80	0.78	0.78	1	0.90	0.83
14	0.86	0.89	0.71	0.87	0.79	0.87	0.91	0.83	0.79	0.74	0.80	0.77	0.79	0.90	1	0.83
15	0.79	0.79	0.71	0.82	0.79	0.82	0.81	0.80	0.72	0.77	0.77	0.78	0.81	0.83	0.83	1

Table 13: Percentage agreement of pairs of raters

Raters	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0		0.72	0.53	0.58	0.63	0.58	0.59	0.49	0.34	0.36	0.45	0.40	0.56	0.60	0.59	0.46
1	0.72		0.41	0.59	0.54	0.59	0.68	0.47	0.35	0.35	0.44	0.43	0.52	0.58	0.64	0.41
2	0.53	0.41		0.56	0.65	0.55	0.40	0.42	0.27	0.52	0.40	0.50	0.61	0.41	0.39	0.39
3	0.58	0.59	0.56		0.60	0.93	0.62	0.48	0.26	0.51	0.45	0.62	0.59	0.59	0.64	0.54
4	0.63	0.54	0.65	0.60		0.59	0.51	0.50	0.28	0.53	0.49	0.56	0.74	0.54	0.50	0.52
5	0.58	0.59	0.55	0.93	0.59		0.58	0.50	0.34	0.54	0.45	0.62	0.62	0.59	0.64	0.54
6	0.59	0.68	0.40	0.62	0.51	0.58		0.47	0.23	0.36	0.35	0.49	0.45	0.71	0.69	0.47
7	0.49	0.47	0.42	0.48	0.50	0.50	0.47		0.20	0.36	0.45	0.37	0.49	0.54	0.50	0.48
8	0.34	0.35	0.27	0.26	0.28	0.34	0.23	0.20		0.24	0.28	0.21	0.29	0.30	0.27	0.20
9	0.36	0.35	0.52	0.51	0.53	0.54	0.36	0.36	0.24		0.44	0.58	0.54	0.37	0.37	0.49
10	0.45	0.44	0.40	0.45	0.49	0.45	0.35	0.45	0.28	0.44		0.41	0.46	0.44	0.41	0.42
11	0.40	0.43	0.50	0.62	0.56	0.62	0.49	0.37	0.21	0.58	0.41		0.55	0.47	0.42	0.50
12	0.56	0.52	0.61	0.59	0.74	0.62	0.45	0.49	0.29	0.54	0.46	0.55		0.48	0.50	0.57
13	0.60	0.58	0.41	0.59	0.54	0.59	0.71	0.54	0.30	0.37	0.44	0.47	0.48		0.65	0.53
14	0.59	0.64	0.39	0.64	0.50	0.64	0.69	0.50	0.27	0.37	0.41	0.42	0.50	0.65		0.51
15	0.46	0.41	0.39	0.54	0.52	0.54	0.47	0.48	0.20	0.49	0.42	0.50	0.57	0.53	0.51	

Table 14: Kappa agreement for pairs of raters