Reading Time in the Wild: An Assessment of Readability Predictors Based on Naturally-Observed Reading Times

Sijbren van Vaals, Rik van Noord, Malvina Nissim
Center for Language and Cognition Groningen, University of Groningen
Oude Kijk in 't Jatstraat 26, 9712 EK, Groningen, the Netherlands
{s.j.van.vaals, r.i.k.van.noord, m.nissim}@rug.nl

Abstract

Reading time has surfaced as a viable proxy for readability and comprehension. However, most studies used reading times obtained in controlled experimental settings with eye-tracking or self-paced reading tasks, which differs from uncontrolled, more naturalistic reading behaviour *in the wild*. Through a collaboration with a newspaper, we have access to a dataset of Dutch news articles with corresponding clickstream reading times averaged across thousands of readers. To address the issue, we evaluate how well common proxies for readability and comprehension hold on data from online readers. We first group the proxies in four dimensions and compute the correlation between the proxies and the average reading time per token for each dimension. Then we assess if the proxies can meaningfully predict reading time per token. The results are surprising: we find no meaningful correlation between any proxy and the average reading time per token, nor can any proxy be used for reliable prediction. Additionally, we rerun the prediction on corresponding, automatically simplified texts and surprisingly find increased predicted reading times per token. These results imply that clickstream reading time must be considered with caution as a proxy for readability or comprehension.

Keywords: Comprehension, readability, readability proxies, reading time prediction, real-world data, surprisal, text difficulty, text simplification

1. Introduction and Background

A crucial task in natural language processing (NLP) is to automatically establish the difficulty or readability of a given text (Schwarm and Ostendorf, 2005; Vajjala and Meurers, 2012; De Clercq and Hoste, 2016; Deutsch et al., 2020; Xu et al., 2023). The interest for this is mostly driven by their usefulness in real-life applications. For example, the automatic assessment of the readability of a text can serve as a vital instrument for adapting the difficulty level of texts to different audiences (Martin et al., 2020; Hobo et al., 2023; Säuberli et al., 2024). This is a challenging task and has therefore attracted the attention of a lot of previous research. From a high-level perspective, this research can be divided into two directions.

The first direction is conceptually tied to the idea that the **syntactic profile** of a text can serve as a good indicator for the task. Intuitively, for example, longer and more convoluted sentences should hamper readability. In practice, texts are put through some automatic text analysis tool, from which features can be extracted that should correlate well with difficulty or readability (De Clercq and Hoste, 2016; Brunato et al., 2018; Bulté et al., 2018; Deutsch et al., 2020; Martin et al., 2020; Maat et al., 2023; Seidl and Vandeghinste, 2024).

For example, in developing a readability assessment framework for English and Dutch, De Clercq and Hoste (2016) use lexical, semantic, and syntactic features (such as sentence and phrase length, and clause structure) as main predictors of readability. They indeed find that syntactic features are important contributors to readability. Brunato et al. (2018) also find that syntactic features are consistent predictors, in this case of what humans perceive as difficult text.

The picture, however, is not so crystal clear. In a readability metric for Dutch (LiNT) recently introduced by Maat et al. (2023), two of the four features which result in strong predictors of text comprehension are again syntactic features, namely clause length and dependency length, with the other two being lexical-semantic in nature (word frequency and noun concreteness). However, as shown by Beks van Raaij et al. (2024), this metric often misaligned with human judgments, where participants preferred text simplifications that LiNT deemed less readable. Deutsch et al. (2020) explore whether augmenting neural models with syntactic features improves their performance for readability assessment. They find that adding syntactic features to a transformer (BERT) model only barely improves prediction in a low-resource setting and suggest two explanations: one is that neural model representations already implicitly capture syntactic information; the other is that the results might be affected by problems in the readability labels of the corpora they use.

In the second direction, the concept of **surprisal**, which comes from surprisal theory (Hale, 2001; Levy, 2008), comes into play. It is well known that human surprisal, i.e., how surprised readers are to encounter the next word in a text, correlates well with the difficulty, comprehension, and read-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7209–7224
11–16 May 2026.

Text profiling	Readability metrics	LLMs	Metadata
• Profiling-UD (Brunato et al., 2020) • T-Scan (Pander Maat and Dekker, 2016)	• LiNT (Maat et al., 2023) • Flesch-Douma (Douma, 1960) • Brouwer Index (Brouwer, 1963)	• Surprisal scores	• Topic • Newspaper • Day-of-week

Table 1: Overview of the data representation with the four dimensions we study and the features herein.

ing time of that word (Levy, 2008; Smith and Levy, 2008; Wilcox et al., 2023). Large language models (LLMs) are now commonly used to predict the next word and their uncertainty (or surprisal) about the next word also correlates well with human reading behaviour (Xu et al., 2023; Boeve and Bogaerts, 2025). Recent works by Wilcox et al. (2023) and Xu et al. (2023) show that the relationship between LLM surprisal and reading behaviour holds across languages and is mainly a linear or superlinear effect. However, LLMs are vulnerable to changes in size, domain, language, and intended audience that substantially affect the surprisal estimation (Xu et al., 2023; Boeve and Bogaerts, 2025; Oh and Linzen, 2025). Moreover, Huang et al. (2024) and Wang et al. (2024) stress that LLM surprisal seems to only partially explain processing load in syntactically complex sentences, such as garden-path sentences.

Given the challenges of automatic prediction methods for readability, difficulty, and in general comprehension, previous research has also looked into automatically predicting the **reading time** of a text. This has been shown to correlate well with both difficulty (Singh et al., 2016; Hollenstein et al., 2022) and comprehension (Levy, 2008; Wang et al., 2024) and can therefore serve as a valuable proxy. An effective prediction of reading time can then be used to improve automatic text simplification, assuming that people read faster through simpler texts. Better text simplification leads to improved news accessibility for demographics that find news inaccessible, such as native, lower-educated individuals, or nonnatives trying to learn the language.

However, these studies use data collected in controlled experimental studies and do not look at uncontrolled, more naturalistic reading behavior. At the same time, millions of people read the news daily on devices such as desktop, laptop, and mobile phones, leaving behind a gold mine of information on **real-world news reading**. Thanks to a collaboration with a regional newspaper, we have access to such a dataset, consisting of around 25,000 news articles with corresponding reading times averaged across thousands of readers. This puts us in an excellent position to investigate the research question at the heart of this paper: How well do the findings of readability and comprehension research on reading time hold on data from reading in uncontrolled environments? Providing an answer to this question would enable the refinement of existing, or the development of new metrics and strategies to model readability in a more realistic and, hopefully, usable way.

Findings To address this research question, we systematically assess whether known proxies for readability are indeed good indicators of reading time in data from online readers, using Dutch news articles. Additionally, we test the correlation between LLM surprisal and reading time and combine surprisal with measures of syntactic complexity. We group the features into four feature dimensions and compute the correlation between the features and the average reading time per token for each dimension. Our results are quite surprising, as we find no meaningful correlation between any of the well-established proxies for readability and the reading time. This calls for caution for any future research that wants to use clickstream reading time as a proxy for readability or comprehension.

2. Method

In this section, we outline our dataset, describe how we compute or retrieve all the features associated with each feature dimension, and explain our experiments.[1]

2.1. Dataset Description

Through our collaboration with a regional newspaper belonging to the Mediahuis publishing group, we have access to a dataset of all digital articles in Dutch released by the publisher across its outlets in 2024, which adds up to 24,946 news articles. Given that articles are reused between newspapers, can be republished, or can be updated, this

[1]All code is available at: https://github.com/sijbrenvv/Reading-time-in-the-wild. The final dataset will be available upon request under license and non-disclosure agreement.

		Profiling-UD				T-Scan			
Domain	**Feature**	Mean	SD	Min	Max	Mean	SD	Min	Max
Length	Avg. sentence length (words)	16.5	3.2	6.9	38.3	15.1	3.0	6.0	36.8
	Avg. clause length (words)	9.6	3.5	5.8	105.0	9.5	2.0	5.3	36.8
	Max dependency length	8.9	2.1	3.7	28.5	6.1	1.7	0.0	17.6
	Avg. prepositional chain length	1.1	0.1	0.0	2.0	–	–	–	–
Lexical	Content word ratio	0.5	0.0	0.4	0.9	0.5	0.0	0.4	0.8
	Proper nouns (%)	6.8	5.0	0.0	48.7	7.6	5.3	0.0	65.5
	Broad concreteness (nouns)	–	–	–	–	0.6	0.1	0.1	1.0
	Strict concreteness (nouns)	–	–	–	–	0.3	0.1	0.0	0.9
PoS	Adjectives (%)	6.5	2.0	0.5	14.6	7.4	2.0	0.0	16.8
	Nouns (%)	17.8	3.4	1.4	30.9	19.6	3.3	5.4	33.3
	Pronouns (%)	5.5	3.3	0.0	20.1	10.1	4.1	0.0	25.1
	Personal/possessive pronouns	–	–	–	–	42.5	30.4	0.0	184.2
	Person references density	–	–	–	–	96.8	44.9	0.0	476.2
Syntactic	Avg. arguments verb	3.0	0.4	0.7	4.4	–	–	–	–
	Max tree depth	3.3	0.5	1.4	5.7	–	–	–	–
	Subordinate clauses per sent	–	–	–	–	0.5	0.3	0.0	2.6
	Nominal modifiers per clause	–	–	–	–	1.2	0.6	0.1	7.8

Table 2: Descriptive statistics of linguistic features extracted with Profiling-UD and T-Scan.

dataset contains duplicates and needs cleaning. We ensure that each article has a body of at least 100 words, a reading time above zero (see below), and 25 views or more, where the reader could actually read the article and was not hit by a paywall. Subsequently, we drop duplicates based on the publication date, title, and body.

Metadata and Reading Time Each article comes with a set of metadata features. Two of these are closely related to the readers' engagement in the article: the total number of views (views) and aggregated reading time in seconds (reading time sec). Thus, for each article we take the average reading time by dividing the aggregated time by the total number of views, and then obtain the **average reading time per token**, which we use in our experiments. We take reading time per token rather than per article to account for article length, which plays a major role in the total reading time of a given article. Then, we retain only articles with an average reading time per token of 0.15s or higher to reduce the influence of skimmed articles. This threshold is in line with previous findings, which set the average reading time per token specifically for online news in a range between 0.10s and 0.20s (Mitchell et al., 2016; Lehmann et al., 2017). The average reading time distribution is explained in Appendix A.2. As a result, the final dataset contains 2,929 news articles from Mediahuis Noord (regional newspapers only) that will be used for correlation analysis, reading time prediction, and simplification.

Each article additionally has the following metadata available: publication date, the newspaper in which the article is published, title, and the article topic, which is based on automatic assignment by the publisher using embeddings.

Other Existing Datasets Three important reading time corpora in Dutch are RaCCooNS (Frank and Aumeistere, 2024), MECO (Siegelman et al., 2022), and GECO (Cop et al., 2017). These corpora contain reading times obtained via eye tracking, and each target a different reading level. For the RaCCooNS corpus, the 37 participants read 200 sentences from the SONAR-500 corpus (Oostdijk et al., 2013) with sentences between 5 and 30 tokens, cultivating word-level reading times for 2,783 words. MECO, on the other hand, focuses on the paragraph level. For this corpus, the 45 participants read 12 paragraphs similar to Wikipedia, where paragraphs were between 7 and 12 sentences, and 161 and 213 words long. The corpus contains paragraph-level reading times for 2,231 words in Dutch (Boeve and Bogaerts, 2025). In contrast, the GECO corpus contains document-level reading times for 59,716 words through eye tracking of 19 participants reading half of the *The Mysterious Affair at Styles* novel by Agatha Christie. The average sentence length of the documents is 12 words.

How Is Our Data Different? The main difference between the described corpora and our data is that our data was obtained under uncontrolled conditions and stems from online readers, whereas the reading times in the aforementioned corpora are gathered in a controlled experimental setting. Data from a controlled experimental setting can potentially impede the generalisability of the findings to a real-world setting. Given that the news articles

in our corpus have a minimum length of 100 words and can run up to over 2,000 words, the corpus can be used for both paragraph and document-level analysis, allowing potential comparison to the MECO corpus in future work.

2.2. Reading Time Calculation

The publisher measures reading time as time on page with periodic attention checks (i.e., scrolling, mouse movement, or keyboard hit). In digital reading, there are no guarantees that a reader truly reads or finishes the read. The system therefore measures active attention/reading time as a reasonable estimation. The system measures active attention in blocks of 5 seconds with a bounce filter of 3 seconds, a maximum of one minute for idling, and a session timeout of 15 minutes. Appendix A.1 further explains the calculation process and possible disturbing factors like videos that influence the calculation are discussed in Section 4.

2.3. Representation and Features

2.3.1. Text Profiling

For text profiling, we use a publicly available and a closed access toolkit. One that works on various languages (Profiling-UD, closed) and one that was specifically developed for Dutch (T-Scan, open).

Profiling-UD Profiling-UD (Brunato et al., 2020) is a text analysis tool developed to assist in research on language variation using the Universal Dependencies (UD) framework (Nivre, 2015). The system is useful for retrieving features across multiple linguistic layers and allows for parallel profiling in different languages due to the UD framework. Profiling-UD uses a two-step approach in which it first performs linguistic annotation using a preprocessing pipeline specifically for the UD framework, and then performs linguistic profiling over these annotations.

T-Scan T-Scan (Pander Maat and Dekker, 2016) is a Dutch text difficulty and genre analysis tool that, similar to Profiling-UD, provides features across nine linguistic layers, including sentence complexity, lexical complexity, lexical diversity, relational coherence, and concreteness. T-Scan can be used to get feedback on written texts, in which the system informs about text difficulty, highlighting which aspects of the texts need revision specifically.

Feature Alignment and Selection Since Profiling-UD does not always compute the same set of features for each article (87-130), we ensure that each article has the same set of features. T-Scan always produces the same set of features, covering 472 different features. We remove features that have a constant value or where the system cannot compute a particular feature in more than 10% of instances.

Metric	Mean	SD	Min	Max
Brouwer Index +	0.50	0.08	0.15	0.79
Flesch–Douma +	0.62	0.07	0.27	0.83
LiNT (score 1) −	43.0	9.8	0.0	91.3
LiNT (score 2) −	42.6	9.6	0.0	91.8

Table 3: Descriptive statistics for the readability metrics. A '+' indicates a positive slope - higher score means higher readability - whereas a '−' indicates a negative slope.

Statistics To get a general idea of what the texts look like, we compute basic descriptive linguistic properties using both profilers. From Table 2, we observe that both profilers show relative alignment in length and lexical features, whereas a slight divergence is visible in part-of-speech features. On average, the texts in the dataset have a sentence length of 15 to 16.5 words, in which the average clause length is 9.5 words. The percentage of adjectives and nouns is 6.5 to 7.4 and 17.8 to 19.6, respectively, where roughly half of the nouns are, to some degree, concrete, according to T-Scan.

2.3.2. Dutch Readability Metrics

For determining the correlation to reading time for the readability dimension, we calculate readability metrics specifically developed for Dutch.

Flesch-Douma The readability formula introduced by Douma (1960) is based on Flesch's (Flesch, 1948) reading ease formula, in which the constants are updated to Dutch based on agricultural magazines and a multiplication is added to the average sentence length: $207 - 0.93 * avgsentlen - 77 * avgnumsyl$. For interpretability purposes, we scaled the score between 0 and 1.

Brouwer Index Brouwer Index (Brouwer, 1963) is also based on Flesch's reading ease formula and is similar to the Flesch-Douma formula. In contrast, the constants are based on the readability of Dutch prose rather than agricultural magazines: $195 - 2 * avgsentlen - 67 * avgnumsyl$. We also scaled the Brouwer Index between 0 and 1.

LiNT LiNT (Leesbaarheidsinstrument voor Nederlandse Teksten, Maat et al., 2023) is a Dutch readability formula and tool that focuses specifically on lexical and syntactic complexity. The metric relates comprehension to the following four features: Word frequency, noun concreteness, clause length, and dependency length (Maat et al., 2023).

Model	↓ Mean	SD	↓ Min	↓ Max
EuroLLM-9B	81.5	15.6	44.5	208.4
Fietje-2	64.0	15.7	23.5	210.9
mGPT	77.4	18.2	32.7	225.9
Tweety-7B (Dutch v24a)	66.5	15.2	27.5	223.8
Qwen3-1.7B	114.4	24.2	60.2	267.6
Qwen3-4B	145.7	31.3	71.4	342.5

Table 4: Average surprisal scores across sentences in our dataset, computed with various language models. The lower the better.

The LiNT metric is divided into two formulas, using a regression analysis based on cloze tests by 2700 Dutch high-school students, both using word frequency and dependency length as their foundation. LiNT score 1 focuses more on noun distribution, whereas LiNT score 2 focuses more on noun concreteness and content words.

Statistics The readability metrics are computed following their original formulas, with the help of the Pyphen library[2] (Berendsen, 2013) to calculate the number of syllables. In Table 3, we see that the texts are already relatively readable, with average scores close to 0.50 or 50. For Flesch-Douma and Brouwer Index, this means that most articles use shorter words or sentences. For LiNT, this implies the articles use more frequent words and concrete nouns, and shorter clauses. Nonetheless, there are articles with extremely low readability, as depicted by the maximum scores for LiNT, as well as the minimum scores for Flesch-Douma and Brouwer Index.

2.3.3. Large Language Model Surprisal

We use the Minicons library[3] (Misra, 2022) to compute surprisal scores from different LLMs on the texts. First, we extract all sentences from a given article. Then, we compute the surprisal scores from a given LLM for all sentences and then average these scores. The set of LLMs for the surprisal computation contains LLMs that are either pretrained on Dutch or adapted for it and is similar to the models used by Boeve and Bogaerts (2025): mgpt (1.3b, Shliazhko et al., 2023), Qwen3 (1.7b and 4b, Qwen Team, 2025), Fietje-2 (2.7b, Vanroy, 2024), Tweety-7b-dutch (7b, Remy et al., 2024), and EuroLLM (9b, Martins et al., 2024).

Statistics From Table 4, we observe that the Qwen models show a higher variation in computed surprisal scores, and much higher values. While the other models have average surprisal scores per sentence between 64 and 82, the Qwen models range from 114 to 146. These higher surprisal scores imply that the next token would likely require greater processing effort during comprehension, given its lower contextual predictability. It does make sense that models which have been specialised for Dutch show the lowest surprisal when given articles written in Dutch.

Direct Assessment We also experiment with asking LLMs directly to estimate the reading time and text difficulty of given texts, based on various prompts and examples. However, we found that the models were unable to do this well at all, even the larger closed models such as ChatGPT (GPT-4o). Therefore, we do not report on these experiments in Section 3. The prompts can be found in Appendix B.

2.3.4. Metadata

Of the metadata features attached to the articles (Section 2.1), we use three that might be relevant in predicting reading times. These are the article's topic, the newspaper in which the article is published, and on which day the article was published (day of week).

2.4. Analysis and Prediction

To understand the relevance for reading time of the various features we consider, we run both a correlation analysis and a prediction task.

Correlation Analysis In the correlation analysis, we compute Spearman's ρ (Spearman, 1904) between the features of a dimension and the average reading time per token.

Reading Time Prediction In this step, we train a random forest regressor from the scikit-learn library (Pedregosa et al., 2011) per dimension to predict the average reading time per token, with default parameters, no hyperparameter tuning, and using an 80/20 train-test split. Aside from its performance, we also want to analyse the feature importances in such a model. In this way, we assess the extent to which the features of each dimension can be used as a proxy for reading time. Such a predictor could be used to approximate text difficulty when gold reading times are unavailable. We evaluate the prediction using explainability of the dimensions in reading time variation (R^2), mean squared error (MSE), mean absolute error (MAE), the correlation between the true and predicted reading time per token (ρ), and Shapley values (Lundberg and Lee, 2017), which reveal feature importances on the test set and can be used to assess which features are most informative for the model.

[2] https://github.com/Kozea/Pyphen

[3] https://github.com/kanishkamisra/minicons

2.4.1. Baseline

We include a simple length-based baseline model to put the predictive power of the feature dimension into perspective. The baseline uses the number of tokens and sentences in each article, as computed by Profiling-UD. Since article length is typically a strong predictor of reading time, this baseline allows us to assess the additional predictive power of each feature dimension beyond article length.

2.5. Simplification Experiment

To further understand how reading time relates to text simplification, we run a simple simplification experiment in which we automatically simplify our data using an LLM. We used GPT-5.2-mini with a generic prompt that can be found in Appendix C.1.[4] We rerun our profilers and recompute the readability and LLM surprisal scores to analyse how the profiles of the simplified articles change and to examine how the predicted reading times per token by our predictor is affected under simplification.

In the examination, we compare predicted reading time distributions, inspect how the Shapley values (feature importances) change, and dissect which features drive change in predicted reading time per token. To analyse how the profiles differ, we compute the mean relative change, and the percentage of texts in which the feature has changed. For the mean relative change, feature differences are normalised by each feature's median across original texts. This way, the relative change measure is robust against near-zero values and comparable across features with different scales.

3. Results

First, we analyse the correlation between each dimension and the observed average reading time per token. We then evaluate the extent to which each dimension can be used to predict the reading time per token using a random forest model. Finally, we examine how the reading time prediction changes after simplification and analyse how the profiles of the text change under simplification.

3.1. Correlation Analysis

Table 5 shows the correlations for each dimension with observed reading time. Quite surprisingly, we observe low and negligible correlations between the dimensions and the reading time per token across the board.

Profiling For the profiling analysis, we show the five best correlating features per profiler. None of these features stand out as correlating highly, although there is a low correlation for some. Interestingly, the best features are quite different: for Profiling-UD they are mostly based on length, while for T-Scan they are mostly about transition (or linking) words and types of clauses. The length features correlate negatively with reading time. This might suggest a side-effect of the data: readers in general are more likely to scroll through longer articles, while shorter articles are more likely to be read in full. While some of the best features of T-Scan are more intuitive, e.g., texts with more guidance (clear transition words) might require less time to process, others are not, as abstract nouns (a higher level of abstractness) seem to correlate with lower reading times, but usually require a higher processing load compared with concrete nouns.

Dimension	ρ
Profiling-UD	
Number of tokens	-0.418**
Number of sentences	-0.384**
Number of prepositional chains	-0.347**
Longest dependency link in number of tokens	-0.186**
Distribution of adverbial modifiers	-0.155**
T-Scan	
Transition words of time per clause	-0.059*
Dep. length noun - head relative clause	-0.054*
Density of transition words of time	-0.053*
Adverbial phrase with generic adverb per clause	-0.052*
Adverbial phrase per clause	-0.050*
Readability metrics	
Flesch-Douma	-0.083**
Brouwer Index	-0.064**
LiNT (score2)	0.009
LiNT (score1)	0.006
LLM Surprisal	
Qwen3-4B	0.037*
Qwen3-1.7B	0.030
EuroLLM-9B	0.006
Fietje-2	-0.004
mGPT	0.001
Tweety-7b (Dutch v24a)	0.001
Metadata	η
Newspaper	0.170
Topic	0.095
Day of Week	0.068

Table 5: Spearman ρ correlation values between features of each dimension and the average reading time per token. Eta η for the metadata features. A single star (*) indicates a p-value below 0.049, and a double star (**) indicates a p-value below 0.001.

Readability The readability metrics only have a very slight correlation to reading time, showing no substantial difference between traditional length-based readability metrics and LiNT, which is based on syntactic complexity. This finding corroborates previous work that readability metrics are not always reliable proxies (van Oosten et al., 2010; De Clercq and Hoste, 2016).

[4] GPT-5.2-mini was queried in December 2025.

LLM Surprisal and Metadata It has been well-established that LLM surprisal correlates well with reading time on the word level, also for Dutch (Wilcox et al., 2023; Xu et al., 2023; Boeve and Bogaerts, 2025). However, also for this dimension, in our real-world data, we find near-zero correlation values with no noticeable difference between monolingual or multilingual, adaptation to Dutch, and size. This curiously contradicts Boeve and Bogaerts (2025) by finding no meaningful correlation between LLM surprisal and real-world reading behaviour. Finally, we find that the article's topic, newspaper, and day of publication do not correlate meaningfully with its reading time per token. This suggests the dataset noise does not simply explain the findings and will be discussed in more detail in Section 4.

3.2. Reading Time Prediction

The reading time prediction results, shown in Table 6, reinforce the correlation findings by showing low explainability (R^2) of the dimensions in the reading time per token variation. The low correlations between the true and predicted average reading time per token show that the dimensions are ineffective for predicting the reading time per token. Interestingly, Profiling-UD achieves a moderate Spearman ρ score of 0.46, accompanied by an R^2 score of 0.22. This means that there is some useful information in the combination of features.

Moreover, we find that the surprisal scores of the monolingual Dutch model Fietje-2 and the model adapted to Dutch Tweety-7b (Dutch v24a) were not more informative to the model than the surprisal scores of the multilingual models. This suggests that using a monolingual model or adaptation to the target language does not improve over a multilingual model, which is consistent with the finding of Wilcox et al. (2023) that multilingual and monolingual may be equally viable.

Finally, the length-based baseline also shows a low fit, which indicates that article length alone, although typically a strong predictor of reading time, explains little of the variance in reading time per token in our data. This, curiously, suggests that longer articles do not necessarily have longer reading time per token.

3.3. Simplification

To explore a first use of reading time for text simplification evaluation in natural news reading, we conducted a small simplification experiment, in which we automatically simplify our data and examine how the predicted reading time of our predictor is affected. Details about the simplification experiment can be found in Section 2.5. The distribution plots of the predicted reading time per token for the

Dimension	Metric / Feature	Value
Prof. UD	R^2	0.215
	MSE	0.002
	MAE	0.033
	ρ	0.461**
	Number of tokens	0.011
	Dist. of verbs with 4 arguments	0.003
	Average verb edges	0.002
	Dist. of infinitive verbs	0.002
	Dep. dist. of nominal subjects	0.002
T-Scan	R^2	-0.084
	MSE	0.002
	MAE	0.039
	ρ	0.007
	Proportion of abstract nouns	0.001
	Transition words of time per clause	0.001
	Adverbial phrase per clause	0.001
	Density of punctuation	0.001
	Finite verbs at start of sentence	0.001
Readability	R^2	-0.139
	MSE	0.002
	MAE	0.039
	ρ	0.057
	Flesch-Douma	0.013
	Brouwer Index	0.010
	LiNT (score2)	0.005
	LiNT (score1)	0.004
LLM Surpr.	R^2	-0.034
	MSE	0.002
	MAE	0.036
	ρ	0.161**
	Qwen3-4B	0.007
	mGPT	0.006
	Qwen3-1.7B	0.006
	Tweety-7b (Dutch v24a)	0.006
	Fietje-2	0.005
	EuroLLM-9B	0.004
Metadata	R^2	-0.157
	MSE	0.003
	MAE	0.039
	ρ	0.097*
	Topic	0.007
	Newspaper	0.006
	Day of Week	0.006
Length Base.	R^2	0.038
	MSE	0.002
	MAE	0.034
	ρ	0.354**
	Number of tokens	0.026
	Number of sentences	0.013

Table 6: Model evaluation metrics and feature importances per dimension. Evaluation metrics (R^2, MSE, MAE, ρ) are listed first, followed by the top SHAP features with their mean absolute SHAP values. A single star (*) indicates a p-value below 0.049, and a double star (**) indicates a p-value below 0.001.

Feature dimension	p	rbc
Profiling-UD	$<$0.001	0.77
T-Scan	$<$0.001	0.63
LLM Surprisal	$<$0.001	0.49
Readability	0.421	0.04

Table 7: Results from a Wilcoxon signed-rank test on the data that shows the magnitude of changes in predicted reading time per token after simplifying.

non-simplified and simplified datasets are shown in Figure 3 (Appendix C.2). Surprisingly, we observe **increased** reading time predictions for the simplified versions when looking at the different feature sets per dimension, except for the Readability dimension, which remains the same. If our reading times accurately reflected difficulty and comprehension, this observation would suggest simplified articles are actually harder to read, which should be impossible. This leaves us with the question: why does our feature-based model predict higher reading times for simpler texts?

Are the Differences Significant? To investigate this, we first want to establish that the increase in predicted reading time is significant. We conducted a Wilcoxon signed-rank test W (Wilcoxon, 1945) on the data and compute the effect size using rank-biserial correlation rbc. Table 7 shows the results of the Wilcoxon signed-rank test and we observe large effects for Profiling-UD and T-Scan, a medium effect for LLM Surprisal, and a negligible effect for Readability. This tells us that the increased reading times are consistent across articles and that it is highly unlikely it can be solely attributed to noise.

Are the Texts Simpler? Second, we want to establish that the LLM did in fact simplify the articles. Tables 9 and 10 (Appendix C.3) show the feature changes observed under simplification. They indeed reveal a consistent shift in the linguistic profiles of the texts, such as lexical choice, verbal morphology, and sentence structure, which aligns with established definitions of less linguistically complex texts (Gibson, 1998; Martin et al., 2020; Harbusch and Steinmetz, 2022; Ranjan et al., 2022; Säuberli et al., 2024; Seidl and Vandeghinste, 2024). Across multiple linguistic levels, simplification consistently made the articles shorter, more explicit, and put in more canonical structures. Information is conveyed with fewer subordinate clauses, adpositions and adverbs, using more finite constructions. The simplified texts use the canonical SVO sentence structure increasingly, indicating lower syntactic complexity. Therefore, we are confident that the LLM simplifications reflect actual simplification rather than summarisation or paraphrasing.

Is Our Predictor Accurate Enough? Perhaps our predictive model is not accurate enough to capture the nuanced differences in writing styles between the original and simplified texts. Indeed, as could be seen in Table 6, the predictive models show low predictive power to accurately estimate reading time. However, the models are applied to both versions of each text. Even for a model with limited predictive power, we would still expect to see (small) directional changes, i.e. a decrease in predicted reading times and not a significant increase in predicted reading times, as we see here.

What Features Influence Prediction? For the Profiling-UD dimension, the increased reading time prediction is predominantly driven by the number of tokens. In the feature importance plot in Figure 2a (Appendix C.2), we see that articles with fewer tokens generally have an increased predicted reading time per token. Although seemingly counterintuitive at first, this is understandable from a real-world perspective. In natural news reading, people often do not read the whole article and start skimming at some point, especially when articles are longer. For T-Scan, it is less clear. The top features shift slightly, but there does not seem to be a clear set of features that drive the increased reading time prediction. The LLM Surprisal dimension is harder to interpret, but interestingly shows the greatest shift in both top features and the difference in importances. Qwen1.7B and EuroLLM are now the new best features, while the Dutch models drop in importance. The feature importances plot in Figure 2c (Appendix C.2) shows that only high surprisal scores show a directional influence, where high surprisal scores from bigger and more recent models increase predicted reading time per token.

4. Discussion

In our study, we investigate whether known proxies for readability and comprehension are good indicators of reading time in data from online readers. Surprisingly, we find that all well-established proxies, if at all, only slightly correlate with reading time under uncontrolled conditions. In this section, we discuss possible causes, what this implies for the findings in this field to date, and the questions future research should answer.

Dataset Noise One explanation for our results is that real-world datasets are noisy. People scroll through articles, are distracted by incoming messages, use different screen sizes, or simply lose interest in the story. However, we do not believe this to be a main limiting factor in this study. We filter our original dataset from close to 25,000 articles to around 3,000 to account for any random

noise introduced, by ensuring a minimal number of views and reading time per token. It is nearly impossible to remove all the noise in a real-world dataset, but the noise left in the data on its own cannot completely explain the results we find.

Topic and Newspaper Our dataset averages across news articles from different newspapers and on different topics. Potentially, these features simply overshadow the other features based more on the profile of the articles. However, our data shows no indication that this is the case. In both Table 5 and 6, the metadata predictors also score poorly. To gain more insight into this, we trained a simple linear regression model with unigrams, bigrams, and trigrams to predict the reading time. We found that at least simple topic indicators such as words also do not explain our results. All in all, the topic of an article does explain *some* part of the data, but it cannot be the only explanation.

Similarity of the Articles Newspapers use guidelines and rules on how articles should be written. Perhaps that partially explains why there is no clear relation between text difficulty and reading time. However, newspapers publish more than factual news, including opinion pieces, columns, interviews, and long reads, all with inherently varying levels of text complexity. And more importantly, we actually do find a large spread in scores for all the features tested in this paper, i.e., they are not all clustered around a single value.

Automatic Processing Linguistic information is derived automatically, using standard tools. These tools are not perfect, of course, and might not capture all linguistic nuances needed to accurately assess text complexity or readability. However, the same tools are used in previous work on readability metrics or difficulty assessment, so the gap must lie in using reading time in natural settings.

Disturbed Reading Time Some factors in news articles potentially disturb measured reading time. For example, playing an in-article video is not detected as activity, and screen size is not accounted for, which can lead to capped reading times on large screens. Nevertheless, such factors are rare in the dataset. Links only occur in live blogs, designed to redirect readers, and videos are infrequent and rarely longer than one minute. As a result, the noise introduced is expected to be minimal and should also average out across the dataset.

Implications For us, this paper is mostly a cautionary tale. Reading times of real-world data do not correlate well with common proxies for comprehension and readability. Dataset noise or the general interest of readers in a topic do not seem to be limiting factors per se, but even if they were, we have no reason to believe that our dataset would be different from any other real-world dataset of news articles. Additionally, in our experiments, reading time does not reflect improved readability or comprehension after simplification. This means that any research focused on investigating, for example, text simplification in such real-world scenarios should be careful with using clickstream reading time as a proxy. That is not to say that reading time is meaningless. In fact, we argue that this finding opens up an avenue of new research directions: What actually drives reading time on real-world data? And what keeps readers interested? How can we automatically measure this? And what would this mean for research on text simplification and comprehension?

5. Conclusion

In this paper, we explored how well the findings of readability and comprehension research hold for data from reading in uncontrolled environments, using a dataset of digital news articles with aggregated reading times from thousands of readers. Through a systematic assessment with correlation analysis and prediction experiments, we assessed the extent to which known proxies for readability and comprehension are meaningful indicators of the average reading time per token in online news reading. We also conducted a small simplification experiment, in which we automatically simplify the articles, and analyse how this affected both reading time prediction and the text profiles of the articles.

Our findings show no meaningful correlation between any proxy and the average reading time per token, nor can any proxy be used effectively to approximate reading times in online news reading. In addition, the simplification findings show that reading time is not a trustworthy proxy for changes in the linguistic profile of a simplified article compared to its original version. These results strongly suggest that clickstream reading time must be considered with caution as a proxy for readability or comprehension. Also, they call for more research involving human participants to discover which (linguistic) factors do influence reading time.

From our experiments, we conclude that known proxies for readability and comprehension should be used with caution when relating them to reading time in natural news fruition, and that reading time is not a viable proxy for automatic simplification evaluation in a naturalistic news reading context.

6. Acknowledgments

This work was supported by the PPS grant of the Ministry of Economic Affairs and Climate Policy through CLICKNL, the Top Consortium for Knowledge and Innovation (TKI) in the Creative Industries (TKI2006-CI23018). We would like to acknowledge the support of Mediahuis Noord, in particular Alwin Wubs, for providing us with the data used for this study. We thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Hábrók high performance computing cluster. We would also like to thank the CLIN community for their rich and fruitful feedback at CLIN35, where this ongoing work was presented. Finally, we thank anonymous reviewers for their insightful and constructive feedback.

7. Bibliographical References

Nadine Beks van Raaij, Daan Kolkman, and Ksenia Podoynitsyna. 2024. Clearer governmental communication: Text simplification with ChatGPT evaluated by quantitative and qualitative research. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 152–178, Torino, Italia. ELRA and ICCL.

Luca Benedetto, Gabrielle Gaudeau, Andrew Caines, and Paula Buttery. 2025. Assessing how accurately large language models encode and apply the common european framework of reference for languages. *Computers and Education: Artificial Intelligence*, 8:100353.

Wilbert Berendsen. 2013. Custom Python library to hyphenate text using existing Hunspell hyphenation dictionaries.

Sam Boeve and Louisa Bogaerts. 2025. A systematic evaluation of Dutch large language models' surprisal estimates in sentence, paragraph, and book reading. *Behavior Research Methods*, 57(266).

R.H.M. Brouwer. 1963. Onderzoek naar de leesmoeilijkheden van Nederlands proza. *Pedagogische Studiën*, 40.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-UD: a tool for linguistic profiling of texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7145–7151, Marseille, France. European Language Resources Association.

Dominique Brunato, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? Do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2690–2699, Brussels, Belgium. Association for Computational Linguistics.

Bram Bulté, Leen Sevens, and Vincent Vandeghinste. 2018. Automating lexical simplification in Dutch. *Computational Linguistics in the Netherlands Journal*, 8:24–48.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, 49(2):602–615.

Orphée De Clercq and Véronique Hoste. 2016. All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics*, 42(3):457–490.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

Ir WH Douma. 1960. *De leesbaarheid van landbouwbladen*. Afdelingen voor Sociale Wetenschappen aan de Landbouwhogeschool.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Stefan L Frank and Anna Aumeistere. 2024. An eye-tracking-with-EEG coregistration corpus of narrative sentences. *Language Resources and Evaluation*, 58(2):641–657.

Edward Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68:1–76.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Karin Harbusch and Ina Steinmetz. 2022. A computer-assisted writing tool for an extended variety of leichte sprache (easy-to-read German). *Frontiers in Communication*, Volume 6 - 2021.

Eliza Hobo, Charlotte Pouw, and Lisa Beinborn. 2023. "Geen makkie": Interpretable classification and simplification of Dutch text complexity. In

Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 503–517, Toronto, Canada. Association for Computational Linguistics.

Nora Hollenstein, Itziar Gonzalez-Dios, Lisa Beinborn, and Lena Jäger. 2022. Patterns of text readability in human and predicted eye movements. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, pages 1–15, Taipei, Taiwan. Association for Computational Linguistics.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

Janette Lehmann, Carlos Castillo, Mounia Lalmas, and Ricardo Baeza-Yates. 2017. Story-focused reading in online news and its potential for user engagement. *Journal of the Association for Information Science and Technology*, 68(4):869–883.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Henk Pander Maat, Suzanne Kleijn, and Servaas Frissen. 2023. LiNT: een leesbaarheidsformule en een leesbaarheidsinstrument. *Tijdschrift voor Taalbeheersing*, 45(1):2–39.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. EuroLLM: Multilingual language models for Europe.

Kanishka Misra. 2022. Minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Amy Mitchell, Galen Stocking, and Katerina Eva Matsa. 2016. Long-form reading shows signs of life in our mobile news world. Technical report, Pew Research Center. Report by the Pew Research Center in association with the John S. and James L. Knight Foundation.

Joakim Nivre. 2015. Towards a universal grammar for natural language processing. In *Computational Linguistics and Intelligent Text Processing*, pages 3–16, Cham. Springer International Publishing.

Byung-Doh Oh and Tal Linzen. 2025. To model human linguistic prediction, make LLMs less superhuman.

Nelleke Oostdijk, Martin Reynaert, Véronique Hoste, and Ineke Schuurman. 2013. *The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch*, pages 219–247. Springer Berlin Heidelberg, Berlin, Heidelberg.

H.L.W. Pander Maat and N. Dekker. 2016. Tekstgenres analyseren op lexicale complexiteit met TScan. *Tijdschrift voor taalbeheersing*, 38(3):263–304.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Qwen Team. 2025. Qwen3 technical report.

Sidharth Ranjan, Marten van Schijndel, Sumeet Agarwal, and Rajakrishnan Rajkumar. 2022. Dual mechanism priming effects in Hindi word order. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 936–953, Online only. Association for Computational Linguistics.

François Remy, Pieter Delobelle, Hayastan Avetisyan, Alfiya Khabibullina, Miryam de Lhoneux, and Thomas Demeester. 2024. Trans-tokenization and cross-lingual vocabulary transfers: Language adaptation of LLMs for low-resource NLP.

Andreas Säuberli, Franz Holzknecht, Patrick Haller, Silvana Deilen, Laura Schiffl, Silvia Hansen-Schirra, and Sarah Ebling. 2024. Digital comprehensibility assessment of simplified texts among persons with intellectual disabilities. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Sarah Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 523–530, Ann Arbor, Michigan. Association for Computational Linguistics.

Theresa Seidl and Vincent Vandeghinste. 2024. Controllable sentence simplification in Dutch. *Computational Linguistics in the Netherlands Journal*, 13:31–61.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2023. mGPT: Few-shot learners go multilingual.

Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, et al. 2022. Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (MECO). *Behavior research methods*, 54(6):2843–2863.

Abhinav Deep Singh, Poojan Mehta, Samar Husain, and Rajkumar Rajakrishnan. 2016. Quantifying sentence complexity based on eye-tracking measures. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 202–212, Osaka, Japan. The COLING 2016 Organizing Committee.

Nathaniel J Smith and Roger Levy. 2008. Optimal processing times in reading: A formal model and empirical investigation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 30.

C. Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Montréal, Canada. Association for Computational Linguistics.

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

Bram Vanroy. 2024. Fietje: An open, efficient LLM for Dutch.

Daphne Wang, Mehrnoosh Sadrzadeh, Miloš Stanojević, Wing-Yee Chow, and Richard Breheny. 2024. How can large language models become more human? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 166–176, Bangkok, Thailand. Association for Computational Linguistics.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Weijie Xu, Jason Chon, Tianran Liu, and Richard Futrell. 2023. The linearity of the effect of surprisal on reading times across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15711–15721, Singapore. Association for Computational Linguistics.

A. Reading Time

A.1. Calculation

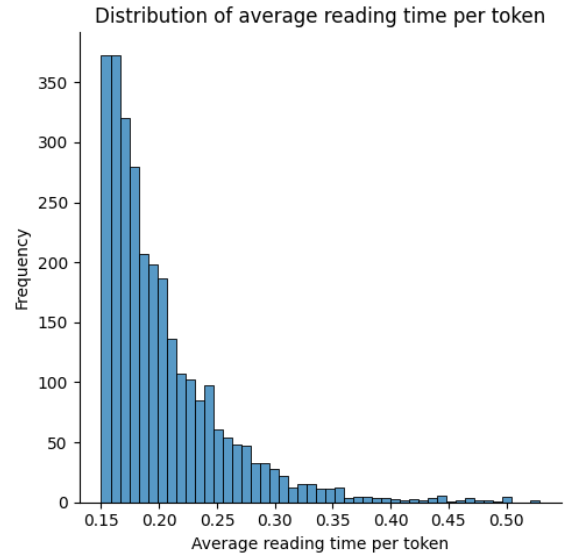
The reading time counter starts after 3 seconds when a reader opens a page to mitigate bouncer influence. The system checks every 5 seconds if a reader is active, and keeps counting when the user shows activity (scrolling, mouse movement, or keyboard hit). When a reader shows no activity (idle), the system counts an extra 60 seconds max before pausing the counter. When a reader leaves the article's page, the counter pauses immediately. The counter resumes when the reader shows activity within 15 minutes, otherwise the session closes. When the article's page is closed, the session closes immediately too.

For example, if a reader spends three minutes reading, pauses for four minutes, and then reads for two more minutes. The recorded reading time is 177 seconds of reading, 60 seconds of idling, and 120 seconds of additional reading, for a total of 357 seconds.

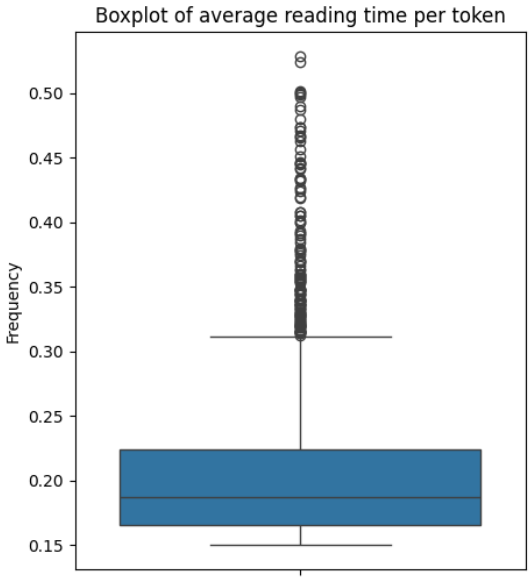
A.2. Distribution

As shown in Figure 1, the average reading time per token distribution is skewed with a right tail, with 50% of the data below 0.20. This is due to our data

cleaning, where we ensure a minimum average of 0.15 to reduce the influence of skimmed data. Table 8 with descriptive statistics of reading time, views, and document length additionally shows the reading time statistics from the GECO corpus for comparison with a reading time corpus from a controlled setting.



(a) Histogram



(b) Boxplot

Figure 1: Average reading time per token distributions by histogram and boxplot.

Feature	Mean	Median	SD	Min	Max
Avg. rt tok (GECO)	0.26	-	0.12	0.10	0.85
Avg. rt tok	0.20	0.19	0.05	0.15	0.53
Views	406	123	1094	25	14022
Nb. of tok	380	297	259	105	2967
Nb. of sen	24	18	19	3	237

Table 8: The descriptive statistics of the average reading time per token, views, number of tokens, and number of sentences. In addition, the reading time per token descriptions from the eye-tracking GECO corpus (Cop et al., 2017).

B. Direct Assessment

In the direct assessment, we explored GPT-4o, Llama-3.3-8b-instruct, and Fietje-2-chat to estimate total reading time and text difficulty as a score between 0 and 1. In this section, we show the different prompts we used to experiment with in zero-shot, one-shot, and few-shot (three examples) settings.

B.1. Reading Time

We started with a generic prompt that explains the task:

```
Your task is to predict the reading
time in seconds of a given text and
to provide an explanation of how you
arrived at the reading time in
seconds. Ignore the length of the
text.
```

Next, we added linguistic information to look at:

```
Your task is to predict the reading
time in seconds of a given text and
to provide an explanation of how you
arrived at the reading time in
seconds. Look at linguistic aspects
of the text, such as sentence
structure and difficult words.
Ignore the length of the text.
```

Finally, we tried role prompting to prime the model:

```
You are an expert linguist and
reading comprehension specialist.
Your task is to predict the reading
time in seconds of a given text and
to provide a clear explanation of
how you arrived at that reading time.
Base your prediction on linguistic
aspects of the text, such as
sentence structure, syntactic
complexity, and vocabulary
difficulty. Ignore the length of the
text when making your prediction.
```

B.2. Text Difficulty

The prompts used for text difficulty estimation follow the same format as the reading time prediction.

We started with the generic prompt:

```
Your task is to predict the
difficulty level of a given text.
Use a value between 0 and 1 and
provide an explanation of how you
arrived at that value. Ignore the
length of the text.
```

Then, we used the additional linguistic information:

```
Your task is to predict the
difficulty level of a given text.
Use a value between 0 and 1 and
provide an explanation of how you
arrived at that value. Look at
linguistic aspects of the text, such
as sentence structure and difficult
words. Ignore the length of the
text.
```

Finally, we tried the role prompting format:

```
You are an expert linguist and
readability analyst. Your task is to
predict the difficulty level of a
given text. Use a value between 0
and 1 and provide a clear
explanation of how you arrived at
that value. Base your assessment on
linguistic aspects of the text, such
as sentence structure, syntactic
complexity, and vocabulary
difficulty. Ignore the length of the
text when making your evaluation.
```

C. Text Simplification

C.1. Prompt

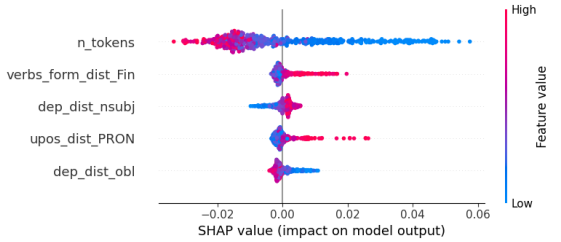
In our simplification experiment, we use the following prompt:

```
I will provide you with a news
article that you must simplify to B1
level. The goal is that people with
limited reading skills can
understand the news article. Return
only the simplified text.
```

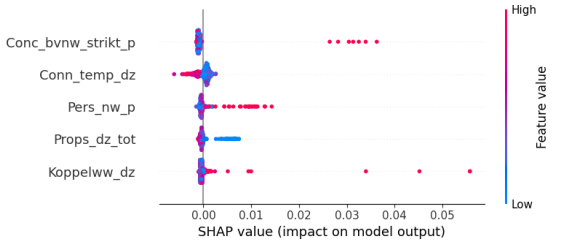
As is known from literature, the prompt conditions text generation and substantially influences the simplification results. We use a generic prompt that prompts for CEFR level B1, while LLMs do not really understand what B1 level entails (Benedetto et al., 2025). When experimenting, we found the LLMs reduce the complexity when prompted for different CEFR levels. We used B1 level, specifically because it is generally recommended as an accessible reading level for people with limited reading skills in Dutch. As the focus of the study was to evaluate reading time change under simplification and not simplification per se, we did not experiment extensively with system prompts and prompt engineering.

C.2. Reading Time Prediction

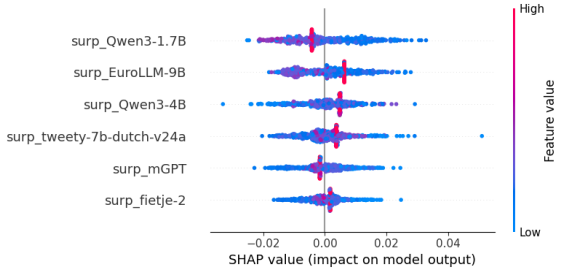
Firstly, this subsection shows feature importances of the top 5 features for Profiling-UD, T-Scan, and LLM Surprisal, as described in Section 3.3. Secondly, it shows overall reading time per token prediction distributions of the original and simplified texts for all dimensions, as described in Section 3.3.



(a) Profiling-UD

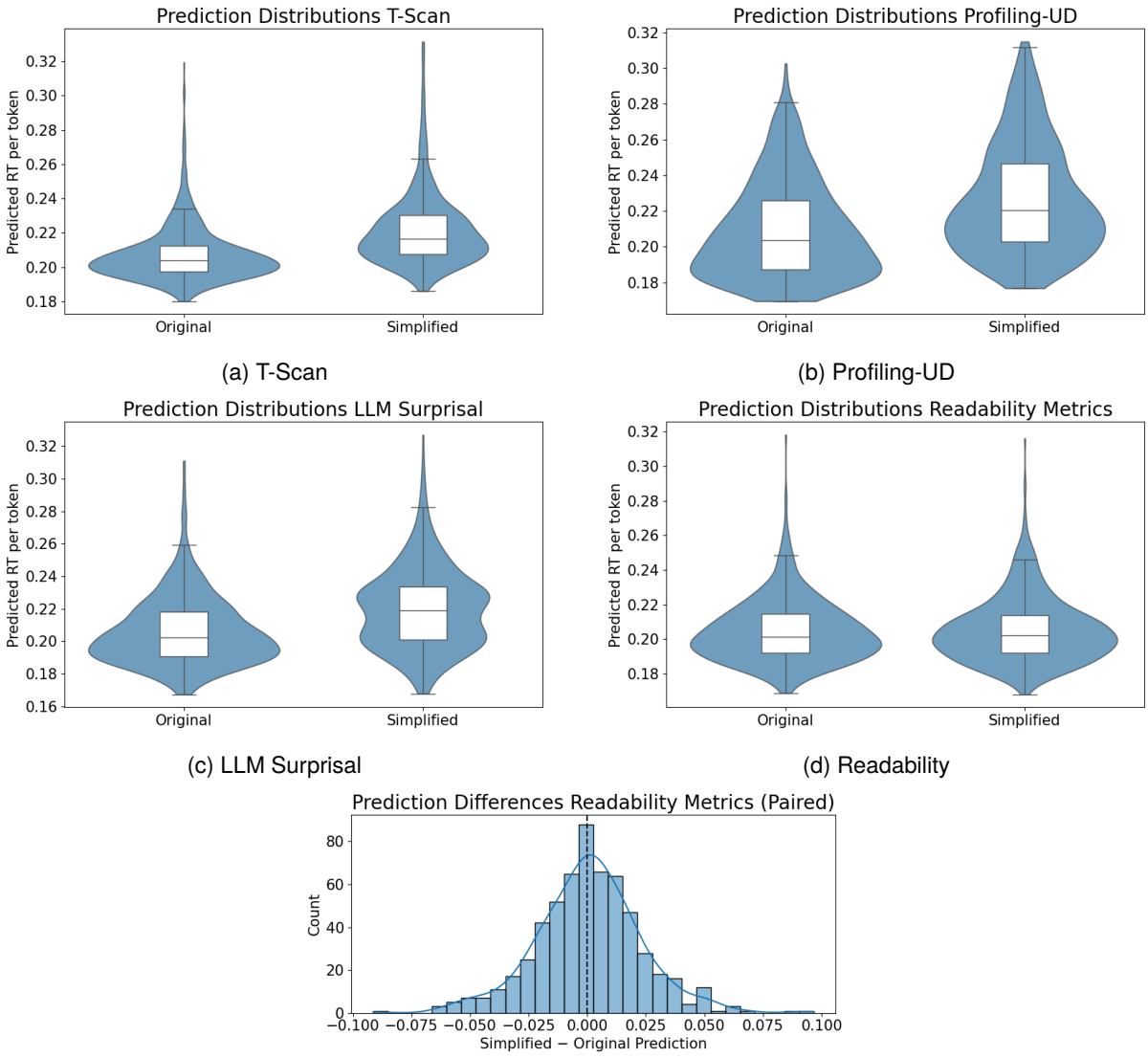


(b) T-Scan



(c) LLM Surprisal

Figure 2: The feature importances for Profiling-UD, T-Scan, and LLM surprisal after simplification.



(a) T-Scan

(b) Profiling-UD

(c) LLM Surprisal

(d) Readability

(e) Prediction differences Readability

Figure 3: Overall prediction distributions of original (left) and simplified (right) texts for all dimensions plus prediction differences and their frequency for the Readability dimension.

C.3. Text Profile Change

This subsection shows two elaborative tables of the text profile change under simplification, which is explained in Section 3.3.

Feature	Mean relative change	% of texts
Dependency length verb - indirect object	-1.46	0.17
Density of 1st person personal and poss. pronouns	-1.13	0.13
Dependency length noun - head of sub clause	-1.03	0.14
Infinitives per clause	-0.94	0.20
Relative clauses per sentence	-0.83	0.16
Clauses in clauses per sentence	-0.79	0.23
Past participles per clause	-0.69	0.22
Proportion of verbs on discussion, reasoning, or causality	-0.69	0.21
Nb. of coordinate clauses	-0.62	0.27
Density of infinitives	-0.61	0.22
Density of present participles	-0.56	0.22
Density of formal words	-0.53	0.24
Nb. of sentences starting with a verb	-0.52	0.18
Surprisal tweety-7b-dutch-v24a	0.50	0.77
Surprisal Qwen3-4B	0.56	0.79
Surprisal fietje-2	0.59	0.78
Surprisal mGPT	0.60	0.80
Density of 3rd person personal and poss. pronouns	0.65	0.59

Table 9: Relative feature changes under simplification for T-Scan and surprisal-based features.

Feature	Mean relative change	% of texts
Nb. of prepositional chains	-0.45	0.22
Nb. of tokens	-0.37	0.29
Dist. of verbs form participle	-0.37	0.25
Dist. of subordinate clauses	-0.30	0.19
Dist. of subordinate clauses preceding main clause	-0.27	0.32
Dependency dist. adverbial clause modifier	-0.22	0.38
Avg. dist. of verbal heads	-0.20	0.21
Dependency dist. nominal modifiers	-0.18	0.35
LiNT score2	-0.15	0.29
Dist. of upos adposition	-0.15	0.25
LiNT score1	-0.14	0.31
Dependency dist. adjectival modifier	-0.14	0.37
Dist. of verbs form infinitive	-0.12	0.40
Avg. token per clause	-0.10	0.30
Dist. of upos verb	0.10	0.67
Dependency dist. objects	0.11	0.58
Dependency dist. copulae	0.17	0.58
Dist. of main clauses	0.20	0.80
Dist. of verbs form finite	0.21	0.75
Dist. of verbs with 3 arguments	0.22	0.68
Dist. of upos numeral	0.23	0.59
Dependency dist. nominal subjects	0.27	0.81
Dist. of objects following the verb	0.29	0.67
Dist. of verbs with 2 arguments	0.45	0.74

Table 10: Relative feature changes under simplification for Profiling-UD and LiNT features.