Towards a Readability Formula for Latin

Thomas Laurs
Georg-August Universität Göttingen
th.laurs@gmail.com

Abstract

This research focuses on the development of a readability formula for Latin texts, a much-needed tool to assess the difficulty of Latin texts in educational settings. This study takes a comprehensive approach, exploring more than 100 linguistic variables, including lexical, morphological, syntactical, and discourse-related factors, to capture the multifaceted nature of text difficulty. The study incorporates a corpus of Latin texts that were assessed for difficulty, and their evaluations were used to establish the basis for the model. The research utilizes natural language processing tools to derive linguistic predictors, resulting in a multiple linear regression model that explains about 70% of the variance in text difficulty. While the model's precision can be enhanced by adding further variables and a larger corpus, it already provides valuable insights into the readability of Latin texts and offers the opportunity to examine how different text genres and contents influence text accessibility. Additionally, the formula's focus on objective text difficulty paves the way for future research on personal predictors, particularly in educational contexts.

Keywords: Readability, Latin, Readability Formula, Linguistic Predictors

1. Introduction

1.1 Readability and Text Comprehension

A method for assessing the difficulty of Latin texts remains a desideratum even though having an objective and precise understanding of the complexity of Latin texts offers numerous advantages in both school and university settings. This knowledge is beneficial for selecting appropriate texts, not only for assessments but also for classroom instruction. It enables textbook authors to craft texts with a steadily increasing level of difficulty, and after the work with the textbook, instructors can use a readability formula to choose suitable texts from authentic Latin authors. The knowledge of text difficulty is especially crucial when it comes to selecting examination texts. This is particularly significant in times of standardized testing, where objective text selection stands as a critical criterion.

Text difficulty, often called readability, is a measure of how smoothly processes of text comprehension can unfold. These processes are determined by both textual features and reader attributes (Friedrich, 2017). Textual features can be divided into two distinct categories. On one hand, texts exhibit a surface structure, encompassing all easily quantifiable linguistic features. On the other hand, texts possess a deep structure, comprising content-related and stylistic features of the text, the translation of which into a numerical value is relatively complex (Groeben, 1982). However, it is essential to note that the boundaries between surface and deep structure are not strictly delineated because some elements of the deep structure can also be calculated objectively. While textual features remain constant within the same text, reader attributes vary, explaining why different readers perceive the same text as more or less difficult. This variation is due to differences in the most important reader attributes, such as intelligence, interest, and prior knowledge (Rost, 2018).

To accurately measure text difficulty, understanding the processes involved in text comprehension is crucial. In general, it can be said that the reader decodes the linguistic information of the text's surface, which includes morphology and syntax, and thus creates a list of propositions at the level of the so-called text base. Subsequently, these propositions are enriched through automatically occurring inferences, resulting in an initial, yet not fully coherent network of propositions. Finally, through actively drawn inferences, reorganization, and reinstatement, a self-contained propositional network is established (the so-called construction-integration model of Kintsch, 1988). Even though the processes of text comprehension for Latin, that might differ from modern languages since being a dead language, have not been extensively researched, this model can be posited for Latin as well due to its generality.

1.2 Phases of Readability Research

In order to develop a metric for predicting the difficulty of texts, readability research has, for about a century, developed various methods, all of which can fundamentally be traced back to the same scheme: (α) Initially, a corpus of texts, whose difficulty has been assessed using a criterion (e.g., a reading test, Cloze test, expert judgment, Common European Framework of Reference for Languages (CEFR)), is gathered. (β) From these texts, linguistic variables are collected. (γ) Finally, the relationships between the predictors and the criterion are statistically modeled (François and Fairon, 2012).

At the beginning of readability research, researchers initially focused on a few linguistic variables, primarily word length as a proxy of vocabulary frequency and sentence length as a proxy for syntactic complexity. Of particular significance in this context are the formulas of Flesch (1948) and Dale and Chall (1948). Both selected a corpus of almost 400 texts. As a criterion, the difficulty was determined through a reading test. Both formulas were established through linear regression and incorporate the two mentioned linguistic variables.

Because these two variables could seem to be too superficial to determine something as complex as the readability of a text, strong criticism of existing

LT4HALA 2024@LREC-COLING-2024, pages 170–175
25 May, 2024. © 2024 ELRA Language Resources Association: CC BY-NC 4.0

formulas has been voiced since 1979 (inter alia Kintsch and Vipond, 1979; Selzer, 1981; Groeben, 1982). Researchers at that time have employed predictors, that were intended to better represent the processes of text comprehension, such as the number of propositions, inferences, or reinstatements, and other deep structural linguistic variables. However, determining these predictors not only requires a considerable effort but is often non-objective. Furthermore, the novel variables and formulas cannot predict text difficulty better than traditional approaches (Kintsch and Miller, 1984).

In recent years, researchers have increasingly turned towards methods of computational linguistics. This allows them to significantly expand the corpus of texts. The difficulty of the texts is usually not assessed by subjects, but often the CEFR is used as a criterion. Machine learning can also be used to rapidly create complex models with numerous linguistic variables (Benjamin, 2012; Vajjala, 2022).

However, there is currently no state-of-the-art readability model for Latin. While some readability formulas exist (e.g., Bayer, 2003 or Gruber-Miller and Mulligan, 2022), their formulas are either based more on theoretical considerations than empiricism or comprise only one linguistic category. In Bayer's formula, a corpus of Latin texts whose difficulty was assessed by a criterion is missing. And Gruber-Miller and Mulligan focused their study only on lexical variables. The goal of this work is to propose a first readability model that follows the established methods of readability research: The difficulty of 67 Latin texts was estimated by students; nearly 200 linguistic variables were calculated using NLP-tools; via stepwise multiple linear regression, a readability model was created to provide a more holistic understanding of Latin text complexity.

2. Empirical Study

2.1 Corpus

There is currently no corpus of Latin texts whose difficulty has been estimated by using an adequate criterion. Since cloze tests and reading tests are not feasible for Latin, we created a questionnaire with a Likert scale, that consisted of 50 items. Bachelor and master students had to read and translate Latin texts and then assessed their difficulty using this questionnaire. They had learned Latin as a historic language in a traditional way. The items of the questionnaire were developed with reference to the theory of the processes of text comprehension presented above and were subsequently analyzed statistically. In total, the 13 best items were retained, which exhibit high discriminatory power and are overall unidimensional, i.e., they all load onto the same factor in the Principal Component Analysis (PCA). All the items are listed in table 1.

In addition to the items based on text comprehension, six additional questions were included to assess the personal knowledge and interests of the participants. After all, personal predictors also influence individual perceptions of difficulty. To eliminate this confounding factor, the same Likert scale was used to gather information about how well the students are versed in vocabulary, grammar, ancient culture and mythology, how well their knowledge is about the given Latin author or literary genre, as well as their level of interest in Latin literature and the duration of their engagement with Latin texts. All six factors exhibited slight correlations with the participants' difficulty assessments, with the strongest correlations observed for knowledge of author and genre ($r = 0.35$) and grammar ($r = 0.27$). As a result, these confounding factors were removed, and, after transforming the modified values onto a 1 to 10 scale, the adjusted difficulty of the texts was obtained. To sum it up, the Latin text of the corpus got their respective difficulty score through the individual difficulty estimations of the students guided through the questionnaire.

#	Question
1	The meanings of most words became clear to me quickly.
2	The sentences had a straightforward syntactic structure.
3	I found it challenging to anticipate how the sentence would continue syntactically.
4	The text contradicted some of the expectations I had formed while reading.
5	I had to frequently backtrack in the text to understand what was being conveyed.
6	Throughout the reading, I had all the necessary information in mind to comprehend the text.
7	At various points, I wished for greater precision in what was meant.
8	Providing a summary of the text would be easy for me.
9	I found it difficult to differentiate between what was important and unimportant in the text.
10	The text was written vividly.
11	I struggled to form a mental image of the content while reading.
12	I found the text to be comprehensible.
13	All in all, the text was easy to understand.

Table 1: Items of the questionnaire

Table 2 includes a selection of five text passages along with their difficulty scores. All in all, 67 Latin texts were assessed by students, 40 prose texts and 27 from poetry, comprising a range of diverse classical authors. The texts had a length of ca. 180 words.

Text passage	Difficulty Score
Pliny 7.19	1.12
Ov. *Met*. 1.283–296	2.54
Verg. *Aen*. 3.147–178	3.29
Livy 44.22.1–8	4.78
Lucan 9.1–33	6.46

Table 2: Difficulty scores of selected texts

2.2 Predictors

Nearly 200 linguistic variables from the areas of Lexicon, Morphology, Discourse, and Syntax were examined. It is not possible to describe all the variables at this point. Therefore, the domains of the

linguistic variables will be outlined briefly, and selected linguistic variables will be described.

2.2.1 Lexicon and Semantics

For Latin, the area of Lexicon and Semantics is particularly crucial. Unlike native speakers, Latin learners must actively acquire vocabulary. If they lack knowledge of the words or cannot retrieve them quickly enough while reading, text comprehension is severely impeded.

The investigation of Lexicon and Semantics is divided into four major categories: (1) word length, (2) word frequency, (3) lexical density, and (4) polysemy.

2.2.1.1 Word Length

Word length is one of the most used variables in readability research. On the one hand, it is easy to calculate, and on the other hand, it serves as a proxy for word frequency (Berendes et al., 2018), because shorter words are more frequent and thus can be understood better by readers (Zipf, 1935). Besides average word length itself, measures like the percentage of monosyllabic words – that can be prepositions, pronouns, verb forms etc. – are added.

2.2.1.2 Word Frequency

Since word length is merely a proxy for word frequency, it is advisable to directly calculate word frequency. Word frequency can be indirectly calculated by examining the percentage of words that do not appear in a list of the most common Latin words (e.g., DCC Latin Core Vocabulary). Alternatively, direct calculations are also possible by determining the number of both lemmas and distinct word forms (i.e. types). In this context, so-called stop words can be excluded, i.e., words that do not significantly contribute to the content of a text, such as conjunctions, etc. (Vogel and Washburne, 1928; McNamara et al., 2014). To ascertain the number of lemmas and the most common Latin words, a corpus comprising texts from Plautus to Augustine was amassed, totaling more than 2 million words. Subsequently, the respective variables of word frequency were computed based on this corpus.

2.2.1.3 Lexical Density

The standard measure for Lexical Density is the Type-Token Ratio (TTR) along with its various calculation methods that aim to minimize the influence of text length (Berendes et al., 2018). Additionally, other measures include the ratio of content words to function words or the curve length R, which is obtained from a rank-frequency distribution by taking the Euclidean distances between adjacent points (Mikros and Voskaki, 2021, following Kubát et al., 2014). This area also encompasses the analysis of Parts of Speech (POS), i.e., examining the ratio of nouns to verbs in a text (Xia et al., 2016).

2.2.1.4 Polysemy

Furthermore, a consideration of polysemy is of paramount importance, especially for Latin, as Latin words are often polysemous and can pose greater difficulties for learners because they may not immediately grasp the meaning, that is correct in each context (McNamara et al., 2014). Polysemy can be determined using the Latin WordNet (LWN). As LWN is not complete, words not covered by the resource were omitted from calculation. Additionally, the number of polysemies can also be determined using the OLD (Oxford Latin Dictionary). The number of meanings given by the OLD of the most important content were stored in a database. From that, the score of polysemy was calculated.

2.2.2 Morphology

As a highly inflected language, Latin, in contrast to English, offers a wider range of difficulties in morphology. Therefore, the occurrence of specific verb forms – ordered by person and number, tense, mood, and voice – as well as the cases of nouns were examined.

2.2.3 Syntax

In the realm of syntax, calculations were carried out in the domains of (1) sentence length, (2) sentence structure, (3) sentence composition, (4) discontinuous noun phrases, and (5) syntactic phenomena.

Sentence length is the traditional measure most frequently used in readability literature (Gray and Leary, 1935; Hancke et al. 2012). In addition to sentence length, the clause length is also significant.

Syntax in Latin places a greater emphasis on word order than in English. This is because the word order in Latin is relatively free. For example, the number of words before the predicate of the main clause or the number of instances where the object precedes the subject of the clause were examined.

Latin prose in particular tends to compose texts in nested complex sentences. One measure to capture this is dependency length, which is also used as a measure of syntactic complexity by Futrell et al. (2015) or Berendes et al. (2018).

Discontinuous noun phrases, also called *hyperbata*, are typical for Latin, especially for Latin poetry, and quite frequent (Haug, 2017). Because of their complexity, they cannot be determined precisely enough by NLP tools, that's why they were calculated manually. The other variables in the syntactic domain were calculated via latinCy, v. infr. Additionally, typical syntactic phenomena such as *Accusativus cum Infinitivo* (AcI) or Gerundive were also manually calculated.

2.2.4 Discourse Variables

In addition to these surface-level text variables, linguistic variables of the deep structure known as discourse-related variables can be considered. The primary goal is to measure the coherence of a text, that means that the text is referring to its own content and connecting the content logically through connectors, pronouns, or co-references. We can calculate that by instances of identical words or lemmas in consecutive sentences (Todirascu et al., 2013; McNamara et al., 2014). Apart from co-reference, latent semantic analysis (LSA) provides another measure of sentence overlap. Essentially, it involves converting the sentences of a text into

vectors and determining their similarity using the cosine measure (François and Fairon, 2012).

2.3 Results

The individual predictors were determined using Natural Language Processing (NLP) techniques. Pre-built tools were employed for this purpose, including the Classical Language Toolkit (CLTK), Stanza, and spaCy (latinCy). However, especially in the realm of syntax, these programs are not yet precise enough (Burns, 2023). Therefore, caution is advised when interpreting the results of the syntactic variables. In addition, some important Latin predictors have been determined manually, including the number of hyperbata (discontinuous noun phrases) and the number of specific syntactic phenomena such as AcI, Ablative Absolute, Gerundives, and so on. The following table 3 contains 20 selected linguistic variables with their correlation coefficients: variables 1–9 come belong to lexicon and semantics, 10–12 to morphology, 13–17 to syntax, and 18–20 to discourse.

#	Description	r
1	Word lengths in letters	.07
2	Percentage of one syllable words	-.33
3	Inverse lemma frequency	.37
4	Frequency of word forms, without stop words, sorted by rank	.23
5	Percentage of words outside a list of the most frequent 750 Latin words	.55
6	Type token ratio, without stop words	.22
7	Ratio of content words to function words	.42
8	Ratio of nouns to all words	.41
9	Average number of polysemes, without stop words, according to the Latin WordNet	-.05
10	Instances of verbs in 3rd singular	.27
11	Instances of verbs in 2nd plural	.25
12	Instances of verbs in pluperfect	-.22
13	Sentence lengths in words	.11
14	Sentence depth, divided by number of t-units	-.05
15	Ratio of finite subclauses to all subclauses	-.30
16	Number of interlaced hyperbata	.54
17	Combination of the easiest syntactic phenomena	-.42
18	Number of connectors	-.25
19	Ratio of pronouns to all words	-.31
20	LSA	-.21

Table 3: Selected linguistic variables with correlation coefficients (r)

The impact on text difficulty is generally greater for lexical variables than for syntax. Word frequency and lexical density, in particular, exhibit a high correlation. Furthermore, these variables tend to yield higher scores in poetic texts. Consequently, it is unsurprising that poetic texts generally receive higher difficulty scores. Contributing to this higher difficulty are also the number of discontinuous noun phrases, which are more prevalent in poetic texts. It is noteworthy that the two standard variables of classical readability studies, word and sentence length, do not exhibit significant correlations with text difficulty in Latin. When examining correlations separately for prose and poetic texts, it becomes apparent that lexical variables exert a greater influence on text difficulty in poetic texts, whereas syntactic variables are more important for computing the difficulty of prose texts.

To model the relationship between linguistic variables and the difficulty of individual texts, a multiple linear regression analysis was conducted as a statistical model. The selection of appropriate variables is not trivial. A stepwise regression analysis was performed: initially, a regression was created with only one parameter, the highest correlated variable (#5). Subsequently, from the remaining variables, the one that resulted in the lowest root-mean-square deviation in a 10-fold cross-validation was added to the model, while all p-values should not fall below the level of significance. This process continued until no significant p-values were obtained. Since the text difficulty here is considered to be a continuous variable, other methods like logistic regression or support vector machines do not work.

Through the described way of selecting variables, the best predictors were 4, 5, 6, 9, 10, 11, 12, 14, and 17. One needs to bear in mind that some of the linguistic variables are highly correlated among each other. Thus, those predictors with smaller intercorrelations were selected, which can have a lower correlation with the criterion. The obtained statistic model has an R^2 of .69, that means it can explain the variance in the students' estimation of text difficulty by about 70%. If one looks at the R^2 obtained in a 3-, 5-, or 10-fold cross-validation, the value gets lower, namely to .54, .50, and .38 respectively.

With these predictors, we get a formula for the readability of Latin literature (the sequence of predictors in the formula corresponds to their inclusion in the statistical model during stepwise linear regression):

$$f(x) = 14.478 + 24.885x_5 + 9.872x_{11} - 0.015x_4 - 9.473x_{12} - 15.215x_{17} + 0.402x_{14} - 0.097x_9 + 2.395x_{10} - 7.141x_6$$

3. Conclusions and Future Work

We have created a readability formula for Latin consisting of nine linguistic factors from various linguistic categories, which can explain the difficulty of Latin texts by about 70%, similar to other models (e.g., François and Fairon, 2012, have created a model with R^2 of .73). The formula presented in this paper could be further improved by adding more text to the corpus. In doing so, one could enhance the slightly lower R^2-values in cross-validation. A reason that those metrics are behind the model of François and Fairon (2012) could be due to the fact that Latin texts, unlike modern schoolbook texts, were composed for a highly educated upper class. All examined texts possess significant literary merit and are not merely instructional or exercise texts. Furthermore, there is the possibility of providing two separate formulas, one for prose and one for poetry texts.

Indeed, if one looks at the correlation between the difficulty of Latin poetry texts and certain linguistic variables, one can find some predictors with much higher correlation, e.g. the percentage of one syllable words correlates with r = -.50, the percentage of words outside a list of the most frequent 750 Latin words correlates with r = .60, and the number of interlaced hyperbata correlates with r = .55. A statistic model based only on poetic text could explain the variance in text difficulty of those text by 87%, but the prognostic power is much lower: one finds R^2 obtained in a 3-, or 5-fold cross-validation of .61, and .21, respectively.

Building upon the final readability model, further investigations can be conducted. By examining the residuals between the model and actual difficulty assessments, insights can be gained into which text genres and contents are generally easier or more challenging for readers to access. It can be expected that narrative passages are easier to understand than, for instance, philosophical treatises.

Since the formula provides a score for objective text difficulty that eliminates the personal characteristics of readers, in a concluding step, investigations can also be conducted on personal predictors. Especially in the context of education, it could be explored what personal prerequisites, particularly in vocabulary and grammar, one should have to understand a text.

4. Bibliographical References

Bayer, K. (2003). Bestimmung des Schwierigkeitsgrades von lateinischen Klassenarbeiten, *Pegasus*, 3(2):1–19.

Benjamin, R.G. (2012). Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty, *Educational Psychology Review*, 24:63–88.

Berendes, K., Vajjala, S., Meurers, D., Bryant, D., Wagner, W., Chinkina, M., and Trautwein, U. (2018). Reading demands in secondary school: Does the linguistic complexity of textbooks increase with grade level and the academic orientation of the school track?, *Journal of Educational Psychology*, 110(4):518–543.

Burns, P.J. (2023). LatinCy: Synthetic Trained Pipelines for Latin NLP. *arXiv preprint arXiv:2305.04365.*

Dale, E. and Chall, J.S. (1948). A Formula for Predicting Readability, *Educational Research Bulletin*, 27:11–20+28.

Flesch, R. (1948). A New Readility Yardstick, *Journal of Applied Psychology*, 32:221–233.

François, Th. and Fairon, C. (2012). An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 466–477, Jejudo, South Korea.

Friedrich, M. (2017). *Textverständlichkeit und ihre Messung. Entwicklung und Erprobung eines Fragebogens zur Textverständlichkeit.* Münster and New York.

Futrell, R., Mahowald, K., and Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages, *PNAS*, 112(33):10336–10341.

Gray, W.S. and Leary, B.E. (1935). *What Makes A Book Readable*. Chicago.

Groeben, N. (1982). *Leserpsychologie: Textverständnis – Textverständlichkeit*. Münster.

Gruber-Miller, J. and Mulligan, B. (2022). Latin Vocabulary Knowledge and the Readability of Latin Texts: A Preliminary Study, *New England Classical Journal*, 49(1):80–101.

Hancke, J., Vajjala, S., and Meurers, D. (2012). Readability Classification for German using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012: Technical Papers*, pp. 1063–1080, Mumbai.

Haug, D. (2017). Syntactic discontinuities in Latin. A treebank-based study, *Bergen Language and Linguistics Studies*, 8:75–96.

Kintsch, W. (1988). The Role of Knowledge in Discourse Comprehension: A Construction-Integration Model. *Psychological Review*, 95(2):163–182.

Kintsch, W. and Miller, J.R. (1984). Readability: A View from Cognitive Psychology. In J. Flood (Ed.), *Understanding Reading Comprehension: Cognition, Language, and the Structure of Prose*. Newark, Del., pp. 220–232.

Kintsch, W. and Vipond, D. (1979). Reading Comprehension and Readability in Educational Practice and Psychological Theory. In L.-G. Nilsson (Ed.), *Perspectives on Memory Research: Essays in Honor of Uppsala University's 500th Anniversary*. Hillsdale, NJ, pp. 329–365.

Kubát, M., Matlach, V., and Čech, R. (2014). *QUITA: Quantitative Index Text Analyzer*. Lüdenscheid.

McNamara, D.S., Graesser, A.C., McCarthy, P.M., and Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. New York.

Mikros, G. and Voskaki, R. (2021). A Modern Greek readability tool: Development of evaluation methods. In A. Pawłowski, J. Mačutek, Sh. Embleton, & G. Mikros (Eds.), *Language and Text: Data, models, information and applications*. Amsterdam and Philadelphia, pp. 163–175.

Rost, D.H. (2018). Leseverständnis. In D.H. Rost, J.R. Sparfeldt, & S.R. Buch (Eds.), *Handwörterbuch Pädagogische Psychologie*. Weinheim and Basel, 5th edition, pp. 494–506.

Selzer, J. (1981). Readability Is a Four-Letter Word, *Journal of Business Communication*,18:23–34.

Todirascu, A., François, Th., Gala, N., Fairon, C., Ligozat, A.-L., and Bernhard, D. (2013). Coherence and Cohesion for the Assessment of Text Readability. In *Proceedings of 10th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2013)*, pp. 11–19, Marseille, France.

Vajjala, S. (2022). Trends, Limitations and Open Challenges in Automatic Readability Assessment Research. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pp. 5366–5377, Marseille, France.

Vogel, M. and Washburne, C. (1928). An Objective Method of Determining Grade Placement of Children's Reading Material, *The Elementary School Journal*, 28(5):373–381.
Xia, M., Kochmar, E., and Briscoe, T. (2016). Text Readability Assessment for Second Language Learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 12–22, San Diego, Cal.
Zipf, G.K. (1935). *The Psycho-Biology of Language. An Introduction to Dynamic Philology*. Boston.

5. Language Resource References

Johnson, K.P., Burns, P.J., Stewart, J., and Todd, C. 2014–2021. *CLTK: The Classical Language Toolkit*. https://github.com/cltk/cltk.
Short, W.M. 2024. *Latin WordNet 2.0*. https://latinwordnet.exeter.ac.uk