Integrating Archaic and Regional Lexicons to Improve the Readability of Old Romanian Texts

Mădălina Chitez[1], **Roxana Rogobete**[1], **Aura-Cristina Udrea**[2,5]
Karla Csürös[1], **Ana-Maria Bucur**[3], **Mihai Dascălu**[2,4,5]
[1]West University of Timișoara
[2]National University of Science and Technology Politehnica Bucharest
[3]Interdisciplinary School of Doctoral Studies, University of Bucharest
[4]Academy of Romanian Scientists, [5]Lucian Blaga University of Sibiu
{madalina.chitez, roxana.rogobete, karla.csuros}@e-uvt.ro
{mihai.dascalu, aura.cojocaru}@upb.ro
ana-maria.bucur@drd.unibuc.ro

Abstract

Access to age-appropriate texts is critical for young readers' literacy acquisition. For limited-resourced languages, such as Romanian, this area remains under-researched. As such, we present ongoing work on improving readability for old Romanian texts by applying Large Language Models (LLMs). First, we compiled and cleaned a comprehensive list of archaic and regional terms from lexicographic sources, including DEX online and printed dictionaries. The cleaning process involved duplicate removal, orthographic normalization, context-based filtering, and manual review. Key challenges included distinguishing archaic forms from rare or poetic ones, resolving polysemous entries, and managing inconsistent labeling across sources. Second, LLMs were utilized to validate the archaic and regional nature of identified terms and replace them with modern equivalents, while also determining the appropriate reading level for both original and modified versions. Results show that through the replacement of archaic and regional terms, the appropriate age for the modified texts decreases by approximately 0.5 years for texts extracted from textbooks and canonical writings.

1 Introduction

Effective text comprehension is intrinsically linked to the reader's familiarity with linguistic and conceptual content. In educational contexts, this relationship is particularly critical as improving the accessibility of reading materials directly enhances student engagement and learning outcomes. Numerous studies have argued that matching texts to learners' cognitive and linguistic development levels enhances reading comprehension, supports motivation, and reduces cognitive overload (Snow and Biancarosa, 2003; Shanahan, 2016).

Traditional readability models, often grounded in linguistic features, have been widely used to assess and improve text readability (Collins-Thompson, 2014). However, such models tend to prioritize surface-level features while overlooking the more subtle, yet impactful, dimensions of linguistic complexity, including vocabulary rooted in historical or regional usage. This limitation becomes particularly clear when working with literary texts in limited-resourced languages, where non-standard vocabulary may hinder comprehension and reduce overall readability, especially for younger readers. In the Romanian context, the issue is accentuated by the widespread use of canonical literary texts in school curricula, which often include archaic and dialectal terms that are no longer in common usage, thus reducing readability for contemporary readers. While previous initiatives have introduced automated readability assessment tools for Romanian - for example, the LEMI platform designed for evaluating children's literature (Chitez et al., 2023)-, existing models have not accounted for diachronic or regional lexical variation.

As such, we extend the LEMI readability platform (Chitez et al., 2024a) by incorporating archaic and regional terms. In addition to defining old Romanian words, we utilize a Large Language Model (LLM) to rephrase archaic and regional words within their context, using the corresponding definitions from the curated lexicons. This contributes to improving readability by simplifying archaic and regional expressions through contextual rewording. We describe the methodology for compiling and integrating these resources and evaluate their impact on text-level readability predictions through a validation study using representative Romanian literary texts.

We release the archaic and regional lexicons with definitions, available at https://huggingface.co/datasets/upb-nlp/lemi_archaic_regional (accessed 30 July 2025).

Proceedings of Recent Advances in Natural Language Processing,pages 240–246
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-029

2 Related Work

This section reviews existing work on text readability and lexical complexity in literary and educational contexts. We first discuss general readability models (Section 2.1), followed by resources for archaic and regional lexicons (Section 2.2).

2.1 Readability Assessment

The assessment of text readability has been a longstanding concern in educational and computational linguistics. Traditionally, readability has been associated with surface-level textual features such as sentence length and word complexity, often operationalized through formula-based models like Flesch-Kincaid, Gunning Fog Index, or SMOG (Martinc et al., 2021; Pitler and Nenkova). These approaches have proven useful in many educational settings, especially for labeling texts with school grade levels and ensuring age-appropriate content for learners (Liu, 2022). Still, even state-of-the-art systems often focus on general-purpose indicators and fail to account for context-specific linguistic variation, such as archaic and regional vocabulary commonly found in children's literary texts. As Liu (2022) argued, children's literature often employs ideational grammatical metaphors and historical lexis that complicate understanding, especially for contemporary readers. Similarly, Imperial and Ong (2020) showed how feature augmentation with lexical variation significantly improves readability prediction in low or limited-resourced languages, highlighting the relevance of specialized vocabulary lists (e.g., rare words, foreign terms) in educational settings. Recent efforts have expanded readability modeling to include such non-standard lexical items.

2.2 Archaic and Regional Lexicons

Computational lexicons with archaic and regional vocabulary are increasingly valuable for historical linguistics and natural language processing (NLP). Archaic lexicons, such as the Old English Core Vocabulary (University of St Andrews, n.d.) and James Orchard Halliwell's Dictionary of Archaic and Provincial Words (Halliwell, 2018), offer digitized lists of obsolete or historical terms that are essential for understanding literary and historical texts. Such resources have been adapted for computational purposes, as seen in projects like the Lexicon for Processing Archaic Slovene, which maps historical Slovene word forms to their modern equivalents to aid tokenization and lemmatization (Erjavec et al., 2011). Similarly, regional lexicons, including Lord Moreton's Glossary of Dialect and Archaic Words Used in the County of Gloucester (Robertson and Moreton, 1890), serve as valuable repositories for studying dialectal variation and language change. These historical and dialectal resources are being integrated into broader computational infrastructures, such as Lexibank, a large-scale lexical database developed by the Max Planck Institute that standardizes word lists across over 2,000 languages (List et al., 2022).

In Romanian linguistics, the study of archaic and regional vocabulary is well-documented through several relevant lexicographic and scholarly resources. Foundational works include The Dictionary of Archaic Lexicon (Dicționarul de arhaisme) (Busuioc, 2005) and The Dictionary of Archaic and Regional Lexicon (Dicționarul de arhaisme și regionalisme) (Bucă, 2008), which remain essential for understanding lexical variation in Romanian literature and dialects. However, these dictionaries exist only in print, limiting NLP applications (Chitez et al., 2024b). The challenges extend to online resources as well: although platforms like DEX online, Wikționar, and Regionalisme.ro provide valuable lexical entries, their tagging systems (e.g., [arhaizant], [regional]) are inconsistently applied, impeding their automated processing. Moreover, regional forms such as *ciojlingar* (a variant of *cioflingar*) exemplify how rare or morphologically variant terms are often missing from official digital lexicons (ibid.). The authors emphasized the stringent need to digitize and standardize these resources not only for computational readability models such as the LEMI index (Chitez et al., 2024a), but also to ensure pedagogical alignment between literary texts and students' comprehension levels.

3 Method

3.1 Corpus

An existing educational corpus, ROTEX (i.e., The Corpus of Romanian School Textbooks; Oraviţan et al., 2023), consisting of all school textbooks approved by the Ministry of Education, was considered for this study. We extracted a focused subcorpus consisting exclusively of literary texts featured in the textbooks for grades 5 to 8. These texts are central to the Romanian national curriculum and represent the primary literary material that students encounter in lower secondary education. The liter-

ary texts were extracted from the ROTEX corpus using Gemini (Team et al., 2023) (see prompt in Table 1). To ensure the relevance and representativeness of the dataset, the literary texts extracted by Gemini were manually filtered to exclude any non-literary content, such as parts of exercise formulations.

The resulting subcorpus includes texts ranging from narrative excerpts and poems to character sketches and allegorical tales, often accompanied by authorial or historical context provided within the textbook framework. In addition to textbook content, we enriched the dataset by incorporating selected works from canonical Romanian authors frequently referenced in classroom reading assignments, such as Ion Creangă[1], Mihail Sadoveanu[2] and Petre Ispirescu[3]. These texts, often included in supplementary school materials or national exams, feature linguistic characteristics, such as archaic or regional vocabulary, that pose potential comprehension challenges for students. The resulting dataset used in this study consists of more than 1 million words (see Table 2).

3.2 Readability Assessment after Replacing Archaic and Regional Terms

In this study, we used a systematic approach (described in Figure 1) to assess the impact of archaic and regional lexis on the readability of text. We began by developing a full database from *Dicționarul de Arhaisme și Regionalisme* (DAR), gathering approximately 11,800 words that can exhibit archaic characteristics and 13,000 words that exhibit regional characteristics with their corresponding definitions. The compiled lists are publicly available and can be accessed at `https://huggingface.co/datasets/upb-nlp/lemi_archaic_regional` (accessed 30 July 2025). We filtered this collection to exclude terms that appeared in a previously compiled common-use lexicon.

Our analysis utilized two distinct text corpora: educational materials from textbooks and literary works from canonical authors. We systematically identified archaic or regional terms in each text. When such terms were detected, we used the Llama 3.3 70B (Grattafiori et al., 2024) language model to determine if the terms were indeed used with their archaic or regional meaning. The model then replaced uncommon vocabulary with accessible equivalents while preserving meaning and tone. We chose this model due to its strong performance in diverse text generation tasks and its flexibility in handling Romanian texts. To measure improvements in readability, we used a comparative evaluation approach. We presented the language model with randomized original and modified versions, evaluated using 25 complexity tags. The full prompt is presented in Table 3.

The categories systematically captured reading difficulty from basic features like simple vocabulary to complex markers including archaic terminology and abstract content. This categorical framework provided the model with a structured set of reasons to justify its age recommendations. This approach not only enhanced the transparency of the model's decision-making process but also enabled us to quantitatively analyze which factors most commonly influenced text difficulty ratings.

4 Results and Discussion

We compiled all the age estimates generated by the model for both the original and modified versions of the texts for corpora of both pedagogical content and canonical literary works. According to the results presented in Table 4, the average estimated reading age for the modified texts (*Age_2*) was consistently, albeit slightly, lower than that of the original texts (*Age_1*).

In addition, we plotted the estimated ages, presented in Figure 2.

We also employed statistical tests to determine whether the text modification pipeline significantly alters reading difficulty. Due to non-normal distribution of age estimates (Shapiro-Wilk $p < 0.001$), we employed the Wilcoxon signed-rank test (Rey and Neuhäuser, 2011), a non-parametric alternative to the paired t-test that evaluates whether the median difference between paired observations significantly differs from zero. The mean age difference was calculated as Modified - Original (with negative values indicating easier reading). Category stability denotes the percentage of texts that retain the same readability category after replacing archaic or regional terms with simpler ones.

The findings revealed statistically significant reading age reductions for canonical texts ($p < 0.001$ for both archaic and regional variants), with mean decreases of 0.50 and 0.59 years respectively. In contrast, textbook modifications showed

[1] https://en.wikipedia.org/wiki/Ion_Creanga
[2] https://en.wikipedia.org/wiki/Mihail_Sadoveanu
[3] https://en.wikipedia.org/wiki/Petre_Ispirescu

Gemini Prompt
You are given a page from a Romanian language and literature textbook written in Romanian. Extract the anchor text or reading passage that is provided for each unit or lesson. Anchor texts might be found in the textbooks under sections such as ”Text de bază”, ”Textul 1”, ”Textul 2”, ”Citește fragmentul de mai jos”, etc. If the reading passage is in a two-column format, usually the left part is the first part of the text, and the right column represents the second part of the text. If the text is divided into two columns, extract the text from both columns and combine them into a single text. Keep in mind that many pages might not contain any reading passage or anchor text, and in that case, the output should be empty. Do not include in the output the text from exercises, questions, or in which theoretical concepts are explained, which are usually numbered or more structured and are found in sections such as ”Explorare”, ”Repere”, ”Aplicații”, ”Explorăm și învățăm”, ”Reținem”, ”Proiect de grup”, ”Pentru început”, etc. Do not include other content or information about the textbook or the editors. Do not include any other text or information that is not part of the reading passage. The output should only contain the text of the reading passage. Usually reading passages are long, so the output should be a long text. Do not shorten or summarize the reading passage. Do not remove any sentence from the reading passage. Join the syllabified words in the reading passage into a single word. Remove any subscripts, superscripts, and footnotes from the reading passage. Remove any special characters or symbols that are not part of the reading passage. Remove any extra spaces or line breaks from the reading passage. Make a JSON file of the output. The output JSON should include the reading passage found on the page in the ”text” field. If the document does not contain any reading passage, leave the JSON file empty.

Table 1: Prompt used for extracting the reading passages from textbooks

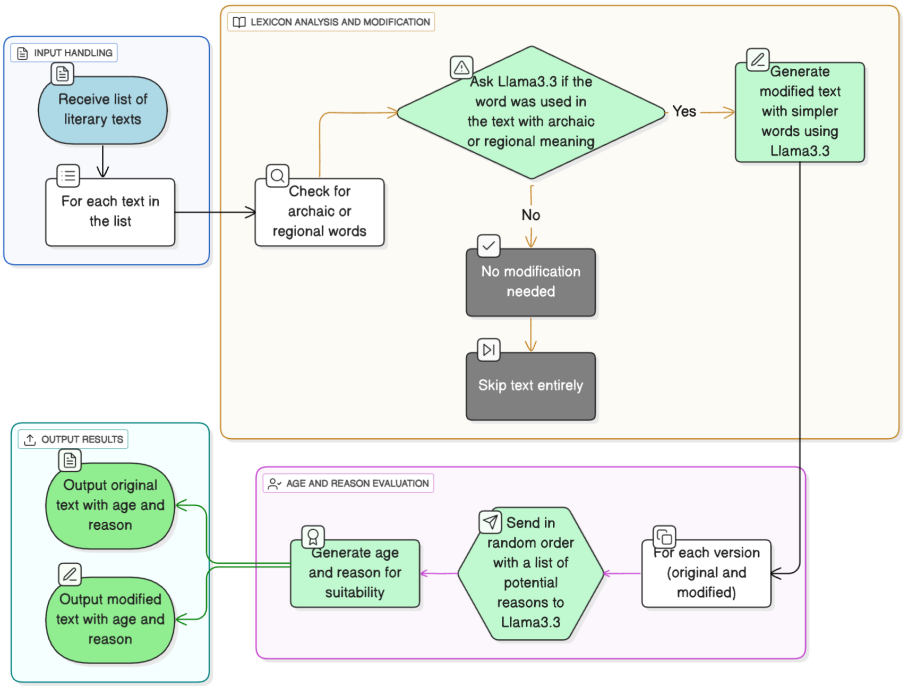

Figure 1: Automated text simplification and age estimation flow

Text Category	**Total Words**
Textbooks	559,148
Canonical Writings	451,993
Total	**1,011,141**

Table 2: Total word count by text category

no significant reading age changes ($p < 0.05$), suggesting that these texts were already optimally calibrated for their target audience. The stability of the categories was high for textbooks ($> 95\%$) but lower for canonical texts, indicating more substantial lexical transformations in literary works.

To analyze the model's age estimates, we identified the top 3 reasons selected by the LLM for each text version (Tables 6 and 7). ”Varied vocabulary but familiar to older children” remained predominant across versions, indicating a deliberate focus on preserving this feature during simplification.

In canonical writings with regional and archaic lexicon, text modifications show a rise in easily recognizable words (from 203 to 377 for archaic lexi-

Llama3.3 Prompts

1. Archaic Word Usage Analysis Prompt

You are a Romanian linguist expert in the Romanian language. You carefully analyze the use of words in context.
Analyze the following text: "{text}"
The word '{word}' can have the following archaic definition: {definition}
Analyze whether in this text the word '{word}' is used with the archaic meaning mentioned above or with a modern/contemporary meaning. Answer ONLY with "YES" if the word is used with archaic meaning, or "NO" if it is used with modern/contemporary meaning. Do not add explanations or justifications. Your answer must be exactly "YES" or "NO".

2. Text Rewriting Prompt

Here is a text to be rewritten: {text}
The word {word} is less commonly used and it means: {definition}. Replace {word} with a modern equivalent that preserves the tone and meaning while being easier to understand. If no modern equivalents exist, rewrite the text without using these words. Do not add explanations or justifications. Respond only with the rewritten text.

3. Text Comparison Prompt

Analyze the following two variants of a text: 1. {texts[0]} 2. {texts[1]}
For each variant, indicate: - the minimum age (in years) at which an average reader could easily understand the text, - the relevant reasons (from the list below) that justify this age.
Respond strictly in the format: [age_of_text_1, age_of_text_2, reason_for_text_1, reason_for_text_2]. Make sure to choose only from the following possible categories:
Very simple and basic vocabulary
Frequent and easily recognizable words
Rarer words, but easily deduced from context
Some passages slow down reading through density
Complex vocabulary, including rare terms
(Note: This is a subset of the complete category list used in the actual prompt)
Decide which text is easier and which is more difficult, then choose a reason for each text. Do not add any explanation. Do not add extra text. Only the list of two numbers and the two reasons.

Table 3: Consolidated Prompt Set for Archaic Analysis, Simplification, and Age Estimation

Text Type	Lexicon	Age_1	Age_2	# texts
Textbook	Archaic	13.79	13.76	261
Textbook	Regional	13.38	13.28	171
Canonical	Archaic	12.74	12.24	1702
Canonical	Regional	12.37	11.77	1451

Table 4: Average estimated reading age for original vs. modified texts by text and lexicon type with number of texts

Table 5: Statistical Analysis Summary: Age Estimates and Category Changes

Text Type	Lexicon	Mean Age Diff.	Wilcoxon p-value	Category Stability %
Textbook	Arch.	-0.031	0.491	96.2
Textbook	Reg.	-0.099	0.046	95.3
Canonical	Arch.	-0.502	<0.001	71.3
Canonical	Reg.	-0.592	<0.001	68.1

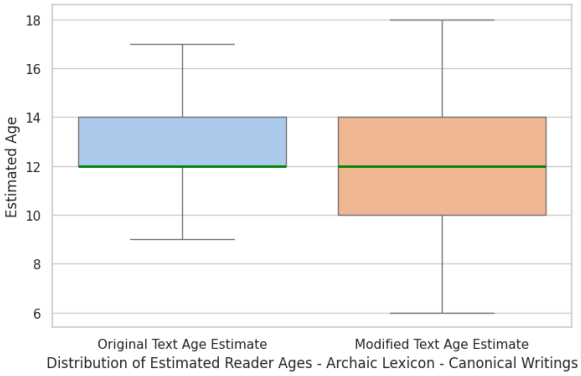

Figure 2: Distribution of Estimated Reader Ages - Archaic Lexicon - Canonical Writings

con and from 200 to 374 for regional lexicon) and a marked decrease in complex vocabulary (from 211 to 110 and from 185 to 71, respectively). These changes indicate a deliberate simplification strategy that increases accessibility while maintaining the variety and conceptual integrity of the core vocabulary.

There were numerous cases where the model was efficient in producing simplified texts without compromising the original meaning. For example, the Romanian word *"conac"* typically means "mansion" or "manor house" in contemporary Romanian, but it also has an old sense of "long distance". When provided with the complete definition of the term, the model was capable of determining the intended meaning based on the context of the source sentence and the estimated reading age was lowered from 14 to 12, demonstrating improved accessibility while preserving narrative integrity.

Original text (Romanian): "Acum e timpul când are să vină la prânz, şi are obicei de aruncă buzduganul cale de un conac şi loveşte în uşă, în masă şi'se pune în cui." (eng. *"Now is the time when he is to come for lunch, and he has the*

Original Text Categories		Modified Text Categories	
Category	**Count**	**Category**	**Count**
Archaic Lexicon			
Varied vocabulary but familiar to older children	822	Varied vocabulary but familiar to older children	845
Complex vocabulary including rare terms	211	Frequent and easily recognizable words	377
Frequent and easily recognizable words	203	Mildly abstract concepts like emotions or rules	110
Regional Lexicon			
Varied vocabulary but familiar to older children	755	Varied vocabulary but familiar to older children	761
Frequent and easily recognizable words	200	Frequent and easily recognizable words	374
Complex vocabulary including rare terms	185	Some passages slow reading due to density	71

Table 6: Top 3 Categories Distribution: original versus modified texts - Canonical Writings

Original Text Categories		Modified Text Categories	
Category	**Count**	**Category**	**Count**
Archaic Lexicon			
Varied vocabulary but familiar to older children	152	Varied vocabulary but familiar to older children	154
Complex vocabulary including rare terms	75	Complex vocabulary including rare terms	74
Some passages slow reading due to density	26	Some passages slow reading due to density	26
Regional Lexicon			
Varied vocabulary but familiar to older children	109	Varied vocabulary but familiar to older children	113
Complex vocabulary including rare terms	43	Complex vocabulary including rare terms	39
Some passages slow reading due to density	17	Some passages slow reading due to density	17

Table 7: Top 3 Categories Distribution: Original versus Modified Texts – Textbooks

habit of throwing the mace the distance of a manor house and strikes the door, the table, and hangs it on a nail.")

— *Făt Frumos cu Părul de Aur*, Petre Ispirescu

Modified text (Romanian): "Acum e timpul când are să vină la prânz, şi are obicei de a aruncă buzduganul pe o distanță considerabilă și lovește în ușă, în masă și se pune în cui."(eng. *Now is the time when he is to come for lunch, and he has the habit of throwing the mace a considerable distance and strikes the door, the table, and hangs it on a nail.*")

Analysis:

Archaic word replaced: *conac*

Definition: conac, conace, s.n. (înv.) 1. casă boierească la țară, pe o moșie. 2. reședință a unui ispravnic sau a unui subprefect. 3. hotel turcesc. 4. loc de popas; stație de poștă; popas. 5. han, gazdă. 6. distanță de la un loc de popas la altul; poștă. (eng. *1. boyar house in the countryside, on an estate. 2. residence of a steward or sub-prefect. 3. Turkish hotel. 4. resting place; postal station; stopover. 5. inn, lodging. 6. distance from one resting place to another; postal route.*)

Readability assessment:

Text Version	Estimated Age	Reason
Original Text	14 years	Complex vocabulary including rare terms
Modified Text	12 years	Frequent and easily recognizable words

We analyzed the cases where the LLM predicted a higher age for the modified text than for the original one. The top reasons chosen by the LLM to justify its age choice for the modified text were: *Complex vocabulary including rare terms*, *Some passages slow down reading through density* and *Long sentences with advanced syntactic structures*, as opposed to the categories chosen for the original text: *Varied but familiar vocabulary for older children* and *Frequent and easily recognizable words*.

Our evaluation showed that despite contextual analysis, the LLM sometimes misclassified archaic or regional words, leading to inconsistencies in readability assessments and underscoring the need for expert review and more context-aware vocabulary evaluation.

5 Conclusions and Future Directions

We present a systematic approach to improving Romanian text readability through lexical modernization using LLMs and a comprehensive database of archaic/regional terms, achieving measurable reading age reductions. Beyond evaluating the quality of text simplification, our approach has potential in educational settings. One promising direction is a user-guided simplification system, where the user specifies the target age group, and the LLM automatically adapts the text accordingly.

Future efforts will focus on developing and testing such a system, incorporating adaptive simplification methods, more accurate control of linguistic properties (e.g., syntactic complexity, vocabulary size), and alignment with the curriculum. We also intend to improve contextual understanding of polysemous words to make the produced texts more faithful and pedagogically valuable.

Acknowledgments

This work was supported by a grant from the Ministry of Research, Innovation and Digitization, CNCS - UEFISCDI, project number PN-IV-P1-PCE-2023-2025 LUMRO, within PNCDI IV. We also thank DEX online (https://dexonline.ro) for providing additional resources related to archaic and regional lexical items.

References

Marin Bucă. 2008. *Dicţionar de arhaisme şi regionalisme*. Vox Cart.

Monica Mihaela Busuioc. 2005. *Dicţionar de arhaisme*. Ed. ALL Educational.

Mădălina Chitez, Mihai Dascalu, Aura Cristina Udrea, Cosmin Striletchi, Karla Csürös, Roxana Rogobete, and Alexandru Oravițan. 2024a. Towards building the lemi readability platform for children's literature in the romanian language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16450–16456.

Mădălina Chitez, Roxana Rogobete, and Karla Csürös. 2024b. De la hârtie la ecran: provocările în utilizarea intrărilor de dicționar de date pentru dezvoltarea instrumentului lemi de stabilire a lizibilității literaturii române. In *RITL – Revista de Istorie și Teorie Literară / The review for Literary History*.

Madalina Chitez, Roxana Rogobete, Alexandru Oravitan, et al. 2023. Designing lemi: the romanian language tool that makes kids love reading. In *Conference Proceedings. The Future of Education 2023*.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

Tomaž Erjavec, Christoph Ringlstetter, Maja Žorga, and Annette Gotscharek. 2011. A lexicon for processing archaic language: the case of xixth century slovene. In *First International Workshop on Lexical Resources*, page 24.

Aaron Grattafiori et al. 2024. The llama 3 herd of models.

James Orchard Halliwell. 2018. *A dictionary of archaic and provincial words: Obsolete phrases, proverbs, and ancient customs, from the XIV century*. Routledge.

Joseph Marvin R Imperial and Ethel C Ong. 2020. Application of lexical features towards improvement of filipino readability identification of children's literature. In *Proceedings of the Philippine Computing Science Congress*.

Johann-Mattis List, Robert Forkel, Simon J Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9(1):316.

Yan Liu. 2022. Readability and adaptation of children's literary works from the perspective of ideational grammatical metaphor. *Journal of World Languages*, 7(2):334–354.

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

Alexandru Oraviţan, Mădălina Chitez, and Roxana Rogobete. 2023. A linguistically-informed assessment model for multidimensional competence building in romanian school writing. *Educatia 21*, pages 85–91.

Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *2008 Conference on Empirical Methods in Natural Language Processing*, page 186.

Denise Rey and Markus Neuhäuser. 2011. Wilcoxon-signed-rank test. In *International encyclopedia of statistical science*, pages 1658–1659. Springer.

John Drummond Robertson and Lord Henry Haughton Reynolds Moreton. 1890. *A glossary of dialect & archaic words used in the county of Gloucester*, volume 25. English dialect society.

Timothy Shanahan. 2016. Relationships between reading and writing development.

Catherine Snow and Gina Biancarosa. 2003. *Adolescent literacy and the achievement gap: What do we know and where do we go from here?* Carnegie Corporation of New York New York.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

University of St Andrews. n.d. Old english core vocabulary.