LREC 2026

2nd Workshop on Evaluating Text Difficulty in a Multilingual Context (DeTermIt! 2026)

Workshop Proceedings

Editors
Giorgio Maria Di Nunzio, Federica Vezzani, Liana Ermakova, Hosein Azarbonyad, and Jaap Kamps

11 May 2026

Proceedings of the 2nd Workshop on Evaluating Text Difficulty in a Multilingual Context (DeTermIt! 2026)

ISBN 978-2-493814-63-0

Preface

Automatic Text Simplification (ATS) is the process of reducing the linguistic and conceptual complexity of a text in order to improve its comprehensibility and readability. ATS plays an important role in supporting clear and accessible communication, while also serving as a useful preprocessing step for tasks such as information extraction, retrieval, and knowledge access. More broadly, ATS has significant societal implications, especially for people with limited literacy, second language learners, and readers who encounter difficulties when dealing with specialized or terminology-rich texts.

One of the main barriers to text understanding is the presence of unfamiliar terminology and domain-specific concepts. In many real-world settings, accessibility depends not only on simplifying syntax or vocabulary, but also on identifying which terms and conceptual relations are likely to create difficulty for different audiences. Lexical simplification, terminology adaptation, and readability assessment therefore represent complementary perspectives on the same challenge: how to make texts more accessible while preserving their informational value. In recent years, these questions have gained further relevance with the increasing use of Large Language Models (LLMs) and Generative AI (GenAI), which are now widely used to generate, rewrite, and mediate access to information.

The *DeTermIt! Evaluating Text Difficulty in a Multilingual Context* workshop explores the theoretical and practical perspectives surrounding the evaluation and modeling of text difficulty in multilingual and terminology-rich contexts. In today's interconnected world, where access to knowledge increasingly depends on language technologies, it is essential to ensure that information remains understandable, faithful, and usable for audiences with different linguistic backgrounds and levels of domain expertise.

From a *theoretical* point of view, the workshop discusses models of text difficulty, lexical and conceptual complexity, and the interaction between terminology and readability across languages. It promotes reflection on how complexity can be identified, described, and evaluated, and on how multilingual and cross-linguistic perspectives can contribute to more robust and inclusive approaches. From a *practical* standpoint, the workshop considers methods and resources for text simplification, lexical complexity prediction, controllable text generation, and the evaluation of LLM- and GenAI-based systems. Particular attention is devoted to multilingual datasets, interpretable frameworks, domain-specific case studies, and evaluation protocols that support reproducible and accessible research.

The central questions addressed by this workshop concern, first, the linguistic and conceptual factors that make a text difficult and, second, the methods that can be used to assess, control, or reduce such difficulty while preserving meaning and domain adequacy.

This first edition of DeTermIt! 2024 was co-located with the LREC-COLING 2024 joint conference and held in Turin, on May 21, 2024.[1] This second edition of DeTermIt! 2026 is co-located with LREC 2026 and held in Palma de Mallorca, Spain, on May 11, 2026.[2]

The submitted papers went through a double-blind review process that required at least three reviews by members of the international scientific committee. We accepted 9 papers out of 12 submissions (75% acceptance rate): 8 long papers and 1 short paper.

The papers included in these proceedings reflect the variety of current research on text difficulty and simplification. Several contributions focus on multilingual readability and lexical

[1] https://determit2024.dei.unipd.it/

[2] https://determit2026.dei.unipd.it/

complexity, including cross-linguistic analyses of translated children's literature, multilingual lexical complexity prediction for language learning, and calibrated interpretable frameworks for multilingual text difficulty prediction. Other papers address simplification and accessibility in specialized settings, such as biomedical text simplification, patient-centered explanations of gut–brain axis concepts, and the use of small language models as evaluators of simplification quality. The program also includes work on translation as augmentation for difficulty assessment, the automatic generation of graded texts in Old Church Slavonic, and terminology-augmented generation for culturally grounded translation. Overall, these contributions highlight the growing convergence of text simplification, terminology studies, multilingual evaluation, and generative AI.

The keynote speaker of the workshop is Horacio Saggion (Universitat Pompeu Fabra, Barcelona, Spain), who will deliver a talk entitled "Text Simplification as a Tool for Facilitating Democratic Participation in the iDEM Project". The talk addresses the role of accessibility and easy language in supporting democratic inclusion, especially for people with language barriers or limited literacy. It will present work carried out in the Horizon Europe iDEM project, including the creation of language resources and the development of text simplification services integrated into a mobile application for deliberative democratic participation.

We would like to thank all authors for submitting their work, the members of the scientific committee for their careful and constructive reviews, and the LREC 2026 workshop organizers for hosting this second edition of DeTermIt!. We hope that these proceedings will contribute to the ongoing discussion on how terminology-aware, multilingual, and evaluation-driven approaches can support more accessible and effective communication in the age of generative AI.

Giorgio Maria Di Nunzio - University of Padua, Italy
Federica Vezzani - University of Padua, Italy
Liana Ermakova - University of Western Brittany
Hosein Azarbonyad - Elsevier, The Netherlands
Jaap Kamps - University of Amsterdam, The Netherlands

Organizing Committee

General Chairs
Giorgio Maria Di Nunzio - Università degli Studi di Padova, Italy
Federica Vezzani - Università degli Studi di Padova, Italy
Liana Ermakova - Université de Bretagne Occidentale, France
Hosein Azarbonyad - Elsevier, The Netherlands
Jaap Kamps - University of Amsterdam, The Netherlands

Scientific Commitee
Florian Boudin - Nantes University, France
Lynne Bowker - University of Ottawa, Canada
Sara Carvalho - Universidade de Aveiro, Portugal
Rute Costa - Universidade NOVA de Lisboa, Portugal
Eric Gaussier - University Grenoble Alpes, France
Natalia Grabar - CNRS, France
Ana Ostroški Anić - Institute of Croatian Language and Linguistics, Croatia
Tatiana Passali - Aristotle University of Thessaloniki
Grigorios Tsoumakas - Aristotle University of Thessaloniki
Sara Vecchiato - University of Udine, Italy
Cornelia Wermuth - KU Leuven, Belgium

Table of Contents

Cross-linguistic Readability and Controllable Difficulty: A Corpus-Based Comparison of Human and LLM Translations of Children's Literature in Romanian

Karla Csuros, Madalina Chitez, Roxana Rogobete
West University of Timisoara
{karla.csuros, madalina.chitez, roxana.rogobete}@e-uvt.ro

Abstract

Translation can systematically alter text difficulty, particularly when moving into morphologically rich languages. This study examines whether readability-constrained Large Language Models (LLMs) can mitigate difficulty shifts observed in English–Romanian translation of children's literature. We construct a paired four-condition corpus comprising English originals, published Romanian translations, readability-constrained LLM translations, and human readability adaptations (12 aligned passages; $\approx 23,000$ words). Readability is assessed using a Romanian grade-level index (LEMI) designed to be educationally comparable to Flesch-Kincaid Grade Level (FKGL), the cross-linguistic LIX metric, and morphologically informed measures derived from spaCy. Published Romanian translations are significantly more difficult than their English originals, showing higher LIX scores, grade-level estimates, and increased morphological variation. Readability-constrained LLM translation substantially reduces difficulty relative to the published versions (median $\Delta \approx 1.46$ grade levels), with significant decreases in LIX, morphological feature density, and lexical diversity (MTLD). Human adaptation yields a smaller reduction (median $\Delta \approx 0.26$). Although the direct comparison between LLM and human adaptation is marginal ($p = .055$, $r = 0.64$), LLM outputs generally produce larger reductions. These findings demonstrate that translation-induced difficulty shifts are measurable and that controllable LLM translation can modulate readability across structural, lexical, and morphological dimensions in multilingual educational contexts.

Keywords: readability assessment, translation-induced difficulty, controllable text generation

1. Introduction

Readability control is increasingly central to educational NLP, particularly in multilingual contexts where translation mediates access to age-appropriate content. While translation aims to preserve semantic content, it may inadvertently alter linguistic difficulty. In morphologically rich languages such as Romanian, translated texts can exhibit increased syntactic density, inflectional marking, and lexical formalization relative to their source texts. These shifts can raise processing demands for young readers, even when narrative content remains unchanged. Additionally, translated literature occupies a central place in the Romanian children's book market, where a substantial proportion of contemporary titles originate from English-language publishers (Sarbu, 2025; Iovanel, 2022; Cocargeanu, 2015). Translation therefore plays a key role in determining the linguistic accessibility of reading materials available to young Romanian readers. Understanding how translation influences readability is therefore both a linguistic and an educational question.

Recent advances in large language models (LLMs) enable explicit control over stylistic and structural parameters, including readability level. However, empirical evaluation of readability-controlled translation remains limited, especially for languages with high morphological complexity. In particular, it is unclear whether LLM-based translation can mitigate translation-induced difficulty shifts, and whether such mitigation operates through identifiable linguistic mechanisms.

This paper presents a controlled, within-text corpus study examining translation-induced difficulty and readability-constrained generation in Romanian children's literature. We construct a parallel four-condition corpus consisting of: (i) English originals, (ii) published Romanian translations, (iii) readability-controlled LLM translations from English into Romanian, and (iv) human readability adaptations of the published Romanian translations. Each passage ($N = 12$) is represented across conditions, enabling paired statistical comparisons.

We evaluate difficulty using both language-specific and cross-linguistic metrics, including LIX, a custom grade-level index (LEMI) for Romanian, and Flesch–Kincaid Grade Level (FKGL) for English. To move beyond surface readability measures, we additionally compute lexical diversity (MATTR, MTLD) and morphological complexity indicators derived from spaCy's Romanian large model, including average morphological feature density and inflectional diversity (wordform–lemma ratio). Statistical analyses employ paired Wilcoxon signed-rank tests with rank-biserial effect sizes.

The study addresses the following research questions:

RQ1: Do published Romanian translations ex-

Proceedings of the 2nd Workshop on Evaluating Text Difficulty in a Multilingual Context (DeTermIt! 2026), pages 1–11
11 May 2026.

hibit higher readability levels than their English source texts?

RQ2: Can readability-constrained LLM translation significantly reduce translation-induced difficulty?

RQ3: How does LLM-based readability control compare to human readability adaptation?

RQ4: Which linguistic dimensions are associated with observed difficulty shifts?

2. Theoretical Background

The intersection of translation studies, readability assessment, and generative AI provides the theoretical scaffolding for this study. We examine how translation naturally shifts text complexity, how these shifts are measured in morphologically rich languages, and how Large Language Models (LLMs) function as agents of controllable text adaptation.

2.1. Translation Universals and Complexity Shifts

Translation is rarely a neutral act of semantic transfer; it inevitably alters the stylistic and structural properties of the text. Corpus-based translation studies have long posited the existence of *translation universals*, i.e, linguistic features typical of translated texts distinct from original production (Baker, 1993). Two universals are particularly relevant to children's literature: *explicitation* and *normalization*.

Explicitation refers to the tendency of translators to spell out implicit information, often adding conjunctions and explanatory phrases to ensure clarity (Blum-Kulka et al., 1996). While intended to aid comprehension, this process frequently inflates sentence length and syntactic density. In the context of English-to-Romanian translation, this effect is compounded by systemic linguistic differences. As a Romance language with a high degree of inflection, Romanian often requires more morphological markers (e.g., definite articles attached to nouns, elaborate verbal agreement) than the analytic English structure (Pirvulescu, 2002; Ivancu, 2019; Giurgea, 2024). Consequently, a standard translation of a Grade 3 English text may inadvertently shift to a Grade 6 difficulty level in Romanian purely through obligatory grammatical expansion and the formal register often adopted by professional translators.

Cross-linguistic studies (Ciobanu et al., 2015) demonstrate that readability metrics are sensitive to the structural properties of different language families. For example, when examining texts translated into English, source language characteristics systematically influence the target text's readability profile, a phenomenon known as source language interference. Specifically, texts translated from Romance languages into English consistently yield higher Flesch-Kincaid complexity scores than original English texts or texts translated from Germanic source languages. While readability features alone may not always perfectly discriminate between original and translated texts, these metric variations underscore how systemic linguistic differences across language families can systematically impact measured text complexity during the translation process(Ciobanu et al., 2015).

2.2. Readability Assessment in Multilingual Contexts

Quantifying these shifts requires robust metrics. Traditional formulas like the Flesch-Kincaid Grade Level (FKGL) rely heavily on surface features such as sentence length and syllable count (Kincaid et al., 1975). While effective for English, these metrics often fail to capture the complexity of morphologically rich languages. The LIX (Läsbarhetsindex) metric (Björnsson, 1968) offers a more cross-linguistically valid alternative by measuring the proportion of long words (> 6 characters) rather than syllable counts, which vary wildly between Germanic and Romance languages.

However, surface metrics miss the "deep" complexity of text, such as cohesion and morphological load. Recent frameworks like ReaderBench (Dascalu et al., 2017) have moved toward multidimensional analysis, integrating lexical diversity, syntactic depth, and discourse structure. For Romanian, assessing readability requires specific attention to inflectional density, i.e., the ratio of functional morphemes to lexical roots, which significantly impacts cognitive processing load for developing readers.

2.3. Controllable Generation and Text Simplification

The advent of Large Language Models (LLMs) has introduced new paradigms for *controllable text generation*. Unlike Statistical Machine Translation (SMT), which prioritizes fidelity to the source, LLMs can be prompted to satisfy specific stylistic constraints, such as "translate for a 9-year-old" or "limit sentence complexity" (Brown et al., 2020).

Recent work in Automatic Text Simplification (ATS) suggests that LLMs can perform translation and simplification simultaneously (Martin et al., 2020). However, this *trans-adaptation* capability remains under-explored for lower-resource or morphologically complex languages. While LLMs excel at fluency, there is a risk of "stylistic flattening," where the model reduces difficulty by stripping away

the narrative voice or unique cultural markers essential to literary fiction (Maddela et al., 2021). This study addresses this gap by evaluating whether LLMs can balance the preservation of narrative intent with the structural constraints required for accessibility in Romanian.

3. Corpus and Experimental Design

3.1. Source Data

We curated a parallel corpus of 12 passages from contemporary English children's literature and their published Romanian translations. Passages were selected to represent narrative prose suitable for primary and lower-secondary readers. The mean passage length is approximately 506 words (English originals), with minor variation across conditions due to translation and editing. Each passage constitutes an independent experimental unit and is represented across multiple conditions, enabling within-text comparisons. The resulting dataset forms a small but tightly controlled parallel corpus ($\approx 23,000$ words) designed to isolate readability shifts introduced by translation and subsequent adaptation. Because the dataset contains 12 passages, the study should be interpreted as a tightly controlled pilot experiment designed to isolate translation-induced readability shifts rather than to estimate population-level parameters.

3.2. Corpus Conditions

For each passage (identified by a shared code), we constructed four aligned conditions: EN_OG (original English text), RO_OG (published Romanian translation), RO_ED (human readability adaptation of the published Romanian translation), RO_LLM (readability-constrained LLM translation from English into Romanian). This design yields a matched quartet per passage, allowing paired statistical analysis across conditions.

The RO_ED condition was created by a small readability-specialized team through manual adaptation of the published Romanian translations, without access to the original English publication. Editors were instructed to preserve narrative content and plot structure, reduce sentence length where possible, simplify lexical choices, minimize excessive formal register and complex morphological constructions,and maintain natural Romanian style. Edits were conservative and focused strictly on readability, not stylistic rewriting. The human adaptation condition was produced by a team of six editors with experience in Romanian children's literature readability. Each passage was edited independently by one editor and subsequently reviewed by another team member to ensure consistency with the readability guidelines. Disagreements regarding lexical or structural simplifications were resolved through discussion. Because the task involved controlled editing rather than categorical annotation, formal inter-annotator agreement metrics were not computed. This condition serves as a human baseline for readability-oriented adaptation.

The RO_LLM condition was generated by prompting the *ChatGPT 5.2* large language model to translate each English passage into Romanian under explicit readability constraints. The prompt specified the target grade level (identical to the original English version), preservation of narrative content and proper nouns, avoidance of unnecessary formal pronouns and politeness inflation, preference for shorter sentences and high-frequency vocabulary, and prohibition of additions, omissions, or summarization. Translations were produced in a single-pass controlled generation setup. Outputs were manually checked for content fidelity (no omissions or hallucinated content) prior to inclusion in the corpus.

3.3. Experimental Design

The study employs a within-text paired design. Each passage is compared across conditions using matched observations. This design minimizes variance introduced by genre, topic, and narrative content, and enables non-parametric paired statistical testing. Primary comparisons include: translation-induced shifts (RO_OG vs EN_OG), readability-controlled translations (RO_LLM vs RO_OG), human adaptation (RO_ED vs RO_OG), and LLM vs human adaptation (RO_ED vs RO_LLM).

Statistical analysis employs paired Wilcoxon signed-rank tests with rank-biserial effect sizes. For each comparison family (e.g., RO_ED vs RO_LLM), p-values across metrics were adjusted using the Benjamini–Hochberg procedure. The Wilcoxon signed-rank test (Taheri and Hesamian, 2013) is a non-parametric paired comparison test appropriate for small samples. Rank-biserial correlation (r) (Cureton, 1956) provides an effect-size estimate for paired ordinal comparisons. Benjamini–Hochberg correction (van Loon et al., 2017) controls the false discovery rate across multiple tests.

All preprocessing and metric computation were fully automated. The pipeline includes tokenization and morphological annotation using spaCy's RO_CORE_NEWS_LG and EN_CORE_WEB_TRF models, syllable counting via *Pyphen* (with language-specific hyphenation dictionaries), readability, lexical diversity, and morphological complexity metrics computed in Python, and statistical analysis conducted in R using the *tidyverse* and *rstatix* packages. Scripts for corpus processing, metric extraction, and statistical analysis are designed for

reproducibility and can be extended to additional passages or language pairs.

3.4. LLM Text Generation Setup

The readability-constrained translations (RO_LLM) were generated using the *GPT-5.2* large language model in a controlled, single-pass prompting setup. The model was instructed to produce Romanian translations of the English source passages while adhering to explicit readability constraints aligned with the target grade level of each text.

A standardized two-part prompting procedure was used to ensure consistency across passages. First, a fixed system-style instruction defined the model's role as a professional translator specialized in children's literature and readability control, and specified both hard constraints and soft readability guidelines. Second, for each passage, a variable prompt segment provided the target readability level and the source text. Hard constraints included preservation of semantic content (no additions, omissions, or summaries), retention of proper nouns and character names, maintenance of dialogue structure and paragraph segmentation, and avoidance of "politeness inflation" (e.g., introduction of formal pronouns or honorifics not present in the source text). The model was also instructed to preserve narrative tone and stylistic coherence.

Readability control was operationalized through soft constraints encouraging shorter sentence structures, reduced syntactic subordination, preference for high-frequency vocabulary, avoidance of nominalizations and abstract phrasing, and maintenance of clear cohesion. These constraints were framed as guidelines rather than strict rules in order to avoid distortion of meaning or narrative flow.

The prompt additionally required the model to output (i) the Romanian translation and (ii) a short list of 3–6 explicit changes applied to improve readability (e.g., sentence splitting, lexical simplification). While these notes were not included in the quantitative analysis, they were used for qualitative inspection of simplification strategies.

All translations were generated in a single pass with deterministic decoding settings (low temperature) and without post-editing. Outputs were manually checked for content fidelity (i.e., absence of omissions or hallucinated content) prior to inclusion in the corpus.

4. Metrics

We evaluate readability and linguistic complexity using a combination of established readability indices, lexical diversity measures, and morphologically informed metrics. All measures were computed automatically using a fully reproducible pipeline.

4.1. Readability metrics

For Romanian texts, we employ the **LEMI grade-level index** developed within the LEMI readability framework for Romanian children's literature (Chitez et al., 2024). The index integrates structural and lexical components, including average sentence length, proportion of complex (polysyllabic) words, proportion of unique complex word types, and overall lexical diversity. The formulation combines these components into a single grade-level estimate calibrated for Romanian educational contexts. While the specific coefficients of the LEMI index will be detailed in a forthcoming work, we utilize the metric here to provide educationally calibrated grade-level estimates distinct from raw complexity scores.

For English source texts, we compute the **Flesch–Kincaid Grade Level (FKGL)** (Kincaid et al., 1975). FKGL estimates U.S. grade-level difficulty based on sentence length and syllable counts. This metric is used primarily to contextualize the difficulty of the original English passages. The LEMI Romanian grade-level index was designed to approximate grade-level interpretation analogous to the FKGL for English. Both indices estimate school-grade readability difficulty within their respective educational systems and are calibrated to map linguistic features onto grade-level expectations. Although developed independently and trained on language-specific data, their outputs are intended to represent comparable educational grade levels rather than arbitrary complexity scores. We therefore visualize raw grade estimates on a shared axis, interpreting them as educationally aligned measures while acknowledging that linguistic scaling may differ across languages.

A potential concern is the temporal distance between the FKGL formula (1975) and the recently developed LEMI index (2024). Educational expectations, reading habits, and exposure to written language have evolved substantially over this period. In the present study, however, FKGL is used primarily as a contextual reference for the original English passages, and the LEMI index estimates Romanian grade-level readability within its contemporary educational context. To also provide a cross-linguistic readability proxy applicable to both English and Romanian, we compute the **LIX (Läsbarhetsindex)** score (Björnsson, 1968). LIX combines average sentence length and the proportion of long words (more than six characters). Its language-independent design makes it particularly suitable for comparing difficulty shifts across translation conditions.

4.2. Lexical diversity

To capture lexical variation independently of text length, we compute two widely used measures: **MATTR (Moving-Average Type–Token Ratio)** with a window size of 50 tokens (Covington and McFall, 2010), which calculates average type–token ratio across sliding windows, reducing sensitivity to overall text length, and **MTLD (Measure of Textual Lexical Diversity)** (McCarthy and Jarvis, 2010), computed using a standard threshold of 0.72. MTLD segments the text into sequential factors based on declining type–token ratio, producing a length-robust diversity estimate. Both measures provide complementary views of lexical richness and are particularly informative when evaluating simplification effects.

4.3. Morphological Complexity

To capture language-specific structural effects in Romanian, we compute morphologically informed metrics using spaCy's RO_CORE_NEWS_LG model. First, **Morphological Feature Density** is defined as the average number of morphological features assigned per token, excluding punctuation (Bentz et al., 2016). Features include grammatical categories such as case, number, gender, tense, mood, and person. This metric approximates inflectional and agreement load within the text. Morphological feature density is used here as an operational proxy for inflectional and agreement load. Additionally, we compute the **Wordform–Lemma Ratio**, as inflectional diversity is approximated by the ratio of unique surface wordforms to unique lemmas (Kettunen, 2014; Lu, 2012). Higher values indicate greater morphological variation relative to lexical base forms.

Together, these metrics allow us to assess readability shifts at multiple linguistic levels: structural (sentence length), lexical (word complexity and diversity), and morphological (inflectional density). This multidimensional approach supports a more fine-grained analysis of translation-induced difficulty and readability-controlled generation.

5. Results

All statistical comparisons use paired Wilcoxon signed-rank tests with rank-biserial correlation (r) as effect size. Reported p-values are Benjamini–Hochberg adjusted within each comparison family. Medians refer to within-passage deltas.

5.1. Translation-Induced Difficulty

We first examine whether published Romanian translations exhibit higher difficulty than their English originals.

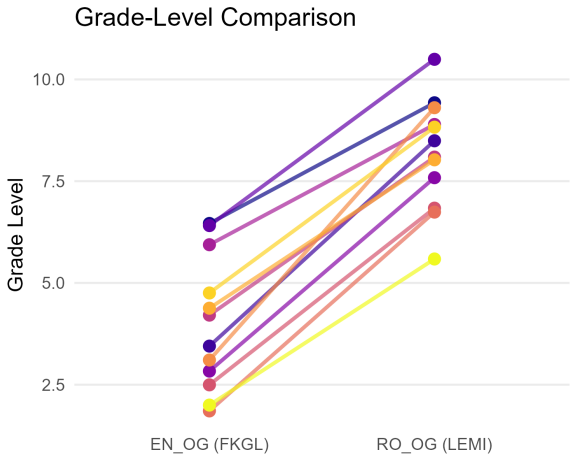


Figure 1: Comparison of English originals (FKGL) and Romanian translations (LEMI)

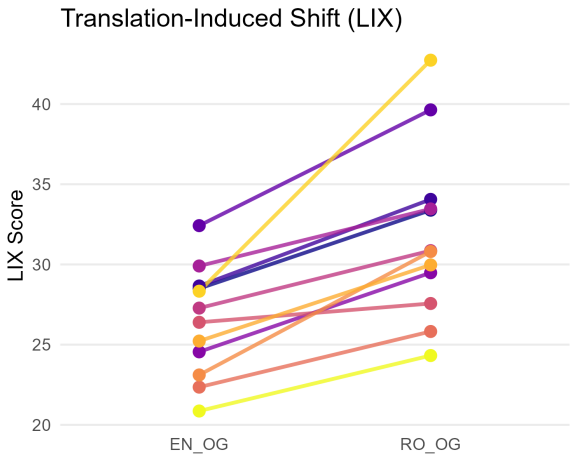


Figure 2: Cross-linguistic comparison of English originals and Romanian translations using LIX

Because FKGL and the LEMI-based index were calibrated independently, comparisons are interpreted in terms of directional and magnitude shifts in educational grade level rather than strict metric identity. As shown in Figure 1, Romanian translations (RO_OG) consistently exhibit higher grade-level estimates than the English originals (EN_OG) across passages, indicating an upward shift in educational readability level.

To provide a strictly cross-linguistic comparison, we examine *LIX* scores, which are computed using the same formula in both languages. Romanian translations show significantly higher LIX values than the English originals (median $\Delta > 0$; $p = .0025$, $r = 1.00$), reflecting increased sentence length and/or a higher proportion of long words. The consistency of this shift is illustrated in Figure 2, where nearly all passages display an upward trajectory from EN_OG to RO_OG.

Morphological complexity patterns further sup-

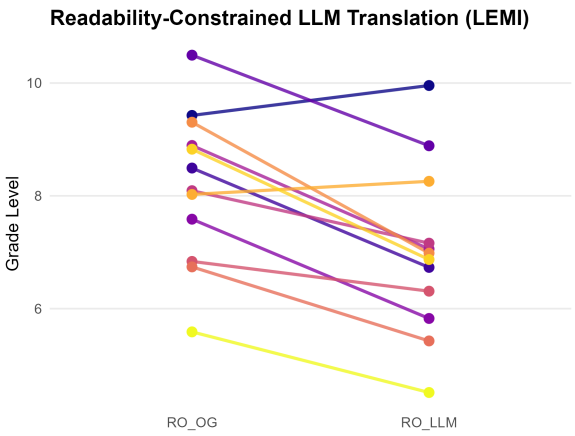


Figure 3: Paired comparison of published Romanian translations and readability-constrained LLM translations using the LEMI grade-level index

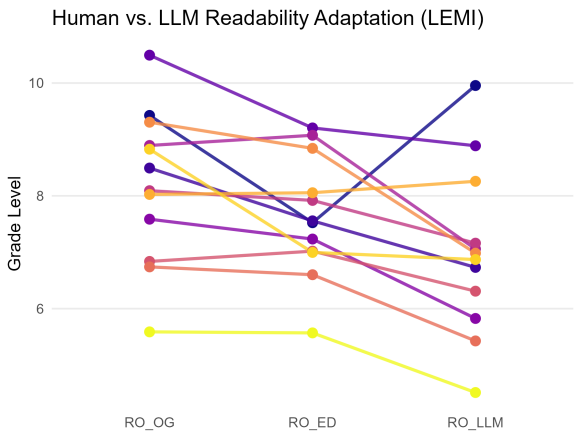


Figure 4: Grade-level comparison across published Romanian translations, human adaptations, and readability-constrained LLM translations

port this finding. Inflectional diversity, measured via the wordform–lemma ratio, is significantly higher in Romanian translations ($p = .0025$, $r = 1.00$), indicating greater morphological variation relative to English source texts.

Collectively, the grade-level indices, cross-linguistic readability scores, and morphological measures converge in showing a systematic translation-induced difficulty shift from English to Romanian in this corpus.

5.2. Readability-Constrained LLM Translation

We next evaluate whether readability-constrained LLM translation reduces difficulty relative to the published Romanian translations.

As illustrated in Figure 3, LLM translations (RO_LLM) consistently reduce grade-level estimates compared to the published Romanian versions (RO_OG). The LEMI-based index shows a median within-passage decrease of $\Delta = -1.46$ grade levels ($p = .0068$, $r = -0.90$), indicating a substantial downward shift in educational readability level.

Cross-linguistic readability patterns converge with this finding. LIX scores are significantly reduced in RO_LLM relative to *ro_og* ($p = .0033$, $r = -0.97$), reflecting shorter sentences and/or a lower proportion of long words.

At the morphological level, LLM translations exhibit significantly lower morphological feature density per token ($p = .0207$, $r = -0.77$), suggesting a measurable reduction in inflectional and agreement load. Lexical diversity, as measured by *MTLD*, is also significantly lower in RO_LLM ($p = .0167$, $r = -0.79$), consistent with vocabulary simplification.

Overall, these results indicate that readability-constrained LLM translation substantially attenuates translation-induced difficulty across structural, lexical, and morphological dimensions. The consistency of effect sizes across all 12 paired samples indicates a robust pattern of difficulty shifts, though larger-scale validation is required to confirm generalizability.

5.3. Human Adaptation vs. LLM Translation

We further compare human readability adaptations (RO_ED) to both the published translations and the readability-constrained LLM outputs.

Relative to the published Romanian translations (RO_OG), human adaptations yield a modest but statistically significant reduction in grade-level estimates (LEMI-based index; median $\Delta \approx -0.26$; $p = .045$, $r = -0.67$). *LIX* scores likewise decrease significantly ($p = .0053$, $r = -0.92$), indicating structural simplification.

However, the magnitude of reduction differs across adaptation types. LLM translations (*ro_llm*) exhibit a substantially larger median decrease relative to RO_OG (median $\Delta \approx -1.46$), as reported in the previous section. Direct comparison reveals that LLM translations yielded lower grade-level estimates than human adaptations (median difference ≈ 0.79 grades). While this difference fell just short of the conventional significance threshold ($p = .055$), the large effect size ($r = 0.64$) suggests a substantive divergence in simplification magnitude, likely limited by the statistical power of the small sample size ($N = 12$).

Figure 4 visualizes the trajectory across conditions. Most passages show a small decrease from RO_OG to RO_ED, followed by a larger decrease in RO_LLM. While individual passages vary, including one clear outlier, the overall pattern indicates that readability-constrained LLM translation tends

to reduce grade-level estimates more strongly than manual editing in this dataset.

Altogether, these findings suggest that controlled LLM translation not only approximates but, in this sample, often exceeds the magnitude of readability reduction achieved through human adaptation. At the same time, the marginal direct comparison ($p = .055$) indicates that differences between LLM and human editing should be interpreted cautiously.

5.4. Linguistic Mechanisms

To examine the linguistic mechanisms underlying the observed difficulty shifts, we analyze morphological and lexical complexity metrics.

At the morphological level, readability-constrained LLM translations (RO_LLM) exhibit a significant reduction in morphological feature density relative to the published translations (*ro_og*) ($p = .0207$, $r = -0.77$). This decrease suggests a measurable reduction in inflectional marking and agreement complexity in the LLM outputs. In contrast, human adaptations (*ro_ed*) do not show a statistically significant reduction in morphological feature density relative to *ro_og*, indicating that manual editing does not systematically reduce inflectional load in the same way.

Lexical diversity patterns show a similar asymmetry. *MTLD* decreases significantly in RO_LLM relative to RO_OG ($p = .0167$, $r = -0.79$), consistent with vocabulary simplification. Differences between RO_ED and RO_LLM on lexical diversity measures are limited and not statistically robust.

These findings suggest that readability-constrained LLM translation reduces difficulty not only through structural simplification (e.g., shorter sentences), but also through systematic reductions in morphological density and lexical diversity. In this dataset, the stronger grade-level reductions observed for RO_LLM appear to be accompanied by measurable changes in inflectional complexity, an effect particularly relevant for morphologically rich target languages such as Romanian.

5.5. Qualitative Analysis of Simplification Strategies

To better understand the linguistic mechanisms driving the observed quantitative shifts, we conducted a qualitative inspection of aligned sentence trios. This analysis reveals distinct simplification strategies between human editors and the readability-constrained LLM.

While human editors occasionally perform structural interventions, they often leave morphological and idiomatic complexities intact. Table 1 illustrates this contrast using an excerpt from *Inkling* (Code: `013-KO-PT`).

In this example, the human editor explicitly attempts to lower difficulty by splitting the long sentence into two shorter ones (*"...de umbre. Era cu ochii..."*). This effectively reduces sentence length, a key component of the LEMI and LIX metrics. However, the editor retains the *Perfect Simplu* tense (*se strecură*), a literary form often challenging for younger readers, and preserves the complex idiom *"cu ochii în patru"* (literally "with eyes in four").

The LLM, conversely, targets all three dimensions of difficulty. Structurally, it condenses the phrasing without splitting. Morphologically, it shifts from the literary *Perfect Simplu* to the standard conversational *Perfect Compus* (*a alunecat*), significantly lowering inflectional density. Lexically, it dissolves the idiom into a direct adjective (*atent* / "careful"). This demonstrates that while human editing in this corpus was primarily *remedial* (fixing length), the LLM generation was *transformative* (altering register and encoding).

6. Discussion

This study investigated translation-induced readability shifts and the effects of readability-constrained LLM translation in Romanian children's literature. The findings yield three main insights regarding the mechanics of automated simplification in morphologically rich languages.

First, published Romanian translations were consistently more difficult than their English originals. This pattern was confirmed across cross-linguistic (*LIX*), language-specific grade-level indices, and morphological variation measures. The convergence of structural and morphological indicators suggests that translation into a morphologically rich language may systematically increase surface complexity, even when semantic content is preserved. In this corpus, translation-induced difficulty appears to be driven not only by longer sentences or lexical choices, but also by increased inflectional diversity, which is a factor often overlooked in standard readability formulas.

Second, readability-constrained LLM translation substantially attenuated this difficulty shift. The median grade-level reduction relative to published translations was approximately 1.46 grade levels, accompanied by significant reductions in *LIX*, morphological feature density, and lexical diversity (*MTLD*). These results indicate that controlled LLM generation operates across multiple linguistic dimensions simultaneously. As observed in the qualitative analysis, the LLM did not merely shorten sentences; it actively normalized literary tenses (e.g., shifting from *perfect simplu* to *perfect compus*) and dissolved complex idiomatic constructions. This measurable decrease in morphological feature density suggests that LLM-based simplification can

Table 1: Comparison of Human vs. LLM adaptation strategies in passage `013-KO-PT`.

Condition	Text Segment	Key Features
English Original	"Along the shadowy hallway, he slid cautiously, keeping an eye out for Rickman, who also kept night hours."	Participial phrase (*keeping an eye out*); idiom; descriptive clause.
Published Translation	*"Se strecură prudent, de-a lungul holului plin de umbre, fiind cu ochii în patru să nu apară de undeva Rickman..."*	**Literary tense** (*Perfect Simplu*: *strecură*); **Complex idiom** (*ochii în patru*); **Syntactic expansion** (added sub-clause *să nu apară...*).
Human Adaptation	*"Se strecură atent, de-a lungul holului plin de umbre. Era cu ochii în patru să nu apară de undeva Rickman..."*	**Sentence splitting**; **Lexical edit** (*prudent* $\rightarrow$ *atent*); Retention of literary tense and idiom.
LLM Translation	*"A alunecat cu grijă pe holul întunecat, atent la Rickman, care era și el treaz noaptea."*	**Tense normalization** (*Perfect Compus*: *a alunecat*); **De-idiomatization** (*ochii în patru* $\rightarrow$ *atent*); **Syntactic compression**.

effectively target the specific inflectional burdens that characterize Romance languages.

Third, human adaptation and LLM simplification differed fundamentally in magnitude and strategy. Manual editing produced a modest median reduction (approximately 0.26 grade levels), whereas LLM outputs exhibited larger shifts. While the direct statistical comparison between human and LLM adaptation approached significance ($p = .055$), the substantial effect size ($r = 0.64$) points to a distinct divergence in approach. Human editors adhered to a *conservative* strategy, prioritizing fidelity and limiting interventions to local sentence splitting or lexical substitution. In contrast, the readability-constrained LLM adopted an *aggressive* structural strategy, rewriting entire syntactic architectures to meet the target grade level. The statistical margin likely reflects the small sample size ($N = 12$) rather than an absence of effect; the qualitative evidence confirms that the LLM performed types of simplification (morphological reduction) that human editors largely avoided.

These findings have implications for readability control in multilingual education. While the specific morphological effects observed here are tied to Romanian, the paired-corpus methodology and readability-controlled generation pipeline can be replicated for other language pairs. In languages such as Romanian, morphological complexity contributes substantially to processing load. Automated simplification systems that modulate inflectional density may therefore have a disproportionate impact on perceived difficulty compared to systems that only target sentence length. At the same time, the aggressive reduction of lexical diversity and morphological marking raises questions about the balance between accessibility and stylistic richness. While the LLM outputs were more readable, they occasionally sacrificed the literary register preserved by human editors.

From an NLP perspective, the results contribute empirical evidence that prompt-based control can achieve consistent difficulty reductions within a translation setting, unlike traditional post-hoc simplification pipelines. This suggests potential for integrated translation-and-adaptation workflows in educational publishing.

These findings have implications for readability control in multilingual education. While the specific morphological effects observed here are tied to Romanian, the paired-corpus methodology and readability-controlled generation pipeline can be replicated for other language pairs. In languages such as Romanian, morphological complexity contributes substantially to processing load. Automated simplification systems that modulate inflectional density may therefore have a disproportionate impact on perceived difficulty compared to systems that only target sentence length. At the same time, the aggressive reduction of lexical diversity and morphological marking raises questions about the balance between accessibility and stylistic richness. While the LLM outputs were more readable, they occasionally sacrificed the literary register preserved by human editors.

Several limitations must be acknowledged. The corpus is small ($N = 12$) and restricted to a single genre. While effect sizes are large and directionally consistent, replication on larger and more diverse datasets is necessary. Additionally, this study evaluates readability through automated indices; it does not include comprehension testing or human judgment of pedagogical appropriateness. Future work should incorporate reader-based validation to determine if the morphological reductions achieved by the LLM correspond to improved comprehension outcomes in young readers.

The findings demonstrate that translation-induced difficulty shifts are measurable in a morphologically rich target language, and that readability-

constrained LLM translation can mitigate these shifts through structural and morphological interventions that exceed the scope of standard human editing.

7. Conclusion

This study examined translation-induced readability shifts and the effects of readability-constrained LLM translation in Romanian children's literature. Using a paired four-condition design, we showed that published Romanian translations are consistently more difficult than their English originals across cross-linguistic (*LIX*), grade-level (FKGL/LEMI), and morphological measures, confirming that translation into a morphologically rich language can increase structural and inflectional complexity. Readability-constrained LLM translation substantially attenuated this shift, yielding a median reduction of approximately 1.46 grade levels relative to the published versions, alongside significant decreases in *LIX*, morphological feature density, and lexical diversity (*MTLD*). Human adaptation produced a smaller median reduction (approximately 0.26 grade levels). Although the direct comparison between LLM and human adaptation was marginal ($p = .055$, $r = 0.64$), LLM outputs generally achieved larger reductions. Methodologically, the study contributes a reproducible framework combining paired experimental design with morphologically informed readability metrics. These findings suggest that controllable LLM translation can meaningfully modulate readability, including inflectional complexity, in multilingual educational contexts.

8. Acknowledgments

We would like to thank **Quarticle Research SRL** for contributing to the technical integration of the automatic readability assessment interface within the LEMI platform (https://www.lemi.ro), which was developed by the authors. The functionality was employed for the automated computation of readability and linguistic complexity metrics in Romanian in the present study.

9. Bibliographical References

Mona Baker. 1993. Corpus linguistics and translation studies — implications and applications. In *Text and Technology*. John Benjamins.

Christian Bentz, Tatyana Ruzsics, Alexander Koplenig, and Tanja Samardžić. 2016. A comparison between morphological complexity measures: Typological data vs. language corpora. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 142–153, Osaka, Japan. The COLING 2016 Organizing Committee.

C.H. Björnsson. 1968. *Läsbarhet:*. Pedagogiskt Utvecklingsarbete vid Stockholms Skolor. 6. Liber; [Solna, Seelig].

Shoshana Blum-Kulka, Shoshana Blum-Kulka, and Julian House. 1996. Shifts of cohesion and coherence in translation.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Madalina Chitez, Mihai Dascalu, Aura Cristina Udrea, Cosmin Striletchi, Karla Csürös, Roxana Rogobete, and Alexandru Oravițan. 2024. 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, Torino, Italia. ELRA and ICCL.

Alina Maria Ciobanu, Liviu P. Dinu, and Flaviu Pepelea. 2015. Readability assessment of translated texts. In *Recent Advances in Natural Language Processing*.

Dana-Mihaela Cocargeanu. 2015. *Children's literature across space and time: the challenges of translating Beatrix Potter's tales into Romanian*. Ph.D. thesis, School of Applied Language and Intercultural Studies, Dublin City University.

Michael A. Covington and Joe D. McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of Quantitative Linguistics*, 17(2):94–100.

Edward E Cureton. 1956. Rank-biserial correlation. *Psychometrika*, 21(3):287–290.

Mihai Dascalu, Philippe Dessus, Maryse Bianco, Stefan Trausan-Matu, and Jean-Luc Nespoulous. 2017. Readerbench: A multi-lingual framework for analyzing text complexity. *Educational Psychology Review*, 29:495–525.

Ion Giurgea. 2024. Article drop and case marking in romanian. *Isogloss J. Var. Roman. Iber. Lang.*, 10(2):1–23.

Mihai Iovanel. 2022. Children's literature in romania. *Transylvanian Review*, 31(1).

Ovidiu Ivancu. 2019. A pedagogical perspective on the definite and the indefinite article in the romanian language. challenges for foreign learners. *Verbum*, 10:1.

Kimmo Kettunen. 2014. Can type-token ratio be used to show morphological complexity of languages? *J. Quant. Linguist.*, 21(3):223–245.

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

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. *Mod. Lang. J.*, 96(2):190–208.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553, Online. Association for Computational Linguistics.

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.

Philip M. McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42:381–392.

Mihaela Pirvulescu. 2002. Morphological paradigms and the role of tense. *Revue québécoise de linguistique*, 31(2):77–88.

Simina Sarbu. 2025. 12 cărți pentru copii premiate: Bestsellers în românia 2026. *Drimoland*.

SM Taheri and Gholamreza Hesamian. 2013. A generalization of the wilcoxon signed-rank test and its applications. *Statistical Papers*, 54(2):457–470.

Wouter van Loon, JJ Goeman, M Van Iterson, and M Fiocco. 2017. The power of the benjamini-hochberg procedure. *Leiden, The Netherlands: Leiden University*.

10. Appendices

10.1. LLM Prompting Protocol

The following base prompt was used to define the model's role and constraints. This instruction remained constant across all passages.

> You are a professional translator (English → Romanian) specialized in children's literature and readability control.
>
> Your task: translate the provided English text into Romanian while targeting a specified Romanian school grade level (e.g., grade 3–4). Maintain meaning, tone, and narrative voice.
>
> Hard constraints:
>
> - Do NOT add new content, explanations, or summaries.
> - Do NOT omit content.
> - Preserve character names and proper nouns exactly as in the source unless a standard Romanian form is conventional.
> - Preserve dialogue structure and paragraph breaks as much as Romanian norms allow.
> - Avoid "politeness inflation": do not introduce honorifics, formal pronouns (e.g., dumneavoastră), or extra politeness markers unless explicitly present in the source.
> - Keep register natural and age-appropriate.
>
> Readability controls (apply gently, without distorting meaning):
>
> - Prefer shorter sentences where possible; avoid heavy subordination.
> - Prefer common, high-frequency Romanian words; avoid rare/Latinate synonyms.
> - Avoid nominalizations and overly abstract phrasing when a simpler verb phrase works.
> - Keep cohesion clear (explicit subjects where needed, but don't over-repeat).
> - Keep the style literary/narrative (not textbook-like).
>
> Output format, no extra commentary, return ONLY:
>
> 1. TRANSLATION: <Romanian translation>

2. NOTES: bullet list of 3–6 concrete changes you made to target readability (e.g., “split a long sentence”, “replaced rare word X with common Y”).

For each passage, the base instruction was combined with the following variable prompt specifying the target readability level and the source text:

> Target readability: Romanian grade [X] (Romanian school system).
>
> Translate the text below from English into Romanian following the constraints.
>
> TEXT: « source passage »

The target grade level [X] was set to match the estimated readability level of the English source passage (FKGL), ensuring alignment between source difficulty and translation constraints.

All translations were generated using the GPT-5.2 model in a controlled, single-pass setup with deterministic decoding (low temperature), without iterative refinement or post-processing. All outputs were manually checked to ensure content fidelity, including the absence of omissions, additions, or hallucinated content. In addition to the translation, the model produced a short list of 3–6 notes describing the simplification strategies applied (e.g., sentence splitting, lexical simplification). These notes were not included in the quantitative analysis but were used to support qualitative inspection of simplification strategies (see Section 5.5).

10.2. Source Texts

The full dataset is avaiable on *HuggingFace* via https://huggingface.co/datasets/karlacsuros/lemitranslations

Table 2: Source texts included in the corpus

Code	Title	Author	Year
011-RG-DT	My Father’s Dragon	Ruth Stiles Gannett	1948
013-KO-PT	Inkling	Kenneth Oppel	2018
018-RT-CS	Who in the World is Carmen Sandiego	Rebecca Tinker	1998
019-PG-VC	Treasure Hunters	James Patterson	2013
020-JK-JP	Diary of a Wimpy Kid: Diper Överlöde	Jeff Kinney	2022
022-JB-DT	The Terrible Two	Jory John & Mac Barnett	2015
023-AD-NK	Ninja Kid #1: From Nerd to Ninja	Anh Do	2019
028-LP-TG	The Brilliant World of Tom Gates	Liz Pichon	2011
029-DW-MI	Awful Auntie	David Walliams	2014
057-JB-CB	Little Book of Bob	James Bowen	2018
060-EC-MR	Mariella, Queen of the Skies	Eoin Colfer	2018
061-SS-FF	Mighty Murphy	Shelley Swanson Sateren	2016

A Benchmark for Overgeneration Detection in Biomedical Text Simplification

Berkay Chakar[†], **Liana Ermakova**[‡], **Jaap Kamps**[†]
[†]ILLC, University of Amsterdam, The Netherlands, berkay.chakar2@student.uva.nl, kamps@uva.nl
[‡]HCTI, University of Brest, France, liana.ermakova@univ-brest.fr

Abstract

Large Language Models deployed for biomedical text simplification frequently produce *overgeneration*: extraneous content appended beyond the faithful simplification, including leaked model instructions, ungrounded medical claims, and repetitive or redundant text. Despite its prevalence, this failure mode remains largely unaddressed. We present a benchmark for document-level overgeneration detection, releasing two resources: **SimpleOG-manual**, 500 abstract-level examples with human-validated positive labels, and **SimpleOG-auto**, over 46,000 automatically labeled abstract-level examples derived from submissions to the CLEF 2025 SimpleText Track. Our method exploits the positional regularity of overgeneration in simplification output through sequence alignment, identifying trailing content that lacks a corresponding segment in the source. Human validation of 117 automatically flagged positives confirms ∼95% precision, with leaked model instructions accounting for 75.7% of confirmed cases. Analysis across teams and models reveals that overgeneration is primarily driven by system-level choices, such as prompting and post-processing, rather than by model architecture. We evaluate three detection paradigms and find that sentence similarity (F1 = 0.732, ROC-AUC = 0.921) surprisingly outperforms both NLI-based and LLM-based approaches, suggesting that overgenerated content occupies distinct semantic regions from source material.

Lay Summary: *We investigate a common failure of AI models used for simplifying or summarizing scientific text: overgeneration, where models add extraneous content such as leaked instructions, ungrounded statements, or repetitive text that is not supported by the source. We introduce a benchmark for document-level overgeneration detection and release two resources: SimpleOG-manual, 500 abstract-level examples with human-validated positive labels, and SimpleOG-auto, over 46,000 automatically labeled abstract-level examples derived from CLEF SimpleText 2025 Track submissions.*

Keywords: Natural Language Generation, Text Simplification, Large Language Models, Overgeneration

1. Introduction

Large Language Models (LLMs) have become the dominant approach to text simplification, offering the ability to transform complex technical content into accessible language while aiming to preserve semantic fidelity (Li et al., 2024). This capability is particularly critical in the biomedical domain, where dense clinical literature remains difficult to understand for patients and the general public. The SimpleText shared task at CLEF (Ermakova et al., 2025) has driven this research forward, with the Cochrane Library (Bakker and Kamps, 2024), a collection of high-quality, evidence-based systematic reviews, as its primary testbed (Bakker and Kamps, 2024, 2025). The 2025 iteration challenged participants to simplify these abstracts using systems ranging from fine-tuned encoder-decoder models (BART, T5) to state-of-the-art LLMs (GPT-4, LLaMA-3, Gemini, Mistral).

However, deploying LLMs for text simplification introduces a critical but understudied failure mode: *hallucination*. While hallucination has been extensively investigated in abstractive summarization (Maynez et al., 2020; Kryscinski et al., 2020) and question answering (Li et al., 2023), its manifestation in text simplification remains largely unexplored. Unlike summarization, where the model must compress and select information, simplification requires preserving the full semantic content while transforming linguistic complexity. This fundamental difference creates a distinct failure profile: rather than fabricating facts within the output, simplification models predominantly exhibit *overgeneration*, appending extraneous content such as leaked instructions, unsolicited advice, or ungrounded elaborations beyond the faithful simplification. While overgeneration falls under the broader umbrella of hallucination, the two differ operationally: hallucination detection typically targets factual contradictions or unsupported claims *within* the generated text, whereas overgeneration detection targets extraneous content *appended beyond* the faithful output. We focus on the latter, which we detect via positional analysis of trailing content (Section 3.2). Figure 1 illustrates this at the abstract level: a model faithfully simplifies all three source sentences but leaks its internal strategy between them.

Through automatic analysis of over 707,000 sentence-level simplification pairs from SimpleText 2025 submissions, we sampled 500 abstract-level examples into a curated test set (SimpleOG-manual), of which 117 were automatically flagged as positive by our trailing-content method. Man-

Proceedings of the 2nd Workshop on Evaluating Text Difficulty in a Multilingual Context (DeTermIt! 2026), pages 12–21
11 May 2026.

Abstract-level overgeneration example

Source: *"~~The~~ functionality ~~of the Internet and the World Wide Web is determined in large part by the standards that allow for interoperable implementations, as a result, the privacy of our online interactions depends on the work done within standard-setting~~ organizations. ~~But how~~ do the ~~organizational structure and processes of these multistakeholder~~ groups ~~affect the engineering of values such as~~ privacy~~?~~ This paper ~~reviews the history of considerations for~~ security ~~and privacy in~~ Internet ~~and Web standard-setting, the impact of Snowden surveillance revelations and reactions to them, and some trends in how we review for privacy in Internet and Web~~ standards."*

Prediction: *'The internet's functionality and privacy depend on standards set by certain organizations. (I chose 'rephrase' as the internal simplification strategy) How do the groups that set Internet standards impact privacy. This paper looks at how privacy and security are considered when creating Internet standards."*

~~Strikethrough~~ = deletion Green = insertion Red = trailing content

Figure 1: Abstract-level overgeneration: a 3-sentence source is simplified correctly, but the model leaks its internal strategy between sentences 1 and 2 (red). Annotations show token-level alignment.

ual review of these 117 positives confirmed 111 as true overgeneration (∼95% precision), revealing four primary categories: leaked model instructions (75.7%), ungrounded information injection, repetitive content, and other failures with leaked instructions as the dominant pattern (discussed in Section 4). We term these extraneous additions the *"chatter"* problem: content that, while often linguistically fluent, violates the fundamental faithfulness requirement of text simplification. We also observed instances of *factual distortion* during qualitative analysis, where models alter numbers or invert relationships within the simplified text itself; this falls outside the scope of our trailing-content method and remains an important direction for future annotation.

In this paper, we present two contributions: (1) A benchmark for detecting overgeneration in biomedical text simplification at the abstract level, released as two resources: SimpleOG-manual (500 examples with human-validated positive labels) and SimpleOG-auto (over 46,000 automatically labeled abstract-level examples from 666 source abstracts across all system runs). We exploit the sentence-level structure of shared task submissions to construct reliable automatic labels via trailing-content detection at the sentence level, then aggregate to the document level: an abstract is labeled positive if any of its constituent sentences contains overgeneration (Section 3). (2) A comparative evaluation of three detection paradigms (sentence similarity, natural language inference, and LLM-based classification), finding that, surprisingly, sentence similarity outperforms both more complex approaches.[1]

2. Related Work

Hallucination in NLG. Factual inconsistency in neural text generation has been primarily studied in abstractive summarization. Maynez et al. (2020) provided a comprehensive analysis of faithfulness errors in summaries, distinguishing intrinsic hallucinations (contradicting the source) from extrinsic ones (introducing unsupported content). Automated detection methods include FactCC, a BERT-based entailment model trained on synthetic data (Kryscinski et al., 2020), and SummaC, which reframes consistency checking as an NLI task applied at the sentence level (Laban et al., 2022). The evaluation of FactCC (Kryscinski et al., 2020), QAGS (Wang et al., 2020), FEQA (Durmus et al., 2020), and FactAcc (Goodrich et al., 2019) measures on the TRUE (Honovich et al., 2022) benchmark, which aggregates 11 such datasets covering summarization, dialogue, paraphrasing, and fact verification, showed low F1 and AUPRC (Vendeville et al., 2025). For LLM outputs more broadly, Li et al. (2023) introduced HaluEval, a benchmark covering hallucinations in QA, dialogue, and summarization. However, all these resources target summarization or general-purpose generation. Text simplification, where rephrasing and synonym substitution are expected rather than erroneous, poses distinct detection challenges that these tools do not address. In particular, overgeneration in simplification often follows a recognisable positional pattern, with extraneous content appearing as trailing material after the faithful output, which enables detection through alignment-based methods that complement the entailment and classification approaches used for general hallucination.

Text Simplification Evaluation. Standard evaluation of text simplification relies on metrics such as SARI (Xu et al., 2016), which measures the quality of lexical edits (additions, deletions, and kept words) against reference simplifications. While SARI captures simplification quality, it does not assess faithfulness to the source. Davari et al. (2024) showed that automatic measures such as BERTScore (Zhang et al., 2020), BETS (Zhao et al., 2023), BLEU (Papineni et al., 2002), SARI (Xu et al.,

[1] Code and data are available at https://github.com/chakarberkay/og-benchmark.

2016), FKGL (Kincaid et al., 1975), LENS (Maddela et al., 2023), and others have low correlation with humanly annotated meaning preservation on 9 datasets. A detailed taxonomy of errors in text simplification were proposed in the SALTED benchmark (Vendeville et al., 2025). The evaluation of FactCC (Kryscinski et al., 2020), QAGS (Wang et al., 2020), FEQA (Durmus et al., 2020), and FactAcc (Goodrich et al., 2019) measures on this dataset showed low F1 and AUPRC (Vendeville et al., 2025). The most directly related work is Devaraj et al. (2022), who introduced a taxonomy of factuality errors in text simplification (insertion, deletion, and substitution) and found that such errors are common yet uncaptured by existing metrics. Our work extends this direction in two ways: we focus specifically on overgeneration (trailing content) rather than within-text factual errors, and we construct an automatically labeled benchmark from a large pool of shared task submissions rather than relying on manual annotation alone.

SimpleText Shared Task. The SimpleText track at CLEF (Ermakova et al., 2025) promotes research on making scientific texts accessible. Task 1 addresses sentence-level and document-level simplification of biomedical abstracts from the Cochrane Library, using the Cochrane-auto corpus (Bakker and Kamps, 2024). Our benchmark is built from the 2025 edition of Task 1 submissions, which introduced a Task 2.3: detecting overgeneration in the simplified outputs. This adds a faithfulness evaluation dimension that complements the existing quality-oriented metrics used in the shared task.

3. Dataset Construction

3.1. Source Data

Our dataset is built from submissions to the SimpleText 2025 shared task (Ermakova et al., 2025) (Task 1.1), which requires sentence-level simplification of biomedical abstracts from the Cochrane Library. The source data consists of 666 abstracts comprising 9,160 sentences drawn from systematic reviews. Participating teams submitted simplified versions using a range of models: LLaMA-3, GPT-4, GPT-3.5, BART, T5, Gemini, and Mistral. In total, we processed 707,898 sentence-level pairs, each consisting of a source sentence and the corresponding simplification produced by one system, from 70 system runs across 15 teams. While reference simplifications are available for the source data (enabling evaluation with standard metrics such as SARI (Xu et al., 2016)), we focus on reference-free overgeneration detection, as our goal is to identify extraneous content rather than assess overall simplification quality. Table 1 summarises the source data.

Statistic	Value
Source abstracts	666
Source sentences	9,160
System runs	70
Sentence-level pairs	707,898
SimpleOG-auto (abstract-level)	∼46,000
SimpleOG-manual (abstract-level)	500

Table 1: Source data and dataset overview.

3.2. Trailing-Content Detection

Our key observation is that overgeneration in text simplification frequently manifests as *trailing content*, text appended after the legitimate simplification that has no corresponding segment in the source. We exploit this positional regularity through a sequence alignment algorithm, noting that our method does not capture mid-text overgeneration (e.g., inserted clauses), which we leave to future work.

Algorithm. Given a source sentence and its simplified prediction, we (1) tokenize both using NLTK[2] `word_tokenize`, (2) align them with `difflib.SequenceMatcher`[3] to find the longest contiguous matching blocks, and (3) identify any unmatched prediction tokens that appear after the final aligned block as trailing content. Tokens that do not have matches are classified as:

- **Deletions** (source tokens absent from prediction): expected in simplification, where complex details are removed.
- **Mid-text insertions** (prediction tokens with further matches ahead): expected, as rephrasing introduces new wording.
- **Trailing content** (prediction tokens after the final alignment match): overgeneration candidate.

Not all trailing content constitutes overgeneration: short trailing spans often arise from alignment noise rather than genuine extraneous content. Manual inspection of 30 randomly sampled trailing segments revealed that spans below 25 characters consisted almost entirely of punctuation differences, formatting artifacts, or a few word additions from rephrasing that do not constitute meaningful

[2] https://www.nltk.org/

[3] https://github.com/python/cpython/blob/3.14/Lib/difflib.py

Example 1: Leaked model instructions

Source: *"~~Thus, implementing~~ crowdsourcing ~~applications~~ while ~~preserving the privacy of~~ workers' ~~location~~ is a key issue ~~that needs to be tackled.~~"*

Prediction: *"A key issue is keeping workers' locations private while using crowdsourcing apps.*

(I chose 'rephrase' as the internal simplification strategy)

Trailing content: 59 chars $\geq$ 25 $\rightarrow$ OVERGENERATION

Example 2: Ungrounded information injection

Source: *"SGLT2 inhibitors ~~with or without other background~~ treatments were compared to placebo, ~~standard care,~~ sulfonylurea, ~~DPP-4 inhibitors,~~ or insulin."*

Prediction: *"Medicines called SGLT2 inhibitors were compared to a fake treatment (placebo), usual treatment, sulfonylurea [...], or insulin (a hormone that controls blood sugar).*

These comparisons were made whether or not the patients were also taking other medicines.

Trailing content: 89 chars $\geq$ 25 $\rightarrow$ OVERGENERATION

~~Strikethrough~~ = deletion Green = insertion Red = trailing content

Figure 2: Overgeneration examples from the dataset. Example 1 shows leaked task instructions; Example 2 shows a plausible but unsupported medical statement. Annotations illustrate the three token classifications from our alignment algorithm.

overgeneration. At or above 25 characters, trailing spans consistently contained recognisable overgeneration closer to full-sentence material, such as leaked instruction fragments or injected claims. We therefore apply a character-length threshold of 25 to separate noise from substantive trailing content. Spans below the threshold are retained in the data but not labeled as overgeneration; spans at or above the threshold are labeled as overgeneration candidates. We note that this threshold was calibrated for English biomedical text and may require adjustment for other languages or domains. Figure 2 illustrates this process on two real examples from our dataset.

3.3. Document-Level Aggregation

Although the shared task submissions are sentence-level, our benchmark targets *document-level* overgeneration detection: given a full source abstract and its simplified version, the task is to determine whether the simplification contains overgeneration. We exploit the sentence-level structure solely for automatic label construction, where detecting trailing content is most reliable. We then stitch sentence-level detections into document-level labels: an abstract is labeled POSITIVE if *any* of its constituent sentences contains overgeneration. This OR-rule is deliberately strict: even a single overgenerated sentence renders the entire abstract positive. While this loses severity gradation (a minor trailing phrase is weighted equally to a full injected paragraph), we argue that any unsupported content in a biomedical simplification is potentially misleading and warrants flagging. We note that sentence-level false positives propagate to the document level under this rule, but the high precision of our trailing-content method ($\sim$95%, Section 4) limits this effect in practice. The benchmark also provides sentence-level annotations, enabling future work on graded or severity-aware detection.

Model	Count	Pos.	Rate
LLaMA-3-8b	179	71	39.7%
LLaMA-3-70b	159	29	18.2%
BART	67	1	1.5%
Mistral	32	3	9.4%
Gemini	27	3	11.1%
GPT-4	21	0	0.0%
GPT-3.5	8	0	0.0%
T5	7	4	57.1%
Total	**500**	**111**	**22.2%**

Table 2: Dataset distribution by model family. Overgeneration rates vary from 0% (GPT-3.5, GPT-4) to 57.1% (T5).

3.4. Sampling and Statistics

The full processing yields automatic overgeneration labels for 666 source abstracts across 70 system runs. Each (abstract, system-run) pair constitutes one example, producing over 46,000 abstract-level examples. We release this data as two resources: **SimpleOG-auto**, the complete pool of automatically labeled abstract-level examples, and **SimpleOG-manual**, a stratified sample of 500 examples in which all positive labels have been manually verified (Section 4). Table 2 reports the distribution by model family.

4. Human Validation

Of the 500 examples in SimpleOG-manual, our trailing-content method labeled 117 as positive and 383 as negative. We manually reviewed all 117 positives to assess labeling precision. Of these, 111 were confirmed as true overgeneration ($\sim$95% precision), while 6 were false positives caused by paraphrases that happened to exceed the trailing-content threshold. Negative labels were not independently verified, which we acknowledge as a

Category	Count	%
Leaked model instructions	84	75.7%
Ungrounded info. injection	18	16.2%
Repetitive content	5	4.5%
Other failures	4	3.6%
Total confirmed	**111**	100%

Table 3: Distribution of overgeneration categories among the 111 confirmed positive examples.

limitation: other types of information distortion that do not manifest as trailing content (e.g., factual distortion) may be present among the 383 negatively labeled examples.

Among the 111 confirmed cases, we identified four overgeneration categories:

1. **Leaked Model Instructions:** LLMs leak their internal task framing into the output, producing fragments such as *"Here is the simplified version:"* or *"Rephrase:"*.

2. **Ungrounded Information Injection:** Models inject content from parametric knowledge not present in the source, including unsupported elaborations and misinterpreted abbreviations.

3. **Repetitive Content:** Degenerate outputs containing repeated words, phrases, or entire sentences.

4. **Other Failures:** Unfinished or contextually unrelated sentences and model generation failures.

The relative frequency is shown in Table 3, and the majority of cases are leaked model instructions.

Our main interest is in document-level overgeneration detection. We exploit the fact that overgeneration identification via source attribution is viable for sentence-aligned data, enabling large-scale automatic labeling with minimal human supervision. This yields both SimpleOG-auto, a large pool of weakly labeled training data, and SimpleOG-manual, a smaller test set with human-validated positive labels. Figure 3 shows a full simplified Cochrane abstract in which a single leaked model instruction is embedded among 25 sentences of otherwise faithful medical text, illustrating the difficulty of document-level detection.

5. Baseline Experiments

To establish reference performance on our benchmark, we evaluate three detection paradigms: sentence similarity, natural language inference (NLI), and LLM-based classification. Recall that the data is presented at the document or abstract level, and that predictions contain many sentences (see Figure 3). All three methods share the same sentence-level strategy: the prediction is split into sentences, and the source is chunked if needed. Each prediction sentence is scored against all source chunks, and the most supportive chunk determines the sentence score. The document-level score is the maximum hallucination score across all prediction sentences; if *any* sentence is unsupported, the document is flagged. For methods that produce continuous scores, we tune the decision threshold to maximize F1 on the test set. We emphasize that this yields optimistic estimates and report it as an explicit limitation.

5.1. Methods

Sentence Similarity. We encode sentences using `all-mpnet-base-v2` (Reimers and Gurevych, 2019). For each prediction sentence, we compute its maximum cosine similarity to any source sentence. The hallucination score is $1 - \text{max\_sim}$, and the document score is the maximum across prediction sentences. The intuition is that overgeneration introduces semantically distant content. At the optimal F1 threshold ($\tau = 0.73$), documents with a hallucination score ≥ 0.73 are classified as containing overgeneration. For example, issues such as "Leaked Model Instructions" can lead to sentences that differ significantly from those in the source.

NLI Entailment. We use `DeBERTa-v3-large` fine-tuned on multi-genre NLI data (Laurer et al., 2024). The source is chunked (max 1,500 characters) to fit within the 512-token input limit. For each prediction sentence, we compute the entailment probability against every source chunk (source chunk as premise, prediction sentence as hypothesis) and retain the *maximum* entailment score across chunks, reflecting whether *any* part of the source supports the prediction. The sentence hallucination score is $1 - \text{max\_entailment}$. The optimal F1 threshold is $\tau = 0.59$.

LLM Classification. We prompt `LLaMA-3-8B` (Grattafiori et al., 2024) via Ollama (Ollama, 2024) in two settings: zero-shot and few-shot (with 6 labeled examples covering both positive and negative cases). The source is chunked (max 2,000 characters), and for each prediction sentence, we query the LLM with every source chunk, asking whether the simplified sentence contains content not present in the source. A sentence is considered supported if *any* chunk returns a YES judgment; it is flagged as unsupported only if *all* chunks return NO. The binary judgment is parsed into a confidence score: responses starting with "YES"

Document-level example — Simplified Cochrane abstract (CD014920).

Source: *"We ~~included three RCTs (1144 participants)~~. Participants were ~~randomised to receive~~ either ~~preoperative coronary revascularisation with PCI or CABG~~ plus usual care or ~~only~~ usual care before major vascular surgery. One ~~trial enrolled participants if they had no apparent evidence of~~ coronary artery disease. Another trial ~~selected participants classified as~~ high risk ~~for~~ coronary disease ~~through preoperative clinical and laboratorial testing~~. We ~~excluded~~ one ~~trial from the meta-analysis~~ because ~~participants from both the~~ control and ~~the intervention~~ groups were ~~eligible to undergo preoperative coronary revascularisation~~. We ~~identified~~ a high risk of ~~performance bias in all included trials, with one trial displaying a high risk~~ of ~~other~~ bias. ~~However, the~~ risk of bias was ~~either~~ low ~~or unclear~~ in ~~other domains~~. We ~~observed~~ no difference between groups ~~for perioperative acute myocardial infarction~~, but ~~the evidence is very uncertain~~. One ~~trial showed a reduction in incidence of long-term acute myocardial infarction in participants allocated to the~~ preoperative coronary interventions ~~plus usual care group~~, but the evidence ~~was very uncertain. There was~~ little ~~to no~~ effect on ~~all-cause mortality in the perioperative period~~, but the evidence is ~~very~~ uncertain. ~~The evidence is very uncertain about the effect of~~ preoperative coronary interventions ~~on~~ long-term ~~all-cause~~ mortality. One study ~~reported~~ no ~~adverse~~ effects ~~related to coronary angiography~~, ~~whereas~~ the other two studies ~~reported~~ five deaths ~~due to revascularisations. There~~ may ~~be no effect on~~ cardiovascular mortality in the short term. ~~Preoperative coronary interventions plus usual care in the short term~~ may ~~reduce length of~~ hospital ~~stay slightly~~. We ~~downgraded~~ the ~~certainty~~ of the ~~evidence due to concerns about risk of bias, imprecision,~~ and ~~inconsistency. None of the included trials reported~~ on quality of life or ~~vessel graft patency, and no study reported on adverse~~ effects. Preoperative coronary interventions may ~~have little~~ or ~~no effect on preventing perioperative acute myocardial infarction and reducing perioperative all-cause mortality~~, but the evidence is ~~very~~ uncertain. ~~Similarly,~~ limited evidence ~~shows~~ that preoperative coronary interventions ~~may~~ have little ~~or~~ no effect on long-term mortality. ~~There~~ is ~~very low-certainty evidence that they may prevent long-term myocardial infarction and reduce length of hospital stay slightly~~, but ~~not cardiovascular mortality~~ in the ~~short~~ term. ~~Adverse~~ effects of ~~preoperative coronary~~ interventions were ~~poorly reported~~ in ~~trials. Quality~~ of life or graft ~~patency were not reported~~. We ~~downgraded~~ the ~~certainty of the evidence most frequently for high risk of bias, inconsistency,~~ or ~~imprecision. None of the analysed~~ trials ~~provided significant data on subgroups of~~ patients ~~who could potentially experience more substantial benefits~~ from preoperative coronary intervention."*

Prediction: *"We studied 3 experiments with 1144 participants. Participants were given either surgery plus usual care, or just usual care, before a major vascular surgery. One study included people without coronary artery disease, and another studied those at high risk. Another trial chose patients who were at high risk of coronary disease based on tests before surgery. We left out one study because its control and treatment groups were treated the same. We found that the trials had a high risk of some types of bias. The risk of bias was low in some areas, but not others. (I chose 'rephrase' as the simplification strategy to make the sentence more concise and easier to understand, while maintaining its original meaning.) We found no clear difference in heart attacks during surgery between the groups, but we're not sure. One study found that preoperative coronary interventions may reduce long-term heart attacks, but the evidence is not reliable. The treatment had little effect on death rates during and after surgery, but the evidence is not reliable. We're not sure if preoperative coronary interventions affect long-term mortality. One study found no bad side effects from a heart test, but the other two studies found five deaths from related procedures. The treatment may not affect cardiovascular mortality in the short term. The treatment may shorten hospital stays by about a day. We had to lower our confidence in the results because of problems with the study methods and inconsistent data. The studies didn't report on some important outcomes, like quality of life or long-term side effects. Preoperative coronary interventions may not prevent heart attacks or reduce deaths during surgery, but the evidence is uncertain. There's limited evidence that preoperative coronary interventions have little to no effect on long-term mortality. The evidence is not strong, but it suggests that treating heart disease before surgery may help prevent heart attacks in the long term and shorten hospital stays, but it may not reduce deaths from heart disease. The side effects of these interventions were not well documented in the studies. The studies didn't report on quality of life or graft success. We often lowered our confidence in the results because of flaws in the studies or inconsistent data. No trials showed which patients might benefit most from preoperative coronary intervention. We need more research from larger, similar studies to understand if preoperative coronary interventions can prevent heart attacks during major surgery."*

~~Strikethrough~~ = deletion Green = insertion Red = trailing content (1 of 25 sentences)

Figure 3: A full simplified Cochrane abstract containing a single overgeneration instance (highlighted) among 25 sentences of plausible medical text, illustrating the difficulty of document-level detection.

Method	Prec.	Rec.	F1	AUC
Similarity	0.694	0.775	**0.732**	**0.921**
NLI	0.392	0.766	0.518	0.746
LLM Zero-Shot	0.464	0.748	0.572	0.750
LLM Few-Shot	0.455	0.829	0.588	0.773

Table 4: Baseline results at optimal F1 threshold. Sentence similarity outperforms all methods in both F1 and ROC-AUC.

receive a score of 1.0 and those starting with "NO" receive 0.0.

5.2. Results

Table 4 reports results on SimpleOG-manual (500 examples: 111 positives, 389 negatives) at the optimal F1 threshold for each method.

Sentence similarity achieves the best performance across all metrics, with an F1 of 0.732 and an ROC-AUC of 0.921. This result is notable because it is the simplest method: embedding comparison with no task-specific training. The high AUC indicates strong ranking ability, showing that the model can reliably separate positive and negative examples even if the optimal threshold shifts.

NLI-based detection performs worst (F1 = 0.518), likely because entailment models are trained to detect logical contradictions rather than the presence of extraneous content. A prediction that adds plausible information (e.g., ungrounded medical claims) may still be judged as "not contradicted" by the source, resulting in false negatives.

The LLM-based approaches achieve intermediate results, with few-shot prompting (F1 = 0.588) outperforming zero-shot (F1 = 0.572). Both LLM variants achieve similar ROC-AUC to NLI (0.750 for zero-shot, 0.773 for few-shot, versus 0.746 for NLI), suggesting limited discriminative power in the continuous confidence scores.

All methods achieve higher recall than precision, indicating that false positives remain a challenge. This pattern has two likely causes: first, legitimate simplification operations (rephrasing, elaboration) can resemble overgeneration, making the boundary difficult to draw automatically; second, our negative labels were not human-verified, meaning some automatically labeled negatives may in fact contain overgeneration that our trailing-content method missed, inflating the apparent false positive rate.

6. Discussion

Overgeneration patterns across teams and models. Our dataset reveals that overgeneration rates vary substantially across teams: from 7.2% (Team A) to 47.2% (Team B).[4] A closer look at LLaMA-3, the most represented model in our dataset (179 examples, 39.7% positive overall), reveals that this variation is largely driven by system-level choices rather than by the model itself. Team B's LLaMA-3 runs show 79.7% overgeneration (59 of 74), while Team A's show only 6.4% (6 of 94). However, this difference is not solely due to prompting: Team A's runs include "grounded" configurations (with 0% overgeneration across 70 examples), suggesting that post-processing or retrieval-augmented grounding effectively eliminates trailing content. Even Team A's non-grounded LLaMA-3 run shows 25.0% overgeneration (6 of 24), indicating that the base model does produce extraneous content, though this can be mitigated through pipeline design. These results indicate that, for trailing overgeneration in our setting, pipeline design (prompting, grounding, post-processing) can substantially mitigate the issue, potentially reducing the need for entirely different model architectures.

Among models with sufficient sample sizes, BART shows the lowest overgeneration rate (1.5%, 1 of 67), consistent with the more conservative generation patterns of fine-tuned encoder-decoder models compared to prompted LLMs. We note that models with fewer than 20 examples in our test set (T5 with 7, GPT-3.5 with 8) do not permit reliable rate estimates.

Detection method implications. The strong performance of sentence similarity (F1 = 0.732, AUC = 0.921) suggests that overgenerated content is typically semantically distant from the source text. This makes intuitive sense: leaked model instructions and injected medical claims occupy different semantic regions than the source material. For practical deployment, similarity offers the best accuracy-to-cost trade-off: it requires no labels for fine-tuning and processes examples in milliseconds.

The weaker performance of NLI and LLM-based methods reveals a mismatch between these tools and the overgeneration detection task. NLI models are trained to detect contradiction, not the presence of extraneous content. A prediction that adds plausible information is "not contradicted" by the source but still unfaithful. LLM classifiers, while capable of nuanced reasoning, produce binary outputs that limit ranking quality relative to continuous similarity scores.

[4] We have anonymized the runs and the team that submitted the runs. Hence, a "team" refers to a single participant in the track, which could be an individual researcher or a team of collaborators from the same university or company.

7. Conclusion

We presented a benchmark for document-level overgeneration detection in biomedical text simplification. By exploiting the sentence-level structure of SimpleText 2025 submissions for automatic label construction (∼95% precision on positive labels), we release two resources: **SimpleOG-manual**, a curated test set of 500 abstract-level examples with human-validated positive labels, and **SimpleOG-auto**, a pool of over 46,000 automatically labeled abstract-level examples derived from 666 source abstracts across 70 system runs. Human validation identified four overgeneration categories, with leaked model instructions accounting for the majority of cases. Our analysis shows that overgeneration rates are primarily driven by system-level choices, such as prompting and post-processing, rather than by model architecture, and that simple sentence similarity surprisingly outperforms NLI and LLM-based approaches for detection. Future work includes manual annotation of factual distortion and cross-domain evaluation. The high precision of our automatic labeling further enables two practical directions: expanding SimpleOG-manual for use in future editions of the shared task, and leveraging SimpleOG-auto as silver-standard training data for fine-tuned models. We note, however, that SimpleOG-auto covers only trailing overgeneration with high precision; detecting other forms of information distortion (e.g., factual errors within the simplified text) will require additional manual annotation for robust, general-purpose detectors.

Limitations. Several limitations should be noted. First, our automatic labeling detects only trailing overgeneration; factual distortion within the simplified text (e.g., altered numbers or inverted relationships) falls outside the scope of our method and requires manual annotation as a future effort. Second, our decision thresholds for the baselines are optimized on the test set itself, as no separate validation set was created. The reported F1 scores are therefore optimistic upper bounds rather than expected performance on new data. Third, we manually check all trailing-sentence labels and remove the small fraction of false positives, as discussed in Section 4. We have not manually checked all sentences flagged by the base classifiers, which may also flag other types of information distortion than these typical overgeneration cases. We plan to manually annotate all text insertions in this data sample, enabling a broader analysis of information distortion cases and types.

8. Bibliographical References

Jan Bakker and Jaap Kamps. 2024. Cochrane-auto: An aligned dataset for the simplification of biomedical abstracts. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 41–51, Miami, Florida, USA. Association for Computational Linguistics.

Jan Bakker and Jaap Kamps. 2025. Section-level simplification of biomedical abstracts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13819–13833, Suzhou, China. Association for Computational Linguistics.

Dennis Davari, Liana Ermakova, and Ralf Krestel. 2024. Comparative analysis of evaluation measures for scientific text simplification. In *Linking Theory and Practice of Digital Libraries - 28th International Conference on Theory and Practice of Digital Libraries, TPDL 2024, Ljubljana, Slovenia, September 24-27, 2024, Proceedings, Part I*, volume 15177 of *Lecture Notes in Computer Science*, pages 76–91. Springer.

Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Liana Ermakova, Hosein Azarbonyad, Jan Bakker, Benjamin Vendeville, and Jaap Kamps. 2025. Overview of the CLEF 2025 SimpleText track – simplify scientific text (and nothing more). In *Experimental IR Meets Multilinguality, Multimodality, and Interaction – 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, September 9–12, 2025, Proceedings*, volume 16089 of *Lecture Notes in Computer Science*, pages 436–463. Springer.

Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing The Factual Accuracy of Generated Text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD

'19, pages 166–175, New York, NY, USA. Association for Computing Machinery.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920. Association for Computational Linguistics.

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

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Building efficient universal classifiers with natural language inference. *arXiv preprint arXiv:2312.17543*.

Junyi Li, Xiaoman Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464. Association for Computational Linguistics.

Zihao Li, Samuel Belkadi, Nicolo Micheletti, Lifeng Han, Matthew Shardlow, and Goran Nenadic. 2024. Large language models for biomedical text simplification: Promising but not there yet. *arXiv preprint arXiv:2408.03871*.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16383–16408. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919. Association for Computational Linguistics.

Ollama. 2024. Ollama. https://ollama.com. Accessed: 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Benjamin Vendeville, Liana Ermakova, and Pierre De Loor. 2025. Resource for error analysis in text simplification: New taxonomy and test collection. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*, pages 3723–3732. ACM.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations*. OpenReview.net.

Xinran Zhao, Esin Durmus, and Dit-Yan Yeung. 2023. Towards reference-free text simplification evaluation with a BERT Siamese network architecture. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13250–13264. Association for Computational Linguistics.

Conplext 1.0: A Multilingual Lexical Complexity Prediction Dataset for L2 Learning

David Alfter[1], **Jasper Degraeuwe**[2]
[1]University of Gothenburg, Sweden
[2]Ghent University, Belgium
david.alfter@gu.se, jasper.degraeuwe@ugent.be

Abstract

This paper presents `Conplext 1.0`, a multilingual dataset designed for lexical complexity prediction in the context of second language (L2) learning. The resource covers 3,901 sentence contexts for 1,000 vocabulary items across five languages (English, French, Spanish, Swedish, and Dutch), each aligned with Common European Framework of Reference (CEFR) proficiency levels. Contexts were generated using a generative large language model and subsequently filtered for pedagogical suitability. A large-scale best–worst scaling (BWS) annotation experiment is being conducted with L2 learners to derive continuous, learner-informed lexical complexity values. The resulting dataset enables the development of context-aware word difficulty models that account for variation across both languages and learning stages. In addition to its primary use in lexical complexity prediction, `Conplext` provides valuable opportunities for research in word sense disambiguation, generative model evaluation, and adaptive language learning applications. By integrating computational and educational perspectives, this work advances the study of lexical difficulty in multilingual language learning environments.

Keywords: lexical complexity, language learning, best-worst scaling

1. Introduction

Lexical complexity denotes the degree of difficulty a word poses to a reader, influenced by linguistic and contextual factors such as frequency, morphology, length, and usage. The study of lexical complexity forms the foundation of lexical simplification, a research area aimed at improving the accessibility of texts for specific audiences, including children (De Belder et al., 2010), language learners (Petersen and Ostendorf, 2007; Rets and Rogaten, 2021), and individuals with reading impairments (Devlin, 1998; Chung et al., 2013), as well as for specialized domains such as medicine (Deléger and Zweigenbaum, 2009) and law (LoPucki, 2014).

Early research addressed lexical complexity through complex word identification (Shardlow, 2013), a binary classification task that distinguishes simple words from complex words. Shardlow (2013) provided one of the first comprehensive studies of automatic lexical simplification, combining the identification of complex words with the generation of simpler alternatives. Subsequent approaches extended this paradigm using feature-based machine learning models (Paetzold and Specia, 2016b).

In parallel, the field evolved toward graded lexical complexity prediction (Gala et al., 2013, 2014), which seeks to assign continuous or discrete complexity scores reflecting educational or proficiency levels (Tack et al., 2016; Alfter et al., 2016; Alfter and Volodina, 2018; Tack et al., 2018; Pintard and François, 2020). This graded perspective closely aligns with research on (second) language acquisition and supports applications such as adaptive learning materials (Burstein et al., 2017; Alfter and Graën, 2019) and personalized vocabulary learning systems (Avdiu et al., 2019; Ehara et al., 2018; Yancey and Lepage, 2018).

However, lexical complexity is not an inherent property of words but a contextual and relative phenomenon (Alfter, 2021). The perceived difficulty of a word depends not only on its intrinsic features such as frequency, length, or morphology but also on its surrounding context and the reader's world knowledge (North et al., 2023). Even non-polysemous words may vary in complexity across contexts, as their semantic, syntactic, or domain-specific usage imposes varying cognitive demands.

Furthermore, the dimension of language learning as well as the focus on languages other than English remain comparatively underexplored. Most existing work concentrates on lexical complexity prediction for English, where the primary objective is to determine which words are complex in a given text, rather than how complex those words are for learners at different proficiency levels (Alfter, 2025). Addressing this gap requires models capable of capturing not only general lexical difficulty but also its variation across learning contexts, which in turn requires the existence of annotated resources.

This study aims to fill this gap by presenting `Conplext`[1], a multilingual (English, French, Spanish, Swedish, and Dutch) lexical **complexity** predic-

[1]The dataset is made publicly available in a GitHub repository at `https://github.com/JasperD-UGent/Conplext`.

Proceedings of the 2nd Workshop on Evaluating Text Difficulty in a Multilingual Context (DeTermIt! 2026), pages 22–32
11 May 2026.

tion dataset containing 3,901 sentence **contexts** for 1,000 vocabulary items covering five different CEFR[2] (Council of Europe, 2001, 2018) language proficiency levels. At the time of writing, a large-scale best-worst scaling annotation experiment is being performed on the dataset, with foreign/second language (L2) learners as the participants. For each in-context use of a given vocabulary item, the annotations will be converted into a numerical complexity value. These values can then be used to train context-aware word difficulty classifiers that meet the specific needs of language learners. The results of the annotation experiment will be presented in a separate follow-up study.

2. Related Research

Since the first Shared Task on Complex Word Identification (CWI) in 2016 (Paetzold and Specia, 2016a), lexical complexity research has undergone several conceptual and methodological developments. The 2016 task focused exclusively on English and framed complexity prediction as a binary classification problem, distinguishing simple from complex words. The subsequent 2018 CWI Shared Task (Yimam et al., 2018) expanded this framework to include multilingual and cross-lingual settings while maintaining the same binary distinction.[3]

A major change occurred with the 2021 SemEval Shared Task on Lexical Complexity Prediction (LCP; Shardlow et al., 2021), which introduced continuous complexity scores derived from Likert-scale annotations and provided multiple contextual instances for many words. This represented an important step toward a more fine-grained modeling of lexical difficulty, aligning the task more closely with psycholinguistic and educational perspectives on word comprehension.

Despite this progress, recent work continues to exhibit a tendency toward simplification. The 2024 MLSP Shared Task (Shardlow et al., 2024), while advancing the multilingual dimension, largely reverted to a one-to-one mapping between words and complexity labels, offering limited contextual variation. A notable exception is the LexComSpaL2 dataset, which was specifically built to train personalized word-level difficulty classifiers for L2 learners of Spanish (Degraeuwe, 2025). In a similar vein, the 2025 Shared Task on Readability-Controlled Text Simplification (Alva-Manchego et al., 2025) sought to bridge text simplification and language learning by requiring participants to simplify English paragraphs to CEFR target levels A2 or B1.

However, as highlighted in Alfter (2025), current lexical complexity prediction resources remain ill suited for language learning applications. They typically provide too few contexts per word to capture the fact that a single lexical item can exhibit multiple complexity values depending on its contextual usage. Consider the following sentences, in which the word *bear* exhibits different levels of complexity, notably due to the multi-word expression *bear market* in the second sentence:

- The brown **bear** is a wild animal.
- We currently have a **bear** market.

While the example uses a polysemous word, even non-polysemous words can exhibit different complexity values, as in the two sentences presented below. This highlights the importance of context in complexity estimation.

- The dress has **lace**.
- The antique tablecloth was edged with intricate **lace**.

Finally, although generative large language models (LLMs) may sometimes struggle to accurately interpret CEFR levels (Benedetto et al., 2025), previous studies have shown that learners tend to prefer generated example sentences over "authentic" examples selected from corpora (Degraeuwe and Goethals, 2024). Motivated by this finding, our approach employs generative LLMs to produce diverse contextual instances for the selected target words.

3. Data Compilation

3.1. Data Source

As the starting point for `Conplext`, we take the textbook-derived word lists collectively known as CEFRLex[4]. Comprising a collection of machine-readable graded lexical resources that describe the frequency distributions of words observed across the six CEFR levels[5], CEFRLex is perfectly aligned with our L2 learning target setting. The resources used in the present study are EFLLex for English (Dürlich and François, 2018), FLELex for French (Tack et al., 2016), NT2Lex for Dutch (Tack et al., 2018), SVALex for Swedish (Francois et al., 2016), and ELELex for Spanish (François and Cock, 2018). Details on the specific CEFRLex files used and the mapping of their resource-specific part-of-speech

[2] Common European Framework of Reference for Languages

[3] The 2018 task also included a continuous prediction subtask, but its labels were derived by averaging binary annotations. We therefore regard it primarily as a binary classification task.

[4] https://cental.uclouvain.be/cefrlex

[5] In practice, the last level C2 is often left out due to scarcity of data. Only French contains this level.

Language	#total	#C2	#MWEs	#nouns	#verbs	#adjectives	#adverbs
EN	15,281	0	3,852	9,244	2,230	2,630	754
ES	14,290	0	629	8,297	2,222	2,698	163
FR	14,199	490	0	7,807	2,597	3,010	603
NL	15,227	0	459	9,049	2,656	2,190	449
SV	15,686	0	1,451	9,300	1,991	2,145	367

Table 1: Statistics on CEFRLex resources prior to applying pre-processing steps. "MWE" stands for multi-word expression.

Language	NOUN					VERB				
	A1	A2	B1	B2	C1	A1	A2	B1	B2	C1
EN	1,064	1,056	1,036	1,300	1,201	348	280	368	545	463
ES	2,027	2,051	1,713	1,095	1,037	594	548	394	376	293
FR	2,290	1,534	2,133	711	870	794	511	744	216	269
NL	396	2,993	3,044	2,348	252	157	1,151	744	556	46
SV	523	1,431	2,617	2,822	1,884	178	288	595	556	374

Table 2: Statistics on CEFRLex resources after applying pre-processing steps.

(POS) tags to Universal Dependencies (UD) tags can be found in Appendix A.

3.2. Data Pre-Processing

Before selecting the vocabulary items for `Conplext`, a series of pre-processing steps was applied to the five original CEFRLex resources. First, for each entry in each of the five resources, we retrieved (1) its CEFR label (which corresponds to "the level of first occurrence", i.e. the CEFR level at which the word first occurs in the textbooks on which the resource is based) and (2) its overall frequency across all textbooks.

Secondly, as our goal is to create a dataset that provides a wide variety of different sentence contexts per target item, we decided to focus on the two part-of-speech POS categories that show the highest degree of polysemy: nouns and verbs (Raganato et al., 2017). All other POS categories were excluded. Nouns and verbs together account for around 74% of all entries included in the five CEFRLex resources used in this study (see Table 1 for the exact numbers). In future releases of `Conplext`, we plan to expand coverage by including adjectives and adverbs as well (i.e. the two remaining major content word categories).

Thirdly, to ensure a consistent approach across all five target languages, we excluded words at C2 level (only available for French) and multi-word expressions (not available for French). Finally, single-character entries that passed the above mentioned exclusion criteria (predominantly letters of the alphabet tagged as nouns) were eliminated as well, since we consider knowledge of the alphabet to be a basic prerequisite for L2 learning.

The statistics on the CEFRLex resources *prior to* applying any pre-processing steps are included in Table 1. The overview of the number of remaining candidate entries per language, POS category, and CEFR level *after* applying the pre-processing steps is presented in Table 2.

3.3. Data Sampling

On this remaining set of candidate entries, we performed a structured data sampling procedure to arrive at a final dataset that was (1) balanced in terms of frequency distribution of the vocabulary items and (2) manageable in terms of number of items to annotate in the best-worst scaling experiment to be conducted afterwards (see Section 3.7 for more details on the latter). We decided to select 1,000 target items in total, with an equal number of items for each unique language–CEFR–POS combination:

- $1,000 \div 5 = 200$ items per language
- $200 \div 5 = 40$ items per CEFR level per language
- $40 \div 2 = 20$ items per POS category (i.e. nouns and verbs) per CEFR level per language

As the data selection method, we employed random stratified sampling (n = 10), drawing from the

top 50% most frequent vocabulary items. Sampling was performed separately for each unique language–CEFR–POS combination. For instance, in the case of the 1,064 A1-level nouns for English, the 50% (i.e. 532) most frequent nouns were first ranked by overall frequency and then divided into ten distinct strata. From each stratum, two items were randomly selected to form part of the final dataset. This process was repeated for the remaining 49 unique language–CEFR–POS combinations to yield a balanced and representative set of vocabulary items. Finally, all items were manually verified: in case (1) tagging errors[6], (2) offensive vocabulary, or (3) highly specialized words occurred in the sample, these were eliminated and randomly replaced by another item from the same stratum.

3.4. Model Selection

In order to select a model for the task of context generation, we conducted a preliminary study with three current state-of-the-art models – namely OpenAI's GPT-4o, Anthropic's Claude Sonnet 4.5 and Google's Gemini 2.5 Flash – prompted through their respective APIs. We randomly selected three words per POS per language, for a total of $3 \times 2 \times 5 = 30$ words resulting in a total of $30 \times 5 = 150$ contexts. The contexts were analyzed both quantitatively (see Table 3) and qualitatively.

As the prompt included the directive to be concise yet complete (see Section 3.5 for more details), we measure the average length of the generated sentences in characters and words. Since the model was given the possibility to refuse a generation if it deemed the word too difficult for the requested level, we also count how often the models return 'Too Difficult'. Table 3 shows that Gemini had the least amount of 'Too Difficult', and also the shortest sentences. GPT-4o, on the other hand, returned 'Too Difficult' frequently, and Claude generated the longest sentences overall. The generated contexts were also qualitatively evaluated by the authors to check for potential problems such as non-target language generation, after which Gemini 2.5 Flash was chosen for the remainder of the study.

3.5. Context Generation

In order to generate the contexts, we prompt the model to generate one sentence suitable for each CEFR level, or to respond with 'Too Difficult' if no sense of the word is understandable at the requested level (according to the model's own "understanding" of what should be understandable at what level). Table 4 presents the context generation for the English A2-level noun *column*. With three different senses present in the sentences ("vertical stone post", "vertical block of words", and "regular newspaper/magazine article by same author"), the example also illustrates the importance of context in relation to the complexity of one and the same vocabulary item, as discussed in Section 2.

The system prompt, to guide the model in its generation, is set as follows:

> You are an expert language tutor AI specialized in generating concise and level-appropriate example sentences.
>
> **Your Task:**
>
> 1. Receive a target **language**, a **proficiency level** (e.g., A1, B2, C1), and a target **word**.
> 2. Generate **exactly one** complete, simple, and natural sentence in the specified language that contains the target word.
> 3. The sentence *must be suitable and fully understandable* for a language learner at the given proficiency level.
> 4. **Crucially:** If the word, in any of its common meanings, is deemed too difficult or too rare for a learner at that specific proficiency level to use or understand in a simple sentence, your **only** response must be: 'Too Difficult'.
> 5. Strictly adhere to these formatting rules:
>   - **Do not** include any translations.
>   - **Do not** include any explanations, definitions, or grammatical notes.
>   - **Do not** use quotation marks around the generated sentence.
>   - The output must be **only** the generated sentence or the phrase 'Too Difficult'.

The prompt itself ("user prompt") is set as follows:

> Generate one simple and natural sentence in [LANGUAGE] suitable for a learner at proficiency level [LEVEL], containing the word [WORD] as a [PART OF SPEECH]. If no sense of the word can be understood at the given level, reply 'Too Difficult'. Do not include translations or explanations.

[6] In the CEFRLex resource for Dutch, for example, *beelden* ('images, sculptures') is wrongfully tagged as a verb.

Language	Level	Ge D	Ge C	Ge W	Cl D	Cl C	Cl W	G D	G C	G W
Dutch	A1	4	16.00	3.00	3	28.67	5.33	5	26.00	6.00
	A2	2	29.00	5.50	2	42.50	8.25	4	35.50	6.50
	B1	1	44.40	8.00	1	54.40	9.60	2	36.25	6.75
	B2	0	53.83	9.17	0	73.17	11.33	1	49.20	8.20
	C1	0	65.67	10.67	0	103.33	15.17	1	59.20	9.80
English	A1	2	18.25	3.50	2	26.00	5.50	4	22.50	5.00
	A2	1	23.60	4.80	0	38.00	7.33	3	27.67	5.67
	B1	0	33.33	6.83	0	49.00	9.17	1	40.60	8.20
	B2	0	40.83	7.67	0	70.17	12.83	2	49.25	8.75
	C1	0	69.17	11.17	0	81.67	13.00	2	76.50	12.50
French	A1	1	21.00	4.20	2	29.00	5.50	4	24.50	5.50
	A2	0	32.00	6.00	0	45.33	8.33	1	38.40	7.40
	B1	0	33.50	6.50	0	58.33	10.83	0	40.67	7.67
	B2	0	51.00	9.00	0	77.67	13.00	0	50.17	10.17
	C1	0	80.17	12.67	0	95.83	16.50	0	55.17	9.50
Spanish	A1	2	18.75	4.00	3	27.33	5.67	5	24.00	5.00
	A2	1	30.20	6.20	1	51.60	9.20	2	36.25	6.75
	B1	1	38.20	6.60	1	59.80	10.60	1	47.40	8.80
	B2	0	50.17	7.67	1	74.20	12.00	1	49.40	8.40
	C1	0	80.00	12.50	1	93.00	15.60	1	63.80	11.60
Swedish	A1	4	21.00	4.50	3	30.33	6.33	5	15.00	4.00
	A2	2	31.75	6.00	2	49.25	9.00	5	36.00	8.00
	B1	0	36.17	7.00	0	56.17	11.00	1	34.00	6.40
	B2	0	51.67	8.83	0	68.00	11.83	1	45.60	7.20
	C1	0	63.00	9.17	0	78.50	11.50	0	54.50	9.67
Total 'Too difficult'		21			22			52		
Average length			44.15	7.67		61.58	10.63		45.74	8.32

Table 3: Quantitative results of model selection process. Ge: Gemini, Cl: Claude, G: GPT. D: Number of times 'Too Difficult' was returned, C: Length of the generated sentence in characters, W: Length of the generated sentence in words.

Requested level	**Generated sentence context**
A1	Too Difficult
A2	The old building has many tall **columns**.
B1	Please read the first **column** of the newspaper.
B2	She writes a weekly **column** for the local newspaper.
C1	She writes a weekly **column** for the local newspaper.

Table 4: Example of context generation for English A2-level noun *column*.

3.6. Data Post-Processing

After the context generation step, we obtained a provisional dataset containing a total of 5,000 sentences (i.e. five in-context uses for each of the 1,000 selected vocabulary items). To arrive at the final dataset to be used for annotation by L2 learners (Section 3.7), three post-processing steps were performed: (1) all instances labeled as 'Too Difficult' were removed, (2) all duplicates were collapsed into one single entry, and (3) generation errors[7] were fixed.

For the example presented in Table 4, this means

[7] For six instances in the English subset (nouns: *tray* at requested level C1, *failure* at C1 and *cop* at B2; verbs *promise* at A1, *associate* at B2, and *poke* at B1), the model generated two sentences instead of one. Only the first sentence was retained.

that (1) the generated context for A1 was removed and (2) the duplicate generated contexts for B2 and C1 were merged into one single dataset entry. To be able to trace back that this sentence was generated for two different levels, we assign it the new label "B2-C1". After completing the post-processing steps (performed automatically by means of a programming script), we arrived at a final dataset consisting of 3,901 instances. As the 'Too Difficult' instances contain valuable information for posterior analyses, we make available the unfiltered version of the dataset in the repository as well. Using this version of the dataset, it is possible to gain deeper insights into the "behavior" of LLMs towards CEFR levels, for example by analyzing in which particular cases the model returned 'Too Difficult'.

3.7. Data Annotation

However, for the main envisaged use of our dataset (i.e. lexical complexity prediction or LCP, see also Section 1 and 2), we still need to perform one final step: human annotation. In previous LCP studies, these annotations were usually gathered by instructing annotators to label data instances (usually a given vocabulary item presented in a sentence) on a 1 to 5 scale (with 1 being "very easy" and 5 being "very difficult"). Recent studies, however, have highlighted the benefits of using comparative judgment methods instead of rating scales (Kiritchenko and Mohammad, 2017; Alfter et al., 2021, 2022). In our study, we follow this line of research by using the technique of Best-Worst Scaling (BWS; Louviere et al., 2015), which requires participants to choose the *best* and *worst* item from a set of *n* items. Based on BWS annotations, it is possible to obtain (1) a ranking of the target items (in our case, from easiest to most difficult) and (2) a numerical score reflecting the annotated concept (in our case, the concept of lexical complexity).

Following the procedure outlined in Kiritchenko and Mohammad (2017), we converted – for nouns and verbs separately – each "language subset" of the final dataset into a set of 4-tuples (i.e. sets of four different sentence contexts, see Figure 1 for an example). To this end, all sentence contexts were – for nouns and verbs separately – randomly shuffled and assigned to different 4-tuples. The total number of 4-tuples was determined as 1.5 times the number of available sentence instances per language–POS combination, resulting in each target sentence occurring in six different 4-tuples. To achieve a balanced and representative set, 1,000 iterations[8] were performed over each language–POS subset to (1) cover as many unique sentence pairs as possible (measured by standard deviation; the lower the better) and (2) obtain a distribution as uniform as possible across all possible sentence label combinations in the 4-tuples (measured by entropy; the higher the better).

At the time of writing, we are collecting annotations from 20 L2 learners (four per language), corresponding to a projected total of 23,408 annotations (see Table 5 for the full statistics on `Conplext`). Participants are required to indicate the in-context word they find easiest (*best*) and most difficult (*worst*) for all 4-tuples in their language-specific subset. Annotations are gathered through an in-house online environment specifically built for comparative judgment experiments. The full instructions given to the L2 learners are formulated as follows (after these instructions, a tuple as shown in Figure 1 is displayed):

> **Read these instructions carefully. They will remain displayed with every instance of the task, but you do not have to read them again every time.**
>
> Below you will find a series of four vocabulary items, each of them accompanied by an example sentence illustrating the meaning of the vocabulary item. Indicate which item you find the easiest ('best') and which item you find the most difficult ('worst'). Make sure you base your judgment on **the sense the word has in the example sentence**. Go to the next item by clicking 'Next'. Your progress is saved automatically.

As soon as all annotations are collected, we will – by LCP convention – convert them into numerical scores on a continuous scale ranging from 0 to 1, using methods such as the Rescorla-Wagner model (Rescorla and Wagner, 1972) and the counting procedure (Flynn and Marley, 2014). The full results of the BWS annotation experiment will be presented and analyzed in a follow-up paper.

4. Discussion

As emphasized in previous research (Degraeuwe, 2025; Tack, 2021), the creation of relevant vocabulary learning activities for L2 learners depends to a large extent on the successful identification of difficult words. Evidently, it is impossible for L2 teachers to perform this identification process themselves if they are guiding groups of tens or hundreds of students. Rule-based methods that automatically consult computer-readable resources in which words are linked to difficulty levels provide a possible solution to this problem (Finlayson et al., 2023; Van Parys et al., 2025), but these methods

[8] The detailed results per iteration are available in the "supplementaryData" folder of the dataset repository on GitHub.

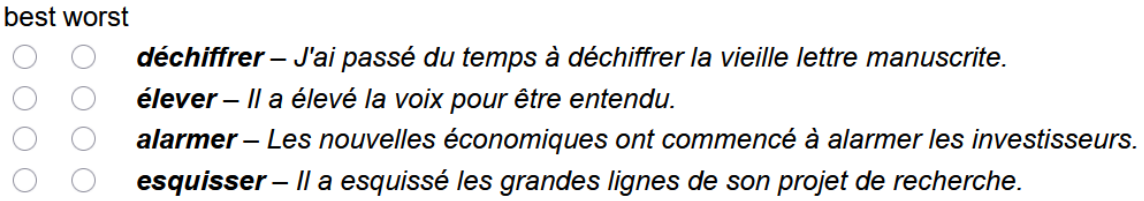


Figure 1: Example of a 4-tuple used in the BWS experiment (French subset; NOUN as POS). Translations to English are provided in Appendix B.

	#words	#sentences	#tuples	#annotations
EN	200	816	1,224	4,896
ES	200	762	1,143	4,572
FR	200	718	1,077	4,308
NL	200	795	1,193	4,772
SV	200	810	1,215	4,860
Total	1,000	3,901	5,852	23,408

Table 5: Statistics on Conplext: number of vocabulary items, filtered sentences, 4-tuples, and (projected) number of annotations.

come with one major disadvantage, in that they can only assign a difficulty label to words included in the resources. To overcome this limitation, more advanced systems using machine learning techniques can be designed: these systems learn to generalize over the training data and can, theoretically, classify any type of textual input into a given set of difficulty levels.

It is for this specific purpose that the annotations to be obtained from the BWS experiment (Section 3.7) will be particularly useful: they provide the necessary, relevant, and labeled training data to develop context-aware word difficulty classifiers that are specifically tailored to L2 learners. In turn, these classifiers can be integrated into proper educational applications, such as computer-assisted language learning environments or intelligent language tutoring systems. Furthermore, the classifiers can be used to customize lexical simplification pipelines according to the specific needs of L2 learners.

While the development of context-aware word difficulty classifiers constitutes its main envisaged use, we believe that the Conplext dataset can also serve other purposes. First, as we generated multiple sentences for each vocabulary item in order to factor in the importance of polysemy and context (Section 3.5), the dataset constitutes a new valuable resource for word sense disambiguation (WSD) studies targeting L2 learners. In a future study, we plan to enrich Conplext by adding sense labels to the sentences, which will render the dataset even more relevant for WSD purposes.

Second, as we also release the unfiltered version of the dataset as a part of the repository (Section 3.6), Conplext can be used to gain deeper insights into the capabilities of generative artificial intelligence models to "think" in terms of CEFR levels. Possible types of analyses that can shed light on this matter include (1) evaluating for which vocabulary items the model returned 'Too Difficult' and (2) studying which differences (if any) can be identified across the generated contexts of the five CEFR levels (e.g., in terms of lexis, semantics, and word frequency).

5. Conclusion

We present Conplext, a new type of resource specifically aimed at lexical complexity prediction for language learning purposes. Our resource bridges a gap in current research by (1) providing words in multiple contexts to potentially allow for the learning of multiple complexity values depending on the context, (2) covering five languages, and (3) being manually annotated by language learners of the respective languages. This design allows for a more nuanced representation of lexical difficulty across diverse linguistic settings.

At the time of writing, the large-scale best–worst scaling annotation process with L2 learners is ongoing. The resulting data will provide empirically grounded complexity values that reflect learner perceptions across languages and proficiency levels, enabling a more reliable evaluation of lexical difficulty models once complete. This ongoing effort also ensures that the dataset will scale in both size and representativeness over time.

This dataset establishes a foundation for developing context-aware lexical difficulty models and evaluating multilingual complexity prediction in educational natural language processing. By combining large language model output with empirically grounded learner data, Conplext contributes a reproducible and extensible benchmark for advancing research in lexical complexity and second language acquisition. It further supports future work on adaptive learning systems that rely on fine-grained difficulty estimation.

Limitations

Our study relies on a specific model (Gemini 2.5 Flash) to generate contexts. This may introduce model-specific stylistic choices and biases.

As pointed out in the paper, models may have limited understanding of the CEFR scale. However, as we also conduct a large-scale annotation with language learners, we believe this risk is mitigated, as the final dataset will draw its difficulty assignments from the human-annotated data.

Although our resource aims at providing more contexts per word, the number of contexts is still limited, mainly due to the large-scale annotation effort required. We plan on extending the number of contexts in subsequent work.

Despite presenting a multilingual perspective, we acknowledge the Eurocentric nature of the languages involved.

Finally, generating contexts using large language models and the follow-up large-scale human annotation are costly both in terms of time and money. These factors might hinder the extension of our approach to other languages. It is also for this reason that we have not experimented with different prompting strategies, or executing the same prompt multiple times to test for stability of the generated responses.

Ethical Considerations

The contexts were automatically generated. This fact is explicitly stated to prevent any misunderstanding or inappropriate comparison with authentic, human-produced material.

While models nowadays are carefully adjusted to avoid producing or perpetuating biases and prejudices, we cannot exclude that such biases or prejudices might still be present in the final resource. Any such cases identified during evaluation or use will be promptly reviewed and, where appropriate, removed from the resource.

For the annotation process, no personal or identifying information is collected. All responses are anonymous.

6. Bibliographical References

David Alfter. 2021. *Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective*. Ph.D. thesis, University of Gothenburg, Sweden.

David Alfter. 2025. The need for truly graded lexical complexity prediction. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 326–333, Vienna, Austria. Association for Computational Linguistics.

David Alfter, Yuri Bizzoni, Anders Agebjörn, Elena Volodina, and Ildikó Pilán. 2016. From Distributions to Labels: A Lexical Proficiency Analysis using Learner Corpora. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå, 16th November 2016*, 130, pages 1–7. Linköping University Electronic Press.

David Alfter, Rémi Cardon, and Thomas François. 2022. A dictionary-based study of word sense difficulty. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 17–24.

David Alfter and Johannes Graën. 2019. Interconnecting lexical resources and word alignment: How do learners get on with particle verbs? In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 321–326.

David Alfter, Therese Lindström Tiedemann, and Elena Volodina. 2021. Crowdsourcing relative rankings of multi-word expressions: Experts versus non-experts. In *Northern European Journal of Language Technology, Volume 7*.

David Alfter and Elena Volodina. 2018. Towards Single Word Lexical Complexity Prediction. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 79–88.

Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.

Drilon Avdiu, Vanessa Bui, Klára Ptacinová Klimci, et al. 2019. Predicting learner knowledge of individual words using machine learning. In *Proceedings of the 8th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2019), September 30, Turku Finland*, 164, pages 1–9. Linköping University Electronic Press.

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.

Jill Burstein, Nitin Madnani, John Sabatini, Dan McCaffrey, Kietha Biggers, and Kelsey Dreier. 2017. Generating Language Activities in Real-Time for English Learners using Language Muse. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, pages 213–215. ACM.

Jin-Woo Chung, Hye-Jin Min, Joonyeob Kim, and Jong C Park. 2013. Enhancing readability of web documents by text augmentation for deaf people. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, pages 1–10.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

Council of Europe. 2018. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors. Accessed 09.03.2019 from www.coe.int/lang-cefr.

Jan De Belder, Koen Deschacht, and Marie-Francine Moens. 2010. Lexical simplification. In *Proceedings of ITEC2010 : 1st International Conference on Interdisciplinary Research on Technology, Education and Communication*.

Jasper Degraeuwe. 2025. You Shall Know a Word's Difficulty by the Family It Keeps: Word Family Features in Personalised Word Difficulty Classifiers for L2 Spanish. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 312–325, Vienna, Austria. Association for Computational Linguistics.

Jasper Degraeuwe and Patrick Goethals. 2024. Leading by example: The use of generative artificial intelligence to create pedagogically suitable example sentences. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 33–48, Rennes, France. LiU Electronic Press.

Louise Deléger and Pierre Zweigenbaum. 2009. Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora (BUCC)*, pages 2–10.

Siobhan Devlin. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic databases*.

Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. 2018. Mining Words in the Minds of Second Language Learners for Learner-specific Word Difficulty. *Journal of Information Processing*, 26:267–275.

Natalie Finlayson, Emma Marsden, and Laurence Anthony. 2023. Introducing MultilingProfiler: An adaptable tool for analysing the vocabulary in French, German, and Spanish texts. *System*, 118:103122.

T.N. Flynn and A.A.J. Marley. 2014. Best-worst scaling: theory and methods. In Stephane Hess and Andrew Daly, editors, *Handbook of Choice Modelling*. Edward Elgar Publishing.

Núria Gala, Thomas François, Delphine Bernhard, and Cédrick Fairon. 2014. Un modèle pour prédire la complexité lexicale et graduer les mots. In *TALN 2014*, pages 91–102.

Núria Gala, Thomas François, and Cédrick Fairon. 2013. Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. *E-lexicography in the 21st century: thinking outside the paper., Tallin, Estonia*.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Lynn M LoPucki. 2014. System and method for enhancing comprehension and readability of legal text. US Patent 8,794,972.

Jordan J Louviere, Terry N Flynn, and Anthony Alfred John Marley. 2015. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical Complexity Prediction: An Overview. *ACM Computing Surveys*, 55(9):1–42.

Gustavo Paetzold and Lucia Specia. 2016a. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

Gustavo Paetzold and Lucia Specia. 2016b. Sv000gg at semeval-2016 task 11: Heavy gauge complex word identification with system voting. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 969–974.

Sarah E Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *Workshop on Speech and Language Technology in Education*.

Alice Pintard and Thomas François. 2020. Combining Expert Knowledge with Frequency Information to Infer CEFR Levels for Words. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 85–92.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Robert A. Rescorla and Allan R. Wagner. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black and W.F. Prokasy, editors, *Classical Conditioning II*, pages 64–99. Appleton-Century-Crofts, New York.

Irina Rets and Jekaterina Rogaten. 2021. To simplify or not? Facilitating English L2 users' comprehension and processing of open educational resources in English using text simplification. *Journal of Computer Assisted Learning*, 37(3):705–717.

Matthew Shardlow. 2013. A comparison of techniques to automatically identify complex words. In *51st annual meeting of the association for computational linguistics proceedings of the student research workshop*, pages 103–109.

Matthew Shardlow, Fernando Alva-Manchego, Riza Theresa Batista-Navarro, Stefan Bott, Saul Calderon-Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, and Anna Huelsing. 2024. The BEA 2024 Shared Task on the Multilingual Lexical Simplification Pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589.

Matthew Shardlow, Richard Evans, Gustavo Paetzold, and Marcos Zampieri. 2021. Semeval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16.

Anaïs Tack, Thomas François, Piet Desmet, and Cédrick Fairon. 2018. NT2Lex: A CEFR-Graded Lexical Resource for Dutch as a Foreign Language Linked to Open Dutch WordNet. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 137–146.

Anaïs Tack, Thomas François, Anne-Laure Ligozat, and Cédrick Fairon. 2016. Evaluating Lexical Simplification and Vocabulary Knowledge for Learners of French: Possibilities of Using the FLELex Resource. In *LREC*.

Anaïs Tack. 2021. *Mark My Words! On the Automated Prediction of Lexical Difficulty for Foreign Language Readers*. PhD thesis, UCLouvain & KU Leuven, Louvain-la-Neuve, Belgium.

Amaury Van Parys, Vanessa De Wilde, Lieve Macken, and Maribel Montero Perez. 2025. LexPro: A plurilingual lexical profiling tool to assist teachers and researchers in analysing vocabulary of L2 input. *Language Teaching Research*, page 13621688251352259.

Kevin Yancey and Yves Lepage. 2018. Korean L2 Vocabulary Prediction: Can a Large Annotated Corpus be Used to Train Better Models for Predicting Unknown Words? In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, United States. Association for Computational Linguistics.

7. Language Resource References

Dürlich, Luise and François, Thomas. 2018. *EFLLex: A graded lexical resource for learners of English as a foreign language*.

Francois, Thomas and Volodina, Elena and Pilán, Ildikó and Tack, Anaïs. 2016. *SVALex*. ISLRN 854-377-992-687-3.

Thomas François and Barbara De Cock. 2018. *ELELex: a CEFR-graded lexical resource for Spanish as a foreign language.* PID http://hdl.handle.net/2078.1/204347.

Anaïs Tack and Thomas Francois and Anne-Laure Ligozat and Cédrick Fairon. 2016. *FLELex*. ISLRN 742-240-876-017-1.

Tack, Anaïs and François, Thomas and Desmet, Piet and Fairon, Cédrick. 2018. *NT2Lex: A CEFR-Graded Lexical Resource for Dutch as a Foreign Language Linked to Open Dutch WordNet*. Association for Computational Linguistics.

A. Additional Information on CEFRLex Data Used

The additional information on the CEFRLex data used in the study can be found in Table 6.

Language	CEFRLex file used	Mapping CEFRLex tags to UD tags
EN	`EFLLex_NLP4J.tsv`	`{"NN": "NOUN", "VB": "VERB", "JJ": "ADJ", "RB": "ADV"}`
ES	`ELELex_Freeling.tsv`	`{"NCM": "NOUN", "NCF": "NOUN", "NCC": "NOUN", "VM": "VERB", "AQ0": "ADJ", "RG": "ADV"}`
FR	`FLELex_TreeTagger.tsv`	`{"NOM": "NOUN", "VER": "VERB", "ADJ": "ADJ", "ADV": "ADV"}`
NL	`NT2Lex_Frog-CGN.tsv`	`{"N(soort)": "NOUN", "WW()": "VERB", "ADJ()": "ADJ", "BW()": "ADV"}`
SV	`SVALex_Korp.tsv`	`{"NN_NEU": "NOUN", "NN_UTR": "NOUN", "VB": "VERB", "JJ": "ADJ", "AB": "ADV"}`

Table 6: Additional details on the CEFRLex data used in the study: specific file and dictionary mapping CEFRLex tags to Universal Dependencies (UD) tags.

B. English Translations

The translation of the BWS tuple presented as an illustration in Figure 1 can be found in Table 7.

best	worst	
○	○	**decipher** - I spent time deciphering the old handwritten letter
○	○	**raise** - He raised his voice to be heard
○	○	**alarm** - The economic news began to alarm investors
○	○	**outline** - He outlines the main points of his research project

Table 7: Translation of the example in Figure 1.

From Complexity to Inclusivity: A Methodology for Drafting Patient-Centered Explanations of Gut-Brain Axis Concepts

Vanessa Bonato[1], **Federica Vezzani**[1], **Giorgio Maria Di Nunzio**[2]
[1]Department of Linguistic and Literary Studies, University of Padova
Via E. Vendramini, 13, 35137, Padova, Italy
[2]Department of Information Engineering, University of Padova
Via G. Gradenigo, 6b, 35131, Padova, Italy
vanessa.bonato@phd.unipd.it, federica.vezzani@unipd.it, giorgiomaria.dinunzio@unipd.it

Abstract

Understanding specialized biomedical knowledge can be particularly challenging, posing significant barriers to the acquisition and use of medical information especially by patients. In this study, a methodology for drafting patient-centered explanations of concepts related to the gut-brain axis and related medical conditions is proposed. The explanations are specifically intended for patients affected by neurodegenerative diseases, who are experiencing cognitive decline. The methodology consists of the following steps: 1) the drafting of specialized definitions in the form of intensional definitions, which enable the structured representation of domain-specific knowledge, and 2) the simplification of specialized definitions into patient-centered explanations. In particular, explanations intended for patients are formulated using popular terms and plain language, considered as two complementary strategies aimed at enhancing the comprehension of specialized biomedical knowledge. This work lays the foundation for the future development of a terminology resource specifically designed to collect and systematically represent knowledge related to the gut-brain axis and associated health conditions.

Keywords: Medical Terminology, Patient-Centered Explanations, Intensional Definitions

1. Introduction

Medical language is characterized by a high level of complexity, which can hinder the comprehension of domain-specific knowledge by non-experts (Tercedor Sánchez and Prieto Velasco, 2013; Bernardis et al., 2025). Among the multiple factors underlying this complexity is the use of specialized terms in medical discourse, which can pose barriers to effective physician-patient communication (Giovagnoli et al., 2024; Wahrenbrock et al., 2025). The syntactic structures found in medical language could also be a source of difficulty for non-experts (Vecchiato and Gerolimich, 2013). Moreover, a limited degree of health literacy can affect the comprehension of medical terminology in individuals without domain expertise (Makhmutova et al., 2025).

These factors, which impact the understanding of domain-specific knowledge, are a subset of the broader challenges to consider in the popularization of specialized knowledge related to the gut-brain axis. Indeed, this domain of study, which investigates the potential link between the gut microbiota and different neurodegenerative diseases (Roy Sarkar and Banerjee, 2019; Li et al., 2024), has seen a rising volume of related PubMed biomedical publications over the last years (Martinelli et al., 2026, 2025; Nentidis et al., 2026). From a terminological standpoint, ongoing scientific advances have led to the emergence of new terms and concepts, and the reconceptualization of existing concepts. Thus, related knowledge requires systematic terminological analysis, accurate representation, and effective dissemination.

These objectives are included within the larger aims of the European-supported project HEREDITARY (HetERogeneous sEmantic Data integratIon for the guT-bRain interplaY).[1] The project aims to investigate the gut-brain axis and related medical conditions from a clinical perspective, integrating the use of machine learning and artificial intelligence for multimodal data management, and embedding a social dimension involving citizen science.

Within this social dimension, a central goal at the terminological level is to represent specialized knowledge on the gut-brain axis and related health conditions in a terminology resource. The resource will be publicly released, to enable the effective dissemination of knowledge related to the gut–brain interplay. The intended users are physicians, language professionals, and patients. In particular, inclusivity in specialized knowledge representation must be ensured, as the resource will be aimed at patients affected by different levels of cognitive decline due to neurodegenerative diseases. The varied audience of the resource calls for a tailored representation of specialized knowledge, to address the respective information needs of the users. In light of this, specialized definitions need to be drafted to convey domain-specific knowledge.

[1] https://hereditary-project.eu

Proceedings of the 2nd Workshop on Evaluating Text Difficulty in a Multilingual Context (DeTermIt! 2026), pages 33–41
11 May 2026.

However, patient-centered explanations also need to be formulated, to enhance comprehension of concepts among patients.

This preliminary study lays the foundation for the future development of the terminology resource, by presenting a methodology for drafting patient-centered explanations that builds on a previous related work of Bonato et al. (2025). Specifically, we consider the drafting of specialized definitions as the starting point for the creation of patient-centered explanations. The formulation of explanations will rely on two strategies aimed at promoting comprehension: 1) the use of popular terms, and 2) the adoption of plain language. These strategies aim to meet the needs of the specific category of patients that will be targeted by the resource.

The work is organized as follows: Section 2 provides the theoretical background, focused on health literacy, the adoption of plain language in the medical domain, and studies that examine the linguistic difficulties experienced by patients affected by different neurodegenerative diseases. Section 3 outlines the methodology for drafting patient-centered explanations. Section 4 presents conclusions and future perspectives.

2. Theoretical Background

This section presents the theoretical background on health literacy, the use of plain language in healthcare, and studies on the linguistic patterns of patients affected by neurodegenerative diseases.

2.1. The Concept of Health Literacy

Health literacy represents a central concept in the framework of health communication and patient-centered healthcare. As defined by Berkman et al. (2010), health literacy is "[t]he degree to which individuals can obtain, process, understand, and communicate about health-related information needed to make informed health decisions".

Low health literacy can affect the understanding of medical texts, potentially impacting the safety of medical procedures. Smith et al. (2012) found that individuals with lower health literacy showed reduced comprehension of a colonoscopy preparation leaflet, and noted that inadequate understanding of the document may compromise the safety of the procedure. In relation to colonoscopy bowel preparation, Gwag and Yoo (2022) particularly highlight the need to account for the health literacy level of older patients.

Low health literacy can also influence the emotional sphere of people. As evidenced by Parikh et al. (1996), individuals with limited health literacy can experience feelings of shame, which in some cases lead them to choose not to mention their difficulties in reading medical materials to family members and healthcare professionals. As a result, shame prevents them from seeking help, and they lack the support to comprehend "prescriptions, follow-up appointments, recommended health care instructions, or informed consent documents".

2.2. Plain Language in Healthcare

In the healthcare setting, plain language can promote health literacy and foster immediate comprehension of medical information (Grene et al., 2017; Di Nunzio et al., 2024; Ermakova et al., 2024a,b). In ISO 24495-1, plain language is defined as "communication in which wording, structure and design are so clear that intended readers can easily find what they need, understand what they find, and use that information" (International Organization for Standardization, 2023). At the core of plain language is the aim of making medical knowledge accessible to non-experts, by adopting criteria that take into account the intended readers and their level of knowledge.

At the time of writing this article, ISO standard 24495-3 (International Organization for Standardization, in press), which provides guidelines for the use of plain language in science writing, is in the approval phase. However, ISO 24495-1 provides relevant principles aimed at drafting texts using plain language.

As indicated in the standard, from a linguistic standpoint, terms that are familiar to the reader and unambiguously designate concepts should be consistently used in texts. Specialized terms exclusively have to be used if readers can understand them and find them preferable, and "if readers need to learn them to achieve their goals". However, in the latter circumstance, specialized terms need to be accompanied by an accessible explanation at first mention.

At the syntactic level, sentences should ideally focus on a single notion, favoring unambiguous syntactic structures easily recognized by the reader. The use of the active voice is preferred, along with punctuation considered acceptable by readers.

Plain language has been adopted in the medical domain in numerous studies, also focusing on its positive impact on physician-patient communication (Peter et al., 2024). One of its applications is the creation of plain language summaries (PLSs), which are texts aimed at communicating medical research in a comprehensible way to non-experts, also generated through the use of large language models (Arias-Russi et al., 2025). For example, plain language summaries are provided on the web pages of scientific articles that present research on neurodegenerative diseases, such as Alzheimer's disease (AD) (Frederiksen et al., 2024) and Parkinson's disease (PD) (Chou et al., 2026).

Concerning plain language summaries, a notable example in the medical setting is Cochrane Plain Language Summaries (Pitcher et al., 2022; Whiting and Davenport, 2023). These texts, written in English language, summarize the key information and findings contained in medical systematic reviews published by the Cochrane organization. In these texts, terms that are easy to understand and syntactically straightforward sentences are used to facilitate comprehension of medical information by non-experts, favoring access to domain-related knowledge.

2.3. Language in Neurodegenerative Diseases

Patients affected by neurodegenerative diseases, however, may experience particular difficulties in language-related tasks (Cummings, 2025). As affirmed by Gumus et al. (2024), this category of patients "tend[s] to use simpler vocabulary and syntax; shorter words and fewer prepositional phrases, reflecting cognitive impairment". Thus, these patterns are linked to the underlying health condition.

Pan et al. (2024) evidence that patients affected by Parkinson's disease manifest impairments at different linguistic levels, namely "morpho-syntactic, lexical-semantic, and pragmatic". Some patients find it difficult to understand sentences with complex structures (Angwin et al., 2006).

Complexity in understanding sentences due to syntax is likewise observed in patients affected by Alzheimer's disease (Nasiri et al., 2022). From a terminological viewpoint, it is particularly interesting to note that "attributes which are more salient for the identification of a given concept are also those most resistant to semantic memory degradation in AD pathology" (Perri et al., 2019).

Taken together, these considerations are relevant for the formulation of patient-centered explanations, as the specific linguistic features of these patients need to be reflected in both the linguistic and conceptual dimensions of terminology.

3. A Methodology for Drafting Patient-Centered Explanations

As outlined above, in the terminology resource that will be developed within the HEREDITARY project, knowledge related to the gut–brain axis will be provided to healthcare professionals, language professionals, and patients affected by neurodegenerative diseases. These users manifest different information needs, especially with regard to the level of specialization of the medical information they need to access.

To understand the specific needs of patients, within the HEREDITARY project, interaction with them takes place during the Health Social Labs (Pellegrini et al., 2025), which are events designed to facilitate dialogue among researchers, physicians, patients, patient representatives, and caregivers.

During the meeting held in Padua in 2024, it has been possible to explore the difficulties that patients experience with medical terminology. In that occasion, terminology experts engaged with patients, asking them for examples of terms used by clinicians in communication that were unfamiliar to them, or domain-specific concepts that are difficult to understand. However, understanding the needs of patients requires more than identifying inadequately explained concepts used in physician–patient communication. As a matter of fact, it is fundamental to ask them which information they actually consider useful. For instance, patients emphasized the need for physicians to use terms that makes it easier for them to understand the exact prescribed dosage of medications, given its practical implications for managing symptoms in daily life. However, patients also expressed the necessity to access specialized knowledge on the medical conditions they are experiencing, to gain a clear and exact understanding of their health status. Indeed, patients do not seek an over-simplification of the conceptual dimension of medical terminology; rather, they need to access domain-specific knowledge with the support of explanations that aid the comprehension of concepts, making information clearer and more accessible. Moreover, they wish to acquire knowledge of the specialized terms used by healthcare providers, so that these terms are familiar to them when used in physician–patient communication. More specifically, their key requests are: 1) the use of specialized terms in physician–patient communication, 2) the use of terms that physicians typically use to communicate with patients who are not affected by neurodegenerative diseases, and 3) access to explanations of medical concepts.

Based on this, in the future terminology resource, the specialized definitions of medical concepts will be supplemented by patient-oriented explanations, the latter aimed at facilitating the popularization of knowledge on the gut-brain axis. In particular, providing patients with access to specialized definitions of biomedical concepts enables them to gain insight into specialized knowledge as well as the terms used by healthcare professionals to convey that knowledge. At the same time, patient-oriented explanations, which constitute a simplification of specialized definitions, make it easier to understand specialized information, thereby supporting the dissemination of knowledge. Patient empowerment is therefore achieved through terminology, as specialized knowledge is made inclusive and accessible.

In the following, we present a methodology for drafting patient-centered explanations. We con-

sider intensional definitions as the first step toward the drafting of explanations of concepts, serving systematically as the source of knowledge on which explanations are based. This methodology has been adopted to build a dataset of over 1,200 intensional definitions of concepts related to the gut–brain interplay and related medical conditions. Therefore, intensional definitions for more than 1,200 biomedical concepts have been drafted. The dataset also includes patient-centered explanations of biomedical concepts and is currently being further expanded.

3.1. Drafting of Intensional Definitions

The intensional definition is defined in ISO 704 (International Organization for Standardization, 2022) as a "definition that conveys the intension of a concept by stating the immediate superordinate concept and the delimiting characteristic(s)". Thus, in this type of definition, the first concept mentioned helps to precisely situate the defined concept within the concept system, establishing a hierarchical relation with respect to it. The delimiting characteristics, instead, perform a dual function: 1) they allow the comprehension of the defined concept, and 2) they differentiate the defined concept from other concepts that, within the same concept system, share the same generic concept.

In this study, intensional definitions are used to represent specialized knowledge related to the gut-brain axis. As a matter of fact, in the ISO standard, this type of terminological definition is recommended for the purpose of terminology work, since it enables the identification of the characteristics that distinguish one concept from others.

However, representing knowledge specific to the biomedical domain presupposes that terminologists have already gained this knowledge. To achieve this, it is essential to acquire information about the concepts to be defined relying on specialized resources, within which the specialized knowledge shared by experts is conveyed.

Taking this into account, the process of drafting intensional definitions for each concept involves three distinct sequential steps: 1) collecting multiple definitions from reliable biomedical sources to acquire specialized knowledge about the concept, 2) identifying the immediate generic concept, and 3) determining the delimiting characteristics.

The resources consulted include biomedical ontologies, with particular reference to the NCI Thesaurus (NCIT),[2] the Chemical Entities of Biological Interest (ChEBI) ontology, [3] and the Gene Ontology (GO)[4]. Other sources are the Medical Subject Headings (MeSH),[5] the Unified Medical Language System (UMLS),[6] the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text revision (DSM-5-TR) (American Psychiatric Association, 2022), and the International Classification of Diseases, 11th Revision (ICD-11)[7]. In addition, scientific papers are also considered as a reference, mainly retrieved from PubMed.[8] We specifically prioritize recently published scientific papers, to ensure that the representation of knowledge related to biomedical concepts is aligned with the constant scientific evolution that characterizes the domain of the gut-brain axis.

The consultation of multiple sources is also fundamental to identify the immediate generic concept and the delimiting characteristics of the concept to be defined. As a matter of fact, in many cases, different sources respectively provide different definitions of the same concept, and may each state a different generic concept. Comparing the definitions, thus, proves to be a crucial process for identifying the immediate generic concept. This comparison also allows for the identification of the essential and delimiting characteristics needed for comprehending the concept, distinguishing them, for instance, from non-essential characteristics.

An example of an intensional definition is the following definition of the concept <Bacterium>:

> microorganism that is unicellular, prokaryotic, and reproduces by cell division

In our concept system, the characteristics stated in this intensional definition are sufficient to distinguish bacteria from other types of microorganisms. The definition is specifically formulated according to the guidelines established in ISO 704, as it is “a statement in the form of an incomplete sentence without a full stop”. Moreover, it contains specialized terms, known to experts who are already familiar with the concepts linguistically designated by the terms and used by them in the medical setting.

Another example of intensional definition is the following definition of the concept <Multiple sclerosis>, which is regarded as one of the diseases showing an association with the gut–brain axis:

> demyelinating disease of the central nervous system that is characterized by destruction of myelin by the immune system

In this definition, the immediate generic concept <Demyelinating disease> is followed by the delimiting characteristics of the defined concept. The

[2] http://purl.obolibrary.org/obo/ncit.owl

[3] http://purl.obolibrary.org/obo/chebi.owl

[4] http://purl.obolibrary.org/obo/go.owl

[5] https://www.ncbi.nlm.nih.gov/mesh/

[6] https://uts.nlm.nih.gov/uts/umls/home

[7] https://icd.who.int/en/

[8] https://pubmed.ncbi.nlm.nih.gov

information provided concerns the affected anatomical site, namely the central nervous system, and the fact that the disease involves the destruction of myelin by the immune system. In particular, to formulate the intensional definition, the terminologists considered essential to include the information regarding the anatomical site specifically affected by demyelination in multiple sclerosis, as demyelination may also involve the destruction of myelin located in the peripheral nervous system. The definition conveys the specialized knowledge held by experts, who do not need, for instance, to understand what a demyelinating disease is or the concept of myelin.

In this work, we consider the formulation of intensional definitions as the first crucial step that provides the foundation for the drafting of patient-centered explanations. This reasoning is informed by different considerations. In the first place, to effectively convey specialized knowledge to patients, it is essential to first analyze, understand, and represent domain-related knowledge, particularly by identifying the delimiting characteristics of defined concepts. Intensional definitions precisely allow to unambiguously define a concept and, in doing so, to clearly distinguish it from other concepts that share common characteristics.

Secondly, using intensional definitions as the starting point for drafting explanations ensures that the delimiting characteristics of each defined concept are also represented in the explanations. This can be particularly important for patients, as it can help them gain a clearer understanding of the condition they are affected by, and distinguish it from other conditions.

The distinction between intensional definitions and explanations is grounded in their respective purposes. Intensional definitions are specifically aimed at defining concepts and establishing precise boundaries between them, as each concept is regarded as "a unit of knowledge created by a unique combination of characteristics." (International Organization for Standardization, 2019). The aim of intensional definitions, therefore, is not to simplify specialized knowledge, but rather to convey and represent it faithfully by listing the characteristics that distinguish concepts. Explanations, in contrast, are intended to make specialized knowledge understandable to a non-expert audience, given that their purpose is to explain specialized knowledge and make it more comprehensible. In light of this, the future terminology resource will provide patients with both intensional definitions and patient-centered explanations. The integration of both specialized definitions and patient-centered explanations is intended to accurately represent and disseminate specialized knowledge, thereby enabling patients to become informed patients. Additional information that may support and further enhance the understanding of biomedical concepts by patients will also be provided in the future terminology resource, presented as supplementary notes within the respective concept entries.

3.2. Formulation of Patient-Centered Explanations

As previously mentioned, unlike intensional definitions, patient-centered explanations are designed to facilitate the understanding of the concept. For this reason, it is necessary to simplify the definition, limiting the use of specialized terms and adopting a recognizable syntactic structure. In the work by Bonato et al. (2025), two strategies are used to draft patient-centered explanations. The first strategy is the use of popular terms, that are more easily understood. Secondly, plain language is adopted, to ensure that the text is also structured in a way that allows patients to clearly comprehend medical information.

Based on the knowledge contained in the intensional definition of <Bacterium>, by adopting these strategies, the following explanation is proposed:

> A bacterium is a microscopic organism composed of a single cell. This cell does not have a nucleus. It also does not have organelles that are surrounded by a membrane, that can be considered the organs of the cell. The bacterium can reproduce by dividing into other cells.

As can be observed, several changes have been made compared to the intensional definition. At the textual level, the intensional definition consists of one sentence, whereas the explanation contains four separate sentences. The text is therefore longer; however, it aligns with the guidelines provided in ISO 24495-1, according to which sentences should contain a limited amount of information and focus on a single idea. Notably, this allows for greater clarity in explaining the different characteristics of the concept. The sentences are short and include popular terms, which allow patients to understand the text at the linguistic level, while providing access to specialized knowledge.

In particular, the concept designated by the term "prokaryotic" that appears in the intensional definition is explained through a sentence that makes the concept more understandable: "[t]his cell does not have a nucleus. It also does not have organelles that are surrounded by a membrane, that can be considered the organs of the cell". This sentence partly adopts the wording used in the MSD Manual Consumer Version, a medical resource for non-experts, in which it is stated that organelles "could

be considered the cell's organs".[9] In addition, the term "cell division" is not included in the explanation, replaced by the expression "dividing into other cells", to easily explain the cellular process.

The same strategies are used to draft the patient-centered explanation of the concept <Multiple sclerosis>, based on the information conveyed in the respective intensional definition:

> Multiple sclerosis is a disease. This disease affects the brain, the spinal cord, and the optic nerve. In this disease, the immune system destroys the substance that surrounds nerve fibers.

As exemplified in the previously proposed explanation, the text is composed of more than one sentence, to facilitate the understanding of the different delimiting characteristics of the defined concept. As can be noted, the specialized terms "demyelinating disease", "central nervous system" and "myelin" are excluded from the explanation, as they may be unfamiliar to the patients. In particular, this explanation also adopts the text simplification strategies and wording used in the MSD Manual Consumer Version, where the concept of <Multiple sclerosis> is presented as follows: "In multiple sclerosis, patches of myelin (the substance that covers most nerve fibers) and underlying nerve fibers in the brain, optic nerves, and spinal cord are damaged or destroyed".[10]

In the explanation, the terminologists likewise decided to specify the parts of the central nervous system that are affected by demyelination in multiple sclerosis. The information provided immediately afterward concerns the fact that the immune system destroys myelin. Myelin is referred to as a substance, using the generic concept <Substance> rather than <Myelin>, which may facilitate understanding of the concept. The function of myelin is also mentioned, which is essential for helping the patient understand the specific type of substance referred to, avoiding the use of the specialized term "myelin".

To draft explanations, thus, it is also relevant to consult existing medical resources aimed at non-experts, as they can help in identifying terms that are considered more easily understood.

This methodology is therefore characterized by the contextual consultation of medical resources aimed at both experts and non-experts. This combination allows for the creation of patient-centered explanations that reflect the needs of patients, which concern both the linguistic and conceptual dimensions of terminology.

4. Conclusions and Future Work

In this work, we outlined the methodology for the formulation of explanations of concepts related to the gut-brain axis, targeted at patients affected by neurodegenerative diseases. Building on intensional definitions, explanations can convey specialized knowledge to patients, facilitating the understanding of biomedical concepts.

We plan to further refine the methodology presented. In particular, we will explore the adoption of easy language (Maaß, 2020; Pedrini, 2022) to draft patient-centered explanations, as easy language targets users that present cognitive decline and intellectual difficulties. We will focus on the drafting of different patient-centered explanations tailored to the respective varying levels of cognitive decline experienced by patients. We plan to conduct a study in which patients are presented with the drafted explanations and asked to provide feedback on the syntax and the terms used. For this purpose, questionnaires could be used to collect patient feedback. This process will allow the terminologists to identify any challenges that could hinder the comprehension of medical knowledge. In addition, we plan to engage health professionals in the validation of both intensional definitions and patient-centered explanations. In this context, we consider the evaluation of patient-centered explanations within shared tasks and benchmarking initiatives, such as DETECH 2026,[11] to assess the effectiveness of different strategies for improving accessibility and comprehension of specialized knowledge. Specialized intensional definitions and patient-oriented explanations will be included as terminological data in the future terminology resource. Within the resource, intensional definitions and explanations will be provided in multiple languages, to further support the dissemination of knowledge about the gut–brain axis.

5. Acknowledgements

This work is partially supported by the HEREDITARY Project, as part of the European Union's Horizon Europe research and innovation programme under grant agreement No GA 101137074, and it is part of the initiatives of the Center for Studies in Computational Terminology (CENTRICO) of the University of Padua and in the research directions of the Italian Common Language Resources and Technology Infrastructure CLARIN-IT.

[9] https://www.msdmanuals.com/home/fundamentals/the-human-body/cells?query=cells

[10] https://www.msdmanuals.com/home/brain-spinal-cord-and-nerve-disorders/multiple-sclerosis-ms-and-related-disorders/multiple-sclerosis-ms

[11] https://detech2026.dei.unipd.it/

6. Bibliographical References

American Psychiatric Association. 2022. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5-TR)*, 5th edition, text revision edition. American Psychiatric Association Publishing.

Anthony J. Angwin, Helen J. Chenery, David A. Copland, Bruce E. Murdoch, and Peter A. Silburn. 2006. Self-paced reading and sentence comprehension in Parkinson's disease. *Journal of Neurolinguistics*, 19(3):239–252.

Felipe Arias-Russi, Carolina Salazar-Lara, and Rubén Manrique. 2025. Bridging the gap in health literacy: Harnessing the power of large language models to generate plain language summaries from biomedical texts. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 269–284, Albuquerque, New Mexico. Association for Computational Linguistics.

Nancy D. Berkman, Terry C. Davis, and Lauren McCormack. 2010. Health literacy: What is it? *Journal of Health Communication*, 15(sup2):9–19. PMID: 20845189.

Andrea Bernardis, Federica Vezzani, and Giorgio Maria Di Nunzio. 2025. Pour une simplification de la terminologie médicale multilingue : le cas du projet ExaMode. *mediAzioni*, 47:A214–A234.

Vanessa Bonato, Federica Vezzani, and Giorgio Maria Di Nunzio. 2025. Advancing inclusivity in a medical terminology resource concerning the gut-brain interplay: Insights from the HEREDITARY Project. In *Semmelweis Medical Linguistics Conference 2025 Book of Abstracts*, Budapest, Hungary. Institute of Languages for Specific Purposes, Semmelweis University.

Shantao Chloe Chou, Cen Cong, Rosiered Brownson-Smith, Madison Milne-Ives, and Edward Meinert. 2026. Assessment of mental and behavioural non-motor symptoms of Parkinson's Disease using artificial intelligence (AI): A systematic review. *Communications Medicine*, 6:101.

Louise Cummings, editor. 2025. *The Oxford Handbook of Communication Disorders in Neurodegenerative Diseases*. Oxford University Press.

Giorgio Maria Di Nunzio, Federica Vezzani, Vanessa Bonato, Hosein Azarbonyad, Jaap Kamps, and Liana Ermakova. 2024. Overview of the CLEF 2024 SimpleText Task 2: Identify and Explain Difficult Concepts. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*, volume 3740 of *CEUR Workshop Proceedings*, pages 3129–3146, Grenoble, France. CEUR.

Liana Ermakova, Eric SanJuan, Stéphane Huet, Hosein Azarbonyad, Giorgio Maria Di Nunzio, Federica Vezzani, Jennifer D'Souza, Salomon Kabongo, Hamed Babaei Giglou, Yue Zhang, Sören Auer, and Jaap Kamps. 2024a. CLEF 2024 SimpleText Track. In *Advances in Information Retrieval*, pages 28–35, Cham. Springer Nature Switzerland.

Liana Ermakova, Eric SanJuan, Stéphane Huet, Hosein Azarbonyad, Giorgio Maria Di Nunzio, Federica Vezzani, Jennifer D'Souza, and Jaap Kamps. 2024b. Overview of the CLEF 2024 SimpleText Track. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 283–307, Cham. Springer Nature Switzerland.

Kristian Steen Frederiksen, Xavier Morato Arus, Henrik Zetterberg, Serge Gauthier, Mercé Boada, Julie Hahn-Pedersen, Luis Rafael Solís Tarazona, and Soeren Mattke. 2024. Focusing on earlier diagnosis of Alzheimer's disease: A plain language summary. *Future Neurology*, 19(1):2419271.

Alice Giovagnoli, Federica Vezzani, and Giorgio Maria Di Nunzio. 2024. Doctor-patient communication: terminological and simplification analysis in the domain of female oncology. *Umanistica Digitale*, 8(17):143–164.

Margaret Grene, Yvonne Cleary, and Ann Marcus-Quinn. 2017. Use of plain-language guidelines to promote health literacy. *IEEE Transactions on Professional Communication*, 60(4):384–400.

Melisa Gumus, Morgan Koo, Christa M. Studzinski, Aparna Bhan, Jessica Robin, and Sandra E. Black. 2024. Linguistic changes in neurodegenerative diseases relate to clinical symptoms. *Frontiers in Neurology*, Volume 15 - 2024.

Minju Gwag and Jaeyong Yoo. 2022. Relationship between health literacy and knowledge, compliance with bowel preparation, and bowel cleanliness in older patients undergoing colonoscopy. *International Journal of Environmental Research and Public Health*, 19(5).

International Organization for Standardization. 2019. ISO 1087:2019 — Terminology work and terminology science — Vocabulary. https://www.iso.org/standard/62330.html.

International Organization for Standardization. 2022. ISO 704:2022 — Terminology work —

Principles and methods. https://www.iso.org/standard/79077.html.

International Organization for Standardization. 2023. ISO 24495-1:2023 — Plain language: Part 1: Governing principles and guidelines. https://www.iso.org/standard/78907.html.

International Organization for Standardization. in press. ISO/FDIS 24495-3 — Plain language: Part 3: Science writing. https://www.iso.org/standard/86938.html. Final Draft International Standard (FDIS), under approval.

Songlin Li, Linna Zhao, Jie Xiao, Yuying Guo, Rong Fu, Yunsha Zhang, and Shixin Xu. 2024. The gut microbiome: an important role in neurodegenerative diseases and their therapeutic advances. *Molecular and Cellular Biochemistry*, 479(9):2217–2243.

Christiane Maaß. 2020. *Easy Language – Plain Language – Easy Language Plus*. Frank & Timme.

Liliya Makhmutova, Giancarlo Dondoni Salton, Fernando Perez-Tellez, and Robert J. Ross. 2025. The evaluation of medical terms complexity using lexical features and large language models. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 682–693, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Marco Martinelli, Stefano Marchesin, Vanessa Bonato, Giorgio Maria Di Nunzio, Nicola Ferro, Ornella Irrera, Laura Menotti, Federica Vezzani, and Gianmaria Silvello. 2026. A domain-specific curated benchmark for entity and document-level relation extraction. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5693–5711, Rabat, Morocco. Association for Computational Linguistics.

Marco Martinelli, Gianmaria Silvello, Vanessa Bonato, Giorgio Maria Di Nunzio, Nicola Ferro, Ornella Irrera, Stefano Marchesin, Laura Menotti, and Federica Vezzani. 2025. Overview of GutBrainIE@CLEF 2025: Gut-Brain Interplay Information Extraction. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025)*, volume 4038 of *CEUR Workshop Proceedings*, pages 65–98, Madrid, Spain. CEUR.

Maryam Nasiri, Saeideh Moayedfar, Mehdi Purmohammad, and Leila Ghasisin. 2022. Investigating sentence processing and working memory in patients with mild Alzheimer and elderly people. *PLOS ONE*, 17(11):e0266552.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, Martin Krallinger, Miguel Rodríguez-Ortega, Eduard Rodriguez-López, Natalia Loukachevitch, Andrey Sakhovskiy, Elena Tutubalina, Dimitris Dimitriadis, Grigorios Tsoumakas, George Giannakoulas, Alexandra Bekiaridou, Athanasios Samaras, Giorgio Maria Di Nunzio, Nicola Ferro, Stefano Marchesin, Marco Martinelli, Gianmaria Silvello, and Georgios Paliouras. 2026. Overview of BioASQ 2025: The Thirteenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 173–198, Cham. Springer Nature Switzerland.

Xueyao Pan, Bingqian Liang, and Ting Cao. 2024. A bibliometric analysis of speech and language impairments in Parkinson's disease based on Web of Science. *Frontiers in Psychology*, Volume 15 - 2024.

Nina S. Parikh, Ruth M. Parker, Joanne R. Nurss, David W. Baker, and Mark V. Williams. 1996. Shame and health literacy: the unspoken connection. *Patient Education and Counseling*, 27(1):33–39.

Giulia Pedrini. 2022. Plain and easy language as a means to increase health literacy on COVID-19: A contrastive analysis of english and german texts. *trans-kom*, 15(1):142–155.

Giuseppe Pellegrini, Chiara Lovati, Rute Costa, and Anna Romanovych. 2025. Deliverable 6.2: Health social lab activities. Zenodo.

Roberta Perri, Giovanni Augusto Carlesimo, Marco Monaco, Carlo Caltagirone, and Gian Daniele Zannino. 2019. The attribute priming effect in patients with Alzheimer's disease. *Journal of Neuropsychology*, 13(3):485–502.

Maryke Peter, Stacy Maddocks, Clarice Tang, and Pat G Camp. 2024. Simplicity: Using the power of plain language to encourage patient-centered communication. *Physical Therapy*, 104(1):pzad103.

Nicole Pitcher, Denise Mitchell, and Carolyn Hughes. 2022. Template and guidance for writing a Cochrane Plain language summary. In Julian Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew Page, and Vivian Welch, editors, *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane.

Suparna Roy Sarkar and Sugato Banerjee. 2019. Gut microbiota in neurodegenerative disorders. *Journal of Neuroimmunology*, 328:98–104.

Samuel G. Smith, Christian von Wagner, Lesley M. McGregor, Laura M. Curtis, Elizabeth A. H. Wilson, Marina Serper, and Michael S. Wolf. 2012. The influence of health literacy on comprehension of a colonoscopy preparation information leaflet. *Diseases of the Colon & Rectum*, 55(10):1074–1080.

Maribel Tercedor Sánchez and Juan Antonio Prieto Velasco. 2013. Las barreras en la comunicación médico-paciente: el proyecto VariMed. In Ana Belén López, Isabel Jiménez, and Isabel Martínez, editors, *Translating Culture*, pages 593–605. Editorial Comares, Albolote (Granada).

Sara Vecchiato and Sonia Vanna Gerolimich. 2013. La langue médicale est-elle « trop complexe » ? *Nouvelles perspectives en sciences sociales*, 9(1):81–122.

Taylor Wahrenbrock, Kelly Landry, Dhara P. Amin, Lum Rizvanolli, Ranganathan Chandrasekaran, Mark B. Mycyk, and Joanne C. Routsolias. 2025. Medical jargon is often misunderstood by emergency department patients. *The American Journal of Emergency Medicine*, 96:25–29.

Penny Whiting and Clare Davenport. 2023. Writing a plain language summary. In Jonathan J. Deeks, Patrick M. Bossuyt, Mariska M. Leeflang, and Yemisi Takwoingi, editors, *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*, chapter 13, pages 377–397. John Wiley & Sons, Ltd.

Translation as Augmentation: Effect of Translated Data on Assessment of Difficulty

Yiheng Wu, Jue Hou, Roman Yangarber
University of Helsinki, Finland
first.last@helsinki.fi

Abstract

Reliable Text Difficulty Assessment is a prerequisite for valid text simplification workflows and personalized learning applications. However, the development of robust assessment models is severely hindered by a critical bottleneck: the scarcity of expert-annotated corpora containing fine-grained difficulty levels (e.g., CEFR), particularly for lower-resource languages. This paper addresses this data scarcity problem in the context of a low-resource European language. We propose a cross-lingual data augmentation strategy that leverages machine translation to transfer labeled resources from high-resource languages to the target low-resource language. We train BERT-based regression models to predict difficulty scores and investigate whether synthetic, translated data can effectively supplement native training sets. Our experiments demonstrate that augmenting scarce native data with machine-translated corpora significantly improves the accuracy of difficulty estimation, offering a viable solution for languages lacking extensive expert annotations.

1. Introduction

Assessment of text complexity—Difficulty Assessment—has become increasingly important for accessible communication practice, driven by policy mandates in the United States[1] and the European Union[2]. It also plays an essential role in personalized second-language (L2) instruction (Nahatame and Yamaguchi, 2026). Compliance with these mandates and the personalization of L2 instruction both hinge on this shared technical bottleneck: the reliable estimation of text difficulty. We argue that difficulty assessment is a prerequisite for the task of automatic text simplification: a simplification system cannot be meaningfully guided or evaluated without reliable metrics to determine whether its output meets the target difficulty specification.

Consequently, this paper focuses on the assessment problem. We model this problem as a regression task, where a *difficulty model* predicts a continuous score, which can then be mapped to the CEFR scale.[3] Such models are critical not only for assessing learner texts but also for acting as critics in Large Language Model (LLM)-based simplification pipelines (Hurst et al., 2024). A central obstacle to building robust difficulty models is data scarcity. High-quality assessment requires sizable corpora annotated by domain experts, a resource that is lacking, e.g., in many lower-resource European languages, such as Finnish.

To address this bottleneck, we propose a cross-lingual approach: leveraging existing expert-annotated corpora from higher-resource languages and applying machine translation to generate synthetic labeled training data in the target language. We address two **Research Questions**:

1. Can training data augmented with machine-translated data from a higher-resource language improve the quality of difficulty assessment in a lower-resource language?
2. To what extent does training on translated data improve cross-lingual generalization between the source and target languages?

We train a BERT-based *regression* model to predict difficulty scores on a continuous scale. Our results indicate that using machine-translated data can substantially improve the accuracy and robustness of difficulty assessment in the target language. The paper is organized as follows. Section 2 presents an overview of related work. Section 3 describes the datasets used for training. Section 4 details the experimental setup and results. Section 5 provides conclusions and future directions.

2. Related Work

Estimating the difficulty of written text—variously referred to as readability, proficiency, or grade-level assessment[4]—has a long history in both educational research and NLP. Early formula-based approaches such as Flesch-Kincaid and the Lexile framework provide simple numeric difficulty scores (Kincaid et al., 1975; Stenner, 1996), but rely on surface-level features and fail to capture deeper lexical or syntactic complexity. Subsequent supervised systems incorporated richer handcrafted linguistic features, including parse depth,

[1] Plain Writing Act of 2010, Pub. L. 111-274

[2] European Accessibility Act

[3] Behindertengleichstellungsgesetz (BGG), Federal Republic of Germany, 2002 (amended 2016)

[4] Throughout this paper, we use these terms interchangeably.

Proceedings of the 2nd Workshop on Evaluating Text Difficulty in a Multilingual Context (DeTermIt! 2026), pages 42–50

grammatical constructions, and word-frequency lists (Collins-Thompson and Callan, 2004; Vajjala and Meurers, 2012; Laposhina et al., 2018). More recent neural approaches—from hierarchical attention networks (Azpiazu and Pera, 2019) to fine-tuned BERT models (Martinc et al., 2021)—have substantially outperformed feature-based baselines, and Transformer-based models have been compared against feature-engineered systems for both English and Russian (Sharoff, 2022). Large language models have also been evaluated on readability tasks with competitive results (Imperial and Tayyar Madabushi, 2024).

Difficulty assessment plays an equally important role within text simplification pipelines, both as a signal for identifying complex content (Gasperin et al., 2009; Aluísio et al., 2010) and as an optimization objective in rule-based systems (Woodsend and Lapata, 2011). Recent work has moved toward feedback-driven generation, using readability classifiers with reinforcement learning (Alkaldi and Inkpen, 2023) or controllable generation conditioned on target reading level (Agrawal and Carpuat, 2023)—a paradigm directly relevant to our use of a difficulty model as a critic in an LLM-based simplification loop (Hurst et al., 2024).

A persistent bottleneck across all these approaches is the scarcity of expert-annotated, difficulty-labeled corpora, which is especially acute for lower-resource languages. Even a recent shared task on English simplification provided no training data (Alva-Manchego et al., 2025). For Finnish, we build on existing annotated resources (Dmitrieva and Konovalova, 2023; Katinskaia et al., 2025) and extend them via machine translation of Russian-language corpora (Dmitrieva, 2025), directly targeting this labeled-data bottleneck.

3. Data

We first describe the Finnish- and Russian-language data used for training and evaluating the difficulty models. A major challenge is the scarcity of annotated data in Finnish for prediction of difficulty. To address this, we augment a small collection of texts in Finnish annotated with difficulty levels—"native" data—with a larger collection of texts in Russian that were annotated with difficulty levels, and then translated into Finnish using machine-translation (MT) models.

3.1. Native Data

We compile a dataset for Finnish by combining various native and machine-translated sources, spanning the range of CEFR readability levels (A1–C2). Table 1 provides an overview of the composition of this dataset. Native data is drawn from five main sources:

- *Easy Language* (EL)—a collection of texts from government and NGO websites, written in "Easy Language" for non-native speakers;
- *TextBook*—a collection of texts from textbooks for L2 learning, at various CEFR levels;
- *Helsingin Sanomat* (HS)—a commercial news site with the widest coverage nationally;
- *YLE*—a government news site;
- *YLE Selkouutiset* (Selko)—YLE's simplified news for non-native speakers and L2 learners.

Each source contributes texts at different levels of linguistic complexity and genre. *Easy Language* and *YLE selkouutiset* provide simplified Finnish materials aimed at beginners and intermediate learners, while *YLE* and *Helsingin Sanomat* offer authentic journalistic texts at advanced CEFR levels (C1–C2). In total, the native corpus contains 4544 texts distributed across the CEFR scale.

We assign each text in this collection to one of 11 classes—these correspond to the 6 "principal" CEFR levels, plus 5 *intermediate* levels, i.e., A1+, A2+, B1+, etc. The rationale for introducing the intermediate levels is as follows. Some sources, such as textbooks, provide fine-grained assignment of the texts to the CEFR levels. However, other sources (e.g., news sites) yield only a coarse-grained *estimate* of difficult. Thus, we assume that newspaper texts are on a hypothetical level "C", approximately between C1 and C2.

It is also important to note that in modeling we make the assumption that the CEFR levels are *evenly* spaced on a linear difficulty scale. This is done because an exact spacing of the levels on the CEFR scale is *latent* and not known explicitly. For modeling, in Section 4.1, we map these levels onto a continuous numerical scale ranging from 1 to 6, preserving their relative order while enabling regression-based prediction of text difficulty.

To address the problem of data scarcity in Finnish—especially at the lower readability levels—we augment the corpus with machine-translated (MT) texts, which have CEFR annotations in the original. This process is described in detail in Section 3.2. The translated subset (labeled "MT ← RU" in Table 1) contains 8321 documents, which mirror the CEFR distribution of the native data to support balanced training. Including translated data allows us to examine whether CEFR-labeled content from a high-resource language can improve performance in a low-resource target language.

Across both native and translated sub-corpora, the dataset covers all CEFR levels (A1–C2), with a total of 12,865 texts. Lower levels (A1–A2+) are primarily sourced from *Easy Language*, *TextBook*, and translated materials, while mid-level

Level	EL	TextBook	HS	YLE	Selko	Native total	MT ← RU	Overall total
A1	0	1	0	0	0	1	294	295
A1+	153	0	0	0	0	153	282	435
A2	0	363	0	0	0	363	465	828
A2+	0	0	0	0	0	0	96	96
B1	0	229	0	0	0	229	3301	3530
B1+	0	0	0	0	766	766	1672	2438
B2	0	163	0	0	0	163	834	997
B2+	0	0	0	0	0	0	0	0
C1	0	192	0	0	0	192	484	676
C1+	0	0	715	703	0	1418	0	1418
C2	0	175	0	0	0	175	29	204
Total	153	1123	715	703	766	3460	7457	10917

Table 1: Number of documents in Finnish *native* and machine-translated (MT) datasets by CEFR Level

Source	CEFR	Level	Total # Docs	Average # Words	Average # Sent.
RuFoLa	A1	1.0	301	136	8.8
Encyclop.	A1-A2	1.5	282	31	12.3
RuFoLa	A2	2.0	466	183	10.5
Zlatoust	A2-B1	2.5	96	50	8.2
RuFoLa	B1	3.0	3306	91	12.2
Zlatoust	B1-B2	3.5	1677	54	15.8
Zlatoust	B2	4.0	834	228	12.8
RuFoLa	C1	5.0	485	363	14.9
RuFoLa	C2	6.0	29	385	16.5

Table 2: Annotated documents in Russian.

texts (B1–B2) are drawn from *YLE selkouutiset* and corresponding MT data. The advanced levels (C1–C2) are mainly represented by authentic Finnish news and literary texts from *HS* and *YLE*. This distribution ensures that the dataset reflects both pedagogically simplified and naturally complex usage of Finnish, enabling robust analysis of cross-lingual transfer and translation-based augmentation across readability levels. We compile a comprehensive Finnish dataset by combining multiple native and machine-translated sources, spanning a wide range of CEFR readability levels (A1–C2).

3.2. Translated Data

We use text resources in Russian that have been manually annotated for difficulty, and translate them into Finnish to augment the training dataset. We use a collection of Russian simple-language corpora, introduced in (Dmitrieva, 2025). Two corpora of annotated Russian texts, shown in Table 2:

- the *RuFoLa* corpus (Laposhina, 2020), which contains texts from coursebooks designed for learners of Russian as a foreign language;
- the *RuAdapt* corpus (Dmitrieva and Tiedemann, 2021), a *parallel* corpus of Russian–Simple Russian, with authentic texts adapted for learners of Russian as a foreign language. In this paper, we use only the literary (*Zlatoust*) and encyclopedic sub-corpora (*Encyclop.*).

Split	FI	RU (MT)	Both
Train	2,332	5,961	8,293
Dev	263	748	1,001
Test	865	748	1,613
Total	3,460	7,457	10,917

Table 3: Data splits by source language

We translate the Russian texts into Finnish using models from OpusMT.[5] It is critical to note that machine translation does not *guarantee* that a text in Russian will remain at the same difficulty level after translation into Finnish. This question—under what conditions and to what extent do MT models preserve the difficulty level of the original text—deserves detailed investigation on its own; we would expect that this would depend heavily on how the MT model is trained. However, these particular MT models with which we experiment do seem to exhibit a strong ability to preserve the difficulty level across the translation, as confirmed by manual inspection of a sample by native language experts.

Examining the instances in the Russian corpus, we find that some instances that are too short or too long. Therefore, we removed some outlier instances, and hence the number of *MT ← RU* documents in Table 1 is somewhat lower than the original Russian texts in Table 2. All documents—native and translated—were split into 3 sets: train-

[5] Tatoeba MT model for Slavic–Finnish.

Sample	TTR	Lexical Density	Mean Word Length	Mean Sent. Length	Mean Clause Length	POS Diversity
FI	0.6993	0.6930	7.40	8.65	19.67	1.90
RU	0.8262	0.5934	5.30	12.71	10.76	2.04
FI+MT	0.8598	0.6338	6.61	9.31	10.68	1.85

Table 4: Linguistic features for FI, RU and FI+MT samples.

Exp	Lang	Train & Dev Set	Test Set	MSE (%)	RMSE (%)	MAE (%)	R^2 (%)
1	FI	Native	Native	5.12	22.63	12.17	**97.30**
2	FI	Native	MT	146.84	121.18	102.59	-99.71
3	FI	MT	Native	155.00	124.50	109.25	18.24
4	FI	MT	MT	24.54	49.54	26.80	66.93
5	FI	Native + MT	Native	7.57	27.51	9.43	96.01
6	FI	Native + MT	MT	**4.22**	**20.55**	**8.24**	94.26
7	RU	Native	Native	19.19	43.81	26.04	73.90
8	RU	Native + MT	Native	**16.72**	**40.89**	**19.43**	**77.26**

Table 5: Model performance on difficulty prediction for Finnish and Russian text. MT denotes data autmentation via machine translation (Russian to Finnish). Experiments 1–8 test various training vs. test combinations of native vs. translated datasets.

ing, validation, and test, as shown in Table 3. For the experiments with Finnish, we translated the Russian training and validation sets, and added them to the corresponding Finnish sets, to augment the limited size of the Finnish sets. In contrast, for the experiments with Russian, we translated only the Finnish training set and incorporated it into the Russian training set, as the Russian development set was already sufficiently large.

Table 4 reports six metrics of linguistic complexity, which are commonly used to characterize text difficulty:

- Type-Token Ratio (TTR)—is a measure of vocabulary richness,
- Lexical Density—measures the proportion of content words,
- Mean Word Length—measures morphological variety,
- Mean Sentence Length—reflects syntactic variety,
- Mean Clause Length—reflects the structural complexity of sentences, capturing both the elaboration of clause-internal constituents and the depth of syntactic embedding,
- POS Diversity—indicates part-of-speech variety within the text.

These features roughly reflect the lexical and structural properties of the texts in the dataset. They are not used in modeling (at present), and are presented to give the reader an intuitive description of the data.

The main point in this Section is that the dataset augmented with translated data is 3 times larger than the original "native" dataset—which we hope will help train a more accurate regresion model.

4. Experiments

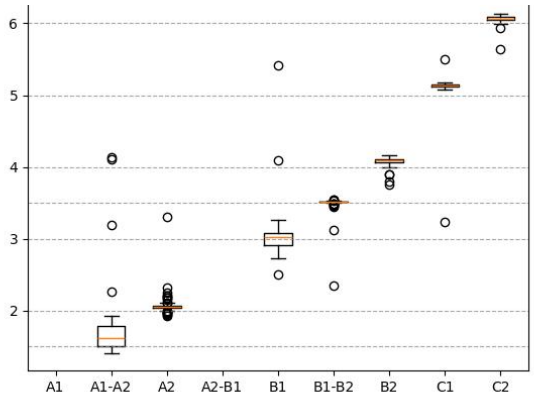
To examine whether machine-translated data can improve text difficulty prediction, we train regression models on native Finnish texts and translated Russian texts. This allows us to explore the research questions: assess how MT-based data augmentation influences model performance and cross-lingual generalization in predicting document-level difficulty.

4.1. Model settings

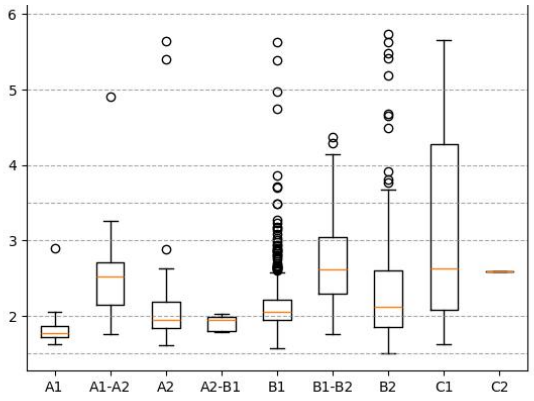
We build a BERT-based regression model to predict text difficulty. For Finnish, we use the TurkuNLP/bert-base-finnish-cased-v1 model; for Russian, we use ai-forever/ruBert-large. Both models are fine-tuned with a regression objective.

4.2. Experiments on RQ1

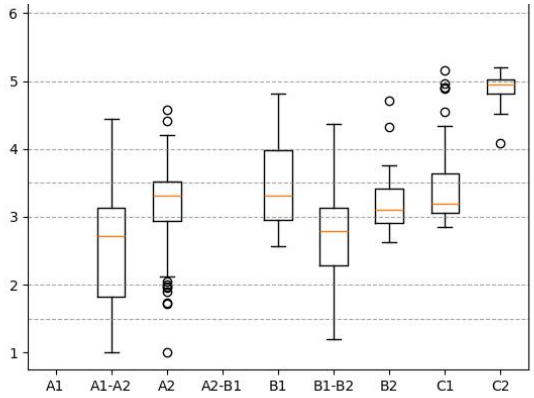
Table 5 presents the results in terms of key performance measures—MSE (mean quared error), RMSE (root mean square error), MAE (mean absolute error) and R^2 (coefficient of determination)—across different training and testing dataset configurations. Overall, the results show that incorporating machine-translated data from Russian improves text difficulty prediction for Finnish, com-



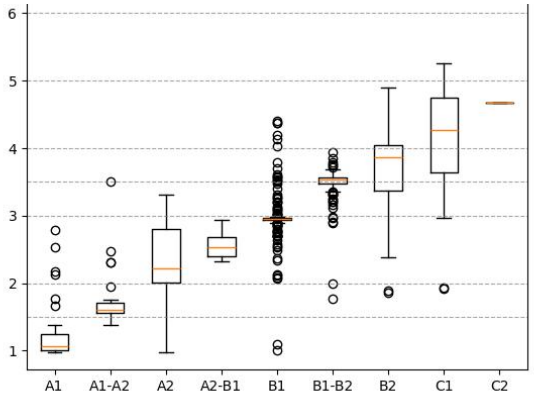
(1) FI train and test on native data



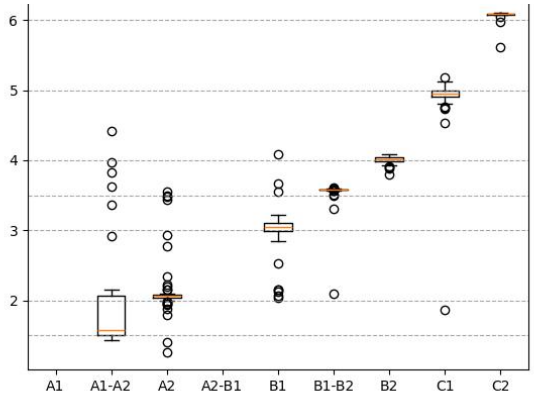
(2) FI train on native, test on MT



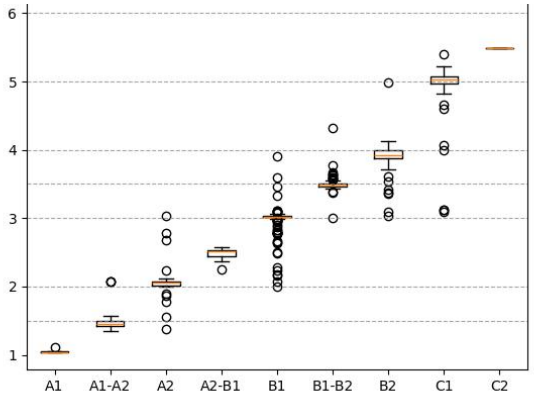
(3) FI train on MT, test on native data



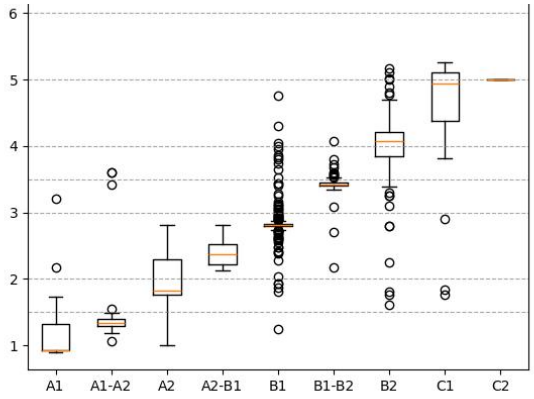
(4) FI train and test on MT



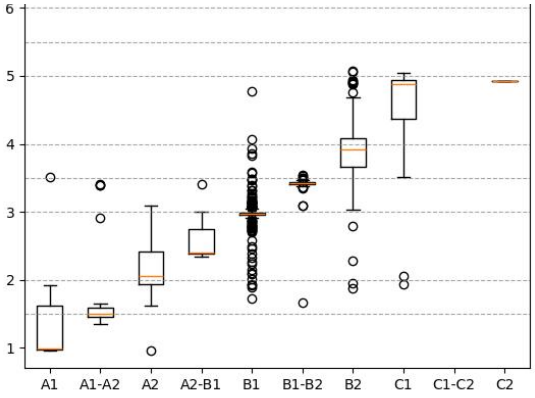
(5) FI train on all data, test on native,



(6) FI train on all data, test on MT,



(7) RU train and test on native data



(8) RU train on all data and test on native data

Figure 1: Prediction of difficulty, corresponding to experiments in Table 5.

Exp.	Lang.	Train Set	Dev Set	Test Set	MSE (%)	RMSE (%)	MAE (%)	R^2 (%)
Label Injection (whole-label inclusion)								
9	FI	Native + MT (A1–A2)	Native + MT	Native	29.29	54.12	41.45	84.55
10	FI	Native + MT (B1–B2)	Native + MT	Native	6.29	25.08	11.40	96.68
11	FI	Native + MT (C1–C2)	Native + MT	Native	16.96	41.18	25.62	91.06
Label-wise Proportional Sampling								
12	FI	Native + 20% MT	Native + MT	Native	7.83	27.97	16.81	95.87
13	FI	Native + 40% MT	Native + MT	Native	7.20	26.84	10.66	96.20
14	FI	Native + 60% MT	Native + MT	Native	8.22	28.68	14.72	95.66
15	FI	Native + 80% MT	Native + MT	Native	8.71	29.51	12.74	95.41
16	FI	Native + 100% MT	Native + MT	Native	7.57	27.51	9.43	96.01

Table 6: Model performance on Finnish text difficulty prediction using machine-translated (MT) Russian data. **Label Injection:** Entire CEFR-level subsets (A1–A2, B1–B2, C1–C2) of MT data are injected into the training set. **Label-wise Proportional Sampling:** Russian MT data are sampled proportionally within each CEFR label (20%–100%) and added to Finnish native data. Note: experiment 16 is exactly the same as line 5 in Table 5 (repeated for clarity).

pared to training on native data alone, or on translated data alone.

When using only translated Russian data (MT RU→FI), model performance remains modest, with MSE of 24.54% and R^2 of 66.93% (Figure 1.4). However, combining translated and native Finnish data—while applying dataset balancing—yields a dramatic performance gain , reaching MSE of 4.22%, MAE of 8.24%, and R^2 of 94.26%. The box plots in Figure 1.6 show notably tighter error distributions in this combined setting, indicating that the model may benefit from both the larger volume and the additional diversity provided by the translated corpus.

The best overall performance is achieved when the model is trained on the native + MT and evaluated on native Finnish data, reaching an R^2 of 96.01% in Figure 1.5. This demonstrates that machine-translated data not only contributes to improved prediction accuracy in a low-resource setting, but also supports generalization to unseen native Finnish texts. By contrast, models trained only on translated data perform poorly when tested on native Finnish data—MSE 155.00%, R^2 18.24%—which suggests that direct transfer without native exposure is insufficient (Figure 1.3).

Training exclusively on native Finnish data yields strong in-domain performance (R^2 97.30% in Figure 1.1), confirming the quality of native annotations. However, the combined model's comparable accuracy, despite including automatically translated material, shows that translation-based augmentation is an effective strategy for low-resource difficulty prediction.

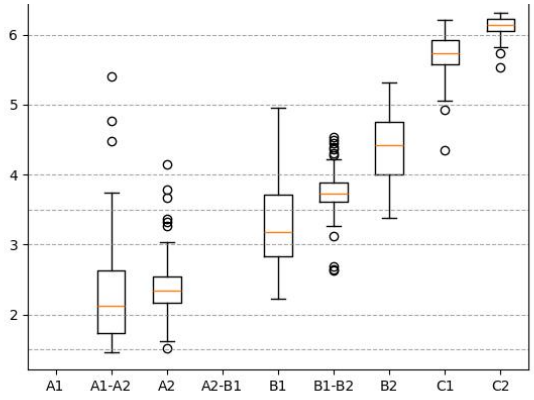
In sum, the box plots and quantitative results jointly confirm that (1) machine-translated data can substantially enhance low-resource Finnish performance.

4.3. Experiments on RQ2

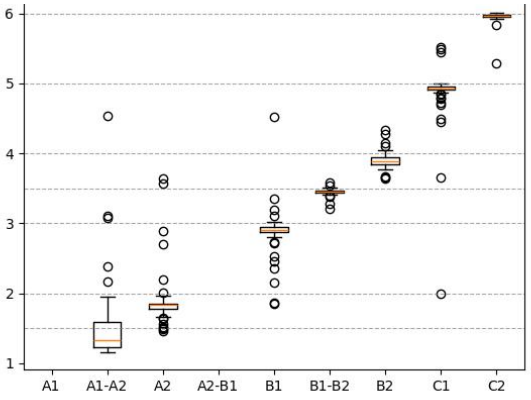
Table 6 presents the results of ablation studies, combining the native Finnish data with varying amounts of machine-translated (MT) data from the Russian corpus. We explored two strategies: label injection, where entire CEFR-level subsets of translated data (A1–A2, B1–B2, C1–C2) were added to the training set, and label-wise proportional sampling, where translated data were added in increasing proportions (20–100%) across all levels.

Under the label injection setting, we observe that including B1–B2-level translated texts yields the strongest improvement, achieving R^2 of 96.68% and the lowest overall error rates (Figure 2.2). This suggests that mid-level translated data contribute most effectively to modeling Finnish text difficulty, possibly because they provide balanced lexical and syntactic diversity without overwhelming the model with extreme examples from beginner or advanced levels. In contrast, adding low-level (A1–A2) (Figure 2.1) or high-level (C1–C2) (Figure 2.3 translated data results in noticeably higher error, indicating limited transferability at the edges of the CEFR scale.

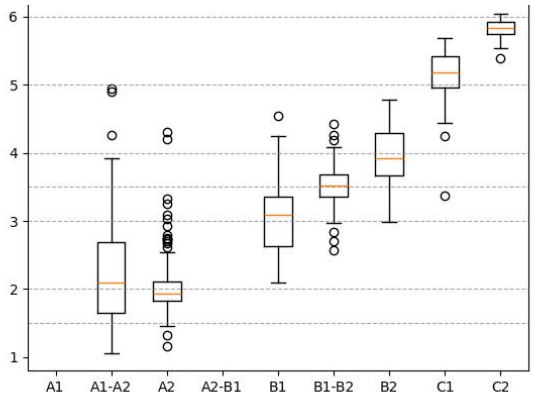
In the experiments with proportional sampling, performance is consistently high across all proportions, with minor fluctuations. The best result (Figure 2.5) is obtained at 40% translated data, reaching R^2 of 96.20%, slightly outperforming full (100%) augmentation ((Figure 2.8). This pattern implies that moderate infusion of translated data effectively regularizes the model, enhancing generalization without introducing domain noise from excessive MT input.



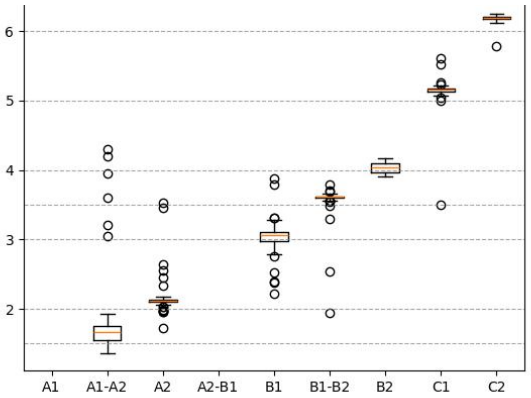
(1) Exp. 9



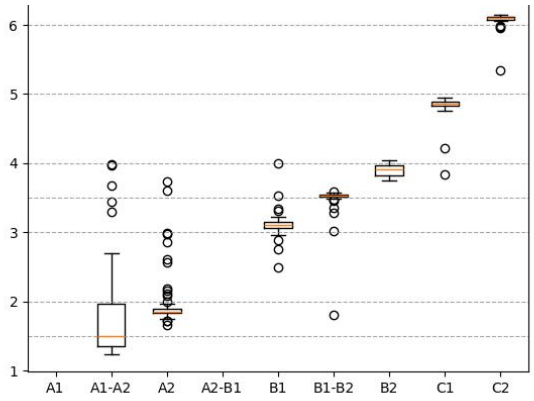
(2) Exp. 10



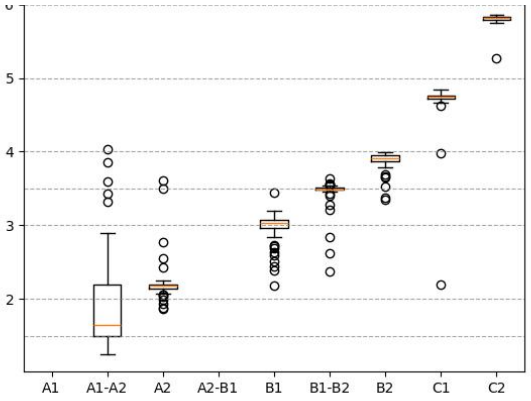
(3) Exp. 11



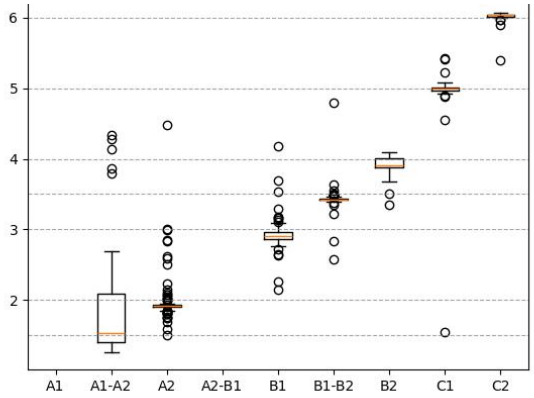
(4) Exp. 12



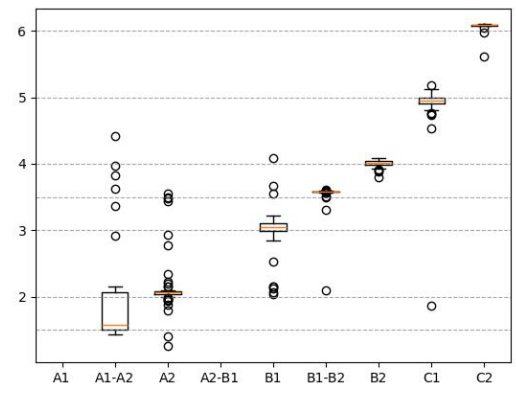
(5) Exp. 13



(6) Exp. 14



(7) Exp. 15



(8) Exp. 16

Figure 2: Box plots for experiments (Exp. 9–16) showing error distributions across models trained with different proportions and CEFR-level subsets of machine-translated data.

5. Conclusion

We set out to explore two research questions: (1) machine-translated data from a high-resource language can substantially improve prediction of text difficulty in a low-resource setting, and (2) the improvement generalizes across languages, especially when the quantity of translated data are well-balanced across linguistic levels.

The results confirm that carefully selected MT data can act as a strong proxy for native material in low-resource language modeling. The experimental findings support both research questions. Data augmentation via translating text annotated with difficulty from another language can indeed improve the performance of the difficulty model. A crucial caveat is that we must assure that the particular MT model we use for translation is able to preserve the CEFR level of its input reasonably well. This is far from a foregone conclusion, since many modern LLMs are trained to "improve" on the text while translating, including simplifying the text or making it more "standard." Care must be taken that the MT model preserves the level reasonably accurately, and additional techniques need to be explored to ensure this in a systematic way.

In future work, we plan to pursue several directions. One plan is to integrate feature-based and Transformer-based models, which could also help with the interpretability of the resulting models in terms of easily understood features. We plan to explore whether can provide reasonable guarantees that MT preserves the levels. This would highlight the importance of our results, since difficulty- and CEFR-annotated data of high quality are very difficult to find, and creating such data is highly resource-intensive. Data augmentation via MT would allow us to grow our training (and test) datasets considerably, to yield substantial improvements in performance on this complex task. At the same time will explore multilingual models of difficulty, to study to what extent the transformer can identify language-independent features that impact on text difficulty.

6. Limitations and Ethical Considerations

While our results show that difficulty models trained on labeled data from multiple languages can be effective, several limitations remain. First, the models are trained and evaluated on small datasets. Working only with Finnish may limit generalizability to other languages or domains, and additional languages should be explored. Second, the mappings that we apply—from continuous regression scores to CEFR levels—introduce discretization errors that may obscure improvements on a more nuanced level.

This work aims at improving language accessibility, particularly for second-language (L2) learners, and seeks to reduce linguistic barriers in education and communication. However, several ethical considerations must be acknowledged. First, automated simplification tools may reinforce biases present in the training data, especially if texts from specific groups or dialects are underrepresented. Second, in general, over-reliance on automated systems may reduce the role of human educators in assessing learner needs, which is not the intent, and is not productive. Lastly, indiscriminate use or misuse of simplification systems—e.g., to manipulate or oversimplify critical content—can have adverse effects. We emphasize that these systems should be used as *assistive* tools, rather than as replacements for human judgment in the context of education or public communication.

References

Sweta Agrawal and Marine Carpuat. 2023. Controlling pre-trained language models for grade-specific text simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12807–12819, Singapore.

Wejdan Alkaldi and Diana Inkpen. 2023. Text simplification to specific readability levels. *Mathematics*, 11(9):2063.

Sandra Aluísio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications*, Los Angeles, California.

Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In *Proceedings of the 4th Workshop on Text Simplification, Accessibility, and Readability*, Suzhou, China.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Kevyn Collins-Thompson and James P. Callan. 2004. A language modeling approach to predicting reading difficulty. In *HLT-NAACL 2004:*

Proceedings of the Human Language Technology Conference of the NAACL, pages 193–200.

Anna Dmitrieva. 2025. *Resources and Tools for Automatic Text Simplification: Cases of Russian and Finnish*. Ph.D. thesis.

Anna Dmitrieva and Aleksandra Konovalova. 2023. Creating a parallel Finnish-Easy Finnish dataset from news articles. In *Proceedings of the 1st Workshop on Open Community-Driven Machine Translation*, pages 21–26, Tampere, Finland.

Anna Dmitrieva and Jörg Tiedemann. 2021. Creating an aligned Russian text simplification dataset from language learner data. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 73–79, Kiyv, Ukraine. Association for Computational Linguistics.

Caroline Gasperin, Lucia Specia, Tiago F. Pereira, and Sandra M. Aluísio. 2009. Learning when to simplify sentences for natural text simplification. In *Proceedings of the Encontro Nacional de Inteligência Artificial (ENIA)*, Bento Gonçalves, Brazil.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2024. Part of the problem: Toward a finer-grained analysis of LLMs for readability assessment. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*.

Anisia Katinskaia, Anh-Duc Vu, Jue Hou, Ulla Vanhatalo, Yiheng Wu, and Roman Yangarber. 2025. Estimation of text difficulty in the context of language learning. In *BEA: 20th Workshop on Innovative Use of NLP for Building Educational Applications*, Vienna, Austria.

J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Benjamin S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Air Station Memphis (Research Branch Report 8-75).

Antonina Laposhina. 2020. A corpus of Russian textbook materials for foreign students as an instrument of an educational content analysis. *Russian Language Abroad*, 6(283):22–28.

Antonina Laposhina, Tatiana Veselovskaya, Maria Lebedeva, and Olga Kupreshchenko. 2018. Automated text readability assessment for Russian second language learners. In *Computational Linguistics and Intellectual Technologies*, pages 403–413.

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

Shingo Nahatame and Katsuyoshi Yamaguchi. 2026. Revisiting text readability and processing effort in second language reading: Bayesian analysis of eye-tracking data. *Language Learning*.

Serge Sharoff. 2022. What neural networks know about linguistic complexity. *Russian Journal of Linguistics*, 26(2):371–390.

A. Jackson Stenner. 1996. Measuring reading comprehension with the Lexile framework. Technical report, MetaMetrics Inc., Durham, NC.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the 7th Workshop on Building Educational Applications Using NLP (BEA)*.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland.

Automatic Generation of Graded Texts in Old Church Slavonic

Iglika Nikolova-Stoupak[1], **Eva Schaeffer-Lacroix**[1]
Gaël Lejeune[1], **Aliona Shestakova-Stukun**[2]
[1]Sens Texte Informatique Histoire, Sorbonne Université, Paris, France
[2]Language School Aspirantum, Yerevan, Armenia
iglika.nikolova-stoupak@etu.sorbonne-universite.fr, {gael.lejeune, eva.lacroix}@sorbonne-universite.fr,
a.szestakowa@gmail.com

Abstract

In the past few decades, graded readers have been valued within language education and have so much as extended onto the so-called classical (or 'dead') languages, such as Latin and Greek. The immersive reading and listening of adapted texts in these languages has been shown to increase students' proficiency, independence and motivation. However, as of now there is only a small number of related resources as well as of classical languages represented. The present study will investigate the current potential for (semi-)automatic generation of adapted classical-language readers while focusing on the Old Church Slavonic language. From a Natural Language Processing (NLP) point of view, work with the language is challenging due to the variety of dialects and diachronic variations it encompasses. The following steps are taken within our study: 1) Representative measurable characteristics of professional classical-language readers, such as the Latin *Lingua latina per se illustrata* and the Greek *Athenaze*, are analysed. 2) Automatic generation of adapted Old Church Slavonic text is attempted through the use of a sequence-to-sequence model (mT5) as well as a Large Language Model (GPT-5) in a one-shot setting. 3) The derived texts' quality is assessed through both human evaluation and a comparison of their textual characteristics with those of professional texts as defined in point 1). The edited versions of the GPT-based texts are shared for future reference and use.

Keywords: historical languages, ChatGPT, automatic adaptation, graded readers, Old Church Slavonic

1. Introduction and Motivation

Language practice with adapted texts in classical languages such as Latin and Ancient Greek (henceforth, Greek) has been shown to offer a number of benefits related to students' proficiency, independence and motivation when used either in isolation or in combination with more traditional methods, such as the grammar-translation method (Diller and Walsh, 1978; McMenamin, 2022; Philips, 1988; Venditti, 2021). An increase in the number and variety of relevant teaching materials that carry the already established ones' qualities would therefore be of significant help in advancing the study of classical and other culturally-significant languages. Currently, technological tools like ChatGPT are commonly used to facilitate the task of teaching professionals to create classroom and self-study materials. In our study, a discrete language will be experimented with, Old Church Slavonic (henceforth, OCS). On one hand, the language shares linguistic similarities as well as common registers (e.g. Biblical) with the classical languages in which established graded readers exist (in particular, Greek). On the other hand, OCS is significantly lower-resourced and less unified in terms of spelling and syntactic rules, thereby presenting a challenge. Little work has been done to date with OCS in the context of NLP. It is worth mentioning BERTislav, a model based on ruBERT and fine-tuned on a corpus of historical Slavic texts (Arkhipov and Trofimova, 2021). In contrast, other historical languages such as Latin and Greek have received notable, though still limited, attention within NLP and the development of digital tools for language learning. Existing datasets and tools in these languages include annotated corpora, interactive reading environments, and NLP-assisted resources (Crane, 1996; Blackwell and Smith, 2009; Johnson et al., 2017).

2. Background

2.1. Graded Readers and Classical Languages

Related to the concept of comprehensible input as put forward by Krashen (Krashen, 1982), graded readers of various modern languages have been in use for decades. They are typically associated with specific proficiency levels, such as per the Common European Framework of Reference for Languages (CEFR). The language within them is learner-friendly, containing accordingly simplified grammar and only a limited number of unfamiliar words. These reading materials have been particularly noted to help reaffirm vocabulary knowledge (Wan-a rom, 2008) and, when used efficiently, to increase students' motivation and sense of com-

Proceedings of the 2nd Workshop on Evaluating Text Difficulty in a Multilingual Context (DeTermIt! 2026), pages 51–62
11 May 2026.

munity (Hill, 2013).

Interestingly, albeit with certain modifications, graded readers have also extended to extinct, classical languages, such as Latin, and have been shown to offer the same benefits (Diller and Walsh, 1978; McMenamin, 2022; Philips, 1988; Venditti, 2021). The primary resources used in the current project have been selected based on the presence of graded text, general quality/reputation and, ultimately, a quest for variety (in the face of different sizes, time frames of composition, and textual genres), so as to increase the robustness of the carried out analysis. The represented classical languages are Latin, Greek and Biblical Hebrew (henceforth, Hebrew); each by two works. It is important to note that these languages differ significantly from one another as well as from OCS, the language of main interest to the project. Hebrew comes as the clearest outlier, as it does not issue from the Indo-European family (rather, it is an Afro-Asiatic language) and does not exhibit some of the other languages' key features (e.g. grammatical cases).

Hans Ørberg's *Lingua latina per se illustrata*[1] (henceforth, *LLPSI*), originally published in 1955, is revolutionary in its sole reliance on the target language in leading the reader from (approximately) zero knowledge to language proficiency. The book comes in two volumes, each of which includes a continuous narrative centred around life in Ancient Rome. The narrative in Volume 1 is authorial, whilst the one in Volume 2 moves through adapted to largely original selections of Roman authors, such as Ovid, Virgil and Cicero. The texts are separated in chapters and accompanied with captioned illustrations, grammatical notes and exercises. In turn, *Fabulae Faciles* (1903)[2] is a reader composed long before comprehensible input was coined as a term. It includes 100 short Latin stories, ordered in ascending difficulty and based on literary works and historical events, linked to Ancient Rome.

Similar to *LLPSI*, *Athenaze*[3] is a two-volume introduction to the Classical (Attic) Ancient Greek language. The first volume features related texts of increasing difficulty describing "the daily life of the ancient Greeks as it was shaped and given meaning by historical developments, political events, and the life of the mind as revealed in mythology, religion, philosophy, literature, and art" (Balme and Lawall, 2003). The second volume consists of mostly unadapted works of classical Ancient Greek authors, such as Homer, Herodotus, and Thucydides. Unlike the case of *LLPSI*, the English language is present in *Athenaze* in the face of cultural information, exercise instructions, and translations of vocabulary items. In many aspects, *Logos*[4] is an even closer equivalent of *LLPSI* for the Greek language; in fact, its subtitle tellingly reads "Logos. Lingua graeca per se illustrata". This reader contains graded texts, accompanied with captioned illustrations, marginal notes, grammatical explanations and exercises - all in Greek. The beginning text is perceptively simpler to the one offered in *Athenaze*, and vowel length signs are excluded so as to facilitate pronunciation. *Logos* features thematically organised discrete stories rather than an uninterrupted narrative line, although narrative elements do emerge at a given point.

Miles Van Pelt and Gary Pratico's *Graded Reader of Biblical Hebrew* (2006)[5] (henceforth, *GRBH*) contains 30 Hebrew texts (202 Bible verses) in increasing levels of difficulty. The texts are compiled rather than adapted, and they come with verb lists, grammatical commentary and parsing exercises. Students are assumed to already have beginner knowledge of the language, and they are guided toward an intermediate level. Finally, *Biblical Hebrew Easy Stories*[6] (henceforth, *BHES*) consists of 52 unrelated stories that gradually increase in difficulty. Most of the stories are based on the Hebrew Bible, while others are authorial. This is not a stand-alone resource but part of a multi-faceted Biblical Hebrew course featured on the YouTube channel *Aleph with Beth* as well as on the associated website. The course is based on comprehensible input and currently contains over 200 video lessons.

2.2. The Old Church Slavonic Language

The term 'Old Church Slavonic' denotes the language of the first Slavic manuscripts, which date from the 9th-11th century AD (Lunt, 2001). The

[1] *Pars I: Familia Romana*, 2nd ed. (Focus [Hackett Publishing Company], 2011).
Pars II: Roma Aeterna, 2nd ed. (Focus [Hackett Publishing Company], 2017).

[2] *Ritchie's Fabulae Faciles: A First Latin Reader*, ed. John Copeland Kirtland (Project Gutenberg, September 2005)

[3] *Athenaze: An Introduction to Ancient Greek. Book I*, 2nd ed. (Oxford; New York: Oxford University Press, 2003); *Book II*, 2nd ed. (Oxford; New York: Oxford University Press, 2003)

[4] *Logos. Hellenike glossa autoeikonographemeni* (Cultura Clásica, 2023).

[5] *Graded Reader of Biblical Hebrew: A Guide to Reading the Hebrew Bible* (Zondervan Academic, August 2006)

[6] *Aleph with Beth*, Betheden Ministries, (2020–). https://freehebrew.online/resources/.

language is characterised with around two centuries of use in a large geographical territory. The writing system is credited almost exclusively to Constantine the Philosopher, a Thessaloniki-born scholar and monk. It initially made use of the Glagolitic alphabet, which then evolved into Cyrillic. The latter system is largely based on Greek letters as combined with additional symbols for typically Slavic sounds. OCS has strong word declension, which includes seven cases, three genders, three numbers, and three simple tenses. Typically, words come in sequences of open and closed syllables, and the reading of the frequently used reduced vowels ъ and ь depends on the type of syllable. In manuscripts, word abbreviation (denoted by a tilde symbol between a word's first and last letters) is common for reasons of both emphasis (words with religious significance) and economy (frequently used words). There is only a limited number of established OCS manuscripts, which present the language's initial and unified characteristics. Most represented, the Gospels appear in five manuscripts, including Codex Zographensis (dated to the 1020s) and Codex Marianos (1030s). Extant OCS texts include Biblical translations, Saints' lives, prayers and sermons (Lunt, 2001).

The most significant challenge for scholars is the large variation within the OCS language. Although the language was initially mostly phonetic in nature, mismatches between spelling and pronunciation as well as alternative spellings started to appear as a result of changes in the spoken language. The differences in dialects, encompassing vocabulary, spelling and grammar, became progressively more significant. Examples of dialect-based variation include the pronunciation of nasal sounds and the use of uncontracted long adjectives. Eventually, OCS gave place to what are now seen as distinct 'Church Slavonic' languages, typical to the country or geographical location in question.

The two texts that are automatically adapted in the context of the current project are the first chapter of the Biblical book of 'Genesis' (henceforth, 'Genesis: 1')[7] and 'The Legend of Saint George and the Dragon' (henceforth, 'Saint George and the Dragon')[8]. The former is the Biblical account of the creation of the world in six days. It is selected as the lower-level (A1) OCS text to be achieved due to its short length, simplicity, and repetitiveness. In contrast, 'Saint George and the Dragon', a hagiographic adventure story with a significant narrative line, undergoes adaption into a higher level (B1). The 'Genesis' text, reconstructed by Tomáš Spevák, follows the norms of the early OCS period. 'Saint George and the Dragon', featured in a scholarly monograph by Alexander V. Rystenko, is normalised based on later forms of Old Church Slavonic (as reminiscent of Russian Church Slavonic).

For the full original OCS texts used, please refer to this GitHub repository.

3. Methods

3.1. ChatGPT and One-Shot Prompting

OpenAI's popular chatbot ChatGPT (as per GPT-5 and the product's official interface) was the large language model (LLM) we used for textual adaptation. Combined with relevant prompt engineering, ChatGPT has been shown to provide high results in tasks linked to textual simplification and summarisation, outperforming alternative models in terms of both automated scores and human preference (Bogireddy and Dasari, 2024; Leroy et al., 2024; Engelmann et al., 2023), including in non-English settings (Nikolova-Stoupak et al., 2024b; Pu et al., 2023) and in relation to literary text (Nikolova-Stoupak et al., 2024b).

One-shot prompting is a setting in which in addition to directions, the user provides the model with an example that illustrates the output's desired qualities. Previous research shows that in the presence of one-shot examples, LLMs tend to output sentences whose linguistic features match more closely those of sentences that have been professionally crafted for the purpose of language teaching (Nikolova-Stoupak et al., 2024a). In particular, compared to zero-shot generation, one-shot generation with ChatGPT offers multilingual literary adaptations that resemble the textual characteristics of human-made adaptations (Nikolova-Stoupak et al., 2024b).

As we discovered no suitable pair of original and learner-adapted OCS text to use in one-shot prompting, we decided to provide examples in another language. We opted for Latin as a high-resourced classical language that shares an alphabet with English, the highest-resourced language overall. More concretely, we selected two stories from *LLPSI* as a critically-acclaimed gold standard to exemplify the two discrete proficiency levels that we aim to produce OCS text in: one from the middle and one from the end of volume 1. The former is 'Litterae latinae' (Ch.18), which features a classroom-setting discussion of the specificities of Latin spelling. And the latter is

[7] Tomáš Spevák, *Old Slavic Library*.

[8] Alexander V. Rystenko, *Legenda o sv. Georgii i drakone v vizantiiskoi i slaviano-russkoi literaturakh* (Odessa: Ekon, 1909).

'De arte poetica' (Ch.34), in which Roman characters introduce Roman literature inside a fictional framework.

3.2. Textual Preprocessing

The selected professional classical readers were first preprocessed for use. All content apart from graded text in the target language was discarded. Optical character recognition was applied for texts that were not already in a machine-readable format as per the proprietary tool 'Pen to Print'[9]. To facilitate the measurement of textual characteristics, each source was converted to full uninterrupted text, devoid of titles, tabs and new lines. Punctuation was standardised (e.g. the Greek ';' was replaced by '?' and the Hebrew end-of-verse ':' was replaced by '.'). The readers that come in two volumes (*Athenaze* and *LLPSI*) were merged into single texts.

As we needed the classical readers to be analysed per proficiency level, we also took steps to divide them accordingly, a task that was not straightforward due to the absence of clear level denotation. For the purpose, we elaborated an additional processing pipeline, at the end of which relevant portions of the texts were extracted whose level closely matches that of the intended adapted OCS texts. Milton et al. (2010) argue that despite not being free of limitations, vocabulary size presents an efficient proxy for CEFR level. In a later study, they go on to estimate the relative vocabulary knowledge (in lemmas) of learners of different languages (English, French and Modern Greek), whose proficiency levels have been previously determined (Milton and Alexiou, 2009). The derived ranges, disregarding a clear outlier group of French learners in the UK, are the following for the first four levels: 894-1492 (A1), 1700-2237 (A2), 2194-3305 (B1), and 2450-4012 (B2).

We applied these conclusions in order to approximate the CEFR level of each of the two *LLPSI* texts we selected as examples in the one-shot prompt to ChatGPT. For the purpose, we calculated the vocabulary size[10] of the portion of the book all the way up to and including each text. Then, we mapped the resulting values to a proficiency level. The first extract was associated with 1385 lemmas, and the second - with 3212. The levels were unproblematically estimated as A1 and B1.

Next, we needed to extract A1- and B1-friendly portions of each of the professional readers in order to be able to analyse their textual characteristics and define gold value ranges per level against which to evaluate automatic output in OCS. Once again, we determined the relevant portions based on the encountered numbers of unique lemmas within them. In cases where a reader's vocabulary does not reach the B1 range, we only extracted the A1 portion[11]. When a given reader does not cover the full level's range, we extracted the text between the lower limit and the end of the reader[12]. *GRBH* demanded a different approach as it contains only 847 unique words, which is below the lower limit for level A1. We assumed the reason for this mismatch to be the reliance of knowledge outside of the presented text, as admitted by the authors. Following the authors' estimation that the text commences at an established beginner level and eventually reaches an intermediate level, we divided it into three approximately equal parts, the first and last one of which we labelled, respectively, as A1 and B1.

3.3. Generation with mT5

We experimented with sequence-to-sequence generation of adapted OCS text with the use of the mT5 model, which has been trained on 101 languages, including modern Slavic languages like Bulgarian, Russian, Serbian and Ukrainian (Xue et al., 2021). First, we performed continued pretraining of mT5-small on monolingual OCS data (a total of 23 282 sentences from the PROIEL[13] and TOROT[14] datasets). The derived model was trained for 5 epochs with a maximum sequence length of 256 and a learning rate of 2e-4.

Then, we moved onto task tuning the model for CEFR-level-based adaptation. For the purpose, we used a database of 36 textual pairs that consisted of an original text and its adapted counterpart in level A1 or B1 (typically, a professional graded reader from a series such as *Oxford Bookworms* or *CLE International*). The texts are in several modern languages[15], and their detailed

[9] https://www.pen-to-print.com/. The tool was selected due to its high-quality output in various languages and alphabets.

[10] based on the number of lemmas as parsed using *UDPipe*, following rule-based preprocessing that eliminates punctuation, proper nouns, macron- or diacritic-based differences, and some OCR-based errors

[11] This was the case with *Fabulae Facilis* (1739 words) and *BHES* (1423 words).

[12] e.g. *Athenaze*'s vocabulary comes at 2535, which is below the upper limit of level B1. Therefore, we took the portion of the reader between vocabulary size 2194 and its end as B1 sample.

[13] https://proiel.github.io/

[14] http://torot.korpus.cz/

[15] English, French, Italian, Japanese, Spanish, Russian

descriptions are available in the project's GitHub repository (due to copyright reasons, the full texts are not made available). As the full textual pairs are too large to directly use in model training, we resorted to fragment alignment. We divided the adapted texts into chunks of approximately 300 words (while keeping sentences whole) and the full texts by the same number of chunks. Then, we performed forward and backward passes through the paired texts until the highest average cosine similarity[16] was achieved per chunk. In each pass, sentences were taken and added incrementally for neighbouring chunks until there was no net improvement in cosine similarity. This alignment method is similar to the one brought forward by Kajiwara and Komachi (2016). As some of the original texts were very large in comparison to their adapted counterparts, we imposed the following conditions: the maximal allowed length for the chunks pertaining to original texts was 4000 words (9000 characters in the case of Japanese); in cases where the corresponding adapted texts contained fewer sentences than the required number of chunks, the textual pair was disregarded. We also excluded chunk pairs whose similarity was below the threshold of 0.35 (which we determined through experimentation). In addition, we composed two discrete datasets to base our training on: a 'cleaner' one (104 excerpts) that included only pairs where the original text is up to three times larger than the adapted one; and a 'noisier' one (1172 excerpts) that included all texts. We used the two CEFR levels as a control token at both training and generation. All training was performed using one GPU Nvidia L40S 45GB.

3.4. Evaluation

3.4.1. Quantitative

The professional classical-language graded readers, as presented in Section 2.1 and grouped into two separate CEFR levels (A1 and B1) as per Section 3.2, were analysed with the use of shallow characteristics that have been established as relevant to readability (i.e. the measurement of a text's complexity as associated with the determination of its suitable audience, typically in terms of age or grade level) (DuBay, 2007). The specific features selected are highly language- and format-independent and, where relevant, rely on computational resources that are readily accessible, including for low-resourced languages. The features pertain to the following general categories: 'length-based', 'vocabulary-related', 'syntax-related' and 'discourse-related' (see Table 1).

As proficiency level is already taken into consideration within the quantitative analysis, we decided against further, length-based regularisation of type-to-token ratio (TTR). Content and function words are defined as, respectively, Universal Dependencies (UD) tags NOUN, PROPN, VERB, ADJ, ADV and AUX, ADP, DET, PRON, PART, CCONJ, SCONJ. 'Punctuation variety' denotes the ratio of non-full-stop over full-stop punctuation. For the purposes of lemmatisation and part-of-speech (POS) tagging, the open-source pipeline *UDPipe*, which makes use of models trained on UD treebanks and covers all concerned languages, was employed via its web-based API.[17] This feature selection is not meant to offer an exhaustive analysis; rather, it serves as a basis for the comparison of various relevant aspects of the investigated texts.

Feature type	Selected features
Length-based	avg # letters/word avg # words/s-ce
Vocabulary-related	word-based TTR lemma-based TTR
Syntax-related	avg # verbs/s-ce avg % function words/s-ce
Discourse-related	avg # pronouns/s-ce punctuation variety

Table 1: Textual features selected for the quantitative analysis of classical-language readers.

3.4.2. Qualitative

The output texts' qualitative evaluation consisted in an in-depth analysis, focused on the following textual aspects: understandability (level-appropriate vocabulary and grammar), correctness (absence of mistakes at the level of vocabulary, grammar, and punctuation), consistency (both internal consistency, such as verb tense usage, and with respect to diachronic and geographical variation), textual coherence (natural textual flow, easy anaphora resolution, absence of unnecessary redundancy) and aesthetic appeal (a more subjective measure involving the text's overall literary quality, length and register).

[16] per PARAPHRASE-MULTILINGUAL-MINILM-L12-v2

[17] https://lindat.mff.cuni.cz/services/udpipe/. The specific models used are: *latin-perseus-ud-2.15-241121*; *ancient_greek-perseus-ud-2.15-241121*; *ancient_hebrew-ptnk-ud-2.15-241121* and later *old_church_slavonic-proiel-2.15-241121*. In the case of multiple models being available for a language, the choice was made based on the quality of lemmatisaton as verified manually.

On the basis of this analysis, the text was manually improved so as to correct errors, remove inconsistencies and increase understandability and aesthetics. The analysis was performed by the authors. We noted common as well as differing observations compared to those resulting from the quantitative analysis.

4. mT5 Experiment

Unfortunately, the result of our experiment with mT5 and adaptation of the two source OCS texts into CEFR levels A1 and B1 was largely negative: the quality of the output was too low to allow for further analysis, and it started with the token <EXTRA_ID_0>, which implies that the associated task was not learned. Typical output ranged from repetition of the same few OCS letters to a string of words that resemble the input text (e.g. бъ: ѩко добро· и бысть тако).

In order to determine whether it was the inclusion of OCS that impeded the model's performance, we also prompted the model (in its original version as well as the one further trained for OCS) to adapt a French textual extract[18]. In the case of the base model, the output still started with <EXTRA_ID_0> while curiously, the further trained model directly performed text generation, whose quality was, however, also poor. When trained on the full ('noisy') adaptation dataset, the model output French text of better quality, in that it included a number of actual French words rather than pseudo-French, which was dominant when the 'clean' dataset was used.

We therefore concluded that the main problem within the task was the insufficiency or non-suitability of the utilised adaptation dataset rather than the implication of OCS. We moved on to evaluation of solely ChatGPT's output.

5. Results

5.1. Quantitative Evaluation

Please refer to Table 2 for the detailed results of all texts' quantitative evaluation.

Firstly, we explored ChatGPT's output for deviations from the gold standard per feature for the intended proficiency level[19]. The values for 'number of letters per word' and 'number of words per sentence' are lower than the established standard for both OCS texts. For 'Saint George and the Dragon', TTR (for both words and lemmas) is higher than the standard, speaking of higher lexical variety. We also calculated the ratio between the two types of TTR for each text, as an estimation of the weight of inflection in lexical variety. In this respect, the two OCS texts do not deviate from the standard. The number of verbs per sentence has a reduced value for both texts, thereby indicating syntactic simplicity. The number of pronouns per sentence (which relates to the need for anaphora interpretation) is also below the baseline for the B1 text. Finally, punctuation variety, which in related to the variety in sentence types, is low for both automatically generated texts.

What follows are observations concerning the relationship between the different variables' values at the A1 versus B1 proficiency level in the context of the same graded reader, where applicable[20]. The average numbers of both 'letters per word' and 'words per sentence' tend to be higher for the higher proficiency level[21], and the two OCS texts fall neatly within this trend. The same goes for the two types of TTR as well as the numbers of verbs and pronouns per sentence[22]. Concerning 'percentage of function words' and 'punctuation variety', no clear trends are established in relation to the professional texts.

Althought it would be interesting to explore the connection between 'language' and the utilised metrics, only limited conclusions can be reached given the presence of solely two texts per language within the corpus. Hebrew is associated with the smallest 'number of letters per word', which is natural given the language's *abjad* (consonant-based) writing system. The gap between word- and lemma-based TTR is highest for Latin texts (the former being larger by 0.13-0.17).

5.2. Qualitative Evaluation

See Figures 1 and 2 for extracts of the output texts as juxtaposed to their manually edited versions. See Appendices A and B for the full output and edited texts.

The language in the adapted 'Genesis: 1' text strikes as beginner-friendly. The sentences are short and the text makes effective use of the repetitive structure present in the original. The verses are numbered and a line is skipped following each narrative unit, facilitating reading. Yet, complexity

[18] from an authorial book; present in the associated repository

[19] Due to the corpus' limited size, we did not go on to calculate the size of the deviations.

[20] Due to the small corpus size, only overall trends are noted, i.e. which level tends to be associated with a higher value for the variable, rather than exact ratios.

[21] with the exception of *Athenaze* in the case of the latter

[22] In both cases, *GRBH* is an outlier at exhibiting smaller values in relation to B1 text.

Text	*LLPSI*	*LLPSI*	*Fabulae Faciles*	*Athenaze*	*Athenaze*	*Logos*	*Logos*	*GRBH*	*GRBH*	*BHES*	'Genesis: 1'	'Saint George and the Dragon'
Level	A1	B1	A1	A1	B1	A1	B1	A1	B1	A1	A1	B1
Language	Latin	Latin	Latin	Greek	Greek	Greek	Greek	Hebrew	Hebrew	Hebrew	OCS	OCS
Avg letters/word	5.23	5.35	5.66	4.78	5.20	4.58	4.97	4.42	4.53	4.37	3.70	4.37
Avg words/s-ce	12.51	14.49	18.68	35.01	25.22	12.53	17.42	13.43	13.70	9.08	7.54	9.00
TTR (words)	0.36	0.42	0.40	0.40	0.52	0.33	0.40	0.50	0.57	0.28	0.42	0.69
TTR (lemmas)	0.21	0.25	0.27	0.30	0.40	0.25	0.30	0.43	0.50	0.20	0.37	0.57
Avg verbs/s-ce	1.43	1.84	3.19	4.08	4.20	1.33	2.29	2.48	2.10	2.46	1.23	1.65
Avg % funct. words/s-ce	27.89	30.38	34.54	33.11	33.11	31.31	33.78	30.86	22.34	32.50	33.51	33.49
Avg pronouns/s-ce	0.44	0.59	0.75	1.46	1.65	0.63	0.84	1.26	1.16	1.29	0.35	0.62
Punctuation variety	3.02	3.23	2.02	6.66	3.19	3.71	3.46	2.83	2.95	2.13	0.97	1.65
Total s-ces	905	1317	366	198	107	458	487	92	98	728	40	29

Table 2: Statistics pertaining to the professional texts and the automatically generated OCS texts. For the latter, the *intended* level is noted, and the values are underlined when they fall outside of the range established by the professional texts for the variable and level. All values are rounded to the second digit after the decimal point.

is occasionally high due to close reliance on the original language. For instance, in (21) ("И сътво_ ри бъ человѣка: мужа и жен8 сътвори и."[23]), the use of *и* as both a conjunction and a personal pronoun is likely to confuse a beginner learner. We propose a shorter alternative: "И сътвори бъ чло_ вѣка: мужа и жен8."[24] Similarly, anaphora resolution may be difficult in the following verb-less construction: (23) "И виде бъ вся, ѩже сътвори. И се, добро зѣло."[25].

We noted a few stylistic issues within the text. The letters оу and 8, which are equivalent in representing the sound /u/, are both present in the text, unlike in its original counterpart. The same goes for the pair з and ꙁ (/z/). In order for beginner students not to be led to wrongfully attribute specific reasons to the choice of letters, we would suggest the use of a single letter in these cases.

Occasionally, the adapted text's spelling speaks of a later variety of OCS than that of the original text. For instance, the letter *е* sometimes replaces ѥ and ѣ (e.g. единъ[26]; посредѣ[27]). We deem that it would be more fitting to make use of the language's earlier spelling conventions as laid out in established grammar materials. Also, prepositions and pronouns are occasionally not separated from the word they modify (e.g. съберетъ_ ся[28]; the original reads съберетъ ся). Whilst such spelling does sometimes occur, consistent separation of these words at level A1 would increase readability and help students achieve notions of word formation.

Moving towards stricter mistakes rather than stylistic choices, we noted the use of the letter я (e.g. вся[29]), which is not typical to OCS but to later Slavic languages[30]. The Russian word меншее[31] is used in place of the OCS equivalent мьнѣѥ[32], as found in the original. Several errors are noted in relation to cases or declension types e.g. (9) во_ ды[33] (correct: водъ[34]); (13) дн8[35] (correct: дни[36]); (21) жен8[37] (correct: женѫ[38]). Finally, there is a spelling mistake in the word четв8ртыи[39](correct: четврьтыи[40]).

The adapted story 'Saint George and the Dragon' is given the title `Чюдо сѣаго Гєѡргіа'[41]. The text includes relatively short and simple sentences and clearly discernible dialogue. In our opinion, the rendition may be a little too short, thereby limiting the story's action and losing some of its aesthetic appeal. Influences of later Slavic languages are more frequent than in 'Genesis: 1'. Apart from the letter я (змия)[42], the letters ё and й, non-existent in traditional OCS, appear (нёмъ[43]; твой[44]). Occasionally, е replaces the traditional ѣ (наконецъ)[45]. The letters ъ and ь are sometimes confused, such as in the word царъ[46],

[23] i stvori bog chloveka: mǫzja i zhenu stvori i "And God created the human: man and woman he created them."

[24] "And God created the human: man and woman."

[25] i vide bog vsya, yazhe strovi. i se, dobro zelo "And God saw everything that He created. And it [was] very good"

[26] edin, 'one'

[27] posred 'in the midst'

[28] sberetsę 'gather-REFL'

[29] vsya 'all'

[30] The OCS equivalent is ѩ.

[31] menshee 'smaller'

[32] menyeye

[33] vodi

[34] vod 'water-ACC.PL.F '

[35] dnu

[36] dni 'day-DAT.SG.M'

[37] zhenu

[38] zhenǫ 'woman-ACC.SG.F'

[39] chetvurtiy

[40] chetvretiy 'fourth'

[41] chudo svetago georgiya 'Saint George's Miracle'

[42] zmiya 'snake'

[43] nyom 'it-LOC.SG.M'

[44] tvoy 'your-ACC.SG.M'

[45] nakonets 'finally'

[46] tsar 'king'

Original	12 И рече ѣъ: да бѫджѫтъ свѣтила на небеси, свѣтити *земльѫ*. 13 И сътвори ѣъ два свѣтила: свѣти_ло *вє́лико* дн8, и свѣтило меншее но_щи, и *звѣзды*. 14 И виде ѣъ, ѩко добро. И бѣ вечерь и бѣ *оутро*, дьнь четв8рты́и.
Edited	12 И рече ѣъ: да бѫджѫтъ свѣтила на небеси, свѣтити ꙁемльѫ. 13 И сътвори ѣъ два свѣтила: свѣти_ло велико дни, и свѣтило мьнѥѥ нощи, и ꙁвѣꙁды. 14 И виде ѣъ, ѩко добро. И бѣ вечерь и бѣ 8тро, дьнь четвертыи.

Figure 1: ‘Genesis: 1’: an extract of original vs. edited output. Legend: underline = mistake; *italics* = stylistic choice (the distinction may be ambiguous).

although this is not an uncommon occurrence in later OCS manuscripts. Accents are sometimes placed above vowels (e.g. что́[47]) in an unpredictable way, possibly due to interference from Greek. Shorter verb conjugations of a more recent nature are also opted for, particularly in the case of imperfect forms: живѣше[48]; метаху[49] (in place of живѣаше; метааху). We consider that a learner would benefit from use of the established imperfect forms, which are more easily recognisable due to the distinctive presence of two adjacent vowels.

An inconsistency comes in the face of the spelling of the word гражане[50], which is neither South nor East Slavic in nature, as the former would call for the presence of жд /zhd/ instead of ж /zh/, and in the latter, the root would read город[51] rather than град[52]. A declension mistake is found in the adjective лютъ[53], which should be люто[54] in agreement with the neuter noun, змии_ще[55]. Two cases of wrong word choice also came to our attention. Saint George addresses the princess as госпоже[56], whilst appropriate words for a young unmarried woman would be дѣвица[57] and отроковица[58]. Also, after the hero has slayed the dragon, he is said to be ‘leading’ (веди)[59] it to the city. A more suitable word would be влеци[60].

Original	Бысть градъ Ласиа, и в нёмъ царъ именемъ Соломонь. Близъ града *бѣ_ше* озеро велико, и в томъ озерѣ *жи_вѣше* змиище лютъ.
Edited	Бысть градъ Ласиа, и в ньмъ царь именемъ Соломонь. Близъ града бѣа_ше озеро велико, и в томъ озерѣ жи_вѣаше змиище люто.

Figure 2: ‘Saint George and the Dragon’: an extract of original vs. edited output. Legend: underline = mistake; *italics* = stylistic choice (the distinction may be ambiguous).

6. Discussion

The Transformer model, mT5, failed at the CEFR-guided textual adaptation task when trained on a corpus of professional graded readers in multiple languages, even when prompted in a more high-resourced language that it already has knowledge of. In contrast, GPT-5’s output demonstrates strong linguistic and pedagogical qualities, such as short sentences, simple grammar and, where applicable, dialogue. The undergone quantitative analysis shows that the difference between the generated texts in terms of level matches all trends established in relation to professional texts. In turn, the qualitative analysis helps confirm some quantitative observations. A general ‘lack of action’ is observed in relation to the B1 text, and the value for the feature ‘number of verbs’ is indeed lower than the gold standard established for the level. The qualitative analysis also helps uncover possible microscopic problems that are not obvious at the quantitative level: for instance, a couple of verses in the A1 text are of high difficulty, whilst the quantitative analysis speaks, if anything, of excessive simplicity.

The LLM also demonstrates a level of specific proficiency in OCS. Words and declensions as well as alternative characters that are not present in the provided unadapted texts are used in the output. Still, the resulting text is not mistake-free, in particular in relation to grammatical cases and to word choice when it comes to more complex vocabulary. There is a perceptible tendency for later

[47] chto ‘why’
[48] zhiveshe ‘live-IMPF.PST.3SG’
[49] metahu ‘throw-IMPF.PST.3PL’
[50] grazhdane ‘citizens’
[51] gorod
[52] grad
[53] lyut ‘fierce’
[54] lyuto
[55] zmiishte ‘snake, dragon’
[56] gospozhe ‘Madam’
[57] devitsa
[58] otrokovitsa
[59] vedi
[60] vletsi ‘to drag’

versions of OCS or even modern Slavic languages to interfere with the output, which can be seen as a natural consequence of these languages' (in particular, Russian's) higher resourcedness. The original text that uses later language conventions resulted in higher interference with modern languages. Also, when a rarely used letter, such as $o\gamma$, was present in the original text, it also appeared in the output. These tendencies demonstrate significant reliance of the model on the input OCS text, which in turn would make automatic (such as rule-based) correction and/or standardisation of the issuing text inconvenient. Manual editing took us about an hour per text and remains the method we currently recommend in order for usable text to be achieved.

7. Conclusion

While a Transformer model is not up to the task of adapting OCS text, GPT's output is promising in granting economy in time and effort to language and teaching professionals. However, the work of specialists in terms of error correction and linguistic standardisation is currently indispensable in order for the resulting text to be made usable for learners. The deliverable graded texts may be used as part of a formal academic course or for personal study based on cultural, liturgical or academic interests. The described methods of human and automatic evaluation are largely applicable to related tasks of textual generation.

Ethics Statement

When dealing with historical languages, it is important to acknowledge their role as cultural heritage, as well as to note that text issuing from language models has no 'authenticity' as such.

Limitations

It is worth noting that only two primary OCS texts are used within the project, and they are reconstructed. Therefore, the original language in its fullness is far from being represented. Also, the limited number of classical readers available complicates the tasks of defining their linguistic features. In addition, as demonstrated by the diversity in these resources, there is no clear formula as to what makes up an ideal graded reader e.g. adapted text in isolation or as accompanied with explanatory notes, translations and exercises. Concerning the mT5 experiments, it remains uncertain whether a larger corpus of adaptation examples has the potential of leading to a significant improvement of results.

References

Mikhail Arkhipov and Maria Trofimova. 2021. Bertislav: Pre-trained language model for slavic languages. In *Proceedings of the International Conference on Analysis of Images, Social Networks and Texts (AIST)*, Cham. Springer.

Maurice Balme and Gilbert Lawall. 2003. *Athenaze: An Introduction to Ancient Greek. Book I*, second edition. Oxford University Press, Oxford; New York.

Christopher Blackwell and Neel Smith. 2009. Alpheios: Open source tools for reading ancient languages. In *Proceedings of the ACM Symposium on Document Engineering*. ACM.

Srinivasa Rao Bogireddy and Nagaraju Dasari. 2024. Comparative analysis of chatgpt-4 and llama: Performance evaluation on text summarization, data analysis, and question answering. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7.

Gregory Crane. 1996. The perseus digital library. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, New York, USA. ACM.

Karl C. Diller and Thomas M. Walsh. 1978. "living" and "dead" languages: A neurolinguistic distinction. In *Actes du 5e Congrès de l'Association Internationale de Linguistique Appliquée*, Montréal. Les Presses de l'Université Laval.

William H. DuBay. 2007. *The Classic Readability Studies*. ERIC Clearinghouse.

Björn Engelmann, Fabian Haak, Christin Katharina Kreutz, Narjes Nikzad Khasmakhi, and Philipp Schaer. 2023. Text simplification of scientific texts for non-expert readers.

David R. Hill. 2013. Graded readers. *ELT Journal*, 67(1):85–125.

Kyle P. Johnson, Patrick J. Burns, John Stewart, and Todd Cook. 2017. The classical language toolkit: An nlp framework for pre-modern languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (System Demonstrations)*, pages 1–6.

Tomoyuki Kajiwara and Mamoru Komachi. 2016. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1147–1158, Osaka, Japan. The COLING 2016 Organizing Committee.

Stephen D. Krashen. 1982. *Principles and Practice in Second Language Acquisition*. Pergamon Press, Oxford.

Gondy Leroy, David Kauchak, Philip Harber, Aabhas Pal, and Anirudh Shukla. 2024. Text and audio simplification: Human vs. chatgpt. *AMIA Joint Summits on Translational Science Proceedings*, 2024:295–304.

Horace G. Lunt. 2001. *Old Church Slavonic Grammar*, 7 edition. Mouton de Gruyter.

Conor McMenamin. 2022. Greek club: Resurrecting dead languages in secondary schools. *Journal of Classics Teaching*, 23(46):121–123.

James Milton and Thomaï Alexiou. 2009. Vocabulary size and the common european framework of reference for languages. In Brian Richards, Michael H. Daller, David D. Malvern, Paul Meara, James Milton, and Jeanine Treffers-Daller, editors, *Vocabulary Studies in First and Second Language Acquisition: The Interface Between Theory and Application*, pages 194–211. Palgrave Macmillan, London.

James Milton, J. Wade, and N. Hopkins. 2010. Aural word recognition and oral competence in a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, and M. Torreblanca-López, editors, *Further Insights into Non-Native Vocabulary Teaching and Learning*, pages 83–98. Multilingual Matters, Bristol.

Iglika Nikolova-Stoupak, Serge Bibauw, Amandine Dumont, Françoise Stas, Patrick Watrin, and Thomas François. 2024a. LLM-generated contexts to practice specialised vocabulary: Corpus presentation and comparison. In *Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position*, pages 472–498, Toulouse, France. ATALA and AFPC.

Iglika Nikolova-Stoupak, Gaël Lejeune, and Eva Schaeffer-Lacroix. 2024b. Contemporary llms and literary abridgement: An analytical inquiry. In *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024)*, pages 39–57, Sofia, Bulgaria. Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences.

F. C. Philips. 1988. The language laboratory and the teaching of "dead" languages. *The Classical World*, 82(2):105–108.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *ArXiv*, abs/2309.09558.

Erica Venditti. 2021. Using comprehensible input in the latin classroom to enhance language proficiency. *Journal of Classics Teaching*, 22(43):22–28.

Uthai Wan-a rom. 2008. Comparing the vocabulary of different graded-reading schemes. *Reading in a Foreign Language*, 20(1):43–69.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A. ChatGPT Output: Graded Texts in Old Church Slavonic

A.1. Genesis: 1

БЪТИЮ, ГЛАВА 1

1 Въ нача́лѣ сътвори ѥ̈ъ небо и землѫ̀.

2 Земл҃а бѣ пуста и тьма бѣ на водахъ. Д8хъ бж̀ии хожда́ше на водахъ.

3 И рече ѥ̈ъ: да бу̀детъ свѣ́тъ. И бѣ свѣ́тъ.

4 И виде ѥ̈ъ свѣ́тъ, ꙗ҃ко добръ. И разл8чи ѥ̈ъ свѣ́тъ отъ тьмы.

5 И нарече ѥ̈ъ свѣ́тъ дьнь, а тьм8 нарече нощь. И бѣ вечерь и бѣ оу̀тро, дьнь еди́нъ.

6 И рече ѥ̈ъ: да бу̀детъ тврьдь посредѣ водъ. И бѣ тврьдь.

7 И нарече ѥ̈ъ тврьдь небо. И бѣ вечерь и бѣ оу̀тро, дьнь втори́и.

8 И рече ѥ̈ъ: да съберетъс҃а вода, и да ꙗ҃витьс҃а с8ша. И бѣ тако.

9 И нарече ѥ̈ъ с8ш8 землѫ̀, а воды нарече мо_ р҃а. И виде ѥ̈ъ, ꙗ҃ко добро.

10 И рече ѥ̈ъ: да прорастить земл҃а трав8 и древо плодоносно. И бѣ тако.

11 И прорастѣ земла трав8 и древо. И ви_
де бъ, ꙗко добро. И бѣ вечерь и бѣ оутро, дьнь
трети́и.

12 И рече бъ: да бжджтъ свѣтила на небеси,
свѣтити землѫ.

13 И сътвори бъ два свѣтила: свѣтило вел́ико
дн8, и свѣтило меншее нощи, и звѣзды.

14 И виде бъ, ꙗко добро. И бѣ вечерь и бѣ
оутро, дьнь четв8рты́и.

15 И рече бъ: да изведѫтъ воды рыб8, и да
летаютъ птици надъ землѫ.

16 И сътвори бъ рыб8 и птиц8. И благослови
ꙗ бъ: растите сѧ и мъножите сѧ.

17 И бѣ вечерь и бѣ оутро, дьнь пяты́и.

18 И рече бъ: да изведетъ земла скотъ и
звѣрь.

19 И сътвори бъ скотъ и звѣрь. И виде бъ,
ꙗко добро.

20 И рече бъ: сътворимъ человѣка по образ8
наш8.

21 И сътвори бъ человѣка: мжа и жен8 съ_
твори и.

22 И благослови ꙗ бъ: растите сѧ и плодите
сѧ, и наполните землѫ.

23 И виде бъ вся, ꙗже сътвори. И се, добро
зѣло.

24 И бѣ вечерь и бѣ оутро, дьнь шести́и.

A.2. The Miracle of Saint George

Чюдо сѣаго Гевргіꙗ

Бысть градъ Ласиа, и в нёмъ царъ именемъ
Соломонь. Близъ града бѣше озеро велико, и в
томъ озерѣ живѣше змиище лютъ.

И сходяше змиище на землю, и всадневно по_
жираше люди. И рекоша гра́жане къ царю:

— Царю, аще не дадимъ змиищу человѣка, вси
погибнемъ.

И метаху жребии, и по жребию давали дѣти
своѧ змиищу. Наконецъ жребий паде на дщерь
цареву. Царь плакаше горько и рече:

— Чадо моѥ, иди нынѣ къ змиищу.

И обрѣтена бѣ дѣва у озера, плачющиꙗ сѧ.

И прииде сѣыи Гевргіи, воинъ христіанинъ.
И рече къ дѣвѣ:

— Что́ плачеши, госпоже?

Рече дѣва:

— Здѣ обрѣтаетъ мя змиище, и погыбну. Бѣ_
жи скоро, добрый человѣче, да и ты не умреши.

Рече сѣыи Гевргіи:

— Не бойсѧ, дѣво. Поможетъ намъ Христо́съ.

И се змиище изыде изъ воды, велико и страш_
но. Сѣыи же Гевргіи сотвори крьстное зна́мение
и рече:

— Господи Іисусе Христе, помози ми.

И устреми коня своего, и копіемъ порази змия
въ главу, и низверже его на землю.

И рече сѣыи къ дѣвѣ:

— Възложи полсъ твой на змиище и веди во
градъ.

И видѣша людие чюдо велие. Удивиша сѧ и
устрашиша сѧ. Рече же сѣыи Гевргіи:

— Не бойте сѧ. Вѣруите во Господа нашего
Іисуса Христа, и крестите сѧ, и избавитъ васъ
отъ всякаго зла.

И вѣрова царъ и вѣрова людие, и крестивши_
сѧ. И бѣ радость велика во граде.

B. Corrected Output

B.1. Genesis: 1

БЫТИѤ, ГЛАВА 1

1 Въ началѣ сътвори бъ небо и землѫ.

2 Землꙗ бѣ пуста и тьма бѣ на водахъ. Д8хъ
бжии хождаеше на водахъ.

3 И рече бъ: да бждетъ свѣтъ. И бѣ свѣтъ.

4 И виде бъ свѣтъ, ꙗко добръ. И разл8чи бъ
свѣтъ отъ тьмы.

5 И нарече бъ свѣтъ дьнь, а тьм8 нарече нощь.
И бѣ вечерь и бѣ 8тро, дьнь еди́нъ.

6 И рече бъ: да бждетъ тврьдь посрѣдѣ водъ.
И бѣ тврьдь.

7 И нарече бъ тврьдь небо. И бѣ вечерь и бѣ
8тро, дьнь вторы́и.

8 И рече бъ: да съберетъ сѧ вода, и да ꙗвить
сѧ с8ша. И бѣ тако.

9 И нарече бъ с8ш8 землѫ, а воды нарече
морѥ. И виде бъ, ꙗко добро.

10 И рече бъ: да прорастить землꙗ трав8 и
древо плодоносно. И бѣ тако.

11 И прорастѣ земла трав8 и дрѣво. И виде
бъ, ꙗко добро. И бѣ вечерь и бѣ 8тро, дьнь тре_
ти́и.

12 И рече бъ: да бжджтъ свѣтила на небеси,
свѣтити землѫ.

13 И сътвори бъ два свѣтила: свѣтило велико
дни, и свѣтило мьнѥе нощи, и звѣзды.

14 И виде бъ, ꙗко добро. И бѣ вечерь и бѣ
8тро, дьнь четверты́и.

15 И рече бъ: да изведжтъ воды рыб8, и да
летаѭтъ птица надъ землѫ.

16 И сътвори бъ рыбѫ и птицѫ. И благослови
ꙗ бъ: растите сѧ и мъножите сѧ.

17 И бѣ вечерь и бѣ 8тро, дьнь пяты́и.

18 И рече бъ: да изведжтъ земла скотъ и
звѣрь.

19 И сътвори бъ скотъ и звѣрь. И виде бъ,
ꙗко добро.

20 И рече бъ: сътворимъ человѣка по образ8
наш8.

21 И сътвори бъ человѣка: мжа и женѫ.

22 И благослови ꙗ бъ: растите сѧ и плодите
сѧ, и наполните землѫ.

23 И виде ѣ҃ъ всѧ, ѩже сътвори. И се бысть добро зѣло.

24 И бѣ вечерь и бѣ ѕ҃тро, дьнь шести́и.

B.2. The Miracle of Saint George

Чюдо сѣ҃таго Гєѡргіѧ

Бысть градъ Ласиа, и в ньмъ царь именемъ Соломонь. Близъ града бѣаше озеро велико, и в томъ озерѣ живѣаше змиище люто.

И сходѩше змиище на землю, и всѧдневно по_жираше люди. И рекоша граждане къ царю:

— Царю, аще не дадимъ змиищу чловѣка, вси погибнемъ.

И метааху жребии, и по жребию давали дѣти своѧ змиищу. Наконѣцъ жребии паде на дщерь цареву. Царь плакаше горько и рече:

— Чадо моѥ, иди нынѣ къ змиищу.

И обрѣтена бѣ дѣва у озера, плачющиѧ сѧ.

И прииде сѣ҃тыи Гєѡргіи, воинъ христіанинъ. И рече къ дѣвѣ:

— Что плачеши, отроковице?

Рече дѣва:

— Здѣ обрѣтаетъ мѧ змиище, и погыбну. Бѣ_жи скоро, добрыи чловѣче, да и ты не умреши.

Рече сѣ҃тыи Гєѡргіи:

— Не бои сѧ, дѣво. Поможетъ намъ Хри_сто́съ.

И се змиище изыде изъ воды, велико и страш_но. Сѣ҃тыи же Гєѡргіи сотвори крьстное знамение и рече:

— Господи Іисусе Христе, помози ми.

И устреми конѧ своего, и копіемъ порази змиѧ въ главу, и низверже его на землю.

И рече сѣ҃тыи къ дѣвѣ:

— Възложи поѧсъ твои на змиище и влечи во градъ.

И видѣша людие чюдо велие. Удивиша сѧ и устрашиша сѧ. Рече же сѣ҃тыи Гєѡргіи:

— Не боите сѧ. Вѣруите во Господа нашего Іисуса Христа, и крестите сѧ, и избавитъ васъ отъ всѧкаго зла.

И вѣрова царь и вѣрова людие, и крестивши сѧ. И бѣ радость велика во граде.

A Calibrated and Interpretable Framework for Multilingual Text Difficulty Prediction

Voula Giouli, George Tsoulouhas, Athina Sioupi, Stamatia Michalopoulou
Aristotle University of Thessaloniki
Thessaloniki, Greece
pgiouli@del.auth.gr, george.tsoulouhas@athenarc.gr, sioupi@del.auth.gr, smichalo@del.auth.gr

Abstract

We present a framework for automatic text difficulty prediction, centered on a linguistically enriched German dataset comprising texts that are aligned with the levels defined by the Common European Framework of Reference for Languages, with its Greek counterpart currently under development. The dataset bears annotations and is used for training and evaluating a system that integrates: (i) a lexicon of lexical profiling enriched via existing openly available lexical resources, (ii) large-scale frequency modeling using a 1.5M-word corpus, (iii) syntactic complexity pre-computation, (iv) percentile-based rule calibration, and (v) feature-based machine learning classifiers with feature selection. Our framework integrates BERT-based contextual modeling and augments it with SHAP-driven interpretability mechanisms and rigorous analysis of confidence calibration. The framework is designed to be in principle language-agnostic, aiming to facilitate multilingual portability.

Keywords: readability prediction, text difficulty, Computer-Assisted Language Learning

1. Introduction

Automatic text difficulty prediction is central to computer-assisted language learning (CALL), adaptive educational systems, curriculum development, and automated content recommendation. In European contexts, text difficulty is typically assessed through the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) which defines six proficiency levels (A1-C2). Although recent research increasingly adopts transformer-based architectures, educational applications require interpretability and pedagogical justification. This is particularly true in language learning contexts, where model outputs must be transparent, explainable and aligned with established proficiency standards.

This paper presents work aimed at developing a workbench for CEFR-based text difficulty prediction. The proposed platform comprises three main components: (i) a tool for CEFR-aligned dataset preparation incorporating a pipeline for documenting, processing, and enriching textual data, (ii) CEFR-aligned datasets, and (iii) three alternative modeling approaches, namely a rule-based baseline, a feature-based Machine Learning (ML) classifier and a fine-tuned BERT model. Our approach integrates linguistically informed feature engineering with data-driven modeling techniques, thereby balancing transparency and predictive performance.

The proposed workbench has been designed within the EmoBot project (Kallipolitis et al., 2026) as a language-agnostic infrastructure that can be extended to any language. In its current implementation, it has been applied to the creation of a German CEFR dataset, while its Greek counterpart is currently under development.

This work makes three main contributions: (i) an extensible infrastructure for building interoperable CEFR-aligned corpora in multiple languages; (ii) a CEFR-annotated German dataset with a Greek one currently underway to evaluate portability and extensibility; and (iii) a comprehensive text difficulty prediction framework that integrates a calibrated rule-based approach, a ML classifier, and BERT modelling for CEFR-based text difficulty assessment, currently evaluated on German.

The paper is structured as follows: Section 2 reviews related work on text difficulty assessment. The infrastructure, including the CEFR-annotation tool, the CEFR-aligned dataset and lexicon, is described in Section 3, while Section 4 details the predictive modeling approaches. Sections 5 and 6 present and discuss the empirical findings. We conclude and outline future research in Section 7.

2. Related work

Readability assessment - defined as the degree of "legibility, interest, or ease of reading" associated with a text (Dale and Chall, 1949) - has long been a central topic in applied linguistics, education, and computational language processing. In language pedagogy, automated readability assessment is valuable not only for native speakers but is particularly important for second/foreign language (L2/FL) learners, as it supports the systematic alignment of texts with learners' proficiency levels and instructional objectives. In European educational contexts, CEFR provides a standardized scale for describing language proficiency and guiding curriculum

Proceedings of the 2nd Workshop on Evaluating Text Difficulty in a Multilingual Context (DeTermIt! 2026), pages 63–73

design.

Early approaches to text difficulty assessment relied on surface-level readability formulas designed to approximate cognitive processing difficulty using shallow linguistic indicators. These formulas typically combine sentence and word length or syllable count to estimate educational grade level, as for example, the Flesch Reading Ease score (Flesch, 1948) and its derivative, the Flesch–Kincaid Grade Level (Kincaid et al., 1975). However, they have been proved to be useful for larger pieces of language (texts) rather than shorter texts or dialogues usually used in current educational CALL applications (Rooein et al., 2024).

Additional formula-based indices such as the Coleman–Liau Index (Coleman and Liau, 1975) and the Automated Readability Index (ARI) (Smith and Senter, 1967) rely on character-level metrics rather than syllable counts, making them easier to compute programmatically.

Language-specific formulas e.g., for German, namely the Wiener Sachtextformel (Bamberger and Vanecek, 1984) are tailored to specific linguistic properties. Readability formulas, however, exhibit several limitations in that they rely on shallow surface features rather than explicitly modeling syntactic or discourse structure. From an educational point of view, these formulas are not directly aligned with CEFR descriptors while they generalize poorly across typologically distinct languages without adaptation.

Over time, the field has expanded to incorporate lexical frequency, syntactic complexity, discourse structure, and, more recently, machine learning and neural modeling approaches.

2.1. Machine Learning approaches

Supervised feature-based models were among the first to be used in this direction. Vajjala and Meurers (2012) advanced readability classification by incorporating insights from L2 acquisition research, combining measures of lexical richness with syntactic complexity features (including parse-tree-based ones), and demonstrating significant improvements over traditional formulas through experiments with Support Vector Machines (SVMs). For German specifically, Hancke et al. (2012) developed a readability classifier using lexical, syntactic, and morphological features on a graded corpus of school and web texts, establishing an early benchmark for German text difficulty assessment.

SVM classifiers are very common in CEFR-based classification. Xia et al. (2016) attained 80.3% accuracy on the WeeBit corpus utilizing an SVM classifier with lexical and syntactic features supplemented by discourse indicators. Meanwhile, Pilán et al. (2016) illustrated that weakly lexicalized features improve generalization across unseen data for CEFR-level assessment in Swedish L2 reading materials. In addition, Random Forest classifiers have shown strong results in text classification tasks, and Imperial et al. (2025) reported good results for document-level CEFR classification.

The emergence of neural network methodologies has established novel paradigms in readability evaluation. Azpiazu and Pera (2019) proposed Multiattentive Recurrent Neural Networks for multilingual readability prediction, achieving 84.7% accuracy on the VikiWiki dataset. More recently, transformer-based architectures have shown promising results: Deutsch et al. (2020) showed that BERT embeddings enhance readability prediction beyond traditional feature-based methods alone, while Imperial (2021) demonstrated that hybrid models combining BERT sentence embeddings with handcrafted linguistic features typically outperform either approach in isolation. Fine-tuned BERT models have also proven effective for direct CEFR classification, as evidenced by Santos et al. (2021). A consistent theme across studies is that feature-based models provide superior interpretability and robustness with limited training data, whereas transformer-based models deliver higher discriminative performance when ample data are available (Martinc et al., 2021).

2.2. LLMs for Text Difficulty Assessment

The emergence of LLMs has opened new pathways for assessing text difficulty. Yancey et al. (2023) evaluated GPT-4 for rating short L2 essays on the CEFR scale; few-shot calibration was found to approach the accuracy of state-of-the-art automated writing evaluation systems, though agreement with human ratings varied by learners' native languages. Trott and Rivière (2024) showed that GPT-4 Turbo's zero-shot readability estimates are very similar to human judgments (r = 0.76) and work better than traditional formulas on the CLEAR corpus.

While LLM-based approaches show clear performance gains—Rooein et al. (2024) demonstrated that prompt-based LLM metrics improve difficulty classification over traditional measures such as Flesch-Kincaid, particularly for shorter texts—measurement stability remains a concern. Uchida (2024) reported inconsistencies in ChatGPT-generated CEFR ratings in repeated evaluations, suggesting that single-pass LLM assessments may lack the reproducibility required for high-stakes educational applications. Recent research on the Ace-CEFR dataset (Kogan et al., 2025) provides a thorough assessment, contrasting linear models, BERT-based classifiers, and LLM-based methodologies. The results demonstrate that hybrid techniques, which integrate BERT embeddings with LLM scoring, can surpass the performance of individual human expert raters. Nevertheless, LLM

inference is substantially slower than feature-based or BERT-based classification, rendering it more suitable for offline labeling; this efficiency trade-off, coupled with the limited transparency of LLM-internal reasoning, highlights the ongoing value of interpretable feature-based approaches in educational contexts where pedagogical justification is essential.

3. Text Difficulty Assessment Workbench

In this section, we present the infrastructure developed, namely the tool for (semi-) automatically creating a CEFR-aligned corpus, the CEFR-aligned dataset that has been developed as a proof-of-concept and the tools for text difficulty prediction.

3.1. CEFR-annotation tool

The annotation tool is a web-based platform designed to support the full lifecycle of CEFR-aligned corpus construction, from text ingestion and collaborative annotation to quality assurance and multi-format export. The tool is built around a modular architecture comprising several interconnected subsystems described below.

Text management and metadata annotation. The tool provides a structured interface for creating, editing, and managing textual entries. Each text record stores the raw content alongside rich metadata including the assigned CEFR level (A1–C2), genre classification, thematic domain, register, and source attribution. Selected texts follow a status-based workflow progressing through draft, pending review, approved, and rejected stages, ensuring that only quality-controlled entries enter the final corpus. Upon ingestion, texts are automatically analyzed to extract a comprehensive set of linguistic features, including readability indices, lexical diversity measures, and CEFR-aligned vocabulary profiles, which are stored as structured metadata for downstream use in both rule-based and machine learning pipelines.

Text segmentation. Approved texts exceeding a minimum length can be split into smaller segments through a manual segmentation interface. Segments are stored as child records linked to the parent text, preserving the parent's core metadata while maintaining independent CEFR annotations and segment ordering. This feature supports the creation of discourse-level sub-corpora from longer documents.

Activity creation and management. The tool supports the creation of structured CEFR-leveled language learning activities linked to annotated texts.

Role-based access control and annotation workflow. The tool implements a role-based permission system with three user roles: administrator, reviewer, and annotator. Administrators have full system access including user management and model training (see Section 4.2.2). Reviewers manage the annotation workflow by assigning texts to annotators, reviewing submitted assessments, and setting final CEFR levels upon approval. Annotators provide independent CEFR assessments on assigned texts, with the system hiding existing level assignments to prevent bias. This multi-stage workflow — creation, assignment, independent assessment, and expert review — is designed to produce reliable annotations suitable for supervised learning.

Inter-annotator agreement. To assess annotation reliability, the tool computes inter-annotator agreement statistics. It employs Cohen's Kappa for pairwise comparisons and Fleiss' Kappa when three or more annotators assess the same text. Weighted Kappa with linear weights is used to account for the ordinal nature of the CEFR scale. Agreement scores are interpreted according to the Landis and Koch scale and are displayed alongside per-text annotation summaries, enabling reviewers to identify texts with divergent assessments and prioritize them for adjudication.

Statistics and analytics dashboard. A dedicated statistics dashboard provides aggregated views of the corpus, including CEFR-level distributions for texts, words, and sentences; content distributions by genre, topic, and register; quality metrics per proficiency level (average word count, readability scores, lexical diversity); inter-annotator agreement summaries; and a confusion matrix comparing predicted and final CEFR levels. Per-annotator performance metrics are also available. All statistics can be exported to a formatted Excel workbook.

Multi-format corpus export. Beyond ML-oriented dataset generation, the tool supports corpus export in multiple standard formats: JSON, CSV, TSV, XML, RDF/OWL (using Dublin Core metadata terms and a custom CEFR namespace), formatted Excel workbooks, and the Universal-CEFR (Imperial et al., 2025) JSON schema for interoperability with shared CEFR datasets. Exports can be filtered by CEFR level, annotation status, genre, register, topic, and dataset split. Both full exports (including prediction metadata) and minimal variants are available.

REST API and programmatic access. The tool exposes a REST API for programmatic interaction. Key endpoints include CEFR prediction (combining rule-based and machine learning outputs with full feature breakdowns), text and activity export with filtering, and corpus statistics retrieval.

3.2. CEFR-aligned dataset

For the purposes of the present study, a structured and pedagogically oriented corpus of authentic written texts was systematically compiled from a broad spectrum of digital sources, including newspapers, general-interest magazines, and electronic media outlets operating within the German-speaking context (e.g., television networks providing written news reports). The inclusion criteria were guided by principles of authenticity, communicative relevance, and representativeness of contemporary language use. Only openly accessible sources were considered.

In order to ensure alignment with standardized language assessment practices and established proficiency benchmarks, the corpus was supplemented with sample examination materials issued by officially recognized certification bodies. These included open access materials from the State Certificate of Language Proficiency in Greek (KPG), as well as publicly available sample papers provided by the Goethe Institut and the Österreichisches Sprachdiplom Deutsch (ÖSD) for proficiency levels A–C. Furthermore, pedagogically graded resources *from the platform of Deutsche Welle* were systematically integrated, given their explicit alignment with proficiency descriptors and their didactic scaffolding.

The classification and annotation of the texts with respect to thematic domains and genre typology were conducted in accordance with the proficiency levels and thematic specifications defined for each level (A1–C2) by the CEFR. The categorization process followed a level-sensitive and descriptor-informed approach, ensuring coherence between linguistic complexity, communicative function, textual genre, and thematic scope. The corpus encompasses a diverse range of genres, including advertisements, journalistic articles, blog posts, and dialogic interactions, not only from examination material but also from online newspapers and magazines. The thematic spectrum is correspondingly broad, covering domains such as biographical narratives, education, culture, and entertainment, thereby facilitating exposure to varied discourse types, registers, and communicative contexts. Annotation was performed by two trained linguists followed by a review/adjudication process. Pairwise inter-annotator agreement yielded a linearly weighted Cohen's κ of 0.75, indicating substantial agreement on the Landis and Koch scale (Landis and Koch, 1977).

Currently, the dataset comprises a set of CEFR-aligned texts along with CEFR-leveled activities. In specific, the dataset amounts to c.920 texts, 854K tokens and 59K sentences. Each text in the corpus is stored together with its raw textual content, assigned CEFR level, a comprehensive set of linguistic feature metadata, pre-computed syntactic complexity measures, and a CEFR-aligned vocabulary profile. In terms of size and depth of linguistic annotation, the dataset represents one of the more extensive manually curated CEFR-aligned German resources currently available for research in automated text difficulty assessment.

The CEFR-labeled subset covers all six proficiency levels (A1–C2). Although the number of texts is distributed across levels, advanced stages (C1 and C2) account for a substantial proportion of the total word count, reflecting the greater length, lexical density, and structural complexity characteristic of higher-level materials. This distribution supports the modeling of proficiency progression from beginner to intermediate levels, provides a rich representation of advanced syntactic phenomena, and enables systematic analysis of confusion patterns between adjacent CEFR levels. Table 1 gives an overview of the CEFR-aligned corpus.

CEFR	TC	AvgWC	AvgR	AvgLDiv	AvgSL
A1	148	77.81	73.60	0.7804	7.71
A2	170	158.36	64.36	0.7463	9.79
B1	164	162.40	58.13	0.7617	11.89
B2	155	245.10	52.41	0.7359	14.44
C1	187	719.82	49.39	0.6169	17.25
C2	97	1061.45	43.92	0.5864	18.05

Table 1: Descriptive statistics per CEFR level for the German corpus: Text Count (TC), Average Word Count (AvgWC), Average Readability (AvgR; Flesch Reading Ease, Amstad German adaptation), Average Lexical Diversity (AvgLDiv; type–token ratio), and Average Sentence Length (AvgSL, in tokens).

3.3. Word complexity lexicon

The lexical backbone of the system consists of a structured resource centered on a comprehensive German word list that depicts lexical complexity in terms of CEFR levels. As a starting point, we used officially published CEFR-aligned word lists as seeds, ensuring reliable level assignments for core vocabulary. These seed lists were imported via the DWDS API[1] This resource enables the computation of CEFR-specific vocabulary distributions within texts (A1–C2 percentages) as well as derived indicators such as basic-to-advanced vocabulary ratios.

Since official lists provide limited coverage at higher proficiency levels, we implemented a frequency-based expansion strategy using approximately 1.5M German lemmas derived from the OpenSubtitles corpus (Lison and Tiedemann,

[1] https://www.dwds.de/d/api

2016)[2] ranked by frequency[3]. OpenSubtitles was selected for its large scale and broad genre coverage; however, its conversational register may underrepresent academic and formal language typical of higher CEFR levels. The resulting frequency-based CEFR assignments should therefore be understood as corpus-derived approximations of proficiency level rather than pedagogically validated CEFR annotations in the strict sense. The raw frequency ranks were normalized to a 1–1000 scale to ensure uniform comparability and computational stability. We then partitioned the ranked vocabulary into frequency bands aligned with CEFR levels, assigning the 1–2,000 most frequent items to A1, 2,000–5,000 to A2, 5,000–10,000 to B1, 10,000–20,000 to B2, 20,000–35,000 to C1, and items beyond 35,000 to C2. This stratification reflects the well-established correlation between lexical frequency, acquisition sequence, and proficiency development, thereby providing a principled mechanism for approximating CEFR levels for out-of-list lexical items. While this mapping is heuristic rather than prescriptive, it offers a scalable and corpus-derived baseline that can be recalibrated as additional CEFR-annotated data become available.

To incorporate morphosyntactic complexity beyond frequency, lexical entries are enriched with information extracted from Wiktionary (Wiktionary contributors, 2024), including irregular verb status, separable verb properties, compound formation, and international word marking. These annotations capture structural properties associated with acquisition difficulty in German and provide linguistically interpretable signals for modeling. We note that the effect of internationalisms on perceived difficulty is L1-dependent: cognates may facilitate comprehension for learners with Romance L1 backgrounds while offering less support to learners from typologically distant language families. In the current implementation, international word status is treated as a uniform feature; future work could condition its contribution on learner L1. The resulting vocabulary resource directly feeds into the machine learning pipeline.

4. Text difficulty assessment

4.1. Syntactic complexity modeling

All texts undergo automatic dependency parsing using spaCy (Honnibal et al., 2020) models, allowing the extraction of measures of syntactic complexity on a scale. Pre-computed features include subordinate, relative, and infinitive clause ratios, passive voice frequency, modal verb density, nominalization ratio, clause embedding depth, average dependency distance, and noun–verb ratio. These indicators capture structural phenomena associated with proficiency progression, particularly the increasing use of clause embedding and nominal structures at higher CEFR levels. Syntactic features are computed once and stored in the database, ensuring reproducibility and efficient retraining without repeated parsing. This pre-analysis design reduces computational overhead during model experimentation and facilitates integration with both rule-based and machine learning approaches. Together with lexical and readability measures, the syntactic layer provides complementary structural signals that improve level discrimination, as shown in the confusion analysis in Section 6.

4.2. Baseline and Machine Learning

4.2.1. Rule-Based Baseline

As a transparent and pedagogically interpretable baseline, we developed a rule-based scoring system that combines three independent prediction components: (a) CEFR-aligned vocabulary distribution analysis, (b) readability metric voting, and (c) linguistic complexity scoring.

The **vocabulary component** computes the cumulative CEFR level coverage for each text using the lexical resource described in Section 3.3. After lemmatization with spaCy and filtering of proper nouns, each token is assigned to its CEFR level. A text is assigned to a given level if the cumulative percentage of words at or below that level exceeds a threshold (e.g. $\geq$85% for A1, $\geq$80% for A2, decreasing to $\geq$65% for C1). Texts with more than 30% unknown vocabulary default to C2; this threshold serves as a fallback for out-of-vocabulary texts. In practice, unknown-word rates are much lower across all levels, with C2 texts averaging 8.1% and C1 texts 6.4%. Confidence is derived from the margin above the threshold, adjusted downward when the unknown-word percentage is high.

The **readability component** employs weighted voting across seven readability metrics: Wiener Sachtextformel (weight 1.0), Flesch Reading Ease in the Toni Amstad German adaptation (0.9), Flesch–Kincaid Grade Level (0.7), Gunning Fog Index (0.6), SMOG Index (0.6), Automated Readability Index (0.65) and Coleman–Liau Index (0.65). Each metric independently maps its score to a CEFR level using predefined range tables. The final readability prediction is determined by weighted majority voting, with confidence modulated by a spread factor that penalizes disagreement among metrics.

The **linguistic component** computes a composite score (0–100) that aggregates four feature cate-

[2] http://www.opensubtitles.org/

[3] https://github.com/hermitdave/FrequencyWords

gories: syntactic complexity (subordination depth, mean dependency distance, clauses per sentence), grammatical features (passive voice, subjunctive mood, genitive case, complex tenses), lexical sophistication (type–token ratio, average frequency rank, compound word ratio) and discourse markers (proportion of advanced and sophisticated connectives). The composite score is mapped to CEFR levels using calibrated boundaries.

The three components are combined via weighted index averaging. Default weights (vocabulary: 0.30, readability: 0.40, linguistic: 0.30) are dynamically adjusted based on vocabulary coverage quality and linguistic signal strength. Ensemble confidence is boosted when components agree and reduced when all three predict different levels, in which case the median prediction is selected.

Calibration is data-driven: using the approved subset of the corpus, the system extracts per-level feature distributions and computes percentile-based thresholds for syntactic features (e.g., mean dependency distance boundaries at 2.5, 2.75, 3.09, 3.38 for the A1/A2 through C1 transitions). Component weights are optimized based on individual prediction accuracy on the calibration set, yielding readability: 0.424, vocabulary: 0.31, linguistic: 0.266.

4.2.2. Feature-Based Machine Learning

Building on the linguistic features used in the rule-based system, we train supervised classifiers using an expanded feature set of 89 dimensions organized into seven categories: (1) basic text statistics, (2) readability metrics, (3) CEFR vocabulary profile, (4) derived vocabulary ratios, (5) morphosyntactic features, (6) syntactic complexity, and (7) TF-IDF features (see Appendix 11.1). All features are standardized using z-score normalization fitted on the training partition. We evaluate four classifiers: Random Forest (RF), Gradient Boosting (GB), Support Vector Machine with RBF kernel (SVM), and XGBoost. Class imbalance is addressed through balanced class weighting for RF and SVM. Model selection employs grid search with 3-fold stratified cross-validation, optimizing for macro F1. Data are split into training (70%), development (15%), and test (15%) partitions with stratified sampling.

Four feature selection strategies are available: importance-based filtering (retaining features with Random Forest importance ≥ 0.01), correlation-based filtering (removing features with Pearson $r > 0.85$), recursive feature elimination (RFE), and a combined approach applying correlation filtering followed by importance-based selection.

Model interpretability is supported through SHAP (SHapley Additive exPlanations) analysis (Lundberg and Lee, 2017). For tree-based models, exact SHAP values are computed via `TreeExplainer`; for SVM, approximate values are obtained via `KernelExplainer`. The system provides global feature importance rankings, per-class feature contributions, and local prediction explanations for individual texts.

Evaluation metrics include accuracy, macro and weighted F1 scores, per-class precision, recall, and F1, and adjacent accuracy (predictions within ± 1 CEFR level). Probability calibration is assessed via Expected Calibration Error (ECE), Maximum Calibration Error (MCE), and the multi-class Brier score.

4.3. BERT modelling

We fine-tune a German BERT model (`deepset/gbert-base`) (Chan et al., 2020) for direct CEFR classification. The model consists of 12 transformer layers with a hidden size of 768 dimensions. A classification head comprising a dropout layer ($p = 0.3$) followed by a linear projection ($768 \rightarrow 6$) is added on top of the `[CLS]` token representation.

Training uses the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 2×10^{-5}, weight decay of 0.01, and a linear warmup schedule (10% of total steps) followed by linear decay. Sequences are tokenized using WordPiece tokenization and truncated or padded to a maximum length of 512 tokens. Class imbalance is handled through inverse-frequency weighting in the cross-entropy loss, combined with label smoothing ($\epsilon = 0.1$) to mitigate overconfident predictions. Early stopping monitors macro F1 on the validation set with a patience of 3 epochs.

Mixed-precision training (FP16) is enabled on CUDA devices via PyTorch's Automatic Mixed Precision, with gradient clipping at a maximum norm of 1.0. The system automatically detects the available hardware, supporting NVIDIA GPUs (CUDA), Apple Silicon (MPS), and CPU fallback. Four training presets are provided to accommodate different computational budgets: a *default* configuration (batch size 8, 10 epochs, learning rate 2×10^{-5}), a *fast* preset (batch size 16, 5 epochs, max length 256), an *accurate* preset (batch size 4, 15 epochs, learning rate 1×10^{-5}, dropout 0.2, patience 5), and a *CPU-friendly* preset (batch size 4, max length 256, no mixed precision).

Data partitioning follows the same stratified 70/15/15 split as the feature-based models, ensuring comparable evaluation. At inference, the model outputs a softmax probability distribution over all six CEFR levels, enabling both point predictions and uncertainty-aware decision-making.

5. Results

We evaluated three modeling approaches on the German CEFR dataset described in Section 3.2: the rule-based baseline, the best-performing feature-based ML classifier, and the fine-tuned BERT model. The rule-based and ML models were assessed on the same held-out test partition ($n = 134$), and the BERT model on a comparable partition ($n = 136$), both using stratified sampling to preserve class distribution.

5.1. Rule-Based Classification

The rule-based system combines three signal sources through weighted voting: vocabulary profile analysis (weight 0.30), readability metrics from multiple formulas (weight 0.40), and linguistic feature analysis including syntactic complexity (weight 0.30). Weights are dynamically adjusted based on input quality—for instance, vocabulary weight increases when unknown-word percentage is low, indicating good dictionary coverage.

Table 2 presents the per-class performance. The system achieves an overall accuracy of 43.3% and a macro F1 of 0.385. Adjacent accuracy reaches 88.1%, indicating that most errors involve neighboring levels rather than large classification jumps.

Level	Precision	Recall	F1	Support
A1	0.833	0.526	0.645	19
A2	0.560	0.560	0.560	25
B1	0.439	0.720	0.545	25
B2	0.286	0.522	0.369	23
C1	0.286	0.143	0.190	28
C2	0.000	0.000	0.000	14
Macro	**0.401**	**0.412**	**0.385**	134

Table 2: Per-class performance of the rule-based classifier on the test set (accuracy: 43.3%, adjacent accuracy: 88.1%).

Performance is strongly skewed toward lower proficiency levels: A1 achieves the highest precision (0.833) and a reasonable F1 (0.645), while A2 and B1 reach F1 scores of 0.560 and 0.545 respectively. In contrast, the system struggles with higher levels—B2 drops to 0.369, C1 to 0.190, and C2 is never correctly predicted (F1 = 0.000). The confusion matrix reveals the primary failure mode: a systematic downward bias, where C1 texts are most frequently misclassified as B2 (21 out of 28), and all 14 C2 texts are assigned to levels B1–C1. This pattern reflects the limitations of readability formulas and vocabulary frequency analysis for distinguishing advanced proficiency levels, where textual complexity manifests through discourse-level features (argumentation structure, hedging, register) rather than surface-level statistics. These results establish a clear motivation for the machine learning approaches that follow: while the rule-based system provides a functional baseline for lower levels (A1–B1), it fails to discriminate among upper levels (B2–C2), precisely the range where learner placement decisions are most consequential.

5.2. Feature-Based Classification

We trained four classifiers using 3-fold stratified cross-validation with grid search over hyperparameter grids: Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM), and XGBoost. Table 3 reports the cross-validation macro F1 and test set macro F1 for each model.

Classifier	CV F1	Test F1	Acc.	Adj.Acc.
RF	**0.602**	0.597	0.612	0.955
GB	0.567	**0.602**	0.612	0.963
SVM	0.576	0.588	0.597	0.978
XGBoost	0.601	0.600	0.612	0.963

Table 3: Classifier selection via 3-fold stratified cross-validation with grid search (macro F1). Random Forest is selected based on highest CV macro F1.

All four classifiers perform within a narrow range (test macro F1: 0.588–0.602), with Random Forest achieving the highest CV score (0.602) and Gradient Boosting the highest test F1 (0.602). Given this near-parity, we selected Random Forest as the final model based on its best cross-validation performance and its support for feature importance analysis. Its best hyperparameters are: `max_depth`=7, `n_estimators`=200, and `min_samples_split`=2.

Table 4 summarizes the performance of the feature-based classifier trained with 40 selected features (including syntactic and TF-IDF components). The model achieves an overall accuracy of 61.2% and a macro F1 of 0.597. Notably, adjacent accuracy reaches 95.5%, indicating that the vast majority of misclassifications involve adjacent CEFR levels — a pattern consistent with the inherent difficulty of fine-grained ordinal classification on a six-level scale.

Level	Precision	Recall	F1	Support
A1	0.722	0.684	0.703	19
A2	0.645	0.800	0.714	25
B1	0.579	0.440	0.500	25
B2	0.571	0.522	0.545	23
C1	0.606	0.714	0.656	28
C2	0.500	0.429	0.462	14
Macro	**0.604**	**0.598**	**0.597**	134

Table 4: Per-class performance of the feature-based ML classifier on the test set (accuracy: 61.2%, adjacent accuracy: 95.5%).

Performance is strongest at the lower end of the scale: A2 achieves the highest F1 (0.714), followed by A1 (0.703) and C1 (0.656), while C2 and B1 prove most challenging (F1 of 0.462 and 0.500, respectively). This pattern reflects the well-documented difficulty of discriminating between adjacent proficiency levels, where linguistic features overlap substantially.

Calibration analysis reveals an Expected Calibration Error (ECE) of 0.242 and a Brier score of 0.634. The model exhibits overconfidence: average confidence for incorrect predictions is 74.1% compared to 81.0% for correct ones, with 91.5% of errors made with confidence above 0.5. These findings underscore the importance of calibration-aware evaluation in educational applications where prediction confidence informs downstream decisions.

5.3. BERT-Based Classification

The fine-tuned German BERT model was trained using the *accurate* preset (batch size 4, 15 epochs, learning rate 1×10^{-5}, dropout 0.2). Training converged at epoch 14 based on validation macro F1 (0.708). Table 5 presents the test set performance.

Level	Precision	Recall	F1	Support
A1	0.800	0.909	0.851	22
A2	0.786	0.880	0.830	25
B1	0.636	0.583	0.609	24
B2	0.545	0.522	0.533	23
C1	0.600	0.556	0.577	27
C2	0.571	0.533	0.552	15
Macro	**0.656**	**0.664**	**0.659**	136

Table 5: Per-class performance of the fine-tuned BERT model on the test set (accuracy: 66.9%, adjacent accuracy: 95.6%). Rule-based and ML evaluated on n=134; BERT on n=136 due to different stratified split.

The BERT model achieves an accuracy of 66.9% and a macro F1 of 0.659, representing a relative improvement of 10.4% over the feature-based classifier. Adjacent accuracy reaches 95.6%. The improvement is most pronounced for A1–A2 levels, where BERT achieves F1 scores of 0.851 and 0.830 respectively. The B2 level proves most challenging (F1 = 0.533), followed by C2 and C1. B2 occupies a transitional zone between intermediate and advanced proficiency, sharing surface-level features (sentence length, vocabulary breadth) with both B1 and C1, which makes categorical discrimination particularly difficult.

6. Discussion

The progression from rule-based to BERT yields consistent gains: accuracy improves from 43.3% to 61.2% to 66.9%, while macro F1 rises from 0.385 to 0.597 to 0.659. The most dramatic improvement occurs in the upper CEFR levels where the rule-based system fails entirely—the ML classifier recovers C1 (F1 = 0.656) and C2 (F1 = 0.462), while BERT achieves 0.577 and 0.552 respectively. Table 6 summarizes the three approaches.

Metric	Rule-Based	ML (Feature)	BERT
Accuracy	0.433	0.612	0.669
Macro F1	0.385	0.597	0.659
Weighted F1	0.401	0.605	0.663
Adjacent Acc.	0.881	0.955	0.956
Interpretability	High	Medium	Low

Table 6: Comparison of the three modeling approaches on the held-out test split.

Confusion matrix analysis across all three models reveals a consistent pattern: misclassifications concentrate along the diagonal, predominantly between adjacent levels. The B1–B2 boundary is the most error-prone region, followed by C1–C2. A1 texts are the most reliably classified across all approaches, likely due to their distinctive short sentence length, limited vocabulary, and low syntactic complexity. The high adjacent accuracy across all models ($>$88%) suggests that the six-level CEFR scale introduces inherent ambiguity at level boundaries, and that the models capture the ordinal structure of the proficiency continuum even when exact-level prediction fails.

The rule-based baseline, while the least accurate as a standalone classifier, offers full transparency: every prediction can be traced to specific vocabulary thresholds, readability scores, and linguistic indicators. This interpretability makes it suitable for the semi-automatic annotation workflow described in Section 3.1, where human annotators use the rule-based prediction as a starting point. The feature-based ML classifiers provide a middle ground, achieving competitive performance while supporting SHAP-based post-hoc interpretability. SHAP analysis of the Random Forest model reveals that word count, the first TF-IDF principal component, and average sentence length are the three most influential features globally. Notably, the driving features shift across proficiency levels: at lower levels (A1–A2), vocabulary profile features (A1 vocabulary percentage, basic vocabulary ratio) dominate, while at advanced levels (C1–C2), text length and TF-IDF features become the primary discriminators, reflecting the greater lexical and structural diversity characteristic of higher-proficiency texts. BERT delivers the highest discriminative performance but operates as a black-box model, with prediction confidence as the primary transparency mechanism.

It is worth noting that none of the three ap-

proaches achieves particularly high exact-match accuracy, which may appear surprising at first glance. However, this outcome is entirely expected given the nature of CEFR level assignment. Many texts lie at the boundary between two adjacent levels, and human annotators must make a categorical decision where the underlying proficiency is continuous. This annotation process inherently introduces a degree of subjectivity: two expert annotators may reasonably disagree on whether a text is B1 or B2, for instance. The moderate exact-match accuracy across all three methods directly reflects this inter-annotator ambiguity in the training data. Crucially, the consistently high adjacent accuracy (88.1% for rule-based, 95.5% for ML, and 95.6% for BERT) demonstrates that when the models err, they almost always predict a neighboring level—mirroring the same boundary uncertainty that human annotators face.

Furthermore, the prediction confidence scores provide additional diagnostic value: when a model assigns a text to a given level with low confidence and high probability mass on an adjacent level, this signals that the text genuinely straddles two proficiency levels, offering practitioners actionable information beyond the categorical prediction alone.

7. Conclusion

We have presented a framework for CEFR-based text difficulty prediction that combines a transparent rule-based baseline, interpretable feature-based ML classifiers, and a fine-tuned BERT model. Evaluated on a 920-text German corpus, the three approaches yield progressively higher accuracy (43.3%, 61.2%, 66.9%) while adjacent accuracy exceeds 88% across all models, confirming that errors predominantly involve neighbouring CEFR levels. The B1–B2 boundary remains the most challenging region even for BERT, reflecting inherent ambiguity at mid-proficiency transitions. SHAP analysis reveals that the dominant predictive features shift from vocabulary profile at lower levels to text length and TF-IDF components at advanced levels, offering pedagogically interpretable insights into what distinguishes proficiency stages. In its current implementation, the framework handles German, with its Greek counterpart under development. The cross-linguistic extension will allow systematic investigation of feature transferability across typologically distinct languages.

Future work is planned in two directions: (a) finalising the Greek CEFR-aligned corpus and adapting the classifier for Modern Greek, and (b) leveraging LLMs for controlled text simplification, using the calibrated difficulty predictions and feature-level insights from the current framework as constraints to guide generation toward target CEFR levels.

The corpus, the annotation tools, and the models will be freely available via a GitHub repository. Moreover, the tool for automatic text difficulty prediction will be available via a dedicated API.

8. Acknowledgements

This research was supported by the National Recovery and Resilience Plan (NRRP) “Greece 2000” under the “Clusters of Research Excellence” (CREs) program, SUB1.1, with project code $\mathrm{O\Pi\Sigma\ TA}\ 5180519$ and title “Interactive Agent with Emotional Intelligence for Second/Foreign Language Learning”, Acronym: “EmoBot”.

9. Ethical considerations and limitations

Despite its extensibility, the current implementation remains limited to German, with Greek currently under development. Therefore, cross-linguistic generalization has not yet been empirically validated. The corpus used in this study consists exclusively of publicly available written materials, including open-access media sources and sample examination materials released by officially recognized certification bodies. No personal data or learner-produced texts were collected.

10. Bibliographical References

References

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Richard Bamberger and Erich Vanecek. 1984. *Lesen - Verstehen - Lernen - Schreiben. Die Schwierigkeitsstufen von Texten in deutscher Sprache.* Jugend u. Volk Sauerlaender, Wien.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German’s next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Meri Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Council of Europe, Strasbourg.

Edgar Dale and Jeanne S. Chall. 1949. The concept of readability. *Elementary English*, 26(1):19–26.

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. Association for Computational Linguistics.

R Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 1063–1080, Mumbai, India.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Joseph Marvin Imperial. 2021. BERT embeddings for automatic readability assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 611–618.

Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, et al. 2025. UniversalCEFR: Enabling open multilingual research on language proficiency assessment. *arXiv preprint arXiv:2506.01419*.

Athanasios Kallipolitis, Dionysios Koulouris, Melina Tziokama, Kosmas Pinitas, Argyrios Zafeiriou, Andreas Menychtas, Ilias Maglogiannis, Voula Giouli, Athina Sioupi, Stamatia Michalopoulou, George Tsoulouhas, Michail Katras, Panagiotis Charalampopoulos, and Aristotelis Stamopoulos. 2026. Conversational agent with emotional intelligence for foreign language learning. In *Proceedings of the 14th International Conference on Information and Education Technology (IEEE-ICIET 2026), Koriyama, Japan*. IEEE.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical Report Research Branch Report 8-75, Naval Technical Training Command Millington TN Research Branch.

David Kogan, Max Schumacher, Sam Nguyen, Masanori Suzuki, Melissa Smith, Chloe Sophia Bellows, and Jared Bernstein. 2025. Ace-CEFR – a dataset for automated evaluation of the linguistic difficulty of conversational texts for LLM applications. *arXiv preprint arXiv:2506.14046*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*. ArXiv:1711.05101.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30.

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

Ildikó Pilán, Sowmya Vajjala, and Elena Volodina. 2016. A readable read: Automatic assessment of language learning materials based on linguistic complexity. *International Journal of Computational Linguistics and Applications*, 7(1):143–159.

Donya Rooein, Paul Röttger, Anastassia Shaitarova, and Dirk Hovy. 2024. Beyond flesch-kincaid: Prompt-based metrics improve difficulty classification of educational texts. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 54–67, Mexico City, Mexico. Association for Computational Linguistics.

Brucce Neves dos Santos, Ricardo Marcondes Marcacini, and Solange Oliveira Rezende. 2021. Multi-domain aspect extraction using bidirectional encoder representations from transformers. *IEEE Access*, 9:91604–91613.

E. A. Smith and R. Senter. 1967. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories*, pages 1–14.

Sean Trott and Pamela Rivière. 2024. Measuring and modifying the readability of English texts with GPT-4. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 126–134, Miami, Florida, USA. Association for Computational Linguistics.

Satoru Uchida. 2024. Evaluating the accuracy of ChatGPT in assessing writing and speaking: A verification study using ICNALE GRA. *Learner Corpus Studies in Asia and the World*, 8.

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. Association for Computational Linguistics.

Wiktionary contributors. 2024. Wiktionary, the free dictionary. Accessed: 2025.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22. Association for Computational Linguistics.

Kevin P. Yancey, Geoffrey T. LaFlair, Anthony R. Verardi, and Jill Burstein. 2023. Rating short L2 essays on the CEFR scale with GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 576–584. Association for Computational Linguistics.

11. Appendix

11.1. The sets of features used in Machine Learning

1. **Basic text statistics** (5 features): word count, sentence count, average sentence length, lexical diversity (type–token ratio), and Flesch Reading Ease score.

2. **Readability metrics** (8 features): Wiener Sachtextformel, Flesch–Kincaid Grade Level, Gunning Fog, SMOG Index, ARI, Coleman–Liau, average syllables per word, and average characters per word.

3. **CEFR vocabulary profile** (7 features): percentage of tokens at each CEFR level (A1–C2) and percentage of unknown tokens.

4. **Derived vocabulary ratios** (3 features): basic (A1+A2), intermediate (B1+B2), and advanced (C1+C2) vocabulary proportions.

5. **Morphosyntactic features** (4 features): percentages of irregular verbs, separable verbs, compound words, and international/borrowed words, derived from the Wiktionary-enriched lexicon.

6. **Syntactic complexity** (12 features): subordinate, compound, relative, and infinitive clause ratios; passive voice and modal verb ratios; nominalization ratio; average and maximum clause depth; mean dependency distance; noun–verb ratio; and clauses per sentence.

7. **TF-IDF features** (50 features): unigram and bigram TF-IDF vectors (max 500 terms, sublinear TF scaling, minimum document frequency of 2) reduced to 50 principal components via PCA.

11.2. Confusion Matrices

	A1	A2	B1	B2	C1	C2
A1	**10**	7	2	0	0	0
A2	2	**14**	6	0	3	0
B1	0	3	**18**	4	0	0
B2	0	1	10	**12**	0	0
C1	0	0	3	21	**4**	0
C2	0	0	2	5	7	**0**

Table 7: Confusion matrix for the rule-based classifier. Rows represent true labels; columns represent predicted labels. Note the strong downward bias: 21 of 28 C1 texts are misclassified as B2, and all C2 texts are assigned to lower levels.

	A1	A2	B1	B2	C1	C2
A1	**13**	6	0	0	0	0
A2	4	**20**	1	0	0	0
B1	1	4	**11**	8	1	0
B2	0	1	5	**12**	4	1
C1	0	0	2	1	**20**	5
C2	0	0	0	0	8	**6**

Table 8: Confusion matrix for the feature-based ML classifier (Random Forest). Errors concentrate along the diagonal; the B1 level shows the widest spread, with 8 texts misclassified as B2 and 4 as A2.

	A1	A2	B1	B2	C1	C2
A1	**20**	1	1	0	0	0
A2	2	**22**	1	0	0	0
B1	3	4	**14**	3	0	0
B2	0	1	6	**12**	3	1
C1	0	0	0	7	**15**	5
C2	0	0	0	0	7	**8**

Table 9: Confusion matrix for the fine-tuned BERT model. Tightest diagonal concentration among the three approaches; the largest off-diagonal entries are C1→B2 (7) and C2→C1 (7), both involving adjacent levels.

Terminology-Augmented Generation for Intangible Cultural Heritage: A Controlled LLM-Based Translation Framework

Wanda Punzi Zarino[1], **Pilar Sánchez Gijón**[2]
Department of Economics and Law, University of Naples Parthenope, Italy[1]
Department of Translation, Interpreting and East Asian Studies, Autonomous University of Barcelona, Spain[2]
Via Acton, 38 - 80133 Naples, Italy[1]
Edifici K, Plaça del Coneixement - 08193 Bellaterra, Spain[2]
wanda.punzizarino001@studenti.uniparthenope.it, Pilar.Sanchez.Gijon@uab.cat

Abstract

This study examines the integration of a bilingual Italian–Spanish concept-oriented terminological resource into a controlled large language model (LLM) translation workflow within the domain of Campanian gastronomy. The termbase encodes structured conceptual, linguistic, and translational metadata, including grammatical information, translation strategies, and genre-sensitive usage recommendations. Through a local Model Context Protocol (MCP) architecture, the resource is dynamically connected to locally deployed LLMs, enabling the automatic identification and retrieval of relevant terminological units prior to generation. The system combines in-context terminological injection with deterministic post-processing enforcement: genre-specific policies are injected into the model prompt prior to generation and verified through a rule-based post-processing layer that enforces surface-level terminological consistency in the output. Two open-weight models — Mistral 7B Instruct and Gemma3 4B — are evaluated across three conditions and three discursive genres on a dataset of authentic texts. The findings suggest that the combination of terminological injection and deterministic enforcement can improve terminological compliance in controlled, domain-specific settings, while also highlighting differences in instruction-following behavior across models and genres.

Keywords: Terminology-augmented generation, Bilingual IT-ES termbase, Intangible cultural heritage

1. Introduction

The UNESCO Convention for the Safeguarding of the Intangible Cultural Heritage [UNESCO, 2003] defines intangible heritage as encompassing practices, knowledge, and skills transmitted across generations. Within this framework, gastronomy emerges as a culturally dense domain in which culinary practices intertwine material production, ritual traditions, territorial identity, and linguistic expression. Terminological units in this domain cannot be reduced to mere technical labels; rather, following Cabré [1993, 2003], they can be understood as linguistic realizations of structured conceptual knowledge, encoding production processes, historical trajectories, and socio-cultural values [Chessa et al., 2014, Grimaldi, 2017, Buccheri, 2023].

When such culturally embedded terminology enters translation contexts, additional complexity arises, as translation choices affect not only lexical form but also conceptual integrity and cultural mediation. The degree of communicative constraint further modulates terminological stability across text types [Sabatini, 1990, 1999], while established translation techniques provide an operational framework for addressing conceptual asymmetries in gastronomic contexts [Molina and Hurtado Albir, 2002].

Against this background, this paper explores the application of the Terminology-Augmented Generation (TAG) framework to the domain of Campanian traditional agri-food products. A concept-oriented Italian–Spanish bilingual termbase was developed, encoding structured conceptual and linguistic metadata, including grammatical information, translation strategy labels (e.g., borrowing, established equivalent, literal translation), and genre-sensitive usage recommendations. The resource is integrated into an LLM-based translation workflow through a local Model Context Protocol (MCP) server, which enables the deterministic retrieval of relevant terminological units prior to translation. The system combines in-context terminological injection with a deterministic post-processing enforcement layer, ensuring that both lexical and structural genre constraints are applied consistently. Two open-weight models — Mistral 7B Instruct and Gemma3 4B — are evaluated in parallel, allowing for a cross-model comparison that sheds light on the relationship between instruction-following capacity and the effectiveness of terminology-augmented generation.

Three research questions guide this study: RQ1: Can a concept-oriented bilingual termbase regulate terminological compliance across discursive genres in LLM-based translation? RQ2: Does the combination of in-context terminological injection and deterministic post-processing enforcement improve control over terminological realization compared to either component alone? RQ3: To what extent do differences in instruction-following capac-

Proceedings of the 2nd Workshop on Evaluating Text Difficulty in a Multilingual Context (DeTermIt! 2026), pages 74–82
11 May 2026.

ity across open-weight models affect the effectiveness of terminology-augmented generation?

The remainder of the paper is structured as follows. Section 2 reviews related work on RAG and TAG architectures. Section 3 introduces the corpus and terminological data. Section 4 describes the case study and the MCP-based integration architecture. Section 5 presents and discusses the results. Section 6 concludes with reflections on future research directions.

2. Background and Related Work

In recent years, the rapid development of Large Language Models (LLMs) has led to growing interest in integrating structured terminological resources into generative AI systems. One of the most influential architectural paradigms in this context is Retrieval-Augmented Generation (RAG), introduced by Lewis et al. [2020], which combines neural text generation with the dynamic retrieval of external documents. While RAG improves factual grounding by retrieving contextually relevant passages at inference time, it is primarily designed for unstructured text retrieval based on dense vector similarity and does not inherently support deterministic access to structured, concept-oriented terminological data. Subsequent studies have highlighted both the strengths and limitations of RAG, particularly with regard to noise in retrieved content and limited control over fine-grained domain-specific knowledge [Gupta et al., 2024, Lackner et al., 2025b].

Within specialized domains, researchers have begun exploring terminology-enhanced variants of RAG. Martín-Chozas et al. [2025], for instance, show that combining neural retrieval with curated lexical resources can improve retrieval effectiveness and answer quality in legal corpora. However, these approaches remain embedded within the broader RAG paradigm and continue to rely on similarity-based mechanisms.

In response to these structural limitations, Terminology-Augmented Generation (TAG) has emerged as a complementary paradigm to RAG [Fleischmann, 2025], enabling direct and deterministic access to structured termbases. The theoretical foundations of TAG have been further articulated by Di Nunzio [2025], who frames it as a generative architecture grounded in the dual conceptual and linguistic dimensions of terminology science, advocating a design that includes a terminology access layer, filtering and reasoning components, and human-in-the-loop validation mechanisms. Empirical studies have begun to substantiate its effectiveness. Lackner et al. [2025a] show that TAG significantly improves terminological adherence in LLM-based machine translation, while Lackner et al. [2025b] demonstrate that lightweight structured formats such as YAML and JSON enhance model performance in in-context learning settings. They further observe that Mistral 7B consistently produced unusable outputs in multilingual translation tasks — a finding that informs the comparative dimension of the present study. Taken together, these contributions position TAG as a robust framework for embedding curated terminological knowledge into generative pipelines, particularly in multilingual and domain-sensitive contexts where precision and expert validation are essential.

3. Corpus and Terminological Data

The corpus underpinning this study is derived from the official register of *Prodotti Agroalimentari Tradizionali* (PAT) of the Campania region [Regione Campania, 2025]. At the time of data collection, the regional register comprised 610 officially recognized products. For the purposes of this study, a subset of 28 entries was selected, corresponding to the subdomain of gastronomic preparations. This selection reflects a methodological choice to prioritize gastronomic preparations over raw materials, as culinary terms tend to exhibit greater terminological and translational complexity. Unlike names referring to primary agricultural products, they encode procedural knowledge, preparation methods, seasonal consumption patterns, and communal practices embedded in specific territorial, ritual, and socio-historical contexts.

The repertory displays marked lexical heterogeneity. Some terms include dialectal or regionally marked forms, such as *ciauliello*, whose semantic opacity may challenge non-local readers and translators and whose resistance to straightforward lexical equivalence reflects its function as a marker of territorial identity. Others reveal historically motivated lexical forms: *frittata di scammaro*, for instance, derives from socio-religious dietary practices associated with Lenten observance in the Kingdom of the Two Sicilies, where *scammaro* referred to “lean days” during which conventual cooking excluded eggs and meat. The corpus also includes terms anchored in festive contexts, such as *cicci di Santa Lucia*, which situates the preparation within a specific ritual calendar (13 December) and within the communal practices of Avellino and its surrounding area.

4. Case Study

The termbase, encoded in YAML format in line with recent findings on lightweight structured representations for in-context learning [Lackner et al., 2025a,b], constitutes the knowledge layer underpinning the experiment. Rather than functioning as a static glossary, it encodes formally structured

conceptual and translational data that can be programmatically accessed at run time. Each entry is organized into three layers. The conceptual layer encodes a concept identifier, domain classification, definitional content, production notes, cultural notes, area of production, and a source field referencing the official PAT register. The linguistic layer provides grammatical and lexical metadata for both Italian and Spanish, including part of speech, gender, and registered variants. The translational layer specifies translation strategy labels, preferred target equivalents, and genre-sensitive usage rules.

The experimental dataset consists of 30 authentic Italian texts distributed across three discursive genres: regulatory-administrative, promotional-touristic, and narrative-cultural (10 texts per genre). Rather than constructing artificial sentences designed to contain specific terminological units, the texts were drawn from real-world sources, including official PAT product register entries, gastronomic journalism, promotional food blogs, and literary or narrative culinary writing. Each text contains at least one terminological unit from the PAT-Campania termbase. The three genres were selected to represent distinct points on Sabatini's typological cline of communicative constraint [Sabatini, 1990, 1999], ranging from the highly constrained regulatory-administrative register, characterized by strict terminological stability and institutional formatting conventions, to the more loosely constrained narrative-cultural register, in which voice, perspective, and affective engagement take precedence over formal precision.

Integration into the translation workflow was achieved through a local MCP server, which exposes terminological content as callable tools accessible to the orchestration layer. In line with the conceptual framework proposed by Di Nunzio [2025], the system integrates a terminology access component together with filtering, reasoning, and generation modules, extended with a deterministic post-processing enforcement layer. Two open-weight models were evaluated in parallel — Mistral 7B Instruct and Gemma3 4B — both accessed through the Ollama framework. The entire system was implemented and executed locally within an isolated Python virtual environment, ensuring controlled dependency management, modular deployment, and reproducibility. No external retrieval services or third-party APIs were employed. The overall workflow is illustrated in Figure 1.

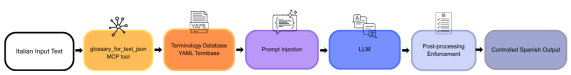


Figure 1: Terminology-augmented generation architecture.

The terminology retrieval phase follows a deterministic sequence prior to generation. Given an Italian source text, the system first invokes the MCP tool *glossary_for_text_json*, which performs automatic identification of relevant terminological units through boundary-aware Unicode-safe matching. The tool retrieves all matching entries from the termbase, including single-word terms, multi-word expressions, and registered variants, together with structured IT–ES correspondences, translation strategy labels, and genre-sensitive usage rules. When genre-specific modulation is required, the system subsequently invokes the *translate_usage* tool to retrieve the translation policy associated with a selected discursive genre. The retrieved policy specifies the output form, the preferred Spanish equivalent, and the field from which the expansion note should be drawn: production notes for promotional-touristic texts, cultural notes for narrative-cultural texts.

The generation module receives a shared system prompt (Appendix A) that defines the translator role and general translation instructions, together with a genre-specific glossary block (Appendix B) injected into the user message. The glossary block is constructed dynamically from the retrieved terminological policies: in regulatory-administrative contexts, the model is instructed to preserve the original Italian term formatted with Spanish angle quotation marks; in promotional-touristic contexts, to use the preferred Spanish equivalent immediately followed by a parenthetical expansion, reformulated based on the context and co-text of the passage; in narrative-cultural contexts, to introduce the preferred Spanish equivalent with an inline cultural gloss introduced by a comma. Crucially, the glossary block is not a static template but a dynamically constructed instruction set that reflects the specific terminological units detected in each source text and the genre rules encoded in the termbase.

Once the output is generated, the same genre-specific policies are applied deterministically through the post-processing enforcement layer. For regulatory-administrative texts, the enforcement layer verifies whether the Italian term is present in the output: if found without angle quotation marks it adds them, and if the preferred Spanish equivalent appears instead of the Italian term it replaces it with the angle-quoted Italian form. For promotional-touristic and narrative-cultural texts, the enforcement layer verifies whether the preferred Spanish equivalent is present: if absent but the Italian source term is found and differs from the Spanish equivalent, it substitutes the latter. The enforcement layer does not intervene on parenthetical expansions or inline glosses — a conservative design choice that avoids the risk of inserting contextually

inappropriate content.

5. Analysis of Results

5.1. Terminology Adherence

Terminology Adherence (TA) was computed automatically for all three conditions across both models and genres. For regulatory-administrative texts, TA measures whether the Italian source term is present in the model output; for promotional-touristic and narrative-cultural texts, it measures whether the preferred Spanish equivalent as specified in the termbase is present. Term presence was verified through boundary-aware, Unicode-safe string matching with apostrophe normalization, and TA was computed as a proportional score per text, averaged across texts within each genre and condition. Results are reported in Tables 1, 2, and 3.

The baseline condition reveals a fundamental limitation of unconstrained LLM-based translation: both models spontaneously preserve the expected terminology in fewer than half of the cases globally — and as rarely as 15% for narrative-cultural texts. Culturally opaque and dialectally marked terms exert no statistical pressure toward preservation in models trained on general multilingual corpora; in the absence of explicit guidance, probabilistic generation tends to resolve terminological opacity through free translation, paraphrase, or domestication. This is illustrated by cases such as *cicatielli con pulieio*, rendered as *cicatelli con pulígaro* by Mistral — a non-existent form resulting from the misinterpretation of the dialectal term *pulieio* — or *frittata di scammaro*, rendered as *frittata de escarola* by Gemma, where *scammaro* is erroneously mapped to *escarola* (chicory), obliterating the socio-religious meaning encoded in the term. The baseline thus operationalizes what the TAG framework is designed to address: the structural inability of unconstrained LLMs to govern their own terminological choices in specialized, culturally dense domains.

A notable asymmetry emerges at the baseline level between the two models on regulatory-administrative texts: Gemma3 4B preserves Italian denominations in 70.0% of cases, compared to 40.0% for Mistral 7B Instruct. This divergence points to a broader pattern observed throughout the evaluation: the two models respond to terminological guidance in qualitatively different ways, and their behavior cannot be reduced to a simple ranking in terms of general translation capability.

The introduction of in-context terminological injection produces consistent improvements across all genres and both models. Nevertheless, the injection condition also reveals a systematic weakness in Mistral's compliance behavior on promotional-touristic texts, where TA reaches only 50.0% against Gemma's 80.0%. This asymmetry is consistent with the findings of Lackner et al. [2025b], who observed that Mistral 7B produced unusable outputs in multilingual translation tasks.

The addition of the post-processing enforcement layer partially compensates for this limitation, raising Mistral's global TA from 70.0% to 85.0% and narrowing the gap with Gemma, which reaches 90.0% under INJ+ENF. The convergence of both models to 100.0% TA on regulatory-administrative texts under the full system condition shows that deterministic enforcement can reliably recover cases in which injection alone fails. The remaining gap on promotional-touristic texts — where Mistral reaches only 65.0% even under INJ+ENF — points to a different failure mode: cases in which the model translates the source term so freely that no recoverable surface form remains, making enforcement ineffective.

Genre	BASE	INJ	INJ+ENF
Regulatory	40.0%	80.0%	100.0%
Promotional	35.0%	50.0%	65.0%
Narrative	15.0%	80.0%	90.0%
Global	**30.0%**	**70.0%**	**85.0%**

Table 1: TA — Mistral 7B Instruct.

Genre	BASE	INJ	INJ+ENF
Regulatory	70.0%	100.0%	100.0%
Promotional	35.0%	80.0%	85.0%
Narrative	15.0%	85.0%	85.0%
Global	**40.0%**	**88.3%**	**90.0%**

Table 2: TA — Gemma3 4B.

5.2. Genre Constraint Compliance: Manual Evaluation

Manual evaluation was conducted on 120 translations produced under the INJ and INJ+ENF conditions across both models and all genres using the Genre Constraint Compliance (GCC; Table 4). GCC evaluates the quality of terminological modulation on a three-point scale: 0 indicates that no modulation is provided; 1 indicates that a modulation is present but consists of a partial or verbatim reproduction of the termbase note, resulting in an expansion that is not fully integrated into the surrounding co-text; and 2 that the modulation is contextually adapted to the passage and its co-text. For regulatory-administrative texts, the score captures typographical compliance rather than reformulation, since no modulation is required in this genre: 0 indicates missing angle quotation marks, 1 partial

Genre	Mistral INJ	Gemma INJ	Mistral INJ+ENF	Gemma INJ+ENF
Regulatory-administrative	80.0%	100.0%	100.0%	100.0%
Promotional-touristic	50.0%	80.0%	65.0%	85.0%
Narrative-cultural	80.0%	85.0%	90.0%	85.0%
Global	**70.0%**	**88.3%**	**85.0%**	**90.0%**

Table 3: TA cross-model comparison: Mistral 7B Instruct vs Gemma3 4B.

compliance, and 2 correct formatting.

The results reveal three main patterns across genres and models: consistent typographical compliance in regulatory-administrative texts, particularly for Gemma; systematic prompt leakage in Mistral across promotional and narrative genres; and a tendency in both models to extend the instructed explicitation behavior beyond the boundaries of the termbase.

Genre	Model	INJ	INJ+ENF
Regulatory	Mistral	0.90	1.40
	Gemma	2.00	2.00
Promotional	Mistral	0.10	0.90
	Gemma	0.60	0.60
Narrative	Mistral	0.00	0.00
	Gemma	0.70	0.50

Table 4: GCC by genre, model, and condition (0–2 scale).

The regulatory-administrative genre yields the clearest results. Gemma3 4B achieves a perfect GCC of 2.00 under both INJ and INJ+ENF, indicating that in every evaluated case the Italian term is correctly preserved with the required angle quotation marks. Mistral 7B reaches 0.90 under INJ and 1.40 under INJ+ENF — a pattern that reflects inconsistent typographical compliance rather than terminological failure: the term is present, but the formatting convention is not reliably applied. These results converge on the same conclusion as Section 5.1: for the most formally constrained genre in Sabatini's typology, Gemma3 4B is a more reliable partner for TAG-based translation.

The promotional-touristic genre reveals a more nuanced picture. Gemma achieves a GCC of 0.60 under both INJ and INJ+ENF, indicating that when parenthetical expansions are produced they tend toward partial rather than full contextual reformulation. Mistral's GCC of 0.10 under INJ reflects a qualitatively different failure mode: in the majority of annotated cases, expansions were either absent or constituted prompt leakage — content appended after the translation as external notes, often reproducing verbatim portions of the glossary block or the background note. This behavior is not captured by TA alone and underscores the importance of manual evaluation for a complete assessment of TAG effectiveness. A related pattern, observable in both models, concerns terminological overgeneralization: when instructed to add expansions for specific terminological units, both models occasionally extend this behavior to gastronomic terms present in the surrounding text but not covered by the termbase. This suggests that the models internalize the structural pattern of the glossary block instruction and replicate it beyond its intended scope. The phenomenon is particularly systematic in Mistral under INJ+ENF, where parenthetical expansions are added for *panzanella* (*ensalada de pan rallado, tomates, cebolla y aceituna*), *pappa al pomodoro* (*sopa de pan rallado y tomate*), and *fusilli alla 'nduja* (*fusilli con salsa de salchichón calabrés*), none of which appear in the termbase. In one particularly noteworthy case, Gemma3 4B applies a contextually integrated gloss to *'O rraù* — a dialectal Neapolitan form of *ragú napolitano* that does not appear as such in the termbase, but which the model appears to recognize as a variant of a known entry, producing *una preparación típica del almuerzo dominical napolitano, considerada un emblema del patrimonio enogastronómico de Campania*. The expansion is semantically grounded in the termbase entry for *ragú napolitano* and is positioned in a contextually motivated way, facilitating the target reader's comprehension of an otherwise opaque dialectal form. While none of these expansions are explicitly instructed by the termbase, they point to a broader tendency of instruction-following models to replicate prompted explicitation patterns beyond their intended scope — a behavior that, under favorable conditions, may produce genuinely useful output.

The narrative-cultural genre produces the sharpest contrast between the two models. Mistral achieves a GCC of 0.00 under both INJ and INJ+ENF — not because no expansions are produced, but because every expansion annotated was classified as prompt leakage or hallucination: content generated outside the translated text, disconnected from the surrounding narrative co-text, and frequently reproducing background note material verbatim or introducing extraneous information. Manual verification confirms that the preferred term is present in 8 out of 10 cases under both conditions, yet the model consistently

fails to integrate the required inline gloss, instead appending expansions as external notes that violate both the structural and the stylistic requirements of narrative prose. Gemma, by contrast, achieves the highest GCC across all genre-model-condition combinations: 0.70 under INJ. A representative case is *frittata di scammaro* (narrative, INJ), where Gemma produces an inline gloss — *un plato típico de la cocina pobre napolitana* — genuinely adapted to the surrounding narrative context, earning a GCC of 2. In one promotional-touristic case, Gemma positions the parenthetical expansion after the anaphoric *menestra ebolitana* rather than immediately after *Ciauliello* — a co-textually motivated placement that, while technically non-compliant with the glossary block instruction, produces a more natural and readable result. The decrease to 0.50 under INJ+ENF may reflect cases where the enforcement layer's term substitution disrupted the natural integration of the gloss — a tension between deterministic correction and contextually sensitive generation that points to a fundamental design challenge for TAG systems operating on loosely constrained text types.

Taken together, these results demonstrate that manual evaluation captures dimensions of TAG effectiveness that automatic TA cannot access. High TA scores are a necessary but not sufficient condition for genre-compliant terminological treatment: what ultimately determines the quality of the output is not merely whether the correct term is present, but whether it is integrated in a way that respects the communicative conventions of the target genre.

6. Conclusions and Future Work

This study has examined whether the integration of a concept-oriented bilingual termbase into an LLM-based translation workflow can effectively regulate terminological compliance and genre-specific realization in the domain of Campanian intangible cultural heritage.

The results suggest that terminology-augmented generation can improve terminological adherence across genres and models, with global TA scores reaching 85.0% for Mistral and 90.0% for Gemma under the full system condition. The combination of in-context injection and deterministic enforcement tends to outperform either component in isolation: injection alone increases adherence but fails to guarantee genre-specific realization, while enforcement cannot operate on terms that have not been properly introduced.

The findings allow us to directly address the research questions. Terminology injection proves effective in regulating terminological compliance across discursive genres (RQ1), with global TA scores increasing consistently from baseline across all genre–model combinations. Injection combined with deterministic enforcement produces uneven effects: while the full system condition consistently outperforms injection alone on TA, the added value of enforcement is strongly genre- and model-dependent — improving terminological realization on regulatory-administrative and promotional-touristic texts for Mistral, but showing limited or no benefit in others, and in some cases interfering with contextually appropriate realization (RQ2).

However, terminological adherence is only one dimension of TAG effectiveness. Manual evaluation reveals that a model can achieve high TA while systematically failing to integrate terms in a genre-appropriate way, as demonstrated by Mistral's narrative-cultural outputs, where the preferred term is present in 8 out of 10 cases yet consistently realized as an external appendix rather than an inline gloss.

The cross-model comparison highlights the central role of instruction-following capacity in TAG effectiveness. Gemma3 4B consistently outperforms Mistral 7B Instruct not only in quantitative terms but also in qualitative behavior, demonstrating a greater ability to integrate terminological modulation into the surrounding discourse. This gap becomes more pronounced in loosely constrained genres, where successful output depends on context-sensitive generation rather than formal compliance alone (RQ3).

Future work should extend the evaluation to larger datasets, additional language pairs, and more capable open-weight models, in order to enable a more systematic investigation of the relationship between model capacity and TAG effectiveness. From a terminological perspective, enabling more flexible access to different types of conceptual metadata — allowing context-driven selection rather than genre-based assignment — may improve the contextual adequacy of terminological modulation in loosely constrained text types. Finally, the development of automatic metrics for contextual reformulation remains an open challenge for scalable evaluation of terminology-aware generation in culturally sensitive translation settings. In this regard, further investigation is needed into how standard MT evaluation metrics, such as BLEU [Papineni et al., 2002] and COMET [Rei et al., 2020], behave under controlled, terminology-augmented generation conditions, where improvements in terminological compliance and genre-sensitive realization may not be fully captured by reference-based similarity.

7. Bibliographical References

L. Buccheri. *Parole del cibo in Campania. Cento voci del lessico gastronomico regionale*. Franco Cesati Editore, Firenze, 2023.

M.-T. Cabré. *La terminología. Teoría, metodología y aplicaciones*. Antártida/Empúries, Barcelona, 1993.

M.-T. Cabré. Theories of terminology: Their description, prescription and explanation. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 9(2):163–199, 2003.

F. Chessa, C. De Giovanni, and M. T. Zanola. *La terminologia dell'agroalimentare*. FrancoAngeli, Milano, 2014.

G. M. Di Nunzio. Terminology-augmented generation (TAG): Foundations, use cases, and evaluation paths. *Journal of Digital Terminology and Lexicography*, 1(1):97–104, 2025.

K. Fleischmann. Terminologiemanagement: Die Schlüsselkomponente für effiziente Kommunikation in Unternehmen. *Information – Wissenschaft & Praxis*, 76(4):169–176, 2025.

C. Grimaldi. *Il prodotto agroalimentare campano. Tra lingua, cultura e tradizione*. Aracne, Roma, 2017.

S. Gupta, R. Ranjan, and S. N. Singh. A comprehensive survey of retrieval-augmented generation (RAG): Evolution, current landscape and future directions. *arXiv preprint arXiv:2410.12837*, 2024.

A. Lackner, A. Vega-Wilson, and C. Lang. Terminology-augmented generation: A systematic review of terminology formats for in-context learning in LLMs. In *Proceedings of the 4th International Conference on Multilingual Digital Terminology Today (MDTT 2025)*, volume 3990 of *CEUR Workshop Proceedings*, 2025a.

A. Lackner, A. Vega-Wilson, and C. Lang. An evaluation of terminology-augmented generation (TAG) and various terminology formats for the translation use case. *Journal of Digital Terminology and Lexicography*, 1(2):31–47, 2025b.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Roktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

P. Martín-Chozas, P. Calleja, and C. R. Limón. Terminology Enhanced Retrieval Augmented Generation for Spanish Legal Corpora. In *Proceedings of the 5th Conference on Language, Data and Knowledge*, pages 147–152, 2025.

L. Molina and A. Hurtado Albir. Translation techniques revisited: A dynamic and functionalist approach. *Meta: Translators' Journal*, 47(4):498–512, 2002.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Regione Campania. Prodotti tradizionali – elenco dei prodotti agroalimentari tradizionali. https://agricoltura.regione.campania.it/tipici/prodotti_tradizionali.htm, 2025.

R. Rei, C. Stewart, A. C. Farinha, and A. Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2685–2702, 2020.

F. Sabatini. Analisi del linguaggio giuridico: Il testo normativo in una tipologia generale dei testi. In M. D'Antonio, editor, *Corso di studi superiori legislativi (1988–1989)*. Cedam, Padova, 1990.

F. Sabatini. "Rigidità-esplicitezza" vs "elasticità-implicitezza": Possibili parametri massimi per una tipologia dei testi. In G. Skytte and F. Sabatini, editors, *Linguistica testuale comparativa*. Museum Tusculanum Press, Copenaghen, 1999.

UNESCO. *Convention for the Safeguarding of the Intangible Cultural Heritage*. UNESCO, Paris, 2003.

A. System Prompt

The following system prompt was shared across all models and conditions (BASELINE, INJ, INJ+ENF):

```
You are translating from Italian into Spanish texts belonging to
the domain of traditional Campanian gastronomy.

##ROLE AND OBJECTIVE##
You act as a professional translator specialized in traditional
gastronomy and cultural heritage (IT->ES). Each term encodes
production processes, historical memory, ritual practices, and
territorial identity, and cannot be reduced to a simple technical
label. Your objective is to produce a complete and faithful
translation that preserves the conceptual integrity of culturally
embedded terms and reads naturally in Spanish.

##TRANSLATION INSTRUCTIONS##
- Translate the ENTIRE text from Italian into Spanish.
- Preserve the tone, register, and voice of the original.
- When terminological instructions are provided, follow them strictly.
- When an expansion or gloss is required, integrate it inline
  within the sentence, adapting its formulation to the co-text.

##REQUIRED OUTPUT##
Return ONLY the final translation in Spanish, complete and ready
for use, with no intercalated comments.

##QUALITY CHECK BEFORE DELIVERING##
- You have translated the entire text with no omissions.
- You have strictly followed the terminological instructions.
- The translation reads naturally in Spanish.
```

B. Glossary Block

The following examples illustrate the genre-specific glossary blocks injected into the user message prior to generation, constructed dynamically from the terminological policies retrieved via the MCP server. No glossary block was injected under the BASELINE condition; the system prompt in Appendix A was the only instruction provided to the model.

Regulatory-administrative

```
##TERMINOLOGICAL INSTRUCTIONS##
The text belongs to the regulatory-administrative discursive genre.
Preserve the following Italian denominations unchanged and format
them with Spanish angle quotation marks. No expansions needed.

- '[IT term]': keep in Italian, write exactly:
  «[IT term]»
```

Promotional-touristic

```
##TERMINOLOGICAL INSTRUCTIONS##
The text belongs to the promotional-touristic discursive genre.
For each term, use the Spanish equivalent immediately followed
by a parenthetical expansion inline: term (expansion).
Reformulate the background note based on the context and co-text.

- '[IT term]': translate as '[ES equivalent] (your expansion here)'
  Background note to reformulate: [term-specific production note
retrieved from the termbase at runtime]
```

Narrative-cultural

```
##TERMINOLOGICAL INSTRUCTIONS##
The text belongs to the narrative-cultural discursive genre.
For each term, use the Spanish equivalent immediately followed
```

```
by an inline cultural gloss introduced by a comma: term, gloss,
Reformulate the background note based on the context and co-text.

- '[IT term]': translate as '[ES equivalent], your gloss here,'
  Background note to reformulate: [term-specific cultural note
retrieved from the termbase at runtime]
```

Assessing Small Language Models as Text Simplification Evaluators

David Carranza Navarrete, Jan Bakker, Jaap Kamps
Institute for Logic, Language and Computation (ILLC)
University of Amsterdam
Amsterdam, The Netherlands
david.carranza.navarrete@student.uva.nl, j.bakker@uva.nl, kamps@uva.nl

Abstract

Text simplification requires reliable automatic evaluation, yet existing learnable metrics such as LENS and LENS-SALSA are specialized and costly to develop. Moreover, it remains unclear how these metrics compare to using large language models (LLMs) as evaluators. Exploring this question is important because LLM-based evaluation could make simplification research and deployment more flexible and easier to adapt than training new task-specific metrics for each setting. In this work, we empirically compare several small, open-weight instruction-tuned LLMs with LENS and LENS-SALSA in both reference-based and reference-free evaluation settings. We measure their alignment with human judgments across multiple datasets. Our results provide insight into when small LLMs can serve as effective evaluators and when specialized metrics remain preferable, informing the design of future evaluation pipelines for text simplification and related text generation tasks.

Lay Summary: *The evaluation of text simplification output remains a great challenge in terms of building reusable evaluation corpora and evaluation measures. Traditional reference-based evaluations are imprecise as references only cover one or a few possible simplifications. Learned measures still require extensive labeled data for training, and may not generalize to new domains. LLM-based evaluation presents a pragmatic alternative in case no extensive references are available. In this paper, we systematically compare these evaluation approaches against human ratings.*

Keywords: LENS, Automatic Evaluation, LLM-as-a-Judge

1. Introduction

Text simplification aims to make content more accessible while preserving its original meaning. Reliable evaluation of simplification quality is therefore essential for both system development and deployment. Traditionally, evaluation relies on human judgments, which are expensive and time-consuming. To address this, automatic metrics such as LENS (Maddela et al., 2023) and its reference-free variant LENS-SALSA (Heineman et al., 2023) have been proposed as learnable metrics for text simplification. These metrics attempt to approximate human judgments using supervised models trained on annotated data.

Recent advances in large language models (LLMs) have used generative models as evaluators (Gao et al., 2025). Instead of relying on task-specific learned metrics, LLMs are prompted to act as judges and directly assign quality scores. This raises an important question: how well do relatively small instruction-tuned LLMs perform as judges of simplification quality compared to specialized learnable metrics like LENS?

In this work, we investigate whether small, open-weight LLMs can serve as reliable simplification judges. We focus on three popular open-source models: Microsoft's `Phi-3-mini-4k-Instruct`, Alibaba Cloud's `Qwen2.5-7B-Instruct`, and Meta's `Llama-3.1-8B-Instruct`. Specifically, we investigate LLM's performance against LENS in both reference-based and reference-free settings. We evaluate: (1) LENS-SALSA (metric without references) against LLM judges without references and (2) LENS (metric with references) against LLM judges with access to references. Finally, we compute correlations between LLM-assigned scores and human judgments to assess alignment with human evaluation.

2. LLMs as Judges

We investigate whether small instruction-tuned large language models (LLMs) can serve as automatic judges of text simplification quality. Unlike supervised evaluation metrics that are explicitly trained for simplification assessment, these models perform evaluation through prompting at inference time. Specifically, the models are asked to assess key aspects of simplification quality, including fluency, meaning preservation, and simplicity.

We experiment with the following open-weight instruction-tuned LLMs:

- `Phi-3-mini-4k-Instruct`[1]

[1] https://huggingface.co/microsoft/Phi-3-mini-4k-instruct

Proceedings of the 2nd Workshop on Evaluating Text Difficulty in a Multilingual Context (DeTermIt! 2026), pages 83–86

- `Qwen2.5-7B-Instruct`[2]
- `Llama-3.1-8B-Instruct`[3]

These models range from approximately 3.8B to 8B parameters and were selected to represent compact open-weight models that remain practical for research and deployment settings.

3. Methodology

Non-learnable evaluation metrics are commonly used to assess text without additional training. For example, BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) measures n-gram overlap between candidate and reference texts, focusing on precision but lacking semantic sensitivity. SARI (System Output Against References and against the Input sentence) (Xu et al., 2016) evaluates the quality of added, deleted, and retained words, better capturing meaning preservation and simplicity. BERTScore (Zhang et al., 2019) uses contextual embeddings to measure semantic similarity, making it more robust to paraphrasing. Readability metrics such as FKGL (Flesch–Kincaid readability tests) (Kincaid et al., 1975) estimate the education level required to understand the text.

In this work, we use two evaluation settings to compare LLM judges with existing learnable and non-learnable automatic metrics. We use the LENS[4] and LENS-SALSA[5] checkpoints released on HuggingFace to evaluate simplification performance using the top-k outputs, reporting results for $k = 3$. We use the corresponding code[6] for our experiments and present our results in the same manner as Maddela et al. (2023), additionally including results from our LLM-based models.

Reference-free. In the reference-free setting, the model receives the original complex sentence and the simplified candidate. The model is prompted to assess the quality of the simplification by assigning scores based on three criteria: meaning preservation, fluency, and simplicity. This setup mirrors the reference-free evaluation paradigm used by metrics such as LENS-SALSA.

Reference-based. In the reference-based setting, the model receives the original complex sentence, the simplified candidate, and one or more human-written reference simplifications. The model is asked to evaluate the quality of the candidate simplification while considering the reference simplifications as guidance for expected outputs. This setup corresponds to traditional reference-based evaluation used by metrics such as LENS.

3.1. Scoring and Correlation

Each setting produces scalar quality scores. We compute Pearson correlation for datasets with continuous human ratings and Kendall τ-like correlation for ranking-based datasets. Alignment with human judgments is used as the primary measure of evaluation quality.

4. Experiments

4.1. Datasets

We evaluate across three benchmark datasets:

- **SimpEval**$_{2022}$: Human rankings of simplifications evaluated using a Kendall τ-like correlation. (Maddela et al., 2023)
- **Wiki-DA**: Direct assessment scores for fluency, meaning, and simplicity. (Alva-Manchego et al., 2021)
- **Newsela-LIKERT**: Human Likert-scale ratings across grammaticality, meaning preservation, and simplicity. (Maddela et al., 2021)

These datasets cover both ranking-based and direct scoring evaluation paradigms.

4.2. Evaluation Settings

We compare:

1. Reference-free LLM judges vs LENS-SALSA.
2. Reference-based LLM judges vs LENS.

All correlations are computed against human ratings provided in the respective datasets.

5. Results

Table 1 presents the correlation between automatic metrics and human judgments in the reference-based setting. LENS achieves the highest correlations overall, particularly on Wiki-DA and Newsela-LIKERT. However, several LLM judges obtain competitive results on SimpEval and Wiki-DA, suggesting that instruction-tuned models can approximate dedicated evaluation metrics when provided with references.

Table 2 reports the reference-free evaluation results. LENS-SALSA shows strong performance

[2] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[3] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[4] https://huggingface.co/davidheineman/lens

[5] https://huggingface.co/davidheineman/lens-salsa

[6] https://github.com/Yao-Dou/LENS

	SimpEval$_{2022}$			Wiki-DA			Newsela-Likert		
Metric / Model	Para	Split	All	Fluency	Meaning	Simplicity	Fluency	Meaning	Simplicity
FKGL	-0.397	-0.318	-0.331	0.084	0.185	0.037	0.169	0.293	-0.053
BLEU	0.048	-0.054	-0.033	0.460	0.622	0.438	0.333	0.261	0.121
SARI	0.206	0.140	0.149	0.335	0.534	0.366	0.234	0.122	0.101
BERTScore	0.238	0.085	0.106	0.642	0.699	0.622	0.389	0.295	0.206
LENS (*k*=3)	**0.429**	0.333	0.331	**0.807**	0.660	**0.750**	**0.621**	**0.431**	**0.362**
LLM Judges (with references)									
Qwen2.5-7B-Instruct	**0.818**	**0.358**	**0.457**	0.630	0.716	0.680	0.458	0.393	0.265
Llama-3.1-8B-Instruct	0.333	0.320	0.358	0.618	**0.761**	0.639	0.379	0.313	0.208
Phi-3-mini-4k-Instruct	0.130	0.347	0.325	0.605	0.689	0.639	0.506	0.371	0.259

Table 1: Comparison between traditional automatic metrics from the LENS paper and LLM judges (with references). Pearson correlations with human ratings are reported. Higher values indicate better alignment with human evaluation. Best results are in bold and second best are underlined.

	SimpEval$_{2022}$			Wiki-DA			Newsela-Likert		
Metric / LLM	Para	Split	All	Fluency	Meaning	Simplicity	Fluency	Meaning	Simplicity
LENS-SALSA	0.263	0.212	0.229	0.701	0.676	0.640	0.497	0.356	0.284
Qwen2.5-7B-Instruct	-0.353	0.117	-0.028	**0.648**	**0.802**	**0.682**	**0.510**	**0.572**	**0.272**
Llama-3.1-8B-Instruct	**0.333**	**0.500**	**0.083**	0.586	0.740	0.620	0.428	0.520	0.233
Phi-3-mini-4k-instruct (**)	0.000	0.000	0.000	0.021	0.026	0.051	0.000	0.000	0.000

Table 2: Correlation between automatic evaluators and human judgments across simplification datasets. LENS-SALSA is a dedicated reference-free metric. LLM judges operate without references. Best values are in bold; second best are underlined.

across datasets, while LLM judges demonstrate varying levels of correlation with human judgments. These findings suggest that LLM-based evaluation may provide a flexible alternative to specialized metrics, although performance depends on both the dataset and evaluation dimension.

6. Conclusion

In this work, we investigated whether small instruction-tuned large language models can serve as automatic evaluators for text simplification. We compared three open-weight LLM judges with dedicated simplification evaluation metrics, including LENS and its reference-free variant LENS-SALSA, across three benchmark datasets.

The experiments show that LLM judges can achieve moderate correlations with human judgments and, in some cases, approach the performance of traditional metrics, particularly in reference-based evaluation settings. However, specialized metrics such as LENS and LENS-SALSA remain more consistent and generally achieve stronger correlations with human evaluation.

These findings suggest that while small LLMs can provide flexible and lightweight evaluation signals, dedicated metrics still offer advantages in reliability and stability. Future work could explore improved prompting strategies, calibration methods, or hybrid approaches that combine LLM-based evaluation with learned metrics to further improve automatic evaluation for text simplification.

7. Limitations and future work

This study focuses on sentence-level simplification in English but does not address document-level evaluation or other languages. Additionally, we restrict our analysis to relatively small open-weight LLMs and do not compare against larger proprietary models. Future work could explore strategies for LLM judges, pairwise preference prompting, or training smaller evaluation models using LLM-generated supervision.

Acknowledgments

Jan Bakker and Jaap Kamps are supported by the Netherlands Organization for Scientific Research (NWO NWA # 1518.22.105). Jaap Kamps is also supported by the University of Amsterdam (AI4FinTech program) and ICAI (AI for Open Government Lab). Views expressed in this paper are

not necessarily shared or endorsed by those funding the research.

8. Bibliographical References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. LLM-based NLG evaluation: Current status and challenges. *Computational Linguistics*, 51:661–687.

David Heineman, Yao Dou, Mounica Maddela, and Wei Xu. 2023. Dancing between success and failure: Edit-level simplification evaluation using SALSA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3466–3495, Singapore. Association for Computational Linguistics.

Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Research Branch Report 8-75.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553, Online. Association for Computational Linguistics.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics (TACL)*, 4:401–415.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A. Appendices

A.1. LLM Evaluation Prompts

The following prompts were used to instruct the language models to evaluate the quality of simplified sentences. The Reference-Free prompt is based on the instructions given to human annotators to rate simplifications by (Maddela et al., 2023). The Reference-Based makes use of the simplification, the original text, and the reference to grade.

LLM Evaluation Prompt (Reference-Free)

```
You are an evaluator specialized in text simplification that
scores how well a simplified sentence improves on an original
sentence.
Assign an integer score from 0–100:
100 – Fully simplified, fluent, preserves core meaning
75 – Somewhat simpler, mostly fluent, meaning close
50 – Simpler, somewhat fluent, meaning similar
25 – Equally simple, some fluency, meaning lost
0 – Completely unreadable
Higher scores indicate better meaning preservation, fluency,
and simpler wording.
Return:
1) "score: X" (0–100)
2) Short explanation (1–2 sentences)
Original: """<original>"""
Simplified: """<simplified>"""
```

LLM Evaluation Prompt (Reference-Based)

```
You are an evaluator for text simplification. Rate how well
the simplification improves on the original.
When a REFERENCE is provided, treat it as a gold-standard
simplification and compare the candidate to both the original
and the reference.
Scoring rules:
100 – Fully simplified, fluent, preserves core meaning and
aligns with reference
75 – Somewhat simpler, mostly fluent, meaning close
50 – Simpler, somewhat fluent, meaning similar
25 – Equally simple, some fluency, meaning lost
0 – Completely unreadable
Return:
1) "score: X" (0–100)
2) Short explanation
Original: """<original>"""
Reference: """<reference>"""
Simplified: """<simplified>"""
```

Author Index