Progressive Vocabulary Learning via Pareto-Optimal Clustering

Deepika Verma
Dept. of Artificial Intelligence
Kyungpook National University
Daegu, South Korea
dvdeepika07@gmail.com

Daison Darlan
Dept. of Artificial Intelligence
Kyungpook National University
Daegu, South Korea
daisondarlan33@gmail.com

Rammohan Mallipeddi
Dept. of Artificial Intelligence
Kyungpook National University
Daegu, South Korea
mallipeddi.ram@gmail.com

***Abstract*—Vocabulary acquisition is a foundational component of language learning and is most effective when learners engage with texts that gradually increase in lexical complexity. However, most existing text simplification systems produce only a single simplified version, lacking support for controlled progression in vocabulary difficulty. To address this gap, we propose a framework that generates multiple lexically varied yet semantically accurate text variants from a given text and organizes them into structured reading levels. Our method first creates a Pareto front of optimal variants by balancing two competing objectives: improving readability, measured by the Flesch–Kincaid Grade Level (FKGL), and preserving semantic integrity by limiting the number of word substitutions. The generated variants are organized into three progressively challenging levels, *Beginner*, *Intermediate*, and *Advanced*, using a clustering process that groups texts with similar vocabulary and complexity. Experimental results across diverse inputs show that our approach consistently produces well-stratified, interpretable text variants, supporting personalized and incremental vocabulary development for language learners.**

***Index Terms*—progressive learning, readability optimization, pareto clustering.**

I. Introduction

Vocabulary acquisition is a core pillar of language learning, directly influencing learners' ability to comprehend written text, articulate thoughts, and engage in meaningful communication [1]. Unlike grammatical instruction, which can often be rule-based and explicit, vocabulary learning typically requires repeated and contextual exposure to a broad range of lexical items. Research in second language acquisition, including the comprehensible input hypothesis [2], emphasizes the importance of presenting learners with materials that are slightly beyond their current proficiency. This scaffolding approach allows learners to infer meaning through context while gradually expanding their vocabulary.

Despite this pedagogical insight, existing systems that deliver simplified or adapted reading content are often limited in flexibility and scalability. Traditional approaches rely on handcrafted simplification rules or static readability thresholds, which are not easily tailored to individual learners' progress or lexical gaps [3]. More recently, neural text simplification systems generate single simplified versions of input texts, but they typically perform sentence-level rewriting, which may simultaneously alter vocabulary, syntax, and discourse structure, which may potentially overwhelm learners and hinder comprehension.

To address these limitations, we propose a novel framework that enables progressive, vocabulary-level control over text complexity. The goal is to generate multiple semantically consistent variants of a given paragraph by substituting words with contextually appropriate synonyms. These variants differ in lexical complexity while retaining the syntactic and semantic structure of the original paragraph, allowing learners to focus on mastering a few new words at a time.

Each variant is evaluated using two complementary criteria: (i) the Flesch–Kincaid Grade Level (FKGL) [4] to assess readability, and (ii) the number of word substitutions, which serves as a proxy for semantic preservation. Using non-dominated sorting [5], we construct a Pareto front of optimal text variants that balance simplicity with lexical fidelity. To support structured vocabulary learning, these Pareto-optimal variants are further grouped into distinct levels of lexical complexity using unsupervised clustering. The number of levels can be adjusted based on the length or instructional goals of the input text. In our experiments, we organize the variants into three levels to demonstrate the effectiveness of the approach.

This framework is designed to be language-agnostic; while our experimental study focuses on English due to the availability of large synonym databases such as WordNet, the approach can be extended to any language with sufficient lexical resources. By limiting changes to the word level, our method preserves sentence-level fluency and semantic intent, offering interpretable, scalable, and pedagogically grounded reading material.

Our framework offers a scalable and linguistically grounded solution for vocabulary progression through structured lexical variation. Our experimental results show that the framework consistently produces well-stratified text variants across input paragraphs of different lengths and styles, laying the foundation for adaptive language learning systems that scale with learner proficiency.

II. Related Work

Vocabulary acquisition in second language learning is widely recognized as more effective when new lexical items

are introduced incrementally, allowing learners to encounter unfamiliar words within familiar syntactic and semantic contexts [2], [6]. Rather than relying on rote memorization or abrupt shifts in text complexity, pedagogical theories emphasize gradual exposure to increasingly complex vocabulary while maintaining grammatical consistency [1]. This motivation leads to frameworks for progressive readability optimization, where lexical difficulty is systematically adjusted to align with the evolving proficiency of the learner. Despite its educational significance, the generation of multiple lexically controlled text versions to support staged vocabulary learning remains relatively underexplored in computational research.

Much of the foundational work in this space falls under the broader domain of text simplification. Early approaches apply rule-based transformations, including syntactic restructuring, clause splitting, and lexical substitution [3]. While interpretable, these methods are often constrained by domain specificity and limited scalability. Recent advancements leverage neural architectures, particularly transformer-based models [7], [8], that are fine-tuned on large simplification datasets such as WikiLarge [9] and Newsela [10]. These models demonstrate fluency in producing simplified outputs, but typically optimize toward a single rewritten version per input.

To introduce more explicit readability control, some works use reinforcement learning to steer generation toward target reading levels [11], [12], while others train models on datasets annotated with grade-level labels [13]. Additionally, multi-objective optimization techniques explore joint improvement of readability and semantic fidelity [14]. However, these systems often generate full-sentence rewrites that simultaneously modify vocabulary, syntax, and discourse structure. While suitable for general-purpose simplification, this approach may overwhelm language learners who benefit from isolating small, controlled vocabulary changes.

In the context of language education, maintaining sentence structure while varying lexical content is especially important for minimizing cognitive load and supporting contextual learning. Lexical substitution has a long history in NLP, with seminal work by McCarthy [15] framing it within word-sense disambiguation, and WordNet [16] serving as a foundational synonym resource. These tools provide a basis for targeted lexical manipulation without altering sentence syntax or discourse-level semantics.

Emerging work also recognizes that excessive rewriting can compromise both meaning and instructional utility. Recent studies highlight that even state-of-the-art simplification models may introduce semantic distortions that affect downstream model predictions [17], [18], reinforcing the need for interpretable and controllable simplification pipelines. To our knowledge, few systems actively generate multiple difficulty-aligned versions of the same input paragraph for the explicit purpose of progressive vocabulary learning. Our work addresses this gap by producing structured lexical variants that vary in complexity while maintaining syntactic consistency, allowing learners to incrementally build vocabulary within a familiar textual scaffold.

III. Background

A central aspect of our work involves estimating and managing textual difficulty through readability metrics. Readability refers to how easily a reader can understand a given text, influenced by syntactic complexity, vocabulary, and sentence structure [19], [20]. Among various metrics, the Flesch–Kincaid Grade Level (FKGL) remains one of the most widely adopted in English-language educational materials due to its interpretability and pedagogical relevance. FKGL combines sentence length and syllable count to estimate the school grade level required to comprehend a passage [4]. Its robustness has been supported by decades of empirical use and recent theoretical analyses, which show that its reliance on word count and syllables contributes to its stability and generalizability [21]. This makes FKGL especially suitable for designing level-aligned learning materials that introduce vocabulary progressively. In our work, we adopt FKGL, described in 1, as one of the scoring functions to guide the selection of optimal text variants.

$$\text{FKGL} = 0.39\left(\frac{W}{S}\right) + 11.8\left(\frac{SY}{W}\right) - 15.59 \qquad (1)$$

where W is the total number of words, S is the total number of sentences, and SY is the total number of syllables.

To organize the generated text variants based on their lexical characteristics, we employ unsupervised clustering. Each variant, generated through controlled synonym replacement, is first vectorized using Term Frequency–Inverse Document Frequency (TF-IDF) [22], a commonly used approach for representing textual data based on word importance. To enable efficient clustering and visualization, we then apply Principal Component Analysis (PCA) [23] to project the high-dimensional TF-IDF vectors into a two-dimensional space. Finally, k-Means clustering [24] is used to group the variants into three clusters, which are interpreted as *Beginner*, *Intermediate*, and *Advanced* levels based on their lexical similarity and complexity. Each cluster is annotated with its average FKGL score and the number of word replacements to support structured vocabulary-level progression.

While clustering has been widely used in educational data mining [25] and text similarity tasks [26], its application for grouping automatically generated text variants into readability-stratified learning levels is novel. Our framework addresses this gap by generating multiple semantically consistent variants through synonym replacement using the WordNet [27] lexical database. By constraining the number of replacements, we maintain semantic integrity and produce lexically varied content without altering sentence structure. This enables learners to engage with progressively more challenging forms of the same text, supporting incremental vocabulary acquisition with minimal cognitive overload.

IV. Methodology

This section presents a framework for generating lexically controlled text variants to support progressive vocabulary

learning. A synonym replacement module uses WordNet [27] to generate semantically consistent variants with varying lexical complexity. Each is scored by FKGL for readability and by word replacement count as a proxy for semantic preservation. Pareto-optimal variants are selected via non-dominated sorting to balance simplicity and fidelity. These are then clustered using TF-IDF, PCA, and k-Means into three difficulty levels. Fig. 1 illustrates the overall pipeline.

A. Problem Formulation

Given an input paragraph T consisting of n tokens, the objective is to generate a set of lexically varied and semantically consistent text variants T' that support progressive vocabulary learning by varying in lexical complexity. This is formulated as a bi-objective optimization problem, where each candidate variant T' aims to optimize the following:

- Minimize $f_1(T')$: the Flesch–Kincaid Grade Level (FKGL) score, encouraging simpler and more readable text.
- Minimize $f_2(T')$: the number of word substitutions relative to the original paragraph, serving as a proxy for semantic preservation.

Each solution $x \in \{0,1\}^n$ is encoded as a binary vector indicating word-level replacement decisions across the n eligible tokens in T. Synonym substitutions are drawn from the WordNet lexical database, excluding function words and stopwords. The optimization process searches for Pareto-optimal solutions that balance readability improvement with minimal semantic drift. The resulting set of variants forms the basis for downstream clustering and level-based organization.

B. Pareto Front Generation

To explore a wide range of lexically varied text variants, we generate multiple rewritten versions of the input paragraph by selectively replacing eligible words with synonyms of varying complexity. Each variant is scored using two criteria: its FKGL readability score and the number of words replaced relative to the original text. These objectives represent a trade-off between text simplicity and semantic fidelity.

After generating the complete set of text variants, we apply non-dominated sorting to identify the Pareto-optimal subset (as shown in Fig 1) - those variants for which no other variant simultaneously achieves a lower FKGL score and fewer word substitutions. These Pareto-optimal texts represent the best trade-offs between readability and semantic preservation. The resulting Pareto front captures a spectrum of high-quality solutions that balance improved readability with minimal semantic alteration. These optimal variants serve as the input for the clustering stage, where they are grouped into progressively difficult levels for vocabulary learning.

C. Clustering of Text Variants

To facilitate personalized vocabulary learning, we organize the Pareto-optimal text variants into three distinct levels of lexical complexity—*Beginner*, *Intermediate*, and *Advanced*. While our experiments use three levels, the number of levels can be flexibly adjusted by changing the number of clusters in the pipeline. These levels reflect gradual increases in vocabulary difficulty while maintaining the semantic content of the original paragraph. The stratification is achieved through an unsupervised clustering pipeline that groups lexically similar variants based on their surface features and degree of word replacement.

Let $\{V_1, V_2, \ldots, V_n\}$ denote the set of text variants selected from the Pareto front. Each variant V_i is transformed into a high-dimensional lexical representation using TF-IDF, which emphasizes words that are important within a document but rare across other documents (generated text variants). The TF-IDF weight for a term t in document d is computed as:

$$\text{TF-IDF}(t,d) = \text{tf}(t,d) \times \log\left(\frac{N}{\text{df}(t)}\right) \tag{2}$$

where $\text{tf}(t,d)$ is the frequency of term t in document d, N is the total number of documents (variants), and $\text{df}(t)$ is the number of documents in which term t appears. This produces a document-term matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$, where n is the number of variants and m is the vocabulary size.

Since $\mathbf{X}$ is typically high-dimensional and sparse, we apply Principal Component Analysis (PCA) to reduce it to a two-dimensional space for both visualization and clustering. The transformation is defined as:

$$\mathbf{X}_{\text{PCA}} = \mathbf{X}\mathbf{W} \tag{3}$$

where $\mathbf{W} \in \mathbb{R}^{m \times 2}$ contains the top two principal components of $\mathbf{X}$. The resulting matrix $\mathbf{X}_{\text{PCA}} \in \mathbb{R}^{n \times 2}$ captures the dominant variance in lexical patterns across variants.

We then apply the k-Means clustering algorithm to partition the variants into $k = 3$ clusters, $\{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3\}$, each corresponding to a different level of vocabulary difficulty. The clustering objective minimizes the within-cluster sum of squared distances:

$$\arg\min_{\mathcal{C}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathcal{C}_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 \tag{4}$$

where $\boldsymbol{\mu}_i$ is the centroid of cluster $\mathcal{C}_i$, and $\|\cdot\|$ denotes the Euclidean norm.

D. Cluster Analysis and Representative Selection

After clustering, each group is analyzed to determine its relative difficulty level based on two key metrics: (i) FKGL readability score and (ii) number of word substitutions. Clusters with lower FKGL scores and a higher number of substitutions are labeled as *Beginner*, reflecting simpler lexical constructions and more extensive synonym replacement. In contrast, clusters with higher FKGL scores and fewer substitutions are designated as *Intermediate* or *Advanced*, indicating greater syntactic and lexical complexity with minimal deviation from the original text. This structured grouping allows learners to progress through increasingly challenging variants of the same content, enabling gradual vocabulary acquisition while maintaining semantic consistency.

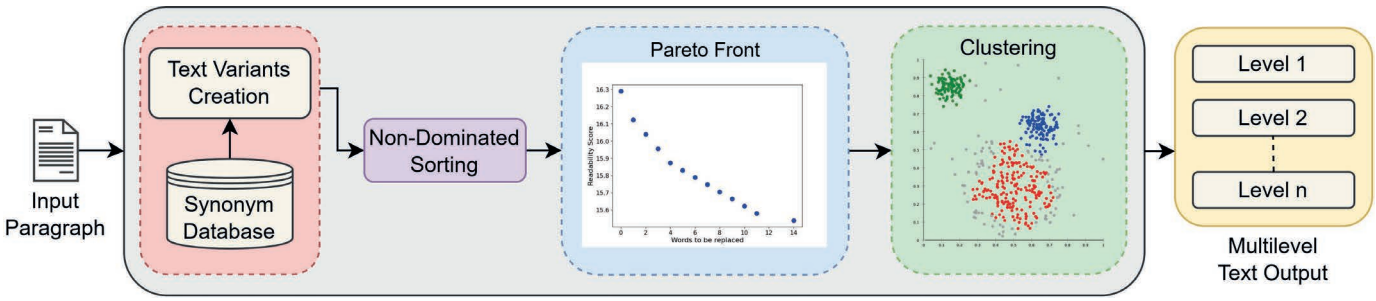


Fig. 1: Overview of the proposed framework for progressive readability optimization and vocabulary-level controlled text generation.

To make this system practical and scalable for learners, we select a single representative variant from each cluster. Specifically, we compute the centroid of each cluster in the reduced vector space obtained via PCA and identify the variant closest to this centroid using Euclidean distance. This ensures that the chosen representative for each level is both lexically central and distinct within its cluster, making it a reliable exemplar for instructional use.

V. Experimental Study

This section presents an empirical evaluation of our proposed framework, focusing on how effectively it generates lexically controlled variants and organizes them into progressive levels for vocabulary learning.

A. Experimental Setup and Metrics

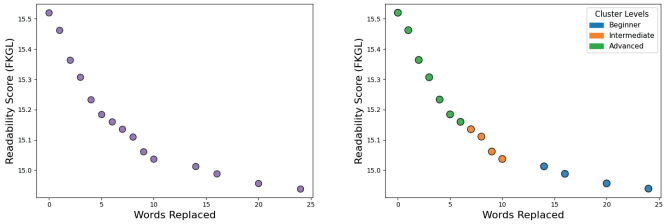
To evaluate our framework, we selected 10 paragraphs from publicly available English reading comprehension texts. The topics of these passages are chosen to reflect those most commonly found in English language learning materials, based on an empirical analysis of ELT coursebooks [28]. Each paragraph was processed using our synonym-based variant generator, producing a diverse set of candidate texts. We then applied non-dominated sorting based on two criteria: (i) the FKGL readability score, and (ii) the number of words replaced, which acts as a proxy for semantic fidelity.

All results shown are for a representative input paragraph. However, our method was evaluated across 10 different English paragraphs from various topics, and the trends reported here remained consistent. All experiments were performed on a standard desktop system running Microsoft Windows 11 (64-bit), equipped with an AMD Ryzen 5 7500F 6-core processor at 3.7 GHz and 16 GB of RAM.

The resulting Pareto-optimal variants were clustered using TF-IDF vectorization, followed by dimensionality reduction via PCA and k-Means clustering ($k = 3$). The three clusters are interpreted as *Beginner*, *Intermediate*, and *Advanced* levels based on their average FKGL and word replacement statistics.

B. Results and Discussion

Figure 2a shows the initial Pareto front, where each point represents a text variant, plotted by its FKGL readability score



(a) (b)

Fig. 2: This figure shows (a) the Pareto front of FKGL versus the number of words replaced and (b) the clustered Pareto front, with each color representing a reading level.

and number of words replaced. Figure 2b plot reveals a trade-off: lower FKGL scores tend to correspond with higher lexical substitution, and vice versa. This validates the use of these two metrics as competing objectives for structuring vocabulary learning paths using non-dominated sorting. This inverse relationship supports the design of our optimization objective, where readability and lexical change act as competing yet complementary criteria for guiding variant generation.

Importantly, incorporating FKGL as an explicit objective also helps constrain the search space of potential text variants. Without this constraint, synonym replacement can lead to a combinatorial explosion of possible outputs, many of which are either redundant or fail to provide meaningful variation in complexity. By evaluating readability alongside substitution count, our method focuses the generation process on variants that are not only lexically distinct but also pedagogically useful.

To further analyze the structure and quality of the clusters, we apply Principal Component Analysis (PCA) to reduce the high-dimensional TF-IDF vectors to two dimensions for visualization. As shown in Figure 3, each point represents a text variant, color-coded by its cluster. The clusters appear well-separated, indicating distinct lexical patterns among groups. Notably, we observe a strong correlation between cluster membership and the number of words replaced—variants with similar substitution patterns consistently fall into the same cluster. This reinforces our hypothesis that lexical substitutions directly influence reading difficulty and validates clustering

as an effective mechanism for organizing text variants by complexity. The observed progression across clusters supports incremental vocabulary learning, where learners engage with increasingly complex vocabulary in a structured, cognitively manageable sequence.

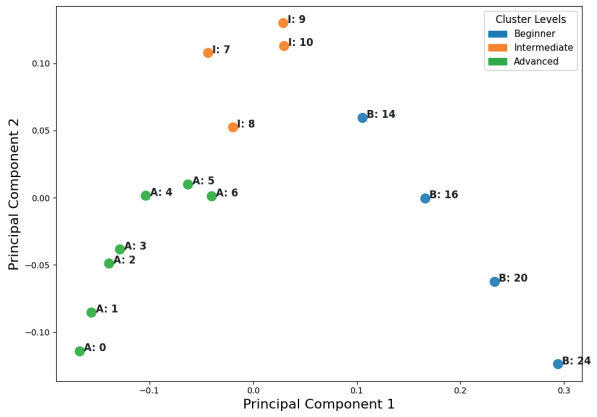


Fig. 3: PCA projection of TF-IDF vectors clustered using k-Means. Points are labeled with cluster ID and words replaced.

Within each cluster, we identify a representative text by computing the centroid and selecting the variant closest to it in vector space. This ensures that the selected texts are prototypical of their respective difficulty levels. Table I summarizes the average FKGL and number of words replaced for each cluster. The trend aligns with expectations: *Beginner* variants exhibit lower FKGL and higher vocabulary transformation, while *Advanced* variants remain closer to the original text.

TABLE I: Cluster-wise Summary of Readability and Lexical Changes

Level	Avg. FKGL	Avg. Words Replaced
Beginner	14.97	18.50
Intermediate	15.09	8.50
Advanced	15.32	3.00

To assess robustness, we repeated the process in 10 diverse paragraphs. In each case, we observed similarly well-formed Pareto fronts and consistent clustering structures. The cluster-wise FKGL and substitution patterns followed the same progressive trend. While figures in this section show results from a single paragraph for space efficiency, the method demonstrates strong generalization across input types. These experiments validate our hypothesis that readability and lexical substitution can be jointly optimized to generate structured sets of text variants. The clustering process organizes the variants into difficulty levels, enabling personalized learning pathways without sacrificing semantic fidelity.

To further interpret these results, we now discuss the consistency and structure of the generated clusters, their relevance for vocabulary progression, and the broader implications of the approach.

Figure 4 presents two representative examples of clustered Pareto fronts derived from distinct input paragraphs. In both

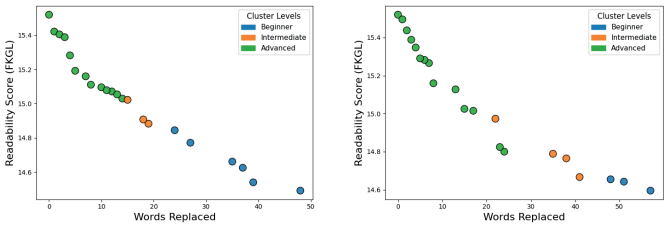


Fig. 4: This figure shows (a) the clustered Pareto front for input paragraph A and (b) the clustered Pareto front for input paragraph B. Each point represents a text variant, positioned by FKGL and the number of word substitutions, with cluster color indicating progression in reading complexity.

cases, the trade-off between FKGL readability and the number of word substitutions is clearly observable, with variants distributed along a frontier that reflects this balance. Variants with lower FKGL scores tend to involve more lexical replacements, while those with fewer substitutions preserve the original wording and register higher on the readability scale. This pattern confirms the suitability of these two metrics as competing but pedagogically meaningful criteria for guiding vocabulary-level text variation. Across all tested inputs, our framework consistently produced three semantically coherent clusters that correspond to increasing levels of reading difficulty. The *Beginner* cluster typically contains variants with the highest substitution counts and the lowest FKGL scores, making them more accessible to novice readers. The *Advanced* cluster, on the other hand, includes variants that closely resemble the original text and maintain more complex lexical structures. The *Intermediate* cluster bridges this gap, enabling a gradual progression in vocabulary exposure. These stratified outputs allow for precise scaffolding of learning materials, enabling educators or adaptive systems to present content aligned with a learner's evolving proficiency.

A key strength of our approach lies in its fully unsupervised design. By combining TF-IDF-based feature extraction with PCA for dimensionality reduction, and clustering via k-Means, our pipeline organizes the search space of text variants into interpretable and well-separated groups. Selecting representative variants based on cluster centroids ensures that learners receive central, non-redundant examples within each tier. Notably, this entire process operates without the need for annotated simplification corpora or external readability labels, making it scalable and domain-agnostic. We also observed that while cluster indices are assigned arbitrarily by the clustering algorithm, post-hoc labeling based on FKGL and substitution statistics allows for consistent semantic interpretation. The qualitative separation between clusters was evident in most runs, reinforcing the utility of our metric-driven generation and clustering strategy.

Although the proposed framework performs reliably across diverse inputs, some cases exhibited less distinct cluster

boundaries where FKGL and substitution counts were closely aligned. These occurrences highlight the inherent limitations of surface-level metrics like FKGL, which, despite their interpretability and educational relevance, do not fully capture nuances such as fluency, cohesion, or idiomatic expression. Additionally, the synonym replacement process operates without context-aware disambiguation, which may occasionally yield semantically suboptimal variants. Future extensions may incorporate semantic similarity constraints, paraphrase scoring methods (e.g. BERTScore), or LLM-based feedback to improve contextual appropriateness and linguistic naturalness.

The results across diverse input texts confirm the consistency, interpretability, and educational relevance of our progressive readability optimization framework. The ability to produce clustered, lexically varied text variants in an unsupervised and controlled fashion makes it a promising candidate for integration into intelligent language learning platforms.

VI. Conclusion

This paper introduced a scalable and unsupervised framework for generating and organizing lexically varied English text variants to support progressive vocabulary learning. By leveraging synonym replacement, readability scoring, and non-dominated sorting, the method constructs a Pareto front of semantically faithful and readability-adjusted variants. These optimal variants are then clustered into pedagogically meaningful tiers using TF-IDF vectorization, PCA, and k-Means clustering. The resulting stratified outputs allow learners to engage with increasingly complex vocabulary in a controlled and gradual manner. Experimental results demonstrate the method's robustness in producing well-separated clusters aligned with substitution and readability patterns. While current metrics such as FKGL and word substitution count provide a strong foundation, future extensions may integrate semantic similarity models and context-aware substitution mechanisms to further enhance text quality. Furthermore, adapting the framework to support multilingual inputs and cross-linguistic synonym resources would broaden its applicability in global language learning contexts. Another promising direction is dynamic curriculum sequencing, where reading levels are automatically adjusted based on the learner's performance over time. This framework lays the foundation for intelligent language learning platforms that adapt to user proficiency through interpretable and lexically controlled content.

Acknowledgement

This work is supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant RS-2023-00251002).

References

[1] Z. Bai, "An analysis of english vocabulary learning strategies." *Journal of Language Teaching & Research*, vol. 9, no. 4, 2018.

[2] S. Krashen, "Principles and practice in second language acquisition," 1982.

[3] A. Siddharthan, "A survey of research on text simplification," *ITL-International Journal of Applied Linguistics*, vol. 165, no. 2, pp. 259–298, 2014.

[4] R. Flesch, "Flesch-kincaid readability test," *Retrieved October*, vol. 26, no. 3, p. 2007, 2007.

[5] N. Srinivas and K. Deb, "Muiltiobjective optimization using nondominated sorting in genetic algorithms," *Evolutionary computation*, vol. 2, no. 3, pp. 221–248, 1994.

[6] I. S. Nation and I. Nation, *Learning vocabulary in another language*. Cambridge university press Cambridge, 2001, vol. 10.

[7] S. Alissa and M. Wald, "Text simplification using transformer and bert," *Computers, Materials & Continua*, vol. 75, no. 2, pp. 3479–3495, 2023.

[8] C.-O. Truică, A.-I. Stan, and E.-S. Apostol, "Simplex: a lexical text simplification architecture," *Neural Computing and Applications*, vol. 35, no. 8, pp. 6265–6280, 2023.

[9] X. Zhang and M. Lapata, "Sentence simplification with deep reinforcement learning," *arXiv preprint arXiv:1703.10931*, 2017.

[10] W. Xu, C. Callison-Burch, and C. Napoles, "Problems in current text simplification research: New data can help," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 283–297, 2015.

[11] C. Scarton and L. Specia, "Learning simplifications for specific target audiences," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 712–718.

[12] K. North, T. Ranasinghe, M. Shardlow, and M. Zampieri, "Deep learning approaches to lexical simplification: A survey," *Journal of Intelligent Information Systems*, vol. 63, no. 1, pp. 111–134, 2025.

[13] H. Zhang, H. Song, S. Li, M. Zhou, and D. Song, "A survey of controllable text generation using transformer-based pre-trained language models," *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–37, 2023.

[14] J. Martinez-Gil, "Optimizing readability using genetic algorithms," *Knowledge-Based Systems*, vol. 284, p. 111273, 2024.

[15] D. McCarthy, "Lexical substitution as a task for wsd evaluation," in *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions*, 2002, pp. 89–115.

[16] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[17] M. Anschütz, E. Mosca, and G. Groh, "Simpler becomes harder: Do llms exhibit a coherent behavior on simplified corpora?" *LREC-COLING 2024*, pp. 185–195, 2024.

[18] D. Fang, J. Qiang, Y. Zhu, Y. Yuan, W. Li, and Y. Liu, "Progressive document-level text simplification via large language models," *arXiv preprint arXiv:2501.03857*, 2025.

[19] G. H. Mc Laughlin, "Smog grading-a new readability formula," *Journal of reading*, vol. 12, no. 8, pp. 639–646, 1969.

[20] G. R. Klare, "Assessing readability," *Reading research quarterly*, pp. 62–102, 1974.

[21] Y. Ehara, "An analytical study of the flesch-kincaid readability formulae to explain their robustness over time," in *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, 2024, pp. 989–997.

[22] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, no. 1. Citeseer, 2003, pp. 29–48.

[23] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.

[24] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[25] R. L. Peach, S. N. Yaliraki, D. Lefevre, and M. Barahona, "Data-driven unsupervised clustering of online learner behaviour," *npj Science of Learning*, vol. 4, no. 1, p. 14, 2019.

[26] B. Du, N. Su, Y. Zhang, and Y. Wang, "A two-stage progressive intent clustering for task-oriented dialogue," in *Proceedings of The Eleventh Dialog System Technology Challenge*, 2023, pp. 48–56.

[27] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[28] A. Arikan, "Topics of reading passages in elt coursebooks: What do our students really read?." *Online Submission*, vol. 8, no. 2, 2008.