Modelling Children’s Language: Vikidia-based Resources to Support Italian Text Simplification

Modelling Children's Language: Vikidia-based Resources to Support Italian Text Simplification

Andrea Tirico[1*], Marco Braga[1], Emiliana Murgia[2] and Gabriella Pasi[1]

***Abstract*— With the rapid advancement of Large Language Models, their role in educational contexts has gained increasing attention due to the text generation capabilities and their emerging value as tools for supporting students and teachers. One relevant application is the task of Text Simplification (TS), which aims to reduce text complexity to enhance readability and accessibility for users with limited linguistic abilities, such as children. Many approaches to Text Simplification focus on the English language with children as target user group, while they are limited for the Italian language. In this paper, to fill this gap, we introduce new resources designed for children aged 8 to 11, which will support the definition of the TS approach. We build our dataset by collecting texts from both the Italian Wikipedia and Vikidia, which is an online encyclopedia written for a young audience. Subsequently, we fine-tune a pre-trained language model on our proposed resource to generate simplified versions of Wikipedia pages that align with Vikidia's linguistic style. The simplification process is further guided by a vocabulary built from Vikidia. This post-processing step enables the replacement of potentially complex terms in the output text with synonyms in the vocabulary that are accessible to children. The effectiveness of the system was evaluated through a qualitative human evaluation conducted by elementary school teachers, showing general positive assessments, alongside some divergent views and suggestions for refinement before the introduction of the educational context. We release code, prompts and the proposed dataset at https://github.com/marostiri/italian-children-text-simplification.**

***Index Terms*— Italian Text Simplification, Text Accessibility, Large Language Models, Human Evaluation**

I. Introduction

Large Language Models (LLMs) [1] have significantly transformed the educational field, particularly due to their abilities in generating fluent text to support both teachers and students [2]. These models have been deployed in a variety of educational tasks, including chatbots to facilitate learning activities [3], [4] and automated question grading [5]. Among these, Text Simplification (TS) is an educational task that involves modifying lexical and syntactic structures to enhance students' comprehension and accessibility to the text [6], [7]. Recent works on English text simplification have increasingly relied on LLMs, also targeting primary school students [8], [9]. By contrast, for the Italian language, only few resources [10], [11] and approaches [11], [12], [13] have been proposed for this target group. Terence [10] is an Italian parallel dataset designed for children's text simplification. However, the limited size of Terence makes it difficult to perform fine-tuning on pre-trained LLMs. Moreover, the Italian version of the Newsela corpus [14] provides a larger amount of data via automatic translation of manually simplified English news articles for different age groups, at the cost of introducing noise and potential translation errors. Concerning TS approaches, a recent work [13] proposes a model to perform text simplification for children with language disorders through lexical and syntactic control. However, this approach or relying on ChatGPT [15], [16] does not guarantee a lexicon suitable for children, thereby potentially reducing the comprehension of the text. To fill this gap, we propose a two-stage pipeline that splits simplification into distinct processing stages. First, we construct a parallel dataset, named VITIC-D, by extracting content from both the Italian Wikipedia and Vikidia, an online encyclopedia for young audiences. The first stage relies on a pre-trained language model, which is fine-tuned on the curated dataset to generate simplified Italian texts suitable for children aged 8 to 11. In the second stage, we apply lexical simplification to replace potentially complex terms in generated simplified texts by leveraging VITIC-V, a vocabulary built from Vikidia and enriched with other curated resources, such as Terence and ItAoA [17]. The second stage aims to replace complex terms with words that are easily understandable to children. The system identifies terms not in VITIC-V and replaces them with candidate synonyms generated by an instruction-tuned LLM. The selection of the most appropriate synonym is performed through a ranking mechanism that regards contextual coherence and its occurrence in children's vocabulary. To evaluate the effectiveness of the proposed approach, we begin by assessing the first stage to identify the best TS model to employ. Subsequently, we assess our TS pipeline through a human evaluation with primary school teachers, who evaluate the system outputs by filling out structured questionnaires and qualitative feedback. Our simplification approach receives a generally positive evaluation, and teachers explain their opinions about the system, highlighting its pedagogical potential, while also revealing points of disagreement and aspects requiring refinement prior to any classroom deployment. The remainder of the paper is organized as follows. In Section II, we review related works about text simplification, with a focus on approaches tailored to Italian children. We then provide a systematic methodology

* He represents the main author.

[1]They are with the Department of Informatics, Systems, and Communication, University of Milano-Bicocca, Viale Sarca 336, Milan, Italy {a.tirico, m.braga}@campus.unimib.it and gabriella.pasi@unimib.it

[2]Emiliana is with the Department of Education Sciences, University of Genoa, Corso Andrea Podesta' 2, Genoa, Italy emiliana.murgia@unifg.it

This work was partially supported by the European Union – Next Generation EU within the project NRPP M4C2, Investment 1.,3 DD. 341 - 15 march 2022 – FAIR – Future Artificial Intelligence Research – Spoke 4 - PE00000013 - D53C22002380006.

for building the resources and the TS pipeline in Section III. In Section IV we introduce our research question and the experimental setup. According to the experiment results, we answer the research question in Section V and conclude the work in Section VI.

II. Related Works

Text simplification involves rephrasing a text into a simpler version by modifying its syntax and lexicon [6]. To this purpose, language models have become the main paradigm [18], and recent evidence shows that LLMs can perform this task effectively [8], [19]. Text simplification should not be considered only a way to reduce cognitive load; rather, it aims to lower linguistic processing demands, thereby helping readers manage conceptual complexity [20], [21] by supporting active engagement and deeper comprehension of the text. This position finds theoretical grounding within the Zone of Proximal Development (ZPD) [22], the range in which learners are able to perform with the support of an expert. Text simplification enables learners to operate within their ZPD, operating through the reduction of lexical and syntactic complexity while preserving semantic and informational integrity. The scaffolding function of text simplification assumes particular significance for heterogeneous learner populations characterized by different literacy profiles, which include: second language learners, individuals with dyslexia, and children. Nevertheless, text simplification entails the potential reduction of exposure to diverse lexical items, thereby limiting opportunities for vocabulary acquisition. Nevertheless, this trade-off is pedagogically justified when the learner's primary objective is content comprehension rather than explicit linguistic development [23], [24]. For the English language, a wide range of resources and methodologies have been developed for text simplification tailored to children [8], [9], [14]. By contrast, for the Italian language, research on text simplification for young readers is limited. Terence [10] is a parallel dataset built for children, comprising sentences extracted from fable books alongside their simplified versions, produced by linguists and psycholinguists. However, the limited size of Terence poses a challenge when fine-tuning large language models. Moreover, the Newsela corpus [14], a collection of manually rewritten English news articles for different age groups, has been automatically translated into the Italian language [11]. In this work, the authors have extracted sentence pairs from Newsela articles and translate them by exploiting a machine translation tool. Although this resource provides a larger number of instances than Terence, the automatic translation introduces information loss for a subset of sentences, characterized by lower linguistic complexity. To provide a resource that exceeds Terence in size while avoiding the noise introduced by automatic translation, we propose a new TS dataset for children, built from the general descriptions in Italian Wikipedia and Vikidia. Recent works rely on Italian Vikidia to build corpora [25] and parallel datasets aligned at the document [26] or sentence level [27]. However, document and sentence alignment may be noisy, since Italian Wikipedia and Vikidia articles often differ in length, structure, and textual content, making one-to-one alignment difficult to guarantee. Therefore, we focus on aligning general description paragraphs to establish a parallelism between the corresponding content in Italian Wikipedia and Vikidia. Regarding text simplification systems for the Italian language, its application tailored to children remains largely unexplored [12], [13]. Among them, ERNESTA [12] is a rule-based system that applies psycholinguistically motivated syntactic transformations for children aged 7-11 and includes anaphora handling. This recent work [13] proposes an encoder–decoder architecture to perform text simplification for children with language disorders using explicit control over lexical and syntactic complexity. This model is fine-tuned on the augmented corpus built from multiple Italian text simplification datasets and further expanded via back-translation from English language. However, these approaches do not guarantee a fully accessible lexicon for target readers, which may limit the comprehension of the text. To address this issue, the lexical simplification task involves identifying words not suitable and replace them with simpler synonyms, while preserving the original meaning of sentences [28], [29], [30], [31]. To generate simplified text that emphasizes a child-appropriate vocabulary, we propose a two-stage pipeline. First, our fine-tuned model performs child-oriented text simplification. Second, the lexical simplification module selects appropriate synonyms, guided by a child-appropriate vocabulary automatically extracted from Vikidia and other resources.

III. Methodology

This section outlines the proposed approach, encompassing the construction of novel resources based on Vikidia. We also provide the description of the two-step pipeline, which is summarized in Fig. 1.

A. The VITIC Resources

To mitigate the issues related to the Terence and Italian Newsela datasets, we build a new dataset, VITIC-D, which includes aligned general descriptions from Italian Wikipedia and Vikidia from 2012 to 2024.

First, we remove HTML tags from the text and eliminate pages where either Vikidia or Wikipedia lack a general description. Afterwards, we align the texts using their URLs and titles, obtaining approximately 8,000 instances for the dataset. Furthermore, we build VITIC-V, a vocabulary based on all words in the Vikidia pages, further enriched through Terence and ItAoA [17]. The latter is a dataset originated from a Web-based survey in which adult participants reported subjective estimates of the age at which they acquired each word. Initially we filter the text from Vikidia and Terence, by removing Italian stopwords, special symbols and words containing non-Italian characters. The resulting set of term entries is extended with words from the ItAoA dataset whose mean age of acquisition is 11 years or younger, in order to align with the target age group for 8-11 years old. After

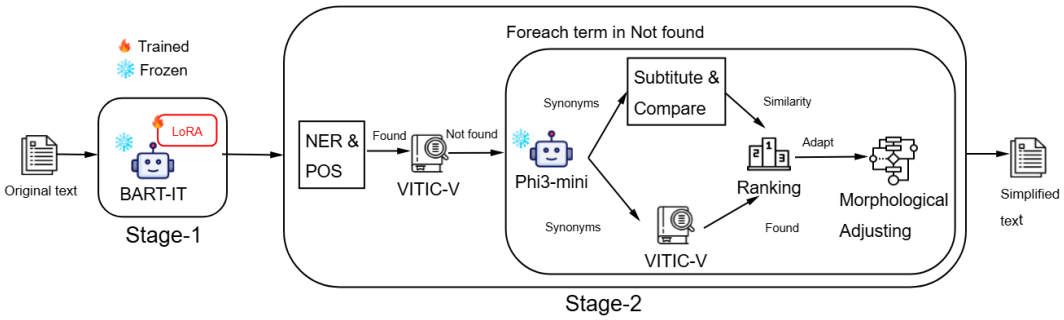

Fig. 1. Simplification pipeline with two stages. Stage-1 performs Text Simplification. Stage-2 enables the replacement of complex terms.

deduplication, this process yields a vocabulary of 59,973 distinct word types.

B. Text Simplification

Stage-1 is finalized to address the text simplification task for children aged 8–11 years. To this purpose, we fine-tune an encoder-decoder pre-trained model on VITIC-D by applying LoRA [32], which is a Parameter-Efficient Fine-Tuning (PEFT) method. PEFT methods [33], [34] allow to train a small set of parameters while adapting the model to a downstream [35], outperforming full fine-tuning in low-resource settings. The aim of the model is to generate the simplified version of a given input text. In Fig. 1, Stage-2 tackles the challenge of generating terms suitable for children aged 8 to 11, through the lexical simplification module applied to the output of the previous step. Initially we apply a Part-of-Speech (POS) tagger to identify syntactic role of the words contained in the input paragraph, classifying them as nouns, adjectives, adverbs or verbs. Subsequently, we employ a Named Entity Recognition (NER) algorithm to identify and remove proper names, names of locations and organizations, which appear in the reference text. This process ensures that the named entities are excluded from the candidate synonym generation step, as no substitutions are needed for these terms. After this preparation phase, for each word identified as a noun, adjective, adverb, or verb, we check its correspondence within VITIC-V, to assess its suitability to children. Inspired by a recent work [29], we rely on an instruction-tuned LLM to generate candidate synonyms if a term is not in the children vocabulary. During the synonyms generation phase, we rely on a structured prompt that explicitly specifies the contextual sentence, the terms not suitable for children, and their grammatical category, which is their syntactic role, such as adverb, verb, adjective and noun. To ensure the suitability of the generated synonyms, we verify whether each candidate occurs in VITIC-V and we assign a binary score, $\mathbf{1}_{found}$, set to 1 when the item is in the lexical resource and to 0 otherwise. Coherence is ensured by replacing synonyms in the original sentence. Accordingly, we quantify the semantic similarity between the original and modified text using a metric based on a multilingual encoder model, referring to this measure as $score_{similarity}$. Next, to choose the best synonym, we rank the candidates by applying the following formula:

$$score = K * \mathbf{1}_{found} + (1 - K) * score_{similarity} \quad (1)$$

The value *score* is computed as a linear combination that considers both the synonym's contextual coherence and its presence in the children's vocabulary. The synonym with the highest *score* is selected to replace the inappropriate term. Then, if needed, we apply the morphological adjusting algorithm, which, based on the synonym's inflectional properties, changes both the gender and the number of dependent elements, such as articles, partitives, prepositions, adjectives and pronouns. This procedure is repeated until all terms that are not in VITIC-V have been processed.

IV. Experimental Setup

In this Section, we address the following research question (**RQ**): Are the simplifications generated by our TS pipeline judged appropriate by teachers, and could this system be integrated into educational practice? To address the effectiveness of the proposed pipeline, we first rely on evaluation metrics applied to Stage-1 outputs to select the best performing model for the subsequent stage. Then, we involve elementary school teachers to assess the system's outputs by also gathering their opinions on its potential classroom use.

To set up the pipeline, we employ `BART-IT` [36] for the Stage-1 and `Phi-3-mini` [37] for the Stage-2. `BART-IT` is an Italian encoder–decoder model based on `BART` [38], which is pre-trained on large-scale Italian corpora. Additionally, `BART` and its variants are commonly adopted in TS approaches, further supporting the appropriateness of this architecture for this task [39], [40]. For the second stage, we employ `Phi-3-mini`, a multilingual instruction-tuned model, which achieves strong generation quality thanks to training on carefully curated web and synthetic data [37]. To perform the TS for Stage-1, `BART-IT` is fine-tuned on the VITIC-D, which we split into training, validation, and test sets with proportions of 64%, 16%, and 20%, respectively, following standard practice. During the pre-processing of Stage-2, we perform POS tagging relying on the SpaCy's large Italian model, and apply an Italian encoder model for the NER task. We assess the performance of Stage-1 by relying on the evaluation metrics BLEU [41], SARI [42], Italian Flesch score [43] and BERTScore [44]. BLEU

and SARI are both n-grams-based metrics for evaluating the quality of text generation. The BLEU metric is based on the comparison of n-gram overlap in tokens from the prediction and one or more reference texts. While, the SARI metric compares the generated simplification with both the original text and one or more reference, assessing how appropriately the model adds, deletes, and retains n-grams. The Italian Flesch score is designed to evaluate the readability of Italian texts, based on the average number of syllables per word and the average words per sentence [45]. We additionally compute the difference between the Italian Flesch scores of the generated and original texts, using as an indicator of readability improvement. Semantic similarity is measured relying on BERTScore, an LLM-based metric. In our experiments, we compute the F_1 value of the BERTScore using `XLM-RoBERTa-base` [46], a multilingual encoder that provides coverage for the Italian language. This metric is also employed to yield the value of $score_{similarity}$. We further investigate how different hyperparameter settings affect the model performance for Stage-1. Specifically, we focus on the number of training epochs and, at inference time, on the beam size used during text generation, since beam search is a key component of sequence modelling and decoding [47]. We train our models with a learning rate of 2×10^{-5}, a batch size of 32, and a weight decay of 0.01. Following the original LoRA parameters [32], we use $alpha = 16$, $rank = 8$ and $dropout = 0.05$.

V. Results and Discussion

In this Section, we show the results of our text simplification analysis and we illustrate, through the teachers-based evaluations, the potential of LLM-based simplification and its possible implementation in educational contexts.

To address our RQ, we first analyze the impact of different hyperparameter settings to identify the configuration that yields the best-performing model for our pipeline. Table I shows the results of our evaluation on the validation set with different hyper-parameter configurations for Stage-1. We note that the optimal configuration is obtained with the `BART-IT` model trained for 60 epochs and using 4 beams, which outperforms all other configurations on Italian Flesch, BLEU and SARI by at least 2.48, 0.04 and 1.06 points, respectively. Overall, Table I indicates that increasing the number of training epochs tends to improve text simplification performance, as reflected by the higher values of the Italian Flesch and SARI scores. In particular, a higher Italian Flesch value indicates that the model generates shorter sentences and fewer syllables per word, while a high SARI score reflects more effective simplification operations, rewarding outputs that appropriately add, delete, and retain terms. However, we can also notice that increasing the search space by using a larger beam size does not lead to better performance. The evaluation of optimal configuration on the test set yields a BLEU score of 0.24, a SARI score of 53.91, and a BERTScore of 0.91 and an Italian Flesch score of 58.40 with a difference of 4.3 points between simplified and original texts. Overall, these outcomes confirm the consistency of the results compared to the validation set. Following these findings, we integrate the best-performing model in Stage-2 using $K = 0.3$ in Equation (1), giving more importance to the $score_{similariry}$ than to the binary score $\mathbf{1}_{found}$. To evaluate our TS pipeline, we perform a human assessment with elementary school teachers, who work with 9- to 10-year-old students, to provide a rigorous evaluation. We restrict our focus to this age group, which corresponds approximately to the fourth grade of Italian primary school, representing a relatively homogeneous developmental window in which foundational decoding skills are typically consolidated [48]. Moreover, the reading development follows a relatively stable trajectory across grades 1–5, as indicated by longitudinal evidence [49] and consultation with a teacher. To gather teachers' opinions, we employ structured questionnaires and qualitative feedback. However, despite our recruitment efforts, participation remains limited due to nine teachers, reflecting a low uptake of the test. Demographic data from the participating teachers reveal that eight out of nine identify themselves as female, half are from Northern Italy, two-thirds are over 50 years old and half have over a decade of experience teaching in elementary schools. To assess teachers' opinions, we ask them to compile a questionnaire based on a 4-point Likert scale [50], a widely used method for detecting attitudes [51]. Notably, this scale does not include a neutral option, requiring participants to express a polarized judgment, ranging from complete disagreement (1) to full agreement (4). As input for our simplification pipeline, we select four texts from upper secondary school books, characterized by textual complexity beyond the usual linguistic abilities of 9-10 year old students. The characteristics of texts are presented in Table II. The selected texts cover history, geography, science, and civic education, ensuring relevant topics for primary school education. We choose a limited number of paragraphs to prevent participant fatigue and maintain the authenticity of the responses. The questions for the assessment address different aspects of

TABLE I
Results on the validation set with different configurations by training BART-IT.

Epochs	Beams	Italian Flesch	BLEU	SARI	BERTScore
10	4	$55.77_{1.22}$	0.11	46.06	0.89
10	8	$54.00_{-0.55}$	0.10	45.14	0.89
60	4	$\mathbf{59.68_{5.13}}$	**0.23**	**52.81**	**0.91**
60	8	$57.20_{2.65}$	0.19	51.75	**0.91**

TABLE II
Length of the four texts for teacher evaluation

Subject	Number of words	Italian Flesch
History	54	26.19
Geography	74	29.46
Science	29	36.56
Civic Education	51	57.26

TABLE III
AVERAGE RESULTS OF THE HUMAN EVALUATION ON THE LIKERT SCALE

	History	Geography	Science	Civic Education
Orig. Comp.	2.11	2.66	2.66	2.66
Simp. Comp.	2.88	3.11	2.11	2.77
Consistency	2.88	2.88	1.66	2.77
Voc. Coher.	2.66	3.11	2.44	2.77
Information	2.33	2.55	1.66	2.33

text simplification including comprehensibility of both the original (Orig. Comp.) and the simplified (Simp. Comp.) texts, content consistency of the simplification process (Consistency), coherent simplified vocabulary for children (Voc. Coher.) and the extent to which key information is preserved (Information). As shown in Table III, teachers generally express agreement regarding the quality of the simplified outputs, except for science-related content, as the removal of information leads to changes in the original meaning. Additionally, we gather opinions on the potential usefulness of an application for text simplification in classroom settings. The average score is 3.00, suggesting agreement about the usefulness of a TS system to support students' comprehension in class. However, the general score about simplification quality is 2.44, suggesting that the generated output is moderately appreciated. To enrich the answer to RQ, we assess opinions on both the overall effectiveness of our TS pipeline and potential resistance to its adoption. The responses are analysed using an interpretive approach [52], which is commonly used for interview analysis [53]. This approach consists of multiple stages: (i) manual identification of recurrent topics in collected feedback; (ii) for both questions, we count the number of answers in which each identified topic is present. Table IV shows a generally positive perception of the proposed approach. However, a possible limitation identified is information loss, as the model omits parts of the text during the simplification process. Although the lexicon appears enough adequate, teachers highlight concerns related to syntax, suggesting further refinement.

TABLE IV
TOPICS EXTRACTED FROM TEACHERS' RESPONSES TO OPEN-ENDED QUESTIONS. INTEGER VALUES INDICATE THE TOTAL NUMBER OF RESPONSES CONTAINING EACH TOPIC.

Topic	Question 1	Question 2
Bad syntax simplification	2	0
Positive general perception	5	0
Good if improved simplification	2	5
Enough good synonyms	2	0
Information loss simplification	4	0
Partially approved simplification	1	0
Positive idea for foreign students	1	2
App and teachers	1	2
Negative perception	1	0
Only teacher without system	1	0
Artificial Intelligence literacy	0	2
Simplification for back-office	0	1
Media education	0	1

As an additional topic, teachers emphasize the importance of Artificial Intelligence (AI) literacy, underscoring the need for education on AI-based tools. These findings show that, while the adoption of an AI-powered simplification tool in schools is not yet fully embraced, teachers express interest, especially if improvements are introduced. Finally, most teachers express that our tool could be highly beneficial, provided that they receive proper training on its application. Different opinions are about a possible support for non-native Italian speaking students, reinforcing its potential as an educational aid.

VI. CONCLUSIONS AND FUTURE WORK

This work focuses on building a text simplification system for Italian children aged 8-11. We propose a new dataset, VITIC-D, aligning general descriptions from Wikipedia and Vikidia. Additionally, we implement a child-specific vocabulary from Vikidia, named VITIC-V. The model is fine-tuned on VITIC-D using LoRA and Stage-2 identifies terms not in VITIC-V and replaces them with candidate synonyms generated by Phi-3-mini. Finally, we evaluate the pipeline through a small-scale teacher-based assessment employing questionnaires, revealing issues with syntax simplification and information loss. Despite these shortcomings, teachers recognize the potential value of the TS system in children's education, suggesting that, with improvements, it could provide valuable educational support. As future work, we plan to apply Llama-3 [54] to perform a step-by-step simplification, supported by our new proposed vocabulary, which can be fine-grained by children-based interaction and assessment.

LIMITATIONS

Our study does not include a direct comparison with other state-of-the-art approaches, as we aim to assess whether the system, built upon these new resources, is perceived as useful by teachers. Second, although the questionnaire received a low response rate, the collected feedback remains valuable for identifying weaknesses and guiding future improvements.

ACKNOWLEDGMENT

During the preparation of this work, the authors used ChatGPT-5.1 for paraphrasing and reformulation, and DeepL for spelling verification. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

REFERENCES

[1] S. Wang, T. Xu, H. Li, C. Zhang, J. Liang, J. Tang, and et al., "Large language models for education: A survey and outlook," *IEEE Signal Processing Magazine*, 2025.

[2] M. Giannakos, R. Azevedo, P. Brusilovsky, M. Cukurova, Y. Dimitriadis, D. Hernandez-Leo, and et al., "The promise and challenges of generative ai in education," *Behaviour & Information Technology*, 2025.

[3] K. Taneja, P. Maiti, S. Kakar, P. Guruprasad, S. Rao, and A. K. Goel, "Jill watson: A virtual teaching assistant powered by chatgpt," in *AIED*, 2024.

[4] V. Kretzschmar and J. Seitz, "Ai-based learning assistants: Enhancing math learning for migrant students in german schools," in *IEEE CAI*, 2024.

[5] J. Pecuchova, Ľ. Benko, and M. Drlik, "Automated grading of open-ended questions in higher education using GenAI models," *International Journal of Artificial Intelligence in Education*, 2025.

[6] S. S. Al-Thanyyan and A. M. Azmi, "Automated text simplification: A survey," *ACM Computing Surveys*, 2021.

[7] N. Pérez-Rojas, S. Calderon-Ramirez, M. Solís, M. Alberto Romero-Sandoval, M. Arias-Monge, and H. Saggion, "A novel spanish dataset for financial education text simplification targeting visually impaired individuals," *IEEE Access*, vol. 13, 2025.

[8] A. Smirnova, K. b. Chun, W. L. Rothman, and S. Sarma, "Text simplification for children: Evaluating llms vis-à-vis human experts," in *CHI EA*, 2025.

[9] K. Mo and R. Hu, "ExpertEase: A multi-agent framework for grade-specific document simplification with large language models," in *EMNLP*, 2024.

[10] D. Brunato, F. Dell'Orletta, G. Venturi, and S. Montemagni, "Design and annotation of the first Italian corpus for text simplification," in *Proc. of the 9th Linguistic Annotation Workshop*, 2015.

[11] A. L. Megna, D. Schicchi, G. Lo Bosco, and G. Pilato, "A controllable text simplification system for the italian language," in *IEEE ICSC*, 2021.

[12] G. Barlacchi and S. Tonelli, "Ernesta: A sentence simplification tool for children's stories in italian," in *International Conference on Intelligent Text Processing and Computational Linguistics*, 2013.

[13] F. Padovani, C. Marchesi, E. Pasqua, M. Galletti, and D. Nardi, "Automatic text simplification: A comparative study in italian for children with language disorders," in *NLP4CALL*, 2024.

[14] W. Xu, C. Callison-Burch, and C. Napoles, "Problems in current text simplification research: New data can help," *TACL*, 2015.

[15] E. Murgia, Z. Abbasiantaeb, M. Aliannejadi, T. Huibers, M. Landoni, and M. S. Pera, "Chatgpt in the classroom: A preliminary exploration on the feasibility of adapting chatgpt to support children's information discovery," in *UMAP Adjunct*, 2023.

[16] E. Murgia, M. S. Pera, M. Landoni, and T. Huibers, "Children on chatgpt readability in an educational context: myth or opportunity?" in *UMAP Adjunct*, 2023.

[17] M. Montefinese, D. Vinson, G. Vigliocco, and E. Ambrosini, "Italian age of acquisition norms for a large set of words (itaoa)," *Frontiers in Psychology*, 2019.

[18] F. Alva-Manchego, C. Scarton, and L. Specia, "Data-driven sentence simplification: Survey and benchmark," *CL*, 2020.

[19] A. Baez and H. Saggion, "LSLlama: Fine-tuned LLaMA for lexical simplification," in *TSAR*, 2023.

[20] S. Stajner, S. Nisioi, and I. Hulpuș, "CoCo: A tool for automatically assessing conceptual complexity of texts," in *LREC*, 2020. [Online]. Available: https://aclanthology.org/2020.lrec-1.887/

[21] E. Sokli, G. Peikos, P. Kasela, and G. Pasi, "Leveraging cognitive complexity of texts for contextualization in dense retrieval," in *EMNLP*, 2025, pp. 27 071–27 084.

[22] L. S. VYGOTSKY, *Mind in Society: Development of Higher Psychological Processes*. Harvard university press, 1978.

[23] L. Javourey-Drevet, S. Dufau, T. François, N. Gala, J. Ginestié, and J. C. Ziegler, "Simplification of literary and scientific texts to improve reading fluency and comprehension in beginning readers of french," *Applied Psycholinguistics*, 2022.

[24] A. Säuberli, F. Holzknecht, P. Haller, S. Deilen, L. Schiffl, S. Hansen-Schirra, and et al., "Digital comprehensibility assessment of simplified texts among persons with intellectual disabilities," in *CHI*, 2024.

[25] E. Marzona, M. Goikhman, A. P. Aprosio, and M. Zancanaro, "Exploring paraphrasing strategies for CEFR a1-level constraints in LLMs," in *EMNLP*, 2025.

[26] M. Trokhymovych, I. Sen, and M. Gerlach, "An open multilingual system for scoring readability of Wikipedia," in *ACL*, 2024.

[27] D. Schicchi, G. Pilato, and G. Lo Bosco, "Deep neural attention-based model for the evaluation of italian sentences complexity," in *IEEE ICSC*, 2020.

[28] K. North, T. Ranasinghe, M. Shardlow, and M. Zampieri, "Deep learning approaches to lexical simplification: A survey," *Journal of Intelligent Information Systems*, 2025.

[29] S. Seneviratne and H. Suominen, "ANU at MLSP-2024: Prompt-based lexical simplification for English and Sinhala," in *BEA*, 2024.

[30] R. Miyata, K. Horiguchi, R. Kondo, Y. Fujiwara, and T. Kajiwara, "EhiMeNLP at TSAR 2025 shared task candidate generation via iterative simplification and reranking by readability and semantic similarity," in *TSAR*, 2025.

[31] D. Aleksandrova and O. Brochu Dufour, "RCML at TSAR-2022 shared task: Lexical simplification with modular substitution candidate ranking," in *TSAR*, 2022.

[32] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and et al., "Lora: Low-rank adaptation of large language models." *ICLR '22*, 2022.

[33] M. Braga, A. Raganato, and G. Pasi, "Adakron: An adapter-based parameter efficient model tuning with kronecker product," in *LREC-COLING*, 2024. [Online]. Available: https://aclanthology.org/2024.lrec-main.32/

[34] ——, "Madakron: A mixture-of-adakron adapters," *Knowledge-Based Systems*, 2026.

[35] M. Braga, "Personalized large language models through parameter efficient fine-tuning techniques," in *SIGIR*, 2024.

[36] M. La Quatra and L. Cagliero, "Bart-it: An efficient sequence-to-sequence model for italian text summarization," *Future Internet*, 2023.

[37] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, and et al., "Phi-3 technical report: A highly capable language model locally on your phone," *arXiv preprint arXiv:2404.14219*, 2024.

[38] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *ACL*, 2020.

[39] L. Cripwell, J. Legrand, and C. Gardent, "Evaluating document simplification: On the importance of separately assessing simplicity and meaning preservation," in *READI*, 2024. [Online]. Available: https://aclanthology.org/2024.readi-1.1/

[40] R. Sun, W. Xu, and X. Wan, "Teaching the pre-trained model to generate simple texts for text simplification," in *ACL*, 2023, pp. 9345–9355.

[41] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002.

[42] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch, "Optimizing statistical machine translation for text simplification," *TACL*, 2016.

[43] V. Franchina and R. Vacca, "Adaptation of flesh readability index on a bilingual text written by the same author both in italian and english languages." *In Linguaggi (3)*, 1986.

[44] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *ICLR*, 2020.

[45] R. L. R. J. Peter Kincaid, Lieutenant Robert P. Fishburne and B. S. Chissom, "Derivation of new readability formulas for navy enlisted personnel." *Research Branch Report, Millington, TN: Chief of Naval Training*, 1975.

[46] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, and et al., "Unsupervised cross-lingual representation learning at scale," in *ACL*, 2020.

[47] A. Graves, "Sequence transduction with recurrent neural networks," *ICML Workshop on Representation Learning*, 2012.

[48] A. Pasqualotto, N. Mazzoni, F. Benso, and C. Chiorri, "Reading skill profiles in school-aged italian-speaking children: A latent profile analysis investigation into the interplay of decoding, comprehension and attentional control," *Brain Sciences*, 2024.

[49] S. Mascheretti, C. Luoni, S. Franceschini, E. Capelli, L. Farinotti, R. Borgatti, and et al., "Development and predictors of reading skills in a 5-year italian longitudinal study," *Infant and Child Development*, 2024.

[50] A. Joshi, S. Kale, S. Chandel, and D. K. Pal, "Likert scale: Explored and explained," *Current Journal of Applied Science and Technology*, 2015.

[51] J. T. Croasmun and L. Ostrom, "Using likert-type scales in the social sciences," *Journal of Adult Education*, 2011.

[52] R. Trinchero and D. Robasto, *I mixed methods nella ricerca educativa*. Mondadori, 2019.

[53] M. B. Miles and A. M. Huberman, *Qualitative Data Analysis An Expanded Sourcebook*. Sage Publication, 1994.

[54] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, and et al., "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.