Automatic Text Simplification with LLMs: A Comparative Study in Italian for Children with Language Disorders

Francesca Padovani[1,2], **Caterina Marchesi**[4], **Eleonora Pasqua**[4,3], **Martina Galletti**[1,3] **& Daniele Nardi**[3,5]

[1]Sony Computer Science Laboratories - Paris, France
[2] University of Trento, Italy
[3] Sapienza University of Rome, Italy
[4]Centro Ricerca e Cura di Roma - Italy
[5]CINI-AIIS - Italy
francesca.padovani98@gmail.com, martina.galletti@sony.com

Abstract

Text simplification aims to improve the readability of a text while maintaining its original meaning. Despite significant advancements in Automatic Text Simplification, particularly in English, other languages like Italian have received less attention due to limited high-quality data. Moreover, most Automatic Text Simplification systems produce a unique output, overlooking the potential benefits of customizing text to meet specific cognitive and linguistic requirements. These challenges hinder the integration of current Automatic Text Simplification systems into Computer-Assisted Language Learning environments or classrooms. This article presents a multifaceted output that highlights the potential of Automatic Text Simplification for Computer-Assisted Language Learning. First, we curated an enriched corpus of parallel complex-simple sentences in Italian. Second, we fine-tuned a transformer-based encoder-decoder model for sentences simplification. Third, we parameterized grammatical text features to facilitate adaptive simplifications tailored to specific target populations, achieving state-of-the-art results, with a SARI score of 60.12. Lastly, we conducted automatic and manual qualitative and quantitative evaluations to compare the performance of ChatGPT-3.5, and our fine-tuned transformer model. By demonstrating enhanced adaptability and performance through tailored simplifications in Italian, our findings underscore the pivotal role of ATS in Computer-Assisted Language Learning methodologies.

1 Introduction

The increasing access of digital information underscores the critical need to ensure universal access to knowledge, regardless of individuals' literacy levels or backgrounds. Automatic Text Simplification (ATS) is the Natural Language Processing (NLP) task aimed at reducing linguistic complexity of texts, while preserving their original meaning (Bott and Saggion, 2014; Shardlow, 2014b). ATS emerges as a promising solution to enhance text accessibility and readability, aiming to transform complex texts into versions that are more comprehensible, thus holding significant potential for fostering communication across diverse audiences and addressing gaps in information accessibility (Štajner, 2021). In recent years, research on ATS has focused on developing approaches to make texts simplified adapted for individuals facing cognitive disabilities or language impairment (Bott and Saggion, 2014; Rello et al., 2013; Aluisio et al., 2010). This development could have a significant impact on computer-assisted language learning (CALL), where adaptive learning technologies can personalize instruction based on individual learner progress and needs, ensuring a tailored and effective educational experience.

The emergence of large language models (LLMs) has significantly advanced automatic text simplification, among other NLP tasks. While their success in many benchmarks and challenges has been demonstrated (Anschütz et al., 2023; Sun et al., 2023; Engelmann et al., 2023; Shaib et al., 2023), it is imperative to ensure that the outputs of these models are truly suitable, especially before deployment in sensitive domains such as education or health (Kasneci et al., 2023). Fur-

Francesca Padovani, Caterina Marchesi, Eleonora Pasqua, Martina Galletti and Daniele Nardi. Automatic Text Simplification: A Comparative Study in Italian for Children with Language Disorders. *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2024)*. Linköping Electronic Conference Proceedings 211: 176–186.

thermore, there is limited research being conducted to investigate how LLMs can specifically be adapted to the needs of each user, including individuals with low literacy levels or cognitive and linguistic impairments, by providing adapted output (Demszky et al., 2023). The training data for large language models (LLMs) primarily comprises text created by individuals without language disabilities. This could potentially lead to a limited exposure to the varied linguistic patterns and communication styles exhibited by individuals with language impairments (Fiora et al., 2024; Guo et al., 2024).

Finally, most of the existing systems, focused on the English language. However, languages like Italian remain relatively under-explored in this domain, primarily due to data scarcity and poor data quality. Despite efforts to address this gap (Brunato et al., 2015, 2016, 2022; Tonelli et al., 2016, 2017), the availability of Italian simplification datasets remains limited, with only a few manually curated datasets and one large corpus assembled through a data-driven approach.

This paper aims to address these gaps by (I) creating a robust corpus by merging and cleaning existing resources (II) training a sequence-to-sequence neural model, (III) incorporating an adaptive component to control simplifications for specific target populations. Our most successful model achieves a SARI score of **60.12** and a BLEU score of **50.30** on the test set. Moreover, we present an experiment evaluating the suitability respectively of our fine-tuned model and Chat-GPT 3.5 for automatic text simplification specifically focused to the disability domain.

2 Related Work

ATS can occur at different levels of granularity: sentence-level, paragraph-level, or even at the level of entire documents and articles. In this work, we focus on a sentence-level automatic text simplification task. Consequently, our attention is solely directed towards existing work related to sentences.

Sentence-level simplification is often approached as a monolingual form of machine translation (MT). For years, attempts have been made to tackle this task using rule-based models capable of handling both lexical simplification and morpho-syntactic simplification. These techniques rely on manually crafted rules (Bott et al., 2012; Shardlow, 2014a; Siddharthan, 2011). Manually curated data offer several advantages. They ensure clear and consistent data labeling, non-redundant metadata recording, and structured presentation of contextual linguistic phenomena associated with text simplification. Nevertheless, constructing such models demands extensive investment of time and resources on experts in language knowledge. Moreover these systems suffer from a notable drawback: limited portability and scalability to new scenarios.

Authors	Description	Approach
Yatskar et al. (2010)	Context similarity to extract simplification rules.	DD
Siddharthan (2011).	Simplification and regeneration from typed dependencies	RB
Biran et al. (2011)	The first data-driven system available for English	DD
Bott et al. (2012).	First model and data for Spanish	RB
Shardlow (2014a).	Errors identification and classification scheme	RB
Glavaš and Štajner (2015)	Based on word vector representations, cased.	DD
Paetzold and Specia (2015)	Modeling words and POS tags.	DD
Nisioi et al. (2017)	Two LSTM layers incorporating global attention.	DD
Zhang and Lapata (2017)	Utilized LSTM, added lexical constraints, and combined with reinforcement learning.	DD
Scarton and Specia (2018)	Enhanced the encoder by incorporating external information.	DD
Zhao et al. (2018)	Transformer-based approach supplemented with a paraphrase database.	DD
Qiang et al. (2020)	Extension to BERT.	DD

Table 1: Models for Sentence Simplification from the least recent to the most recent, along with descriptions of the systems and an indication of whether it's rule-based (RB) or data-driven (DD).

Most sentence simplification models are available for English, primarily due to the availability of extensive supervised training datasets containing pairs of complex and simple sentences that are aligned in structure and meaning (Wubben et al., 2012; Martin et al., 2020). However, efforts have also been made to explore languages beyond English, including Brazilian Portuguese (Aluísio et al., 2008), Spanish (Saggion et al., 2015), (Glavaš and Štajner, 2015), Italian (Brunato et al., 2015; Tonelli et al., 2016), Japanese (Goto et al., 2015; Kajiwara and Komachi, 2018; Katsuta and Yamamoto, 2019), and French (Gala et al., 2020).

Moreover, the emergence of LLMs and, particularly, GPT has brought about a revolution in the field of NLP. Its impressive text generation capabilities, supported by pre-trained knowledge and fine-tuning adaptability, make it a versatile tool for various NLP tasks, including automatic text simplification. Despite their success in many benchmarks and challenges (Anschütz et al., 2023; Sun et al., 2023;

Engelmann et al., 2023; Shaib et al., 2023), it's important to verify that the outputs of these models can be suitable before deployment also in sensitive domains, such as for use with children who have language disabilities.

3 Dataset selection, curation and augmentation

Three main datasets are available for automatic sentence simplification in Italian: (1) Terence & Teacher (Brunato et al., 2015), (2) SIMPITIKI (Tonelli et al., 2016), (3) PaCCSS-IT (Brunato et al., 2016).

Terence & Teacher was introduced as the inaugural Italian Corpus for Text Simplification. Comprising around 1500 sentence pairs, it integrates two sub-corpora: Terence, consisting of 32 simplified children's stories crafted by experts across three linguistic dimensions, and Teacher, which features 24 pairs of texts manually simplified by a teacher targeting L2 students.

In 2016, SIMPITIKI was created by gathering simplification pairs from Wikipedia edits designated as "simplified". The pairs were then manually annotated and filtered, leading to a final set of 575 pairs out of the initially scraped 2,671 pairs. Additionally, employing a similar methodology, a second corpus was created by simplifying documents from the Trento Municipality pertaining to building permits and kindergarten admissions. This corpus, focused on public administration, adhered to the same annotation schema and encompassed an additional 591 pairs.

Finally PaCCSS-IT includes 63,000 pairs of sentences classified by their readability score. The corpus was constructed through monolingual sentence alignment techniques, aligning original sentences with their simplified counterparts using metrics like TF/IDF scores or similar methods assessing word similarity. Each pair includes the cosine similarity, accuracy of automatic classification for predicting sentence alignment, and readability level. Even though the dataset is quite large, the authors gathered a substantial amount of text from the web to initiate the process and reduce costs, which carried the risk of generating occasional errors, repetitions, and other issues.

For this reason, we propose an augmented dataset composed by PaCCSS-IT, SIMPITKI and a translated one. The corpus creation pipeline can be seen in Figure 1. We started by cleaning the larger available corpus, PaCCSS-IT (Brunato et al., 2016), through a preprocessing step similar to the one conducted in Palmero Aprosio et al. (2019). We deliberately retained both capital letters and punctuation within sentences to preserve meaning and convey grammatical and semantic cues. Punctuation was selectively removed primarily at the beginning and end of sentences, and identical pairs of parallel sentences were eliminated to prevent redundancy. However, we retained complex sentences that underwent distinct simplifications to ensure computational models learned the variability in simplifying the same sentence.

Figure 1: The steps we took to construct the Augmented Dataset

	COMPLEX	SIMPLE
PaCCSS-IT	Quale sarebbe allora la soluzione giusta?	È questa la soluzione giusta?
Teacher	I bei tempi finirono nel maggio 1940; prima la guerra, la capitolazione, l'invasione tedesca, poi cominciarono le sventure per noi ebrei.	I tempi felici finiscono nel maggio 1940; dopo la guerra, la sconfitta, e l'arrivo dei soldati tedeschi, cominciano i problemi per noi ebrei.
Terence	Ernesta Sparalesta è una bambina alta, più o meno, un metro e una noce.	Ernesta Sparalesta è una bambina alta poco più di un metro.
SIMPITIKI	Said spiega che questo processo è stato reso possibile attraverso una conoscenza superficiale di ciò che è in effetti l'Oriente.	Said spiega che questo processo si è realizzato mediante una conoscenza superficiale di ciò che è in effetti l'Oriente.
Translated English Sentences	L'orso bruno dell'Himalaya, noto anche come orso rosso dell'Himalaya, orso isabellino o Dzu-Teh, è una sottospecie dell'orso bruno.	L'orso bruno himalayano è una sottospecie dell'orso bruno.

Figure 2: Some examples of the composition of the Augmented Dataset

Additionally, we excluded complex sentences consisting of two tokens or fewer and those with low cosine similarity values compared to their simpler counterparts. More specifically, we disregarded sentences with cosine similarity less than 0.05. This value was chosen after a manual inspection which identified pairs of simple and complex sentences with significantly different meanings.

Lastly, we also addressed the issue of sentences containing numbers with no corresponding counterpart in the simple sentences. This adjustment ensured consistency not only in alphabetical tokens but also in numerical values. After cleaning, the curated version of the PaCCSS-IT corpus comprised 32,650 pairs of complex and simple sentences. Some examples of the senteces in the augmented corpus can be seen in Figure 2.

In a later stage, we integrated the Terence & Teacher (Brunato et al., 2015) and SIMPITIKI (Tonelli et al., 2016) datasets to the curated version of PaCCSS-IT, conducting spe-

cific parsing and pre-processing to allineate with the format in PaCCSS-IT. Our corpus at this stage consisted of 33,891 parallel sentences. The curated version incorporating the three datasets showed an increase in average sentence length due to the inclusion of sentences from Terence&Teacher, and SIMPITIKI datasets.

Finally, we augmented the curated versions by translating sentences from parallel English datasets. This was done with two main goals (I) enhance data variety and (II) improve the model generalization. For doing that, we used the DeepL API to translate around 5000 sentences pairs from a parallel English datasets. We decided to translate the *Human Simplification with Sentence Fusion Data Set* (Schwarzer et al., 2021) and few sentences translated by the first version of the *Wikipedia dataset* (Kauchak et al., 2022). The augmented version exhibited increased linguistic complexity in both complex and simplified sentences compared to the initial PaCCSS-IT dataset or its curated counterpart, as it can be seen in Table 2. The average sentence length slightly increased in the augmented version, with complex sentences averaging 9.14 words and simplified sentences 8.21 words. The use of conjunctions in simplified sentences showed a progressive increase from PaCCSS-IT to the curated and augmented datasets, suggesting greater cohesion in simplified constructs. Overall, both the curated and augmented datasets displayed higher linguistic detail and richer language use compared to the initial PaCCSS-IT dataset. The average length of complex sentences increased to 8.42, and that of simple sentences to 7.63 as it can be seen in Table 2.

Metric	PaCCSS-IT	Curated	Augmented
$AVG\_words\_complex$	8.26	8.42	9.14
$AVG\_words\_simplified$	7.34	7.63	8.21
$SVO\_complex$	0.57	0.54	0.55
$SVO\_simplified$	0.54	0.50	0.52
$CONJ\_complex$	0.23	0.25	0.28
$CONJ\_simplified$	0.26	0.27	0.29
$SUBJ\_complex$	0.03	0.05	0.06
$SUBJ\_simplified$	0.025	0.04	0.05
$stop\_words\_complex$	4.5	4.78	5.08
$stop\_words\_simplified$	2.76	3.02	3.25

Table 2: Normalized metrics for three dataset variations. The *Curated dataset* combines three existing distinct datasets, while the *Augmented Dataset* incorporates the three existing resources together with sentences translated from English parallel corpora. "AVG" stands for average. "SVO" for subject-verb-object. "SUBJ" for subordination conjunctions. "CONJ" for coordination conjuctions".

4 Methods

In this section, we present the architecture details of the two models used in this study, respectively a BERT-based architecture fine-tuned for the task of sentence simplification for Italian and the details of the prompting to Chat-GPT 3.5. In Section 5, we detail the specifics of the BERT-based architecture's fine-tuning and usage used in our experiments.

Proprietary System architecture Our model consists of both an encoder and a decoder component. We employ a BERT-based model fine-tuned for textual simplification tasks. The encoder checkpoints were initialized using pre-trained checkpoints tailored specifically for the Italian language[1] model available in the Hugging Face Hub repository. Conversely, the decoder checkpoints were initialized randomly. When making our architecture choice, it was crucial to consider our target language, namely Italian. At the time of implementing our model, the T5 pre-trained version (Sarti and Nissim, 2022) for Italian was not available. In a second version of our model, we integrated an adaptive component, enabling semi-supervised learning of the model by encoding five numerical values within complex sentences. Following the approach outlined in (Megna et al., 2021; Martin et al., 2019), we incorporated a discrete parametrization mechanism that allows explicit control of the generation. Additionally, we opted to include the Word Ratio parameter proposed by (Sheang and Saggion, 2021). As illustrated in Table 3, these features encompass sentence length (both in terms of characters and tokens), as well as lexical and syntactic complexity. We selected these five parameters because, as highlighted in previous studies, they significantly contribute to the comprehension challenges faced by individuals with reading comprehension deficits (Oakhill and Yuill, 1996; Nation and Snowling, 2000, 2004; Galletti et al., 2023).

LLM architecture To showcase the capabilities of Large Language Models (LLMs), we selected ChatGPT-3.5 (Madaan et al., 2022) due to its proficiency in zero-shot learning scenarios and user-friendly interface accessible through the OpenAI platform, which allows for easy integration and experimentation.

[1] namely the bert-base-italian-xxl-cased

Token	Value	Description
$Word\_Ratio$	0.20	Ratio of words in the complex sentence to words in the simplified sentence.
$Character\_Ratio$	0.20	Ratio of characters in the complex sentence to characters in the simplified sentence.
$Word\_Rank$	0.90	Ranking of words based on frequency or importance.
$Lev\_Similarity$	0.90	Levenshtein similarity between the complex and simplified sentences.
$Dependency\_Tree$	1	Degree of similarity in dependency trees between the complex and simplified sentences.

Table 3: Description of parameters with values used in the adaptive component for simplification.

5 Experiment Settings

This section outlines the parameters for model fine-tuning (Subsection 5.2), and discusses the evaluation metrics (Subsection 5.3) used.

5.1 BERT-based model Fine-Tuning

For the fine-tuning process, we utilized Optuna, an open-source framework for hyperparameter optimization to dynamically build the search space for selecting the optimal parameters for our work. We configured a batch size of 4 for both training and evaluation loops, set a maximum token length of 300, established a learning rate of $3e-4$, configured an Adam epsilon of $1e-8$, implemented a warm-up ratio of 0.10, and conducted 20 epochs. The remaining parameters were kept at their default values from Transformers library. For dividing the three dataset into train and test we used a standard 0.80 split for training and 0.20 for testing. As explained in Section 6, we maintained this fine-tuning parameters for both the two version of our model —the one with the adaptive component and the one without.

5.2 GPT's Prompting

We accessed the *ChatGPT-3.5* model through the open-access model available. For our experiment, we utilized GPT in zero-shot mode. At the time this work was conducted, ChatGPT-3.5 had only very recently been released. As a result, we couldn't fully explore different prompt engineering techniques and we were constrained on relying solely on using -3.5 in a zero-shot mode. Specifically, we presented the model with a list of complex sentences and tasked ChatGPT 3.5 with simplifying them for school children aged 8 to 11 with a reading comprehension deficit. Subsequently, we computed our evaluation scores based on the simplified answers generated by ChatGPT, comparing them to the ground truth provided in our annotated corpus.

5.3 Evaluation Metrics

For assessing the performance of both models, we employed well-established metrics for both automatic machine translation and text simplification evaluations, SARI (Xu et al., 2016) and BLEU (Papineni et al., 2002), on our test corpus. We qualitatively inspected the output data to examine the results from each model. Finally, we involved experts specialised in language disabilities to conduct a human evaluation.

SARI and BLEU were chosen for assessing the performance of both models, because of their use in previous work (Van den Bercken et al., 2019; Monteiro et al., 2022; Cardon and Grabar, 2020). SARI (System-level Automatic Reviewer for Machine Translation) is a metric designed to assess the quality of machine-generated sentences, particularly within the context of machine translation. It centers on evaluating the fluency and preservation of meaning in the generated sentences when compared to reference sentences. In contrast, BLEU (Bilingual Evaluation Understudy) is a widely used metric for evaluating machine-generated sentences, primarily within machine translation contexts. It quantifies the similarity between the generated sentence and one or more reference sentences through an n-gram overlap comparison.

These metrics however have several drawbacks to evaluate text simplification output, as pointed out in the literature (Sulem et al., 2018; Al-Thanyyan and Azmi, 2021). We thus also included qualitative human evaluation of the results by qualitatively inspecting the output data to examine the results from each model. We gathered a panel of experts specialized with domain-specific expertise, i.e. speech and language therapists at a partner specialised center in the rehabilitation of Neurodevelopment[2] to conduct a human evaluation, with a specific focus on young children diagnosed with language disabilities. The criteria for selection was their expertise in language learning and disabilities. All annotators were provided with detailed information regarding the study's purpose, their role in the evaluation and the nature of the data that they were scoring. The experts were not reimbursed financially; however, their participation was voluntary and they were provided with

[2]https://www.crc-balbuzie.it/

informed consent before the beginning of the study.

The evaluation of the quality of the text simplification corpus was made possible through the utilisation of a Google Form available at this link [3]. The form evaluated scales from 0 to 5 (being 0, the lowest and 5, the highest values) concerning grammatical correctness, maintenance of meaning, and level of simplicity gained as similar work in the literature (Xu et al., 2016). We selected 10 sentences to represent both the highest and lowest cosine distances between the sentences generated by ChatGPT-3.5 and our model. Specifically, we selected five pairs with the highest cosine distances and five pairs with the lowest. These sentences have been put at disposal to the ten experts who participated in the users studies. Several considerations prompted this approach: firstly, we needed a manageable sample size feasible for evaluation within our available annotators. Secondly, by including both the most divergent and the most similar cases, we aimed to ensure robustness in extreme scenarios and reduce bias in our evaluation method.

6 Results

In this section we report results on the automatic and human evaluation conducted.

Dataset	SARI	BLEU
Palmero Aprosio et al. (2019)	49.49	N/A
(A) Fine-tuned + Original PACCS-IT Dataset	57.10	46.00
(B) Fine-tuned + Merged and Cleaned Dataset	55.64	49.78
(C) Fine-tuned + Augmented Dataset	51.51	47.40
(D) Fine-tuned + Augmented + Adaptive Component	60.12	50.30
ChatGPT-3.5	40.51	15.00

Table 4: SARI and BLEU scores for all our fine-tuned models with the combinations of the different datasets.

6.1 Automatic Evaluation

In our work, we conducted three different fine-tuning runs using the same fine-tuned model and equivalent hyper-parameters using three different training data, as it can be seen in Table 4. These three models correspond to model (A), (B) and (C) in the table.

The first fine-tuning of the model, i.e. (A), was done using the original version of PaCCSS-IT. It resulted in a SARI score of 57.10 when evaluated on the test corpus. This score was higher than the current state-of-the-art for Italian language Automatic Text Simplification task (Palmero Aprosio et al., 2019). Given the errors manually noticed, it was hypothesized that the high SARI score achieved during fine-tuning resulted from over-fitting to poor-quality data, representing a learning fallacy. To investigate this hypothesis, we fine-tuned our model using the curated version of our dataset, i.e. (B). In this case, SARI fell by two points (55.64). This improvement may be attributed to the inclusion of three merged corpora (Teacher, Terence, and Simpitiki), which provided the model with more diverse material to learn from and thus greater flexibility in the generative phase. The lower SARI value could precisely reflect this behavior and shed light on the previous over-fitting. Following the previous result, we conduct the final fine-tuning with the Augmented dataset,i.e. (C). At this stage, we note that SARI is another 4 percentage points lower than in the last training (51.51). Finally, we fine-tuned an additional model (D), adding the adaptive component detailed in section 4 and using our augmented dataset. Our model obtains a SARI score equal to 60.12 and a BLEU score of 50.30 on the same test set, achieving the best results over the four fine-tuned models.

GPT-3.5 exhibited notably lower performance, yielding a SARI score of 40.51 and a BLEU score of 15.00 on the same test set[4]. GPT-3.5's notably lower performance in comparison to the fine-tuned transformer model within the context of text simplification can be attributed to several factors. Firstly, the fine-tuned transformer model likely benefited from training data and therefore domain adaptation while, GPT-3.5 was used in a zero-shot mode. Secondly, the fine-tuned transformer model's architecture and parameters were carefully adapted to the nuances of text simplification, a step that might have been omitted in GPT's preparation. Finally, the ability to control the level of simplification plays a vital role in effective text simplification (Nishihara et al., 2019; Xu et al., 2016). The nuanced interplay of these factors contributes to the observed performance disparities in the presented context and it pushed us to perform also a qualitative evaluation. All the results are presented in Table 4.

However, despite this good performance on the standard formal metric, a manual and visual evaluation of the generated simplifica-

[3] Click here to access the Google Form

[4] The test corpus is available for exploration via the following link

Input Test Set	Chat GPT	Best Performing Model
A cosa è dovuto ciò What is this due to	Perché è successo? Why did it happen?	A cosa è dovuto? What is it due to?
Beh, questo è tutto un altro discorso *Well, that's a whole other story*	Beh, questa è tutta un'altra storia *Well, that's a whole other story*	*É tutto un altro discorso* *That's a whole other speech*
Ma non ne farei un dramma *But I wouldn't make a drama out of it*	Non dovremmo preoccuparci troppo di questa cosa *We shouldn't worry too much about it*	Ma non faceva un dramma *But it wasn't a drama*
Tutti erano nell'atmosfera della festa *Everyone was in the party mood*	Tutti erano felici e festosi *Everyone was happy and festive*	Era nell ' atmosfera della festa *It was in the party mood*
Ma questo è vero solo in teoria *But this is only true in theory*	Ma questa teoria è vera solo sulla carta *But this theory is only true on paper*	Questo è vero in teoria *This is true in theory*

Table 5: Example of some simplifications on the test corpus given by our model and GPT-3.5.

tions revealed several issues. The simplifications were found to be misleading and contained frequent gross errors. For instance, the reference sentences in the dataset were sometimes not very informative in terms of simplifications and appeared to be alternative versions of the complex sentence but not simplified ones, as shown in Table 6.

Simple	Complex
Questa sarebbe una cosa *positiva* This would be a *positive* thing	Questa è una cosa *gravissima* This is a *very serious* thing
Ma facciamo un passo più *avanti* But let's take a step *forward*	Ma facciamo un lungo passo *indietro* But let's take a long step *backward*

Table 6: The original complex sentences from the test dataset and simplifications produced by the fine-tuned model; highlighting mistakes in italics.

6.2 Human Evaluation

6.2.1 Qualitative Analysis

In a later stage, we inspected the generated simplified sentences given by our models. We found that while the simplification efforts undertaken by ChatGPT-3.5 are generally satisfactory upon close qualitative examination, there are instances where the simplifications verge on being abstract. The generated simplifications at times involve conceptual abstractions that could potentially introduce an unintended layer of complexity as it can be seen in Table 5. This paradoxical outcome could arise because the model simplify, yet occasionally employs abstract concepts that might prove too complex for the intended audience, especially young children or individuals with specific clinical diagnoses. In fact, ChatGPT sometimes seems to capture greater nuances of cause-and-effect or context than an 8- to 11-year-old child who has limited experience of the world and thus may struggle to make such detailed connections, and as a result, the simplification proposed by ChatGPT can sometimes be difficult for children to interpret. For instance, ChatGPT-3.5 might attempt to convey a complex idea by substituting certain words or phrases with simpler alternatives. However, in doing so, it might inadvertently introduce terms that are not within the immediate vocabulary of the target audience or that require a certain level of background knowledge to be fully understood. This kind of simplification could lead to confusion or misinterpretation among individuals who require the content to be presented always in an easily accessible manner.

6.2.2 Experts Evaluation

To complete our qualitative analysis, we asked experts to evaluate the results given by the models. This evaluation yielded mixed results as it can be seen in Figure 3. When we compared the scores of the two models based on the chosen criteria (grammaticality, meaning preservation, and level of simplification), there was not a significant difference between them. This is in contrast to the results of the automatic evaluation, where our fine-tuned transformer model appeared to outperform ChatGPT-3.5 on our test set. This highlighted the fact that we are still lacking an evaluation mechanism that is both objective and aligns closely with human judgment. Without an accurate way to assess the quality of text generated by a simplification model, it becomes challenging to implement effective con-

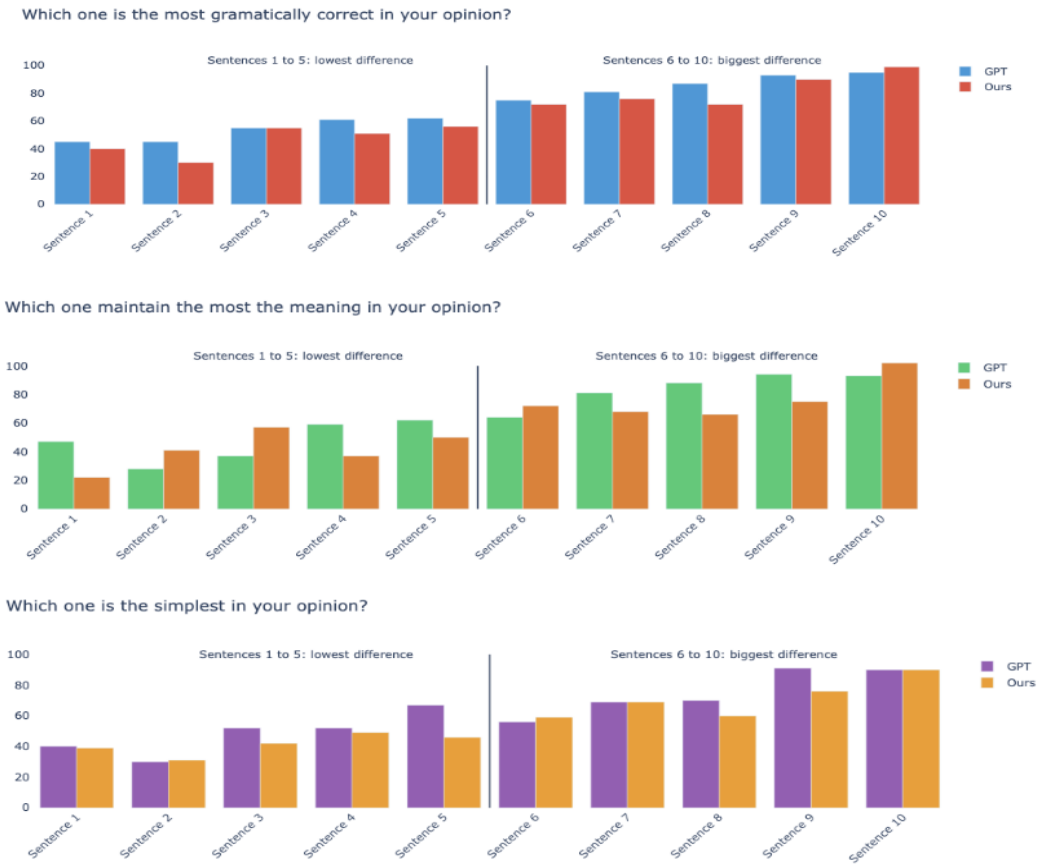

Figure 3: The plots with the form's results. The sentences were re-arranged and their order do not reflects their cosine distances.

trols. This underscores that research in this area is still very much in an experimental stage and is in its early phases.

7 Conclusions and future work

In this paper, we curated a comprehensive corpus by cleaning and combining existing resources, we fine-tuned an adaptive transformer model for the sentence simplification task in Italian, we integrated an adaptable component to tailor simplifications for specific target groups, we evaluated the model by comparing it to ChatGPT-3.5, through both quantitative and qualitative assessments, including expert and automatic evaluations of the simplified text. The automatic evaluation highlighted that the fine-tuned version of BERT model seem the better suited for the task. Moreover the adaptive component increase the State-Of-The-Art (SOTA) results by 11 points. Lastly, LLMs, particularly GPT-3.5, have shown significant advancements in the generation of coherent and fluently articulated text, but a substantial scope for improvement persists concerning the crafting of textual content that aligns effectively with the requisites of individuals possessing particular diagnostic profiles or clinical conditions. This progress can hold promising implications for Computer-Assisted Language Learning, as it can facilitate the creation of tailored educational materials that accommodate the unique learning needs and abilities of diverse student populations. Finally, we believe that there is still much to do to improve the current evaluation metrics for automatic text simplification to understand the nuances and potential biases they may introduce and to make sure they align with human evaluation. Developing and refining new evaluation metrics tailored specifically for populations with diverse linguistic needs and clinical conditions could be a crucial step forward the use of NLP in clinical and educational contexts. Finally, more extensive and robust user studies are required to evaluate the effectiveness of GPT-3.5 in generating text for specific user groups.

References

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 fifth workshop on innovative use of NLP for building educational applications*, pages 1–9.

Sandra M Aluísio, Lucia Specia, Thiago AS Pardo, Erick G Maziero, and Renata PM Fortes. 2008. Towards brazilian portuguese automatic text simplification systems. In *Proceedings of the eighth ACM symposium on Document engineering*, pages 240–248.

Miriam Anschütz, Joshua Oehms, Thomas Wimmer, Bartłomiej Jezierski, and Georg Groh. 2023. Language models for german text simplification: Overcoming parallel data scarcity through style-specific pretraining. *arXiv preprint arXiv:2305.12908*.

Laurens Van den Bercken, Robert-Jan Sips, and Christoph Lofi. 2019. Evaluating neural text simplification in the medical domain. In *The World Wide Web Conference*, pages 3286–3292.

Or Biran, Samuel Brody, and Noémie Elhadad. 2011. Putting it simply: a context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 496–501.

Stefan Bott, Luz Rello, Biljana Drndarević, and Horacio Saggion. 2012. Can spanish be simpler? lexsis: Lexical simplification for spanish. In *Proceedings of COLING 2012*, pages 357–374.

Stefan Bott and Horacio Saggion. 2014. Text simplification resources for spanish. *Language Resources and Evaluation*, 48(1):93–120.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, and Giulia Venturi. 2016. Paccss-it: A parallel corpus of complex-simple sentences for automatic text simplification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 351–361.

Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2022. Linguistically-based comparison of different approaches to building corpora for text simplification: A case study on italian. *Frontiers in Psychology*, 13:707630.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and annotation of the first italian corpus for text simplification. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 31–41.

Rémi Cardon and Natalia Grabar. 2020. French biomedical text simplification: When small and precise helps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 710–716.

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

Björn Engelmann, Fabian Haak, Christin Katharina Kreutz, Narjes Nikzad Khasmakhi, and Philipp Schaer. 2023. Text simplification of scientific texts for non-expert readers. *arXiv preprint arXiv:2307.03569*.

A Fiora, F Piferi, P Crovari, and F Garzotto. 2024. Exploring large language models for the education of individuals with cognitive impairments. In *INTED2024 Proceedings*, pages 4479–4487. IATED.

Núria Gala, Anaïs Tack, Ludivine Javourey-Drevet, Thomas François, and Johannes C Ziegler. 2020. Alector: A parallel corpus of simplified french texts with alignments of misreadings by poor and dyslexic readers. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1353–1361.

Martina Galletti, Eleonora Pasqua, Francesca Bianchi, Manuela Calanca, Francesca Padovani, Daniele Nardi, and Donatella Tomaiuoli. 2023. A reading comprehension interface for students with learning disorders. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, pages 282–287.

Goran Glavaš and Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 63–68.

Isao Goto, Hideki Tanaka, and Tadashi Kumano. 2015. Japanese news simplification: tak design, data set construction, and analysis of simplified text. In *Proceedings of Machine Translation Summit XV: Papers*.

Sichen Guo, François Leborgne, Jun Hu, and Walter Baets. 2024. Can ai bridge the literacy gap? developing a gpt-4 summarization tool for low literacy. *From User to Human*, page 52.

Tomoyuki Kajiwara and Mamoru Komachi. 2018. Text simplification without simplified corpora. *The Journal of Natural Language Processing*, 25:223–249.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh,

Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Akihiro Katsuta and Kazuhide Yamamoto. 2019. Improving text simplification by corpus expansion with unsupervised learning. In *2019 International Conference on Asian Language Processing (IALP)*, pages 216–221. IEEE.

David Kauchak, Jorge Apricio, and Gondy Leroy. 2022. English Datasets resources.

Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. 2022. Language models of code are few-shot commonsense learners. *arXiv preprint arXiv:2210.07128*.

Louis Martin, Angela Fan, Éric De La Clergerie, Antoine Bordes, and Benoît Sagot. 2020. Muss: multilingual unsupervised sentence simplification by mining paraphrases. *arXiv preprint arXiv:2005.00352*.

Louis Martin, Benoît Sagot, Eric de la Clergerie, and Antoine Bordes. 2019. Controllable sentence simplification. *arXiv preprint arXiv:1910.02677*.

Angelo Luigi Megna, Daniele Schicchi, Giosué Lo Bosco, and Giovanni Pilato. 2021. A controllable text simplification system for the italian language. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 191–194. IEEE.

José Monteiro, Micaela Aguiar, and Sílvia Araújo. 2022. Using a pre-trained simplet5 model for text simplification in a limited corpus. *Proceedings of the Working Notes of CLEF*.

Kate Nation and Margaret J Snowling. 2000. Factors influencing syntactic awareness skills in normal readers and poor comprehenders. *Applied psycholinguistics*, 21(2):229–241.

Kate Nation and Margaret J Snowling. 2004. Beyond phonological skills: Broader language skills contribute to the development of reading. *Journal of research in reading*, 27(4):342–356.

Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. Controllable text simplification with lexical constraint loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266, Florence, Italy. Association for Computational Linguistics.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 85–91.

JANE Oakhill and N Yuill. 1996. Reading comprehension difficulties. *Hillsdale, NJ*.

Gustavo Paetzold and Lucia Specia. 2015. Lexenstein: A framework for lexical simplification. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 85–90.

Alessio Palmero Aprosio, Sara Tonelli, Marco Turchi, Matteo Negri, and A Di Gangi Mattia. 2019. Neural text simplification in low-resource conditions using weak supervision. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation (NeuralGen)*, pages 37–44. Association for Computational Linguistics (ACL).

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.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2020. Lexical simplification with pretrained encoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8649–8656.

Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. 2013. Simplify or help? text simplification strategies for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, pages 1–10.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):1–36.

Gabriele Sarti and Malvina Nissim. 2022. It5: Large-scale text-to-text pretraining for italian language understanding and generation. *arXiv preprint arXiv:2203.03759*.

Carolina Scarton and Lucia Specia. 2018. Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718.

Max Schwarzer, Teerapaun Tanprasert, and David Kauchak. 2021. Improving human text simplification with sentence fusion. In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, pages 106–114.

Chantal Shaib, Millicent L Li, Sebastian Joseph, Iain J Marshall, Junyi Jessy Li, and Byron C Wallace. 2023. Summarizing, simplifying, and synthesizing medical evidence using gpt-3 (with varying success). *arXiv preprint arXiv:2305.06299*.

Matthew Shardlow. 2014a. Out in the open: Finding and categorising errors in the lexical

simplification pipeline. In *LREC*, pages 1583–1590.

Matthew Shardlow. 2014b. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Kim Cheng Sheang and Horacio Saggion. 2021. Controllable sentence simplification with a unified text-to-text transfer transformer. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 341–352, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Advaith Siddharthan. 2011. Text simplification using typed dependencies: A comparision of the robustness of different generation strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 2–11.

Sanja Štajner. 2021. Automatic text simplification for social good: Progress and challenges. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Bleu is not suitable for the evaluation of text simplification. *arXiv preprint arXiv:1810.05995*.

Renliang Sun, Wei Xu, and Xiaojun Wan. 2023. Teaching the pre-trained model to generate simple texts for text simplification. *arXiv preprint arXiv:2305.12463*.

Sara Tonelli, Alessio Palmero Aprosio, and Marco Mazzon. 2017. The impact of phrases on italian lexical simplification. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, pages 316–320.

Sara Tonelli, Alessio Palmero Aprosio, and Francesca Saltori. 2016. Simpitiki: a simplification corpus for italian. In *CLiC-it/EVALITA*.

Sander Wubben, Antal Van Den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024.

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.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from wikipedia. *arXiv preprint arXiv:1008.1986*.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. *arXiv preprint arXiv:1703.10931*.

Sanqiang Zhao, Rui Meng, Daqing He, Saptono Andi, and Parmanto Bambang. 2018. Integrating transformer and paraphrase rules for sentence simplification. *arXiv preprint arXiv:1810.11193*.