# Controllable Sentence Simplification in Italian: Fine-Tuning Large Language Models on Automatically Generated Resources

**Michele Papucci**[1,2], **Giulia Venturi**[2], **Felice Dell'Orletta**[2]
[1]University of Pisa
[2]ItaliaNLP @ Institute for Computational Linguistics "A. Zampolli" (CNR-ILC), Pisa
michele.papucci@phd.unipi.it, {giulia.venturi, felice.dellorletta}@ilc.cnr.it

**Abstract**

This paper presents a study on readability-controlled Sentence Simplification for Italian, addressing the scarcity of annotated resources for low-resource languages. We introduce IMPaCTS (Italian Multilevel Parallel Corpus for Text Simplification), the first fully automatically created corpus of 1,444,160 original–simple sentence pairs automatically annotated with readability levels and linguistic features. It was generated using an Italian LLM prompted in zero-shot to produce multiple simplifications per input sentence. Increasing portions of the resource are used to fine-tune mono- and multilingual open-weight LLMs, conditioning them to generate simplifications at a target readability level. Results from automatic and human evaluations show that fine-tuning on IMPaCTS improves performance both in terms of task completion and adherence to the targeted readability levels compared to few-shot baselines.

**Keywords:** Controlled Sentence Simplification, Readability Assessment, Large Language Models

## 1. Introduction

Text Simplification (TS) and Sentence Simplification (SS) aim to rewrite documents or sentences to make them easier to read and understand, typically for a specific target audience, such as language learners, children, or readers with cognitive difficulties. Among the many research directions in TS, a particularly relevant line concerns Controlled Text Simplification (CTS), which focuses on generating text that matches a specified target difficulty level (Scarton and Specia, 2018). In this paradigm, TS models are conditioned during training by feeding parameter tokens as additional inputs, alongside the source input, typically the sentence. Control can be applied at a high level, for example by specifying a target reading grade (Agrawal and Carpuat, 2023; Barayan et al., 2025), or at a more detailed level, by indicating the types of simplification operations to perform (Martin et al., 2020).

Despite these recent advances, progress in CTS remains constrained by the scarcity of large parallel corpora also annotated with reading grade levels or simplification parameters, which limits the accuracy of fully supervised training approaches, particularly for languages other than English. Thus, recently, pre-trained Large Language Models (LLMs) have been explored for zero-shot (Farajidizaji et al., 2024) and few-shot CTS without task-specific training (Barayan et al., 2025; Scalercio et al., 2025), alongside evaluation initiatives aimed at benchmarking LLMs' ability to generate simplified texts (Ryan et al., 2023; Kew et al., 2023).

Building on these premises, this work explores the potential of using *i)* an Italian LLM in a zero-shot setting to automatically generate a parallel resource containing, for each input sentence, multiple simplifications along a gradient of readability levels, and *ii)* the resulting resource to fine-tune six mono- and multilingual open-weight LLMs of different sizes for readability-controlled sentence simplification on increasing amounts of the resource to enhance the models' ability to generate simplifications at target readability levels across different domains. Our main contributions are:

- we introduce **IMPaCTS** (Italian Multilevel Parallel Corpus for Text Simplification), the first Italian **fully automatically created corpus** of 1,444,160 original–simple sentence pairs, including multiple simplifications for the same input sentence with increasing levels of complexity, **automatically annotated with readability levels, and linguistic features**, covering two domains: Wikipedia and administrative texts;
- we demonstrate to what extent **increasing portions of IMPaCTS**, used to fine-tune open-weight LLMs, while conditioning them to generate simplifications at specified readability levels, improves their performance over few-shot in-context learning;
- we provide evidence of alignment between human perception and readability-controlled sentence generation.

## 2. Related Works

In recent years, performance on automatic simplification has advanced considerably with the introduction of CTS. However, several key challenges



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 7178–7191
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

remain open. The first concerns the limited availability of parallel resources explicitly annotated for reading grade levels or simplification types, a limitation that is especially critical for low-resource languages. Existing datasets, such as Newsela (Xu et al., 2015) and the CEFR-based Sentence Profile corpus (Arase et al., 2022), based on the Common European Framework of Reference for Languages, represent notable exceptions, yet are restricted to English. In response to this limitation, research in other languages has explored the automatic creation of parallel resources, often drawing inspiration from paraphrase generation and machine translation. For instance, MUSS (Martin et al., 2022) introduced a controllable sequence-to-sequence model trained on English, Spanish, and French by conditioning on simplification-specific control tokens; for German, transformer-based models have been trained on corpora obtained by automatically aligning pre-existing resources, and conditioned on target CEFR levels (Spring et al., 2021). However, to the best of our knowledge, none of the existing studies systematically investigates the amount of training data required to fine-tune LLMs for achieving reliable CTS performance compared to baseline models. In our opinion, this represents a crucial area of investigation, particularly for low-resource languages, since understanding how data size affects fine-tuning outcomes is key to optimizing both computational effort and resource creation.

Another central issue concerns the definition of the control variable that conditions the generation process (Agrawal and Carpuat, 2023). Most studies have conditioned models on grade levels (Scarton and Specia, 2018; Nishihara et al., 2019; Yanamoto et al., 2022; Kew and Ebling, 2022), while others have employed CEFR levels (Barayan et al., 2025) or focused on grammatical (Martin et al., 2020) and lexical operations (Agrawal et al., 2021; Zetsu et al., 2022; Sheang et al., 2022). Fewer works have explored the use of readability scores as control variables (Farajidizaji et al., 2024).

Concerning the Italian language, which is the main focus of our study, the first work on CTS was conducted by Miliani et al. (2022), who adopted machine translation approaches to build new parallel resources for fine-tuning mT5 (Xue et al., 2021), conditioned on the same control tokens as Sheang et al. (2022), corresponding to grammatical and lexical properties involved in sentence simplification. Machine translation methods were also employed by Palmero Aprosio et al. (2019) to augment existing datasets for training and testing a neural TS model, although without conditioning. In addition, for benchmarking purposes, Nozza and Attanasio (2023) and Russodivito et al. (2024) evaluated the ability of both open and proprietary LLMs to generate simplified sentences in a zero-shot setting, without explicit control mechanisms.

## 3. The Italian Multilevel Parallel Resource for Text Simplification

IMPaCTS was created through an automatic generation process consisting of two steps, building on the previous work on LLM-based generation of multi-level simplifications described by Papucci et al. (2025). In the first step, **we selected the model that most reliably generates a single simplification per input in a zero-shot setting.** We evaluated three models fine-tuned on Italian, differing in architecture and number of parameters: ANITA[1] (Polignano et al., 2024), LLaMAntino-2[2] (Basile et al., 2023), and Italia[3]. Each model was prompted using its system prompt and a shared task-specific prompt to simplify the text while preserving the original meaning[4]. They were tested on a set of 1,212 original–simplified sentence pairs extracted from the test-sets of existing Italian corpora spanning across multiple textual genres and domains. The set includes: 51 pairs from the Wikipedia portion of SIMPITIKI (Tonelli et al., 2016), 994 pairs from PaCCSS–IT (Brunato et al., 2016), 101 pairs from the *Terence* corpus, 17 pairs from the *Teacher* corpus (Brunato et al., 2015), and 49 pairs from ADMIN-it (Miliani et al., 2022). Results were evaluated based on automatic metrics typically used in for TS: BLEU (Papineni et al., 2002), SARI (Xu et al., 2016), BERTScore (Zhang et al., 2020), and SentenceTransformer Similarity (Reimers and Gurevych, 2019, 2020). A further important aspect of text simplification is the difference in readability between original and simplified sentences, which we measured using READ-IT (Dell'Orletta et al., 2011), a tool for automatic readability assessment of Italian documents and sentences that relies on lexical and (morpho-)syntactic parameters of text. Results are reported in Table 1. As it can be seen, **LLaMAntino-2 is the best-performing model according to all evaluation metrics** and was therefore selected for the second step of our approach. Specifically, sentences generated by LLaMAntino-2 obtained higher scores on all metrics, indicating that they are simpler than the original sentences. READ-IT scores, which range from 0 (most readable) to 1 (least readable), show the opposite trend: lower READ-IT values for LLaMAntino-2 generated sentences correspond to less linguistically complex sentences.

The second step of our methodology consisted of prompting LLaMAntino-2 to generate multiple

[1]swap-uniba/LLaMAntino-3-ANITA-8B-Inst-DPO-ITA

[2]swap-uniba/LLaMAntino-2-7b-hf-dolly-ITA

[3]iGeniusAI/Italia-9B-Instruct-v0.1

[4]The full prompt is reported in Appendix A.

| Model | SARI ↑ | Bleu ↑ | BertScore ↑ | SentenceTransformer ↑ | READ-IT ↓ |
|---|---|---|---|---|---|
| ANITA | 39.35 | 0.07 | 0.80 | 0.62 | 54.1 $\pm$ 31.63 |
| LLaMAntino-2 | **40.99** | **0.18** | **0.81** | **0.64** | **53.11** $\pm$ 33.01 |
| Italia | 39.35 | 0.12 | 0.79 | 0.57 | 58.43 $\pm$ 30.16 |

Table 1: Automatic evaluation scores for the zero-shot tested models.

simplifications of varying linguistic complexity for each original sentence, using the same prompt as in the previous step. We employed the Divergent Beam Search decoding technique (Vijayakumar et al., 2016) and, through manual inspection of outputs under different settings, found that using 20 beams divided into 10 groups with a diversity penalty of $\lambda = 0.7$ provided the best balance between diversity of simplifications and text fluency. To ensure linguistic diversity in IMPaCTS, we used two corpora representative of different varieties of Italian as input: a collection of sentences from Wikipedia pages, and the PaWaC – Public Administration Web as Corpus (PaWaC (Passaro and Lenci, 2019)), which contains a wide range of administrative texts. We randomly extracted ~100,000 sentences from Wikipedia, and ~40,000 from PaWaC as input for generating multiple simplifications. At the end of the process, we obtained 10 simplifications for each original sentence that was automatically revised by removing duplicate simplifications and cases where the original and simplified sentences were identical. Consequently, not all input sentences have 10 simplifications; the number of simplifications per sentence ranges from 2 to 10, with an average of 9.6. The final IMPaCTS corpus contains **1,058,960 original–simplified sentence pairs for Wikipedia and 385,200 pairs for PaWaC**[5]. Table 2 reports an example extracted from IMPaCTS. All pairs were enriched with readability scores computed using READ-IT and a set of 144 linguistic features modeling multiple linguistic phenomena as described in Section 3.1. As a general remark, we observed, as expected, that some original sentences were split[6]. However, the average number of splits remains low (1.31 in Wikipedia and 1.47 in PaWaC), suggesting that most simplifications were achieved without splitting, but rather through lexical or (morpho-)syntactic transformations.

## 3.1. Readability and Linguistic Profile

A distinctive feature of IMPaCTS is that it is automatically annotated with readability levels and linguistic features, whose distributions are discussed in the following paragraphs.

**Levels of Readability.** To define the levels, we ordered both the inputs and the corresponding all generated simplifications according to their READ-IT scores. Importantly, this ranking was performed independently within each set of simplifications derived from the same input sentence, rather than globally across the corpus. From these ordered sets, we then selected five representative simplifications: the most simplified (*s_1*) and the least simplified (*s_5*), together with three additional simplifications randomly chosen and positioned according to increasing readability (e.g. *s_2* has a lower READ-IT score than *s_3*). To analyze the resulting distributions, we adopted Kernel Density Estimation (KDE), a probability distribution estimate obtained by smoothing the READ-IT data points to create a continuous probability curve. Results are reported in Figure 1, separately for the Wikipedia and PaWaC sections of IMPaCTS. In both cases, original sentences show a higher frequency of data points with higher READ-IT scores, while the **simplified sentences are generally concentrated towards the lower end of the readability spectrum**. Since the selection of *s_1* and *s_5* was performed independently for each original sentence, this pattern provides empirical evidence that *s_1* sentences are consistently easier to read across IMPaCTS. Moreover, the results suggest the presence of a readability gradient across generated simplifications: sentences in the *s_5* set (least simplified) tend to have higher READ-IT scores, whereas those in the *s_1* set (most simplified) are more concentrated at the lower end of the spectrum. Interestingly, variations of this general trend emerge across the two IMPaCTS sections. As expected, **PaWaC original sentences are more concentrated at the higher end of the readability spectrum**, reflecting the well-known linguistic complexity of administrative language (Cortelazzo, 2021). Consequently, many of the PaWaC simplifications, while still following the overall gradient of decreasing complexity, exhibit higher READ-IT scores than Wikipedia simplifications, suggesting that they remain less accessible on average.

**Variation of Linguistic Profile.** It was computed accounting for the variation of a set of 144 features covering multiple aspects of the linguistic profile of a sentence from raw-text properties such as sentence length, to the distribution of Parts of Speech,

[5]The full resource and a dataset exploration notebook are available on the following repository: `https://github.com/michelepapucci/impacts`

[6]Note that in these cases readability and linguistic profiling were computed as the average across the resulting simplified sentences.

| | Wikipedia | READ-IT |
|---|---|---|
| Original | Gli effetti sulla salute del particolato atmosferico sono opportunamente distinti in effetti a breve termine (acuti) ed a lungo termine (cronici) (The effects on health of atmospheric particulate matter are appropriately distinguished into short-term (acute) and long-term (chronic) effects.). | .66 |
| Simplifications | Gli effetti sulla salute del particolato atmosferico sono distinti in effetti a breve termine (acuti) ed a lungo termine (cronici). (The effects on health of atmospheric particulate matter are distinguished into short-term (acute) and long-term (chronic) effects.) | .60 |
| | La polvere atmosferica ha effetti a breve termine (acuti) e a lungo termine (cronici) sulla nostra salute. (The atmospheric dust has short-term (acute) and long-term (chronic) effects on our health.) | .34 |
| | L'inquinamento atmosferico ha effetti acuti e cronici sulla salute. (The atmospheric pollution has acute and chronic effects on health.) | .07 |

Table 2: Example from the Wikipedia section of IMPaCTS showing three simplifications of the same original sentence, along with their corresponding READ-IT scores.

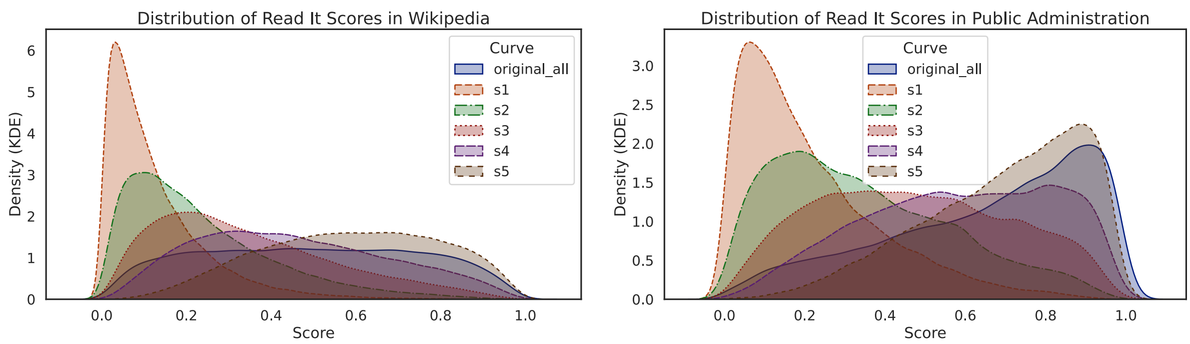


Figure 1: Kernel Density Estimate for READ-IT scores in Wikipedia and PaWaC sections of IMPaCTS.

and to features modeling both local and global syntactic structures, such as syntactic tree depth, syntactic relation length, subordination, the ordering of verb nucleus elements (subject and object)[7]. They were extracted using Profiling-UD (Brunato et al., 2020), a web-based tool designed to linguistically profile multilingual texts using the Universal Dependencies (UD) formalism (De Marneffe et al., 2021). To quantify overall variation, we applied Multivariate Analysis of Variance (MANOVA), which tests whether the mean vectors of multiple dependent variables differ significantly between groups. Results are reported in Figure 2 in terms of Pillai's Trace, one of MANOVA's statistics particularly robust when assumptions such as homogeneity of covariance matrices may be violated[8]. As it can be noted, Pillai's Trace values increase with the distance between compared groups, indicating larger multivariate differences: the **highest values are observed between original sentences and the most simplified ones** (*s_1*), while the lowest occur between original and least simplified sentences (*s_5*), with intermediate values distributed across the other simplification levels. This seem to confirm the presence of a gradient of linguistic variation across simplifications. Despite this general trend, differences emerge across the two IMPaCTS sections: variation is higher for PaWaC, suggesting that **administrative sentences undergo more substantial transformations** during simplification than Wikipedia sentences. Nevertheless, as discussed above, PaWaC sentences remain linguistically more complex on average in terms of readability scores.

For each of the two IMPaCTS sections, Table 3 reports the values of a small subset of these features, which are commonly associated with linguistic complexity in text simplification studies (Saggion, 2017)[9]. As expected, feature values generally decrease from the original to the most simplified sentences (*Simp_1*). For most features, PaWaC original sentences show higher values than Wikipedia ones, and their variations towards *Simp_1* are higher, suggesting that **simplification was more incisive in the administrative domain, even though the simplified sentences still remain more complex** than the Wikipedia simplified sentences. Specifically, sentence length progressively shortens across simplifications; syntac-

[7]The full set of features is detailed in Appendix B.

[8]All comparisons yield statistically significant differences ($p \leq 10^{-4}$).

[9]All comparisons are statistically significant according to the Wilcoxon signed-rank test (with $p < 0.05$).

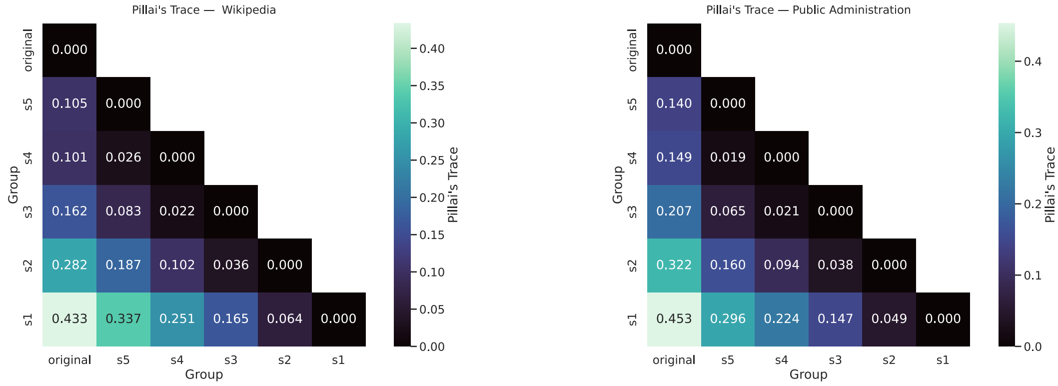


Figure 2: Pillai's Trace of MANOVA on original and simplified sentences in Wikipedia and PaWaC.

| Feature | Original | Simp_5 | Simp_4 | Simp_3 | Simp_2 | Simp_1 |
|---|---|---|---|---|---|---|
| Sentence length | 31.98/48.14 | 27.71/36.10 | 26.31/34.12 | 23.64/30.55 | 20.32/25.92 | 16.82/21.39 |
| Depth of the synt. tree | 5.44/7.48 | 5.01/6.32 | 4.83/6.11 | 4.51/5.72 | 4.11/5.21 | 3.67/4.66 |
| Sub. clauses | 42.64/53.36 | 39.28/46.32 | 35.82/43.35 | 31.21/38.76 | 25.32/32.19 | 19.26/26.11 |
| Post-verbal subjects | 11.66/17.65 | 7.99/12.02 | 8.10/12.14 | 7.24/11.58 | 6.20/9.65 | 4.82/7.57 |

Table 3: Subset of linguistic features involved in simplification (values for Wikipedia/PaWaC).

tic trees (*Depth of the synt. tree*) become shallower, reducing structural complexity (Frazier, 1985); subordinate clauses (*Sub. clauses*) become less frequent; and explicit subjects occur less frequently in non-canonical post-verbal positions, moving toward the canonical SVO order preferred in simplification, as this facilitates comprehension in free word-order languages (Haspelmath, 2006).

Overall, these results show that the generation strategy we devised produces simplifications that undergo systematic lexical and syntactic transformations, progressively reducing linguistic complexity with respect to the original sentences. This suggests that the resulting IMPaCTS corpus captures meaningful linguistic variation across readability levels, making it a suitable resource for fine-tuning models for Readability-Controlled Sentence Simplification.

## 4. Controlled Sentence Simplification

### 4.1. Experimental Settings

**Models.** We tested two families of open-weight models, Minerva (Orlando et al., 2024) and Qwen-3 (Bai et al., 2023), each in three different sizes, for a total of six models. Each model was fine-tuned starting from the base model, i.e. the model that only went through the pre-training. Table 4 reports the model sizes in terms of parameter count (in billions) and the amount of tokens used for pre-training (in trillions). We selected Minerva because it is trained from scratch on Italian, while Qwen-3 was chosen as the latest version of the Qwen family, which has shown remarkable performance in Sentence Simplification among open-weight multilingual models (Scalercio et al., 2025). As shown in the table, while the two families are roughly comparable in terms of parameters, they differ substantially in terms of pre-training resources.

**Baselines.** To directly quantify the performance gains achieved through fine-tuning on the IMPaCTS resource, we included a set of reference models evaluated without any fine-tuning, in a few-shot in-context learning scenario: Minerva 7B and Qwen 4B in their instructed versions, and LLaMAntino-2, which was the model used to generate the IMPaCTS dataset. All reference models were prompted to generate simplified sentences at specific READ-IT scores, using three examples randomly extracted from IMPaCTS[10].

**Fine-tuning.** All models were fine-tuned with increasing amounts of IMPaCTS sentences, conditioning them to generate new sentences at a specified readability level expressed as a READ-IT score. To this end, we introduced a control token `<|readability_value|>` at the beginning of each original complex sentence, so that the models were instructed to simplify the sentence according to the requested readability level. We opted for this single-token format, rather than separating the natural-language control from the numerical value, following Li and Shardlow (2024), who demonstrated that this is the best-performing format for sentence simplification. The models were conditioned for 20 readability levels $(0, 5, 10, \ldots, 100)$, corresponding to READ-IT scores from $0$ to $1$ in bins of $0.05$.

[10]The full prompt used for the baseline models is reported in Appendix C.

| Model | #Parameters | #Tokens |
|---|---|---|
| Minerva | 1B | 0.2T |
| | 3B | 0.66T |
| | 7B | 2.48T |
| Qwen 3 | 1.7B | 36T |
| | 4B | 36T |
| | 8B | 36T |
| Llamantino 2 | 7B | 2T |

Table 4: Models used for fine-tuning.

For each level, we added a dedicated control token (e.g. `<|readability_0|>`, `<|readability_5|>`) to the tokenizer and resized both the embedding layer and the language-modeling head. Since these new tokens had no pretrained representations, we kept the embedding and unembedding layers unfrozen during the fine-tuning with LoRA (Hu et al., 2022), allowing the models to learn embeddings for the new control tokens. Each model was fine-tuned on progressively larger subsets of IMPaCTS: 500, 1k, 2,5k, 5k, 10k, and 25k original sentences. For each split of originals, we considered all their associated simplifications across the 20 readability levels[11].

**Automatic evaluation metrics.** We trained models separately on each IMPaCTS section using the incremental bins, obtaining six versions of each model per section. Training was performed for three epochs, and evaluation was carried out on a test set of 100 original sentences, corresponding to 972 pairs in the Wikipedia section and 971 pairs in the PaWaC section. As evaluation metrics we used BLEU (Papineni et al., 2002) and SARI (Xu et al., 2016) for surface-level properties of writing style, and BertScore (Zhang et al., 2020) computed, following Li and Shardlow (2024), between the IMPaCTS reference simplification and the model-generated sentence, to assess whether the semantic content of the target sentence was preserved. In addition, to assess compliance with the readability constraint, we computed the Root Mean Squared Error (RMSE) between target and generated READ-IT scores, and the Spearman correlation coefficient between the two scores, measuring how well models capture increasing or decreasing readability trends across controls.

## 4.2. Does Fine-Tuning Improve Controlled Simplification?

Results are reported in Figure 3 and illustrate the impact of fine-tuning across models, domains, and evaluation metrics. For both tested domains, **Qwen models consistently outperform both Minerva and reference models on surface-level metrics (BLEU and SARI)**, achieving strong performance even **with only 500 fine-tuning instances**[12] and remaining stable as the training size increases. In contrast, **Minerva models require to be fine-tuned on a larger amount of data** to outperform all the reference models and to reach performance comparable to the Qwen models, showing gradual gains across incremental fine-tuning instances. Note that, among the few-shot models, LLaMAntino-2 consistently outperforms the others, representing, as expected, a particularly challenging baseline, as it is the same model used to generate IMPaCTS and may therefore produce sentences highly similar to those included in the fine-tuning data. However, all results have to be considered in light of the data reported in Table 5, which shows that Minerva models frequently fail to generate any output, particularly when constrained on PaWaC, which may contribute to their unstable performance. Notably, Minerva 1B outperforms the other Minerva models on Wikipedia, whereas on PaWaC it turns out to be the worst-performing Minerva model and declines in performance after 10k training sentences. Overall, BLEU and SARI scores tend to be higher for the administrative domain, possibly because PaWaC originals undergo stronger structural transformations than Wikipedia sentences (see Section 3.1), enabling models trained on PaWaC pairs to generalize better. Similar patterns are observed for semantic similarity (BertScore). Across both domains, all Qwen models achieve higher scores than Minerva and the reference models, even when fine-tuned on only 500 instances, and remain stable across training. However, the improvement over the few-shot reference models is smaller than that observed for BLEU and SARI, suggesting that fine-tuning mainly enhances surface-level simplification. This trend is particularly evident for the Minerva models, which struggle to outperform the reference models on PaWaC, especially 1B, whose performance consistently remains below the few-shot models.

## 4.3. Can Fine-Tuned Models Follow Readability Controls?

According to Figure 4, the **Qwen models are the most effective at following the target READ-IT levels** across both domains, showing the lowest RMSE and high correlation scores. On Wikipedia, they generate sentences with an average deviation of about 0.22 READ-IT points from the target, surpassing all few-shot reference models after

[11]This corresponds to a total of 4,884, 9,766, 24,423, 48,888, 97,806, and 244,363 fine-tuning pairs in the Wikipedia section and of 4,831, 9,433, 23,714, 47,327, 95,006, and 237,953 in the PaWaC section.

[12]Note that for all experiments training set size refers to the number of original sentences, each paired with simplifications at all 20 READ-IT levels.

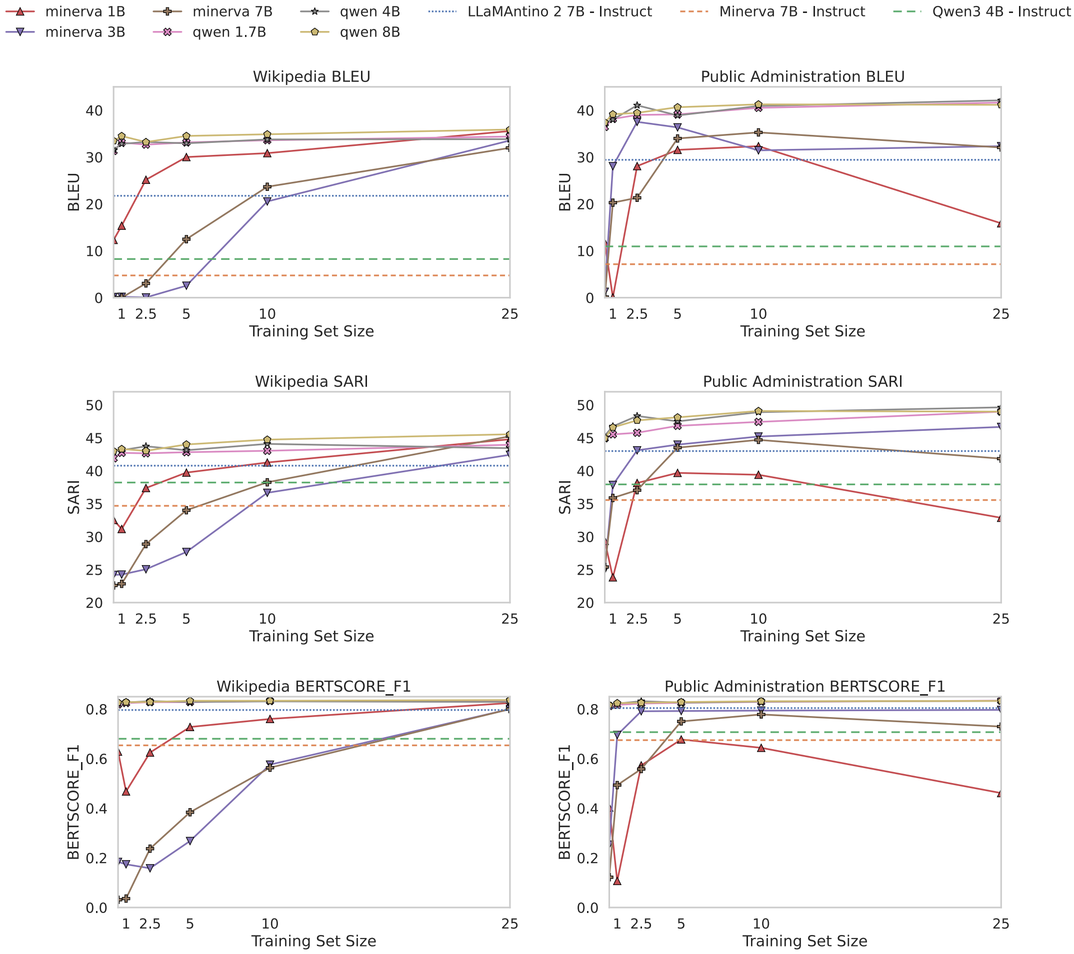


Figure 3: Effect of increasing fine-tuning set size on Controlled Sentence Simplification performance (BLUE, SARI and BertScore) in the Wikipedia and PaWaC test sets.

| | Training set size | | | | | |
|---|---|---|---|---|---|---|
| | 500 | 1000 | 2500 | 5000 | 10000 | 25000 |
| | Wikipedia | | | | | |
| 1B | 44.49 | 86.30 | 29.45 | 12.77 | 17.20 | 38.72 |
| 3B | 14.01 | 2.99 | 0.31 | 0.72 | 0.21 | 0.00 |
| 7B | 84.96 | 39.55 | 31.10 | 8.14 | 4.53 | 8.75 |
| | PaWaC | | | | | |
| 1B | 11.11 | 37.45 | 21.91 | 9.88 | 5.86 | 1.03 |
| 3B | 49.69 | 33.95 | 71.30 | 61.73 | 29.42 | 1.95 |
| 7B | 95.68 | 95.58 | 71.30 | 53.91 | 31.48 | 2.78 |

Table 5: Percentage of empty responses of the Minerva models.

2,5k fine-tuning instances, with Qwen 8B achieving lower error after just 1k instances. Then, RMSE scores remain stable as fine-tuning size increases. **On PaWaC, a larger fine-tuning set is required**, with all Qwen models needing 10k instances to reach comparable error rates and stabilize. **Minerva models, in contrast, struggle more to follow the readability control**. On Wikipedia, their RMSE scores remain above the non-fine-tuned Minerva 7B and Qwen 4B models, with Minerva 1B performing best among them despite frequent generation failures (Table 5). On PaWaC, performance is slightly lower, yet consistently above reference models. Similar to Qwen, Minerva models require about 10k instances to achieve a nearly stable RMSE. Our best performing model, Qwen 3 7B - 25k, achieves an RMSE of just 0.2, meaning that the average error is only about $\pm$ 4 control tokens w.r.t. the asked value. Correlation results reflect these trends. On Wikipedia, Minerva models outperform few-shot references after 2,5k instances, while on PaWaC they struggle to surpass the LLaMAntino-instruct model, with correlation scores fluctuating around LLaMAntino's line. This suggests that the administrative domain is more challenging for maintaining control over readability.

## 5. Human Evaluation

We randomly sampled 150 sentence pairs generated by the Qwen3-8B model fine-tuned on 25k IMPaCTS instances from the Wikipedia section, as it was the best-performing model. The sample included i) 25 pairs composed of 1 original sentence and 1 randomly selected simplification, and ii) 125

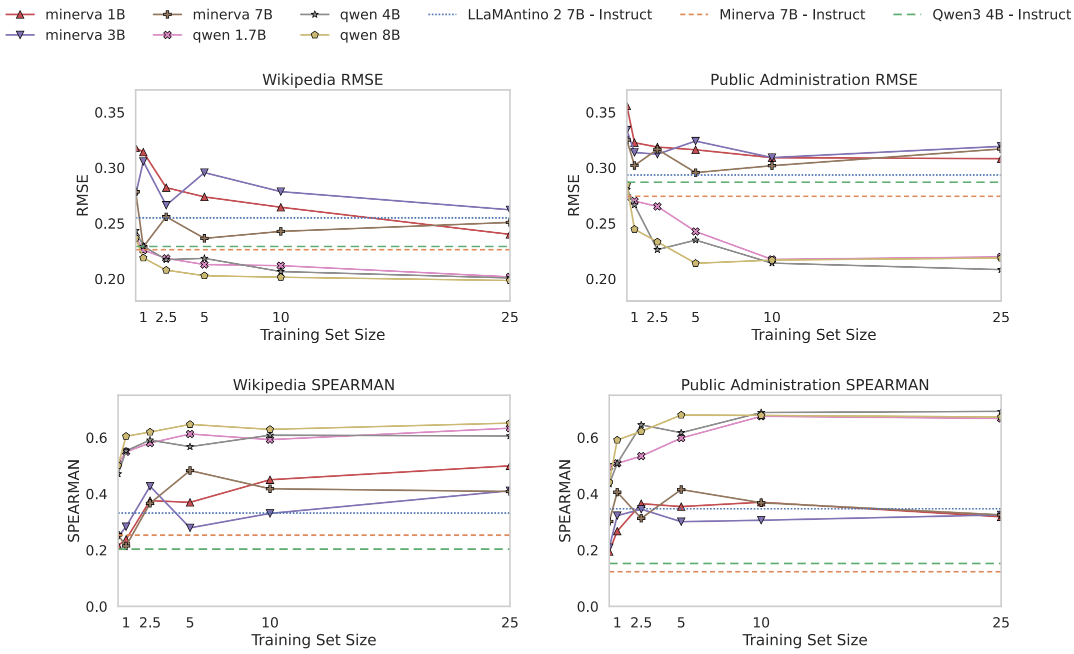


Figure 4: RMSE and Spearman correlation between the READ-IT scores of target and output simplified sentences on the Wikipedia and PaWaC test sets, across increasing fine-tuning sizes.

pairs of 2 simplifications, differing by at least 0.25 READ-IT points, generated from the same original sentence. The 150 pairs were randomly divided into 5 questionnaires of 30 pairs each, administered to 5 distinct annotators recruited via Prolific, all native speakers of Italian, for a total of 25 annotators[13]. For each pair, annotators answered 4 binary questions assessing: 1) which sentence was simpler ("Is sentence 1 simpler than sentence 2?"), 2) whether the two sentences conveyed the same meaning ("Do the two sentences express the same meaning?"), and 3–4) the grammatical correctness of each sentence ("Is sentence 1 grammatically correct? " / "Is sentence 2 grammatically correct?"). The collected annotations were analysed from multiple perspectives.

**Inter-annotator agreement.** It was computed as Krippendorff's $\alpha$ across all annotators for each of the four questions. The highest agreement was observed for Question 1 ($\alpha = 0.50$), which can be considered moderate given the intrinsic subjectivity of the task. A similar agreement was found for Question 2 ($\alpha = 0.46$), suggesting that the generated sentences generally maintain their original content. In contrast, a lower agreement was observed for the grammaticality judgments ($\alpha = 0.32$). A manual inspection suggests that this may be due to surface-level issues, such as sentences beginning with conjunctions or lowercase letters, or lacking proper punctuation, which may have confused some annotators.

**Alignment between human and READ-IT scores.** To assess whether automatic readability values align with human perception of simplicity, we compared the majority vote of annotators' responses to Question 1 with the ranking derived from the READ-IT scores of each pair. Specifically, we considered a match when the READ-IT value of the first sentence was lower than that of the second and the majority vote was "yes", i.e. when the first sentence is perceived as simpler than the second. The comparison yielded an F1-score and accuracy of 0.74 and a Cohen's $\kappa$ of 0.48, thus suggesting that READ-IT scores are consistent with human judgments of simplicity and may be considered as a reliable way to control LLMs in generating sentences at targeted readability levels.

**Sensitivity to readability differences.** Notably, we observed that some sentence pairs were more divisive among annotators regarding Question 1. To investigate this, for each pair, we computed the Spearman's correlation score between the percentage of "yes" answers (indicating that annotators judged the first sentence simpler) with the difference between the two sentences in each pair, both in target and effective READ-IT scores. Our intuition for a positive correlation is that a higher percentage of positive answers, reflecting stronger annotator agreement, should correspond to a larger readability difference between the two sentences. The results support this intuition, with positive correlations in both cases: $\rho = 0.46$ ($p = 0.02$) for controlled READ-IT differences and $\rho = 0.48$ ($p = 0.01$)

[13]Annotators were compensated £7.50/h, with an estimated completion time of 20 min.; the observed average completion time was 17 minutes and 50 sec.

for effective READ-IT differences. This suggests that annotators tended to agree more when the gap in either controlled or effective readability between two sentences was larger. Conversely, when sentences were closer in readability, human perception of simplicity became less consistent.

## 6. Conclusion

We introduced IMPaCTS (Italian Multilevel Parallel Corpus for Text Simplification), the first fully automatically created Italian corpus of original–simplified sentence pairs. The resource is unique in that it provides multiple simplified versions for each original sentence, and is automatically enriched with sentence-level readability scores and linguistic features. We demonstrated that IMPaCTS can be effectively used to fine-tune multiple LLMs of different sizes for readability-controlled sentence simplification. Results show that fine-tuning significantly improves performance over instructed versions of the models in few-shot. This is achieved even with a limited amount of data, especially for larger models, providing insights into how training data size affects performance optimization, which is particularly helpful for low-resource languages. We further showed that our fine-tuned models successfully followed the target readability levels and that human judgments of simplicity align with the automatic READ-IT scores used for control. This confirms the reliability of our methodology for guiding LLMs toward readability-controlled simplification strategies tailored to diverse reader populations. We also found that humans do not perceive small readability gaps, suggesting that in the future, we should condition LLMs using less granular readability controls. Among the many future directions of research, we plan to expand the typology of control, experimenting with specific linguistic features we found to be related to linguistic complexity.

## 7. Limitations

While our work provides insightful results on the reliability of the proposed fully automatic process for constructing a resource for controlled sentence simplification, some limitations must be acknowledged. We leave them as directions for future work.

First, we deliberately limited our evaluation to an extrinsic assessment of IMPaCTS. Specifically, we did not perform a dedicated human evaluation of the automatic process used to build the resource. This choice was motivated by our focus on assessing the models fine-tuned on IMPaCTS, which already provides an extrinsic measure of its quality, as well as by the human evaluation conducted on the sentences generated by the best-performing fine-tuned LLM, offering further evidence of the resource's overall reliability. In future work, we plan to complement these findings with an intrinsic evaluation of IMPaCTS, recruiting human raters to assess the quality of the simplifications more directly.

Second, the fine-tuning experiments were limited to a small set of model families and sizes. While the results clearly show that both larger and smaller LLMs benefit from fine-tuning, an inspection of the outputs revealed that smaller models (i.e., Minerva) occasionally fail to generate sentences. As discussed in the paper, understanding the causes of these failures is crucial for developing methods better suited to smaller models, a key step for real-world scenarios, particularly in low-resource languages and tasks.

Finally, our experiments were limited to two domains (Wikipedia and administrative texts) and a single control variable (readability). As discussed in the conclusions of the paper, extending the approach to additional domains and conditioning LLMs on linguistic features reflecting (morpho)syntactic complexity would further validate the reliability of IMPaCTS for controlled sentence simplification.

## 8. Acknowledgments

This work has been supported by the project "XAI-CARE" funded by the European Union - Next Generation EU - NRRP M6C2 "Investment 2.1 Enhancement and strengthening of biomedical research in the NHS" (PNRRMAD-2022-2376692_VADALA' – CUP F83C22002470001) and by LLMs4EU "Large Language Models for the European Union" project, funded by the European Union through the Digital Europe Programme (DIGITAL-2024-AI-B-06-LANGUAGE - GA 101198470) under the grant agreement 101198470.
Partial support was also provided by the project "Understanding and Enhancing Preference Alignment in Large Language Models Through Controlled Text Generation" (IsCc8_ALIGNLLM), funded by CINECA under the ISCRA initiative, for HPC resource availability and support.

## 9. Bibliographical References

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

Sweta Agrawal, Weijia Xu, and Marine Carpuat. 2021. A non-autoregressive edit-based approach to controllable text simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3757–3769, Online. Association for Computational Linguistics.

Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. Analysing zero-shot readability-controlled sentence simplification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781, Abu Dhabi, UAE. Association for Computational Linguistics.

Michele Cortelazzo. 2021. *Il linguaggio amministrativo: principi e pratiche di modernizzazione*. Carocci.

Asma Farajidizaji, Vatsal Raina, and Mark Gales. 2024. Is it possible to modify text to a target readability level? an initial investigation using zero-shot large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9325–9339, Torino, Italia. ELRA and ICCL.

Lyn Frazier. 1985. Syntactic complexity. In D.R. Dowty, L. Karttunen, and A.M. Zwicky, editors, *Natural Language Parsing*. Cambridge University Press, Cambridge, UK.

Martin Haspelmath. 2006. Against markedness (and what to replace it with). *Journal of Linguistics*, 42(1):25–70.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking large language models on sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics.

Tannon Kew and Sarah Ebling. 2022. Target-level sentence simplification as controlled paraphrasing. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 28–42, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Zihao Li and Matthew Shardlow. 2024. How do control tokens affect natural language generation tasks like text simplification. *Natural Language Engineering*, 30(5):915–942.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

Martina Miliani, Fernando Alva-Manchego, and Alessandro Lenci. 2023. Simplifying administrative texts for Italian L2 readers with controllable transformers models: A data-driven approach. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 303–315, Venice, Italy. CEUR Workshop Proceedings.

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.

Debora Nozza and Giuseppe Attanasio. 2023. Is it really that simple? prompting large language models for automatic text simplification in Italian. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 322–333, Venice, Italy. CEUR Workshop Proceedings.

Alessio Palmero Aprosio, Sara Tonelli, Marco Turchi, Matteo Negri, and Mattia A. Di Gangi. 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*, pages 37–44, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Michele Papucci, Giulia Venturi, and Felice Dell'Orletta. 2025. Generating and evaluating multi-level text simplification: A case study on Italian. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 870–885, Cagliari, Italy. CEUR Workshop Proceedings.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Marco Russodivito, Vittorio Ganfi, Giuliana Fiorentino, and Rocco Oliveto. 2024. AI vs. human: Effectiveness of LLMs in simplifying Italian administrative documents. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 842–853, Pisa, Italy. CEUR Workshop Proceedings.

Michael J Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting non-English text simplification: A unified multilingual benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4898–4927, Toronto, Canada. Association for Computational Linguistics.

Horacio Saggion. 2017. *Automatic Text Simplification*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool.

Arthur Mariano Rocha De Azevedo Scalercio, Elvis A. De Souza, Maria José Bocorny Finatto, and Aline Paes. 2025. Evaluating LLMs for Portuguese sentence simplification with linguistic insights. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24452–24477, Vienna, Austria. Association for Computational Linguistics.

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, Melbourne, Australia. Association for Computational Linguistics.

Kim Cheng Sheang, Daniel Ferrés, and Horacio Saggion. 2022. Controllable lexical simplification for English. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 199–206, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Nicolas Spring, Annette Rios, and Sarah Ebling. 2021. Exploring German multi-level text simplification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1339–1349, Held Online. INCOMA Ltd.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *CoRR*, abs/1610.02424.

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.

Daiki Yanamoto, Tomoki Ikawa, Tomoyuki Kajiwara, Takashi Ninomiya, Satoru Uchida, and Yuki Arase. 2022. Controllable text simplification with deep reinforcement learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 398–404, Online only. Association for Computational Linguistics.

Tatsuya Zetsu, Tomoyuki Kajiwara, and Yuki Arase. 2022. Lexically constrained decoding with edit operation prediction for controllable text simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 147–153, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

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

## 10. Language Resource References

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. CEFR-based sentence difficulty annotation and assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei

Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Pierpaolo Basile, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. 2023. Llamantino: Llama 2 models for effective text generation in italian language.

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, Austin, Texas. Association for Computational Linguistics.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-UD: a tool for linguistic profiling of texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7145–7151, Marseille, France. European Language Resources Association.

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, Denver, Colorado, USA. Association for Computational Linguistics.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ–IT: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

Martina Miliani, Serena Auriemma, Fernando Alva-Manchego, and Alessandro Lenci. 2022. Neural readability pairwise ranking for sentences in Italian administrative language. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 849–866, Online only. Association for Computational Linguistics.

Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Simone Conia, Edoardo Barba, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. 2024. Minerva LLMs: The first family of large language models trained from scratch on Italian data. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 707–719, Pisa, Italy. CEUR Workshop Proceedings.

Lucia C. Passaro and Alessandro Lenci. 2019. PaWaC - Public Administration Web as Corpus (Processed). `http://data.europa.eu/88u/dataset/elrc_1282`. [Data set].

Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. 2024. Advanced natural-based interaction for the italian language: Llamantino-3-anita.

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

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

## A. Prompt Template for Sentence Simplification

Each model was prompted using its default system prompt, as specified in the corresponding Hugging Face model documentation. In addition, we provided a task-specific instruction prompt to guide the model in performing the Sentence Simplification task. The following prompt template was used:

### Istruzione: Semplifica la seguente frase mantenendo il più possibile intatto il significato. (Instruction: Simplify the following sentence while preserving the original meaning as much as possible.)

### Input: original_sentence

### Output:

## B. Linguistic Features Used for Sentence Profiling

The following table presents the full list of linguistic features extracted with Profiling-UD and used to characterize the linguistic profile of sentences in IMPaCTS. For the list of UD POS refer to https://universaldependencies.org/u/pos/index.html, of syntactic relations to https://universaldependencies.org/u/dep/index.html, and of verbal morphological tags https://universaldependencies.org/u/feat/index.html.

| |
|---|
| **Raw Text Properties** |
| Average sentence length in tokens |
| Average word length in characters |
| **Lexical Variety** |
| New Basic Italian Vocabulary (NBIV) for words and lemmas |
| Fundamental/High usage/High availability words of NBIV for words and lemmas |
| Classes of frequency |
| **Morphosyntactic information** |
| Distribution of UD Parts of Speech |
| Lexical density |
| **Dependency Syntactic Relations** |
| Distribution of UD dependency relations |
| **Global and Local Parsed Tree Structures** |
| Average depth of the whole syntactic trees |
| Average and maximum dependency link lengths |
| Total number of prepositional chains and average length and average length |
| Distribution of prepositional chains by depth |
| Average clause length |
| **Order of elements** |
| Relative order of subjects and objects with respect to the verb |
| **Inflectional morphology** |
| Inflectional morphology (tense, mood, verb form, person, number) of lexical verbs and auxiliaries |
| **Verbal Predicate Structure** |
| Distribution of verbal roots |
| Distribution of verbal heads per sentence |
| Average verb arity and distribution of verbs by arity |
| **Use of Subordination** |
| Distribution of principal and subordinate clauses |
| Average length of subordination chains and distribution of chains by depth |
| Relative order of subordinate clauses with respect to the principal proposition |

Table 6: Linguistic features used for linguistic profiling.

## C. Prompt Used for Baseline Models

The baseline models were prompted using an instruction-based prompt combined with a set of in-context examples. Each prompt contains the instruction asking the model to simplify a sentence while preserving its meaning and to produce an output matching the requested READ-IT score and 3 example pairs composed of an input sentence and a target READ-IT score.

The following template was used[14]:

Di seguito è riportata un'istruzione che descrive un'attività, accompagnata da un input che aggiunge ulteriore informazione. Scrivi una risposta che completi adeguatamente la richiesta. (Below is an instruction that describes a task, paired with an input that provides additional information. Write a response that appropriately completes the request.)

### Istruzione: Semplifica la seguente frase mantenendo il più possibile intatto il significato. Ti verrà fornito un valore di Leggibilità da 0 a 100 che dovrà essere rispettato dalla tua frase, usa gli esempi nel contesto per capire come variare la risposta in base a questo valore. Rispondi solamente con la frase semplificata. (Instruction: Simplify the following sentence while preserving the original meaning as much as possible. You will be given a readability value ranging from 0 to 100 that your output sentence should match. Use the examples provided in the context to understand how to adjust the simplification according to the requested readability value. Respond only with the simplified sentence.)

### Input:

Leggibilità - 20: INVIARE la presente delibera, il progetto in narrativa e la bozza di convenzione al Comune di Chiesina Uzzanese. (Readability - 20: SEND the present resolution, the project described in the narrative, and the draft agreement to the Municipality of Chiesina Uzzanese.)

### Output: La presente delibera, il progetto in narrativa e la bozza di convenzione sono inviate al Comune di Chiesina Uzzanese. (The present resolution, the project described in the narrative,

[14]The original prompt and sample sentences are in Italian; English translations are provided in parentheses.

and the draft agreement are sent to the Municipality of Chiesina Uzzanese.)

### Istruzione: Semplifica la seguente frase mantenendo il più possibile intatto il significato. Ti verrà fornito un valore di Leggibilità da 0 a 100 che dovrà essere rispettato dalla tua frase, usa gli esempi nel contesto per capire come variare la risposta in base a questo valore. Rispondi solamente con la frase semplificata. (Instruction: Simplify the following sentence while preserving the original meaning as much as possible. You will be given a readability value ranging from 0 to 100 that your output sentence should match. Use the examples provided in the context to understand how to adjust the simplification according to the requested readability value. Respond only with the simplified sentence.)

### Input:

Leggibilità - 30: Gino Biondi Deliberazione n. 159 del 18 dicembre 2008 "Interventi di riqualificazione lungo il fiume Pescia di Pescia in Comune di Chiesina Uzzanese" - Approvazione progetto esecutivo la deputazione amministrativa premesso che il Comune di Chiesina Uzzanese d'intesa con il Consorzio di Bonifica del Padule di Fucecchio intende valorizzare un'area nel proprio Comune adiacente al Fiume Pescia di Pescia con alcuni interventi di riqualificazione ambientale, partecipando con fondi propri a parte dei lavori; (Readability - 30: Gino Biondi Resolution n. 159 of 18 December 2008 "Redevelopment interventions along the Pescia di Pescia River in the Municipality of Chiesina Uzzanese" – Approval of the executive project. The Administrative Deputation, whereas the Municipality of Chiesina Uzzanese, in agreement with the Padule di Fucecchio Land Reclamation Consortium, intends to enhance an area of its municipality adjacent to the Pescia di Pescia River through several environmental redevelopment interventions, participating with its own funds in part of the works;)

### Output: Il comitato ha approvato un progetto esecutivo per eseguire interventi di riqualifica ambientale lungo il fiume Pescia a Pescia, in comune di Chiesina Uzzanese. Il comune di Chiesina Uzzanese, insieme al Consorzio di Bonifica del Padule di Fucecchio, intende migliorare un'area nel proprio comune vicino al fiume Pescia con alcuni interventi di riqualifica ambientale, utilizzando i propri fondi per contribuire ai lavori. (The committee approved an executive project to carry out environmental redevelopment interventions along the Pescia River in Pescia, in the Municipality of Chiesina Uzzanese. The Municipality of Chiesina Uzzanese, together with the Padule di Fucecchio Land Reclamation Consortium, intends to improve an area of its municipality near the Pescia River through several environmental redevelopment interventions, using its own funds to contribute to the works.)

### Istruzione: Semplifica la seguente frase mantenendo il più possibile intatto il significato. Ti verrà fornito un valore di Leggibilità da 0 a 100 che dovrà essere rispettato dalla tua frase, usa gli esempi nel contesto per capire come variare la risposta in base a questo valore. Rispondi solamente con la frase semplificata. (Instruction: Simplify the following sentence while preserving the original meaning as much as possible. You will be given a readability value ranging from 0 to 100 that your output sentence should match. Use the examples provided in the context to understand how to adjust the simplification according to the requested readability value. Respond only with the simplified sentence.)

### Input:

Leggibilità - 10: VISTO il parere di legittimità rilasciato in data 18 dicembre 2008 dal Direttore Generale Dott. Franco Fambrini; (Readability -10: HAVING SEEN the opinion of legality issued on 18 December 2008 by the Director General Dr. Franco Fambrini;)

### Output: Rilasciato il 18 dicembre 2008, il parere di legittimità del direttore generale Dr. Franco Fambrini. (Issued on 18 December 2008, the opinion of legality of the Director General Dr. Franco Fambrini.)

### Istruzione: Semplifica la seguente frase mantenendo il più possibile intatto il significato. Ti verrà fornito un valore di Leggibilità da 0 a 100 che dovrà essere rispettato dalla tua frase, usa gli esempi nel contesto per capire come variare la risposta in base a questo valore. Rispondi solamente con la frase semplificata. (Instruction: Simplify the following sentence while preserving the original meaning as much as possible. You will be given a readability value ranging from 0 to 100 that your output sentence should match. Use the examples provided in the context to understand how to adjust the simplification according to the requested readability value. Respond only with the simplified sentence.)

### Input:

Leggibilità - 35: La società Strade Scavi Nord Srl, con sede legale in Piazza Brembana (BG), Via Orenghi n. 31, ha predisposto lo studio preliminare ambientale relativo al progetto: (Readability - 35: The company Strade Scavi Nord Srl, with its registered office in Piazza Brembana (BG), Via Orenghi n. 31, has prepared the preliminary environmental study relating to the project:)

### Output:

During inference, the last input block was replaced with the sentence to be simplified and the requested readability value.