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

Michele Papucci[1,2], Giulia Venturi[1], Felice Dell'Orletta[1]

1 - ItaliaNLP Lab @ Institute for Computational Linguistics "Antonio Zampolli" (CNR-ILC), Pisa.

2 - Università di Pisa, Pisa.

LREC2026, 11 - 16 May 2026, Palma

Introduction

Automatic Text Simplification aims at reducing complexity of a text while maintaining the meaning.

- The dominant approach is **data driven**; gold standard is **human-written** simplifications.
- Manually constructed resources are **labor-intensive** and **costly**.

In this work we present:

- **IMPaCTS**: The first fully automatically created Italian corpus with multi-level simplifications for each original sentence;

- How **fine-tuning open-weight LLMs** on **increasing portions of IMPaCTS** improve their capability over their few-shot counterparts in the Controllable Sentence Simplification task;

- A **human perception study** of the generated sentences from our readability-controlled LLMs trained on IMPaCTS;

Dataset Construction

1. Selection of an **Italian fine-tuned LLM** that can reliably simplify texts;
2. Selection of a collection of sentences in the **domains of interest**;
3. Generation with the selected LLM **multiple simplification for each input**;
4. Evaluation of the resulting sentence pairs in terms of **readability** and **linguistic features**;

1. LLM Selection

We tested three Italian fine-tuned LLMs: **Llamantino 2**, **ANITA**, **Italia**

In zero-shot text-simplification on the test-split of Italian Sentence Simplification Dataset: SIMPITIKI, Terence, Teacher, ADMIN-IT and PaCCSS-IT.

We evaluated them with automatic metrics:

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

Llamantino outperforms the other two models on all metrics.

2. Domains Selection

We selected two domains of interest:

- **Italian Wikipedia**;

- **Public Administration** (PaWaC - Passaro et Lenci, 2019);

For both domain we sample randomly around 100.000 original sentences.

3. Creation of the Multi-Level Resource

We employ the **Divergent Beam Search** with a high diversity penalty to generate **10 different simplifications** for each original sentence using Llamantino-2 in a zero-shot setting.

We removed: duplicate simplifications, any simplification that is the same as the original sentence, and low quality responses.

Resulting Resource:

- 1058960 pairs (original/simplification) for Wikipedia, and 385200 for P.A.

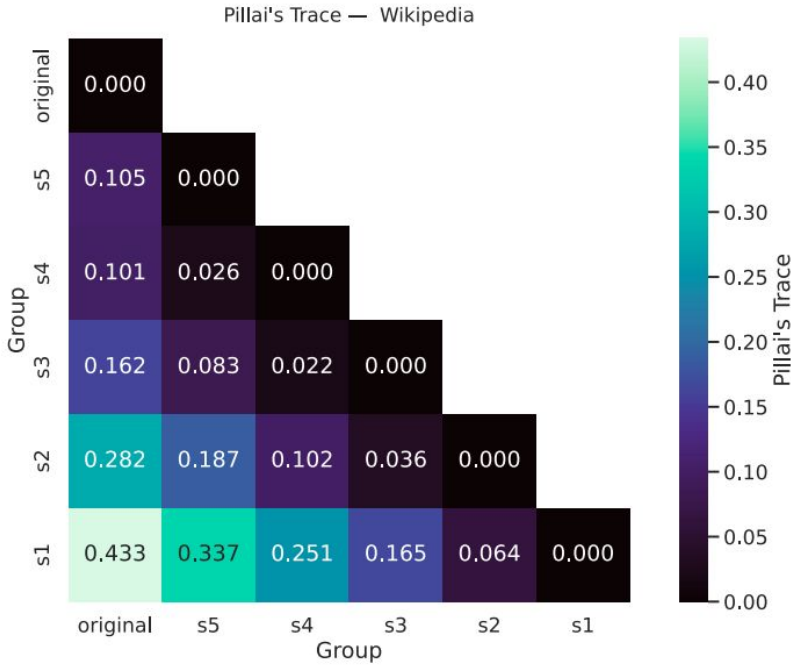
Each of the pairs was then annotated with:

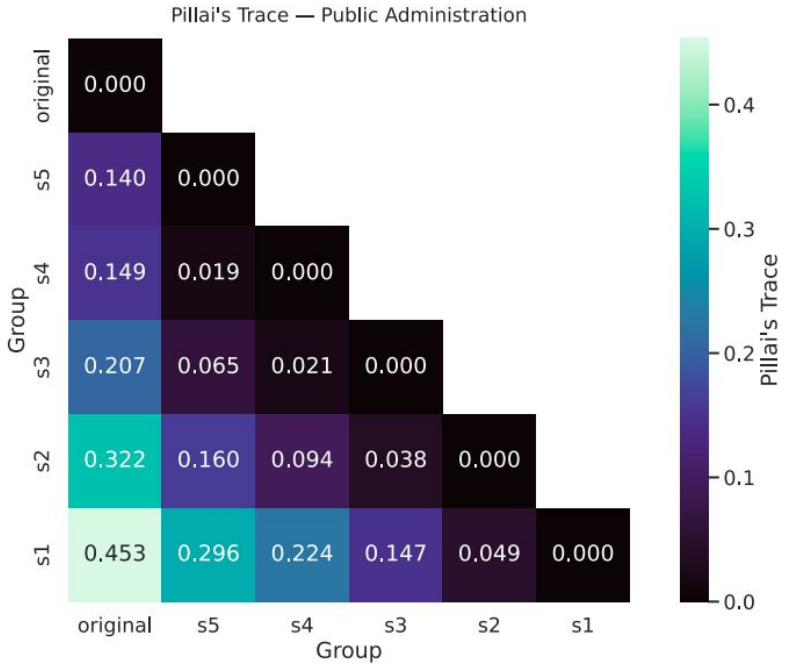
- **Readability score** (Read-IT, Dell'Orletta et al., 2011);
- **Linguistic Profiling** composed of 148 automatically extracted features (Profiling-UD, Brunato et al., 2020);

4. Example of an entry

	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

4. Evaluation of the Resource - Profiling





We find that higher difference in Read-IT scores between original and simplifications results in higher Pillai's Trace between the profiling of the two sentences.

5. Fine-tuning Experiments

Models. Qwen3 Base (1.7B, 4B, 8B) and Minerva (1B, 3B, 7B).

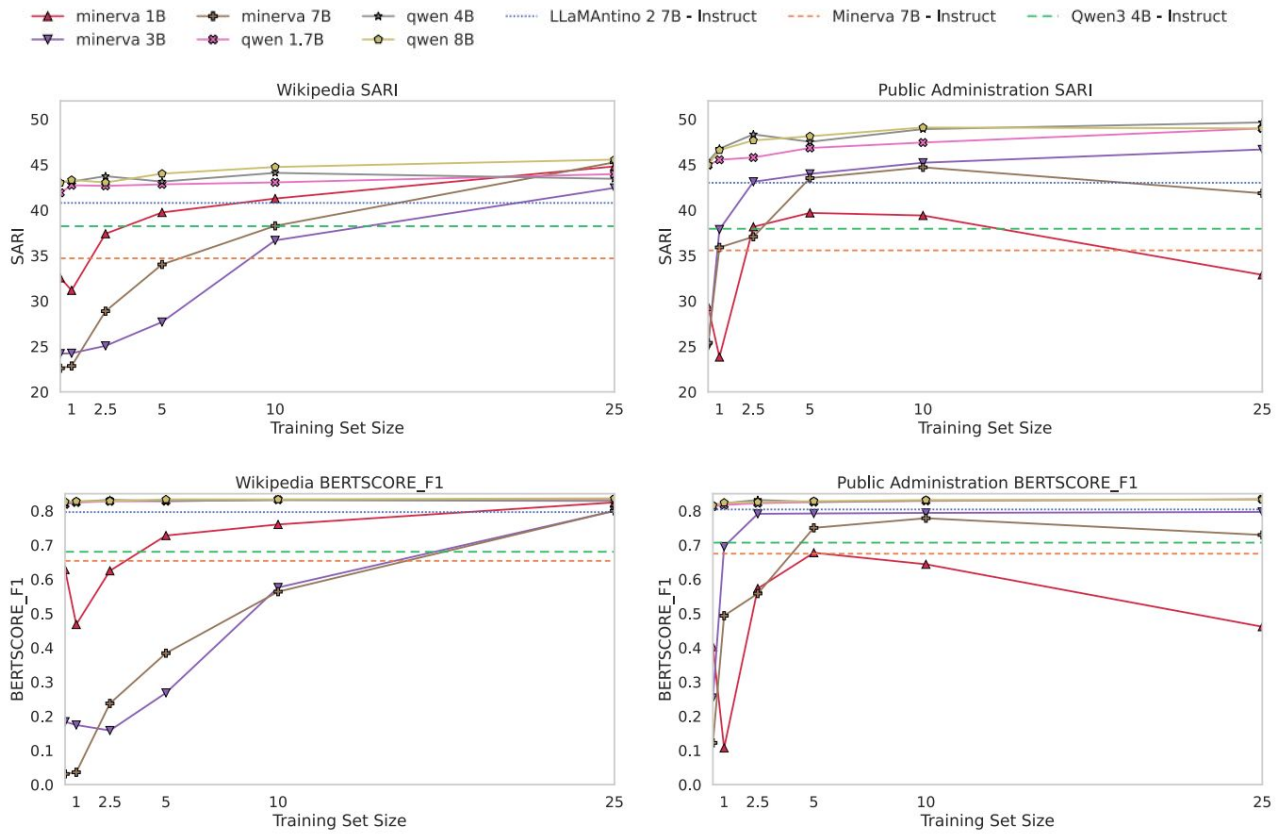
Baselines. Qwen3 4B Instruct, Minerva 7B Instruct, LLaMAntino 2. They were used in In-Context Few Shot Prompting with three examples.

Fine-tuning. The models were fine-tuned used LoRA in 16bit precision.
For solving Controlled Sentence Simplification task we used *control tokens*.
20 new tokens added: `<|readability_0|>`, `<|readability_5|>`, …, `<|readability_100|>`

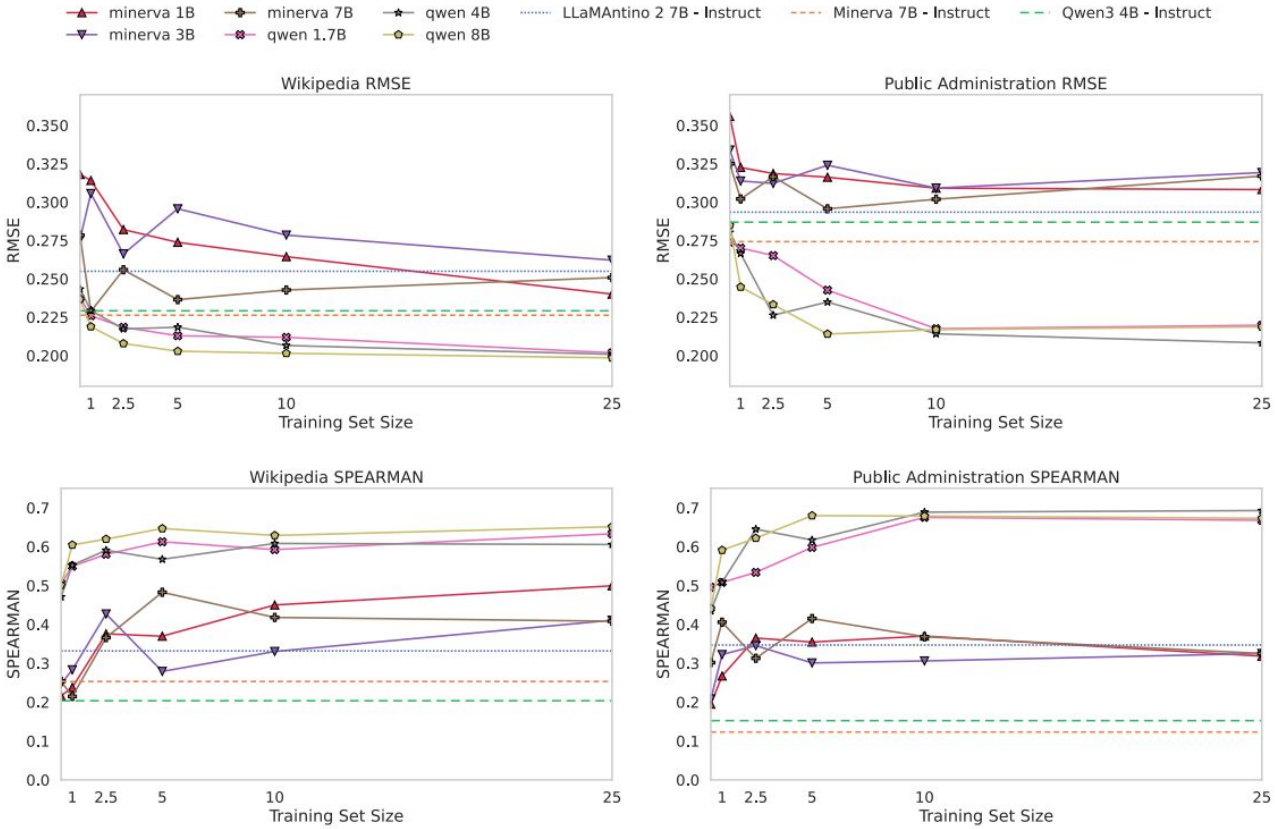
Increasing Portions. We fine-tuned models on 500, 1k, 2.5k, 5k, 10k and 25k original sentences and all their simplifications.

Evaluation. BLEU, SARI and BertScore for the accuracy of the simplifications.
RMSE and Spearman rank for the degree of compliance to the requested readability level.

5. Results - SARI↑ and BertScore↑



5. Results - RMSE↓ and Spearman↑



6. Human Evaluation - Setting

150 randomly sampled sentence couple generated by Qwen3 8B - 25K.
125 pairs: two simplification with different control tokens from the same original sentence.
25 pairs: simplification and original sentence.

Divided in 5 questionnaires of 30 pairs each. 4 binary questions for each pair:

Q1. Is sentence 1 simpler than sentence 2?
Q2. Do sentence 1 and 2 express the same meaning?
Q3. Is sentence 1 grammatical?
Q4. Is sentence 2 grammatical?

Each questionnaire was taken from 5 different annotators selected trough Prolific, screened to be **native speakers of Italian**, totaling 25 different annotators.

6. Human Evaluation - Results

Inter-annotator agreement. Calculated as Krippendorf's α. We found: Q1 α=0.5, Q2 α=0.46, Q3 and Q4 α=0.32.

Alignment with Read-IT. Cohen's k and F1 Score between the majority vote of the annotators on Q1 and whether Read-IT sentence 1 < Read-IT sentence 2.
k = 0.48 and F1 = 0.74

Sensibility to Readability Gap. We find the higher the Readability gap between the two sentences, the higher the agreement between the annotators. True for both the actual Read-IT score (Spearman = 0.48), and the requested one (Spearman = 0.46).

Conclusion

1. We created the first large resource for **Controlled Sentence Simplification** for **Italian** called **IMPaCTS**.

2. We evaluated IMPaCTS both in terms of **linguistic features** and by using it to **fine-tune 6 LLMs** with **increasing portions** of it showing how it outperforms Instruction-tuned versions of the same LLMs both in **simplification solving capability** and in **respecting the requested readability level**.

3. We presented an **human evaluation** that shows how LLMs trained with IMPaCTS **generate sentences that aligns with human perception of simplicity**.

You can find **IMPaCTS** on **GitHub** and **HuggingFace**

Thanks for your attention!

Istituto di Linguistica Computazionale "Antonio Zampolli"

Consiglio Nazionale delle Ricerche