Generating and Evaluating Multi-Level Text Simplification: A Case Study on Italian

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.

CLiC-it 2025 - Cagliari, September 24th - 26th

Introduction

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

- The dominant approach is **data driven**;
- Manually constructed resources are **labor-intensive** and **costly**.

In this work we present:

- An investigation of the capability of Italian fine-tuned LLMs for producing **large resources for Multi-Level Sentence Simplification in Italian**;
- A case-study resource with **multiple simplification at various readability level** for each original human-written sentence;
- An **in-depth linguistic analysis** of the resource.

Our Approach

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

<u>Llamantino outperforms the other two models on all metrics.</u>

2. Domains Selection

We selected two domains of interest:

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

For both domain we sample randomly 10,000 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** from Llamantino-2 in a zero-shot setting.

Examples:

Original: *Alcuni composti aromatici più pesanti, come lo xilene, possono essere utilizzati al posto del toluene ottenendo rese comparabili*.

Least Simplified: *Alcuni composti aromatici più pesanti possono essere utilizzati al posto del toluene ottenendo rese comparabili.*

Randomly-Selected Simplification: *La maggior parte degli aromi più pesanti possono essere utilizzati al posto di toluene.*

Most Simplified: *È possibile utilizzare xilene invece di toluene per ottenere una resa simile.*

3. Creation of the Multi-Level Resource - 2

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

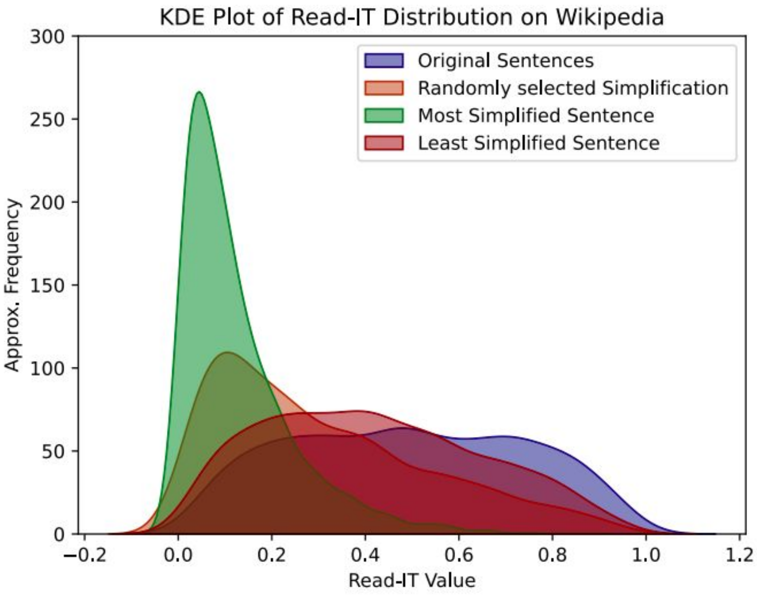
Resulting Resource:

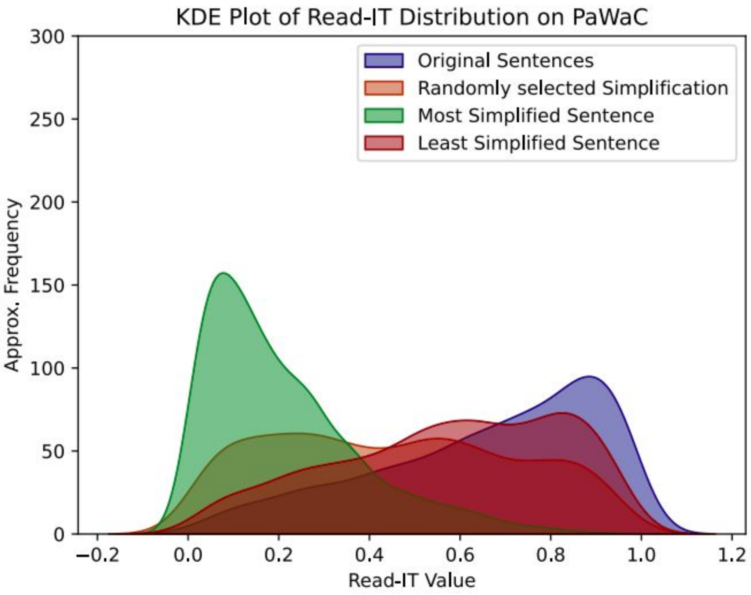
- 71,837 pairs (original/simplification) for Wikipedia, and 78,184 for PaWaC;

Evaluation of a subset of 2000 pairs per domain:

- Readability score (Read-IT) for each pair;
- Linguistic Profiling composed of 148 automatically extracted features (Profiling-UD, Brunato et al., 2020);

4. Evaluation (Readability)





Distribution of Read-IT values of the original sentences, the most and least simplified generations, and a randomly selected simplification.

4. Evaluation (Linguistic Features)

	Wikipedia		Pawac	
	Pillai's Trace	p-value	Pillai's Trace	p-value
Original vs Least Simplified	.12	$\leq 10^{-4}$	.16	$\leq 10^{-4}$
Original vs Randomly-Selected	.18	$\leq 10^{-4}$	.19	$\leq 10^{-4}$
Original vs Most Simplified	.44	$\leq 10^{-4}$	.46	$\leq 10^{-4}$

Multivariate ANalysis Of VAriance (MANOVA) of the linguistic features distribution. It compares the originals against the least, most and randomly selected simplifications.

Pillai's Trace reports a **higher degree of difference in linguistic features** the more the sentence is simplified w.r.t. the original.

4. Evaluation (Linguistic Features) - 2

Wilcoxon signed rank finds feature that change the most after simplification.

- **Raw Text Properties**:
    - Sentence Length;
- **Global Syntactic Structures:**
    - Dependency Tree depth;
    - Max dependency Link Length;
- **Local Syntactic Structures**:
    - Distribution of the subordination clauses;
    - Subordinate position relative to the main clause;
    - Non-canonical subject-object order (pre-verbal objects and post-verbal subjects).

4. Evaluation (Linguistic Features + Readability)

Spearman rank correlation between the simplification's Read-IT score and difference (original - simplification) in feature values finds what impacts readability:

- **Raw Text Properties**: Sentence Length ↑;
- **Lexical variation**: Maximum Frequency class ↑, High Availability words (NVDB)↓;
- **Global Syntactic Structures**: Max dependency Link Length ↑, Number of embedded sequences of prepositional complements ↑;
- **Local Syntactic Structures**: Distribution of subordinative clauses ↑, Recursively embedded subordinate clauses ↑, Subordinate position relative to main clause (post) ↑.

Conclusion

1. Identified the **best performing zero-shot Italian model** for sentence simplification;

2. An automatically created resource for **multi-level sentence simplification**;

3. An **in-depth analysis** in terms of **readability** and **linguistic phenomena** involved in automatic sentence simplification.

Future Work. Create a **Large Dataset** for **Controlled Sentence Simplification** with target readability or linguistic phenomena taken from the multi-level simplification resource.

Check out the **GitHub Repository** with **all the data!**

Thanks for your attention!

Istituto di Linguistica Computazionale
"Antonio Zampolli"
Consiglio Nazionale delle **Ricerche**