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

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

[1]ItaliaNLP Lab @ Institute for Computational Linguistics, National Research Council, Pisa
[2]University of Pisa, Pisa

Abstract

Recent advances in Generative AI and Large Language Models (LLMs) have enabled the creation of highly realistic synthetic content, yet controlling model outputs remains a challenge. In this study, we explore the use of LLMs to generate high-quality synthetic data for Automatic Text Simplification (ATS), evaluating the ability of models fine-tuned on Italian to produce multiple simplified versions of the same original sentence that vary in readability and in their lexical and (morpho-)syntactic characteristics. The approach is tested across two domains, Wikipedia and Public Administration, allowing us to explore domain sensitivity. Additionally, we compare the linguistic phenomena observed in the generated data with those found in ATS resources previously created through manual or semi-automatic methods. Our results suggest that the best-performing LLM can generate linguistically diverse simplifications that align with known simplification patterns, offering a promising direction for building reliable ATS resources, including simplifications suited to varying levels of reader proficiency.

Keywords

Automatic Text Simplification, Large Language Models, Synthetic Data, Linguistic Complexity, Sentence Readability

1. Introduction

Automatic Text Simplification (ATS) aims to reduce the linguistic complexity of a text while preserving its meaning. Given that the dominant approach is data-driven, where models learn simplification operations from examples of complex-simple sentence pairs [1], the availability and nature of resources for ATS play a crucial role in determining the quality of these models.

Traditionally, manually constructed resources have been favored for their reliability and controllability [2]. However, the cost and labor-intensiveness of such efforts limit their scalability, domain coverage, and language diversity. To address these limitations, researchers have explored unsupervised methods for resource construction, including mining sentence pairs from aligned corpora, primarily Wikipedia and Simple Wikipedia [3], or exploiting crowdsourcing approaches [4, 5]. In light of concerns about the suitability of Wikipedia as an ATS resource [6], and to tackle the broader scarcity of parallel simplification data especially for low-resource languages, researchers have also proposed methods to automatically create parallel resources, inspired for example by paraphrase generation [7, 8] or machine translation [9, 10].

More recently, Large Language Models (LLMs) have introduced a new paradigm for ATS, also opening the possibility of generating *synthetic* resources whose quality still requires thorough assessment [2]. This trend aligns with broader efforts to leverage LLMs for alleviating the limitations of real-world data through synthetic data generation [11]. Evaluation initiatives such as BLESS [12] have demonstrated that LLMs, under a few-shot setting, are capable of generating simplified sentences across multiple datasets, languages, and prompts. Yet, research to date has primarily focused on English and has relied on a limited set of evaluation metrics, leaving open questions about model behavior across different domains, languages, and target user needs. Notable exceptions for the Italian language include [13] and [14], who assessed the ability of both open and proprietary LLMs to produce simplified sentences. The former focused on increased sentence readability, while the latter examined both readability and semantic similarity, comparing model-generated simplifications with those written by human simplifiers. Interestingly, both studies targeted the administrative domain.

Starting from these premises, this paper introduces a multifaceted approach to assess the ability of three small LLMs fine-tuned on the Italian language to generate sentence simplifications along a gradient of complexity. After identifying the best-performing model, we examined its output along three main dimensions: *i)* its ability to produce multiple simplifications for the same input sentence with increasing levels of readability; *ii)* the extent to which the linguistic characteristics of the simplified sentences differ from those of the original; and *iii)* the re-

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

michele.papucci@phd.unipi.it (M. Papucci); giulia.venturi@ilc.cnr.it (G. Venturi); felice.dellorletta@ilc.cnr.it (F. Dell'Orletta)
https://michelepapucci.github.io/ (M. Papucci); http://www.italianlp.it/people/giulia-venturi/ (G. Venturi); http://www.italianlp.it/people/felice-dellorletta/ (F. Dell'Orletta)
0000-0003-4251-7254 (M. Papucci); 0000-0001-5849-0979 (G. Venturi)

lationship between the distribution of linguistic features and the readability level. This in-depth linguistic analysis of LLM-generated simplifications aims to achieve two main objectives. First, it investigates whether small, open LLMs can reliably produce multiple simplifications with varying degrees of linguistic complexity, thereby offering a scalable strategy for creating resources tailored to different target populations, which remain scarce [2]. Second, it aims to explore whether specific linguistic patterns observed in original–simplified sentence pairs are influenced by the approach used to construct ATS resources, as discussed in [15].

2. Methodology

The approach we propose for assessing the ability of LLMs to automatically generate sentence simplifications along a gradient of linguistic complexity is articulated in three main steps:

1. selection of an LLM fine-tuned on the Italian language, capable of reliably generating sentences in the target language, and identification of a corpus of human-written sentences to be used as original inputs;
2. prompting the selected LLM to generate multiple simplified versions of each original sentence to obtain diverse outputs per input;
3. evaluation of the resulting sentence pairs in terms of their linguistic feature diversity and variation in readability levels.

The main objective of the first two steps, described in Section 3, is to construct a parallel corpus composed of human-written original sentences and multiple automatically generated simplified versions. This allows for capturing a range of sentence transformations characterized by different linguistic phenomena. In this respect, the proposed methodology is particularly suitable for low-resource languages, where simplified corpora remain scarce, especially those addressing multiple reader profiles, domains, or textual genres.

The evaluation of the generated simplifications, which constitutes the main focus of this study, is presented in Section 4. Our multifaceted evaluation methodology aims to assess not only how readability levels vary across the multiple simplifications and relative to the original sentence, but also how the lexical, morpho-syntactic, and syntactic characteristics of the sentence pairs change. A further contribution of this study lies in a comparative analysis designed to explore whether specific linguistic phenomena observed in the LLM-generated simplifications resemble those found in existing Italian ATS resources, specifically two created manually [16] and one semi-automatically [7].

3. Experimental Settings

LLM selection. To identify the most suitable LLM for the task of generating simplified sentences, we considered three models specifically developed for the Italian language, which differ in terms of architecture and number of parameters: ANITA[1] [17], LLaMAntino-2[2] [18], and Italia[3]. All models were tested in a 0-shot setting. The models' performance was evaluated against the test splits of the following Italian sentence simplification datasets: 51 paired original/simplified sentences from SIMPITIKI[4] [19], 994 sentence pairs filtered from PaCCSS–IT [7], 101 sentence pairs from the *Terence* corpus and 17 from the *Teacher* corpus [16], 49 sentence pairs extracted from ADMIN-it [20], for a total of 1,212 sentence pairs.

As evaluation metrics, we selected a set of complementary measures addressing different aspects of sentence simplification. Specifically, we included *i)* two metrics widely used in the literature that focus on surface-level properties related to writing style, i.e. BLEU [21] and SARI [22], and *ii)* two semantic similarity metrics used to assess meaning preservation, i.e. BertScore [23] and SentenceTransformer Similarity [24, 25]. In addition, we evaluated the simplified sentences in terms of variation in readability computed by READ-IT [26], the first machine-learning-based automatic readability assessment tool developed for Italian, combining traditional surface features with lexical, morpho-syntactic, and syntactic information correlated with linguistic complexity.

All models were evaluated on a single generation for each input. Each model was prompted using its respective system prompt, combined with a shared task-specific instruction to simplify the text while preserving the original meaning.[5]. The results are reported in Table 1, where it should be noted that the evaluation metrics follow an increasing trend, meaning that higher scores correspond to more simplified sentences. In contrast, READ-IT scores exhibit the opposite trend: they range from 0 (most readable sentence) to 100 (least readable sentence), as they reflect the level of linguistic complexity of the input. Notably, LLaMAntino-2 consistently outperformed the other LLMs across all evaluation metrics, generating sentences that are simpler than the original inputs in both surface-level properties and semantic content. Moreover, its outputs had the lowest READ-IT scores, indicating that they are the least linguistically complex among those produced by the tested models. As a result, it was selected for the second step of our methodology.

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

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

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

[4] From SIMPITIKI we took only the Wikipedia sentence pairs and excluded the Administrative domain ones, since those are the same sentences already present in ADMIN-IT.

[5] See Appendix A for more details.

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

Table 1
SARI, BLEU, average BERTScore, average SentenceTransformer similarity score, and average READ-IT scores for the tested models in a zero-shot configuration on the test set.

Textual domains. We tested the full experimental setting on two corpora representative of two Italian language varieties that are widely acknowledged to exhibit significantly different linguistic features. Specifically, we selected a collection of sentences downloaded from Wikipedia pages, as it is the most frequently addressed domain in the literature on ATS [2]. As a counterpart, we included the "PaWaC – Public Administration Web as Corpus" (PaWaC [27]), which contains a wide range of administrative texts (resolutions, circular letters, etc.) and represents the Italian language used in public administration, a language variety well-known for its high level of multilevel linguistic complexity [28]. For both domains, we randomly sampled 10,000 sentences to serve as the original texts for generating multiple simplified variants.

Generation of multiple simplifications. Step two of our methodology was performed by prompting LLaMAntino-2 with the same prompt introduced previously to generate multiple simplified versions for the collection of the original 10,000 sentences for the Wikipedia and administrative domains. To this end, we employed the Divergent Beam Search decoding technique [29] to obtain multiple simplifications for each original sentence. Through manual inspection of the outputs generated under different decoding settings, we found that using 20 beams divided into 10 groups, with a diversity penalty $\lambda = 0.7$, provided the best results in terms of diversity of the simplifications and text fluency.

Using this decoding strategy, we obtained 10 simplifications for each original sentence. The resulting resource was automatically revised by removing duplicate simplifications and cases where the original and simplified sentences were identical. After this clean-up, we obtained 71,837 original/simplified sentence pairs for Wikipedia and 78,184 pairs for PaWaC.

Table 2 reports two examples randomly extracted from the generated resource. Concerning the administrative domain, we can see that the least simplified PaWac sentences (i.e. those with the higher READ-IT scores) are simplified primarily through the deletion of informational content (e.g. *non automaticamente rinnovabili* 'not automatically renewable' is removed). In contrast, the most simplified sentences display linguistic features typically associated with more readable sentence structures while keeping the original information content. For instance, the simplest sentence (i.e. the sentence with the lowest READ-IT score) is characterized by a reduced distance between the nominal subject (*le concessioni* 'the concessions') and the main verb (*devono essere considerate* 'must be considered'). In addition, the main verb undergoes *i)* a lexical simplification since the simpler *considerare* 'to consider' replaces the more complex original verb *interdersi* 'to understand' and *ii)* a morphological simplification since the epistemic future is replaced by a more straightforward present-tense form. Also in the case of the Wikipedia example, the most simplified sentences are the result of structural transformations. Namely, the two versions with the lowest READ-IT scores contain the main at the active voice instead of the passive, and feature shorter syntactic dependency links among words.

Linguistic profiling. Our evaluation step includes a comparative analysis of the distribution of multilevel linguistic features automatically extracted from the original and the LLaMAntino-2–generated simplified sentences. To this end, we adopted Profiling-UD [30], a web-based tool designed to linguistically profile multilingual texts using the Universal Dependencies (UD) formalism [31]. The full set of features is detailed in Table 3. They can be grouped into nine categories, each corresponding to linguistic phenomena possibly related to sentence complexity. These range from raw text properties, such as sentence length, to more complex features, such as the distribution of UD Parts-of-Speech (POS), dependency relations, and verbal inflectional properties (e.g. mood, tense)[6]. Additional features capture global and local syntactic structure, such as the structure of verbal predicates, the order of nuclear sentence elements (subject and object) relative to the verb, and the use of subordination. The set also includes features modeling the lexical variety of sentences, specifically in terms of the *i)* distribution of word forms and lemmas belonging to the New Basic Italian Vocabulary (NBIV) [32], further classified into the three usage categories, and *ii)* distribution of lemmas based on classes of frequency computed using the Italian section of the *MC4* corpus [33, 34] according to the fol-

[6] 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.

	PaWac	READ-IT
Original	Stabilito pertanto che le concessioni rilasciate in base al presente bando dovranno intendersi come decennali e non automaticamente rinnovabili (*It is therefore established that the concessions granted under this call shall be understood decennial and not automatically renewable*)	.75
Simplifications	Stabilito pertanto che le concessioni rilasciate in base al presente bando dovranno intendersi come decennali. (*It is therefore established that the concessions granted under this call shall be understood as decennial.*)	.70
	Stabilito pertanto che le concessioni rilasciate in base al presente bando dovranno intendersi come decennali e non rinnovabili. (*It is therefore established that the concessions granted under this call shall be understood as decennial and not renewable.*)	.61
	Il presente bando stabilisce che le concessioni devono essere considerate decennali e non automaticamente rinnovabili. (*This call establishes that the concessions must be considered decennial and not automatically renewable.*)	.34
	Il presente bando stabilisce che le concessioni devono essere considerate decennali e non rinnovabili automaticamente (*This call establishes that the concessions must be considered decennial and not renewable automatically.*)	.31
	Quindi le concessioni basate sul presente bando devono essere considerate decennali e non automaticamente rinnovabili. (*Therefore the concessions based on this call must be considered decennial and not automatically renewable*)	.29
	Wikipedia	
Original	Alcuni composti aromatici più pesanti, come lo xilene, possono essere utilizzati al posto del toluene ottenendo rese comparabili. (*Some heavier aromatic compounds, such as xylene, can be used instead of toluene obtaining comparable yields.*)	.59
Simplifications	Alcuni composti aromatici più pesanti possono essere utilizzati al posto del toluene ottenendo rese comparabili. (*Some heavier aromatic compounds can be used instead of toluene obtaining comparable yields.*)	.34
	La maggior parte degli aromi più pesanti possono essere utilizzati al posto di toluene ottenendo rese comparabili. (*The majority of heavier aromatics can be used in place of toluene obtaining comparable results.*)	.25
	La maggior parte degli aromi più pesanti possono essere utilizzati al posto di toluene. (*The majority of heavier aromatics can be used in place of toluene.*)	.21
	È possibile utilizzare xilene invece di toluene per ottenere un prodotto finale simile. (*It is possible to use xylene instead of toluene to obtain a similar end product.*)	.16
	È possibile utilizzare xilene invece di toluene per ottenere una resa simile. (*It is possible to use xylene instead of toluene to obtain a comparable yield.*)	.15

Table 2
Cherry-picked examples from the LLaMAntino-2 generated parallel dataset. For each original sentence, multiple simplifications at various readability levels are provided.

lowing function: $C_{cw} = \lfloor \log_2 \frac{freq(MFL)}{freq(CL)} \rfloor$, where MFL is the most frequent lemma in the corpus and CL is the considered lemma.

4. Linguistic Analysis of Simplified Sentences

The evaluation of the LLaMAntino-2–generated simplified sentences was conducted both in terms of readability scores (see Section 4.1) and linguistic profiles (see Section 4.2) in comparison to their corresponding original sentences. In addition, we investigated whether there is a relationship between the changes in linguistic features and the variation in readability levels across original/simplified sentence pairs, with the aim of identifying which linguistic phenomena are most associated with variation in linguistic complexity (see Section 4.3). All evaluations were conducted considering a randomly sampled subset of 2,000 paired original/simplified sentences for each domain[7]. Finally, Section 4.4 presents the results of a comparative analysis designed to examine whether different approaches to the construction of ATS resources influence the linguistic characteristics of simplified texts.

4.1. Sentence Readability

The first evaluation step was conducted by considering, for each original sentence, three representative cases among the multiple automatically generated simplifica-

[7] The dataset is freely available at https://github.com/michelepapucci/multilevel-text-simplification-italian

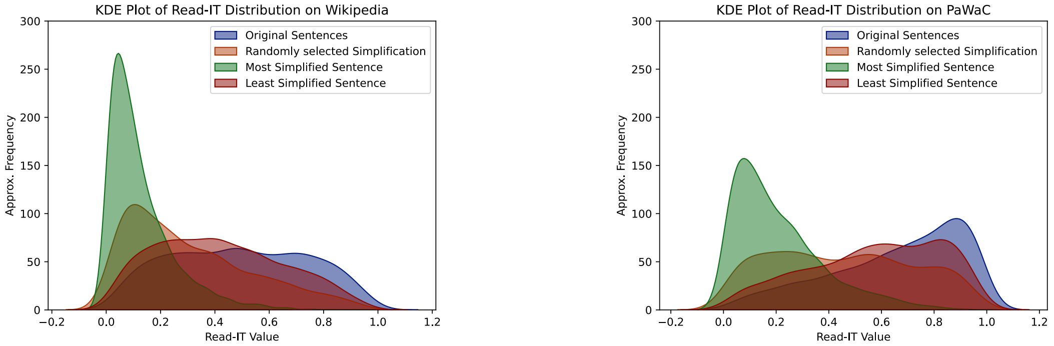

Figure 1: For both Wikipedia on the left and PaWaC on the right, the Kernel Density Estimate for the READ-IT.

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 Part of Speech
Lexical density
Dependency Syntactic Relations
Distribution of 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 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 3
Linguistic features used for linguistic profiling.

tions: the *Most simplified sentence*, i.e. the one with the lowest READ-IT score, the *Least simplified sentence*, with the highest score, and a *Randomly-selected simplification*, selected from the remaining simplifications. The comparison was computed adopting the Kernel Density Estimation (KDE), a probability distribution estimate obtained by smoothing out the READ-IT data points to create a continuous curve. Results are reported in Figure 1, where we can see that for both domains, all three types of simplifications exhibit a higher frequency of data points with lower READ-IT scores, confirming that the simplified sentences are generally easier to read. However, the shape of the distributions indicates that readability improvements vary depending on the source domain. Specifically, Wikipedia original sentences show a more uniform distribution across READ-IT scores, while PaWaC sentences are more concentrated at the higher end of the readability spectrum. This indicates that the simplified sentences in the administrative corpus remain less accessible than Wikipedia simplified sentences, reflecting the intrinsically higher linguistic complexity of administrative texts. Looking at the multiple simplifications, the *Most simplified sentences* exhibit a strongly left-skewed distribution in both domains, indicating that at least one version per original achieves significantly lower READ-IT scores. For the *Randomly-selected simplifications*, the KDE curve for Wikipedia shows a marked shift toward lower scores, suggesting that model-generated simplifications are generally simpler than their originals. A similar trend is observed for the PaWaC domain, although the distribution is flatter and less uniform, indicating greater variability across the simplified outputs.

4.2. Linguistic Features

The linguistic profile–based evaluation is twofold. The first level focuses on analyzing the differences between each of the three types of generated simplifications and

	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	$\leq 10^{-4}$

Table 4
Pillai's Trace reported from a MANOVA test between the linguistic features representing the simplified and original sentences.

Feature	Original	Simplified	*r*
sent_len	31.61 (±15.45)	22.98 (±12.94)	0.89
aux_Sub	2.06 (±12.24)	1.02 (±9.05)	0.87
verbal_head	2.60 (±1.62)	1.96 (±1.34)	0.83
subord_3	1.67 (±11.16)	0.62 (±6.86)	0.83
tree_depth	5.40 (±2.03)	4.48 (±3.22)	0.81
subord_prop	43.38 (±31.29)	29.34 (±31.18)	0.80
verbs_Ind	61.29 (±48.45)	44.36 (±49.52)	0.77
verbs_Fut	0.65 (±6.38)	0.40 (±5.26)	0.77
avg_Schain_len	0.82 (±0.65)	0.57 (±0.63)	0.77
n_prep_chains	1.90 (±1.48)	1.38 (±1.27)	0.76
links_len_max	13.61 (±9.33)	9.96 (±7.27)	0.73
subord_post	59.72 (±46.98)	42.09 (±48.23)	0.73
highest_class	19.40 (±3.87)	18.26 (±4.08)	0.72
principal_prop	54.12 (±31.71)	66.81 (±33.41)	-0.71
dep_iobj	0.13 (±0.84)	0.07 (±0.62)	0.68
verbs_Sing3	48.29 (±48.48)	35.30 (±47.24)	0.68
obj_pre	3.04 (±14.79)	1.92 (±12.27)	0.66
subord_1	59.11 (±47.36)	42.94 (±48.55)	0.65
verbs_Ger	3.97 (±12.60)	2.18 (±9.71)	0.65
dep_aux	1.18 (±2.19)	2.15 (±3.29)	-0.62
upos_AUX	3.30 (±3.36)	5.00 (±4.81)	-0.61
links_len_avg	2.61 (±0.62)	2.36 (±0.63)	0.60
verbs_Plur3	13.52 (±32.02)	9.43 (±28.30)	0.57
avg_Pchain_len	1.07 (±0.61)	0.92 (±0.67)	0.52
aux_Part	3.59 (±12.25)	5.67 (±15.27)	-0.53
subj_post	10.88 (±27.62)	7.23 (±23.66)	0.52
dep_appos	0.70 (±1.66)	0.43 (±1.36)	0.51
obj_post	51.36 (±49.23)	43.53 (±49.21)	0.50
verbs_Pres	26.26 (±38.20)	20.27 (±37.02)	0.49
aux_Pres	37.76 (±46.09)	47.00 (±46.47)	-0.49
verb_edges_5	8.83 (±21.51)	5.38 (±18.50)	0.45
verb_edges_0	0.43 (±4.16)	0.28 (±3.51)	0.44
dep_parataxis	0.14 (±0.70)	0.08 (±0.63)	0.44
verb_edges_1	12.57 (±23.88)	8.75 (±21.70)	0.43
subord_2	8.06 (±24.40)	5.50 (±21.00)	0.42
verbs_Fin	39.87 (±37.92)	31.69 (±39.65)	0.42
aux_Inf	2.54 (±12.99)	1.88 (±10.75)	0.41

Table 5
Mean (and standard deviation) of linguistic feature distribution in original and simplified Wikipedia sentences, ordered by decreasing |*r*| value, with $|r| \geq 0.4$.

Feature	Original	Simplified	*r*
aux_Ger	0.42 (±4.97)	0.17 (±2.73)	0.95
sent_len	48.10 (±30.06)	31.07 (±19.76)	0.89
n_prep_chains	3.47 (±2.66)	2.27 (±1.90)	0.87
verbal_head	2.91 (±2.28)	2.04 (±1.74)	0.78
tree_depth	7.44 (±3.52)	5.79 (±3.11)	0.76
aux_Fut	8.20 (±26.50)	4.73 (±20.34)	0.75
verbs_Sing1	0.48 (±6.15)	0.13 (±2.85)	0.73
aux_Cnd	0.80 (±7.66)	1.40 (±11.07)	-0.73
links_len_max	23.71 (±22.32)	14.48 (±12.95)	0.73
subord_dist	53.54 (±35.68)	38.60 (±36.51)	0.71
verbs_Imp	0.88 (±7.66)	0.48 (±5.83)	0.69
subord_post	62.59 (±45.49)	46.69 (±48.14)	0.66
highest_class	19.20 (±4.09)	17.87 (±4.10)	0.64
aux_Sub	3.17 (±14.83)	2.05 (±12.45)	0.64
avg_Schain_len	0.82 (±0.57)	0.64 (±0.63)	0.64
verbs_Sub	2.18 (±13.01)	1.16 (±9.69)	0.63
subord_1	66.19 (±44.77)	51.20 (±48.79)	0.61
obj_pre	1.13 (±8.89)	0.61 (±6.69)	0.61
dep_iobj	0.04 (±0.34)	0.03 (±0.37)	0.60
dep_list	0.05 (±0.97)	0.01 (±0.23)	0.59
avg_links_len	2.91 (±1.35)	2.57 (±1.72)	0.57
verbs_Sing2	0.27 (±4.80)	0.72 (±8.11)	-0.55
principal_prop	37.26 (±33.51)	47.20 (±38.14)	-0.53
upos_AUX	2.30 (±2.93)	3.38 (±4.37)	-0.52
verb_edges_6	3.25 (±13.70)	1.48 (±9.40)	0.51
subj_post	17.84 (±33.94)	12.24 (±29.83)	0.51
verbs_Fut	2.25 (±12.02)	1.61 (±10.73)	0.49
dep_cop	0.42 (±1.30)	0.77 (±2.20)	-0.49
aux_Imp	1.35 (±9.98)	0.93 (±8.88)	0.47
verbs_Ind	39.77 (±48.41)	31.05 (±46.01)	0.47
verbs_Sing3	30.65 (±44.39)	22.98 (±41.08)	0.45
dep_aux	0.96 (±1.66)	1.34 (±2.49)	-0.44

Table 6
Mean (and standard deviation) of linguistic feature distribution in original and simplified PaWac sentences, ordered by decreasing |*r*| value, with $|r| \geq 0.4$.

their corresponding original sentence, in terms of linguistic profile. To this end, we applied a Multivariate Analysis of Variance (MANOVA), which, unlike traditional ANOVA that considers only a single dependent variable, MANOVA evaluates whether the mean vectors of multiple dependent variables differ significantly between groups, making it well-suited to our multi-feature linguistic profiling. To quantify the degree of difference in each comparison, we report Pillai's Trace, one of the statistics derived from MANOVA. Pillai's Trace is particularly robust, especially in situations where assumptions like homogeneity of covariance matrices may be violated. Higher values of Pillai's Trace indicate greater multivariate differences between groups.

The results, summarized in Table 4, show that all comparisons yield statistically significant differences ($p \leq 10^{-4}$) in both domains. Among the three sets, the *Least Simplified* sentences consistently yield the smallest Pillai's Trace values (.12 for Wikipedia and .16 for PaWaC), indicating the greatest similarity to the original sentences. In contrast, the *Most Simplified* sentences show the highest values (.44 and .46), indicating that the simplification process led to substantial transformations in their linguistic profiles. The *Randomly-Selected* simplifications fall in between, though they are closer to the least simplified set, indicating that they retain a considerable degree of the original sentences' linguistic characteristics. This aligns with the trend observed in Figure 1, where the KDE curve for the *Randomly-Selected* simplifications peaks at lower READ-IT scores, similar to the most simplified set, but also shows a broader tail, indicating that some of these sentences remain close in readability to the originals. This trend is shared across domains, even with some differences that highlight domain-specific characteristics of the simplification process.

Notably, we generally observe slightly higher Pillai's Trace values for the PaWaC dataset. This suggests that, although simplified sentences in the administrative domain tend to have higher READ-IT scores than those from Wikipedia, the MANOVA results indicate that their generation involves more substantial transformations, possibly affecting multiple linguistic features, pointing to more articulated simplification processes in this domain. Consequently, even the *Least Simplified* PaWaC sentences display a more distinct linguistic profile compared to their originals.

Feature-based Analysis. It is focused on the set of *Randomly-selected Simplifications*, which serve as representative examples of typical simplifications, as they were randomly selected from the pool excluding the extremes. Specifically, we applied the Wilcoxon signed-rank test (with $p < 0.05$) to compare the distribution of each feature between the original sentence and its corresponding simplification. In addition, to quantify the strength of the observed differences, we computed their rank-biserial correlation score r [35], which ranges between $+1$ (when the value of the feature occurring in the original sentence is higher than in the simplified sentence) and -1 (in the opposite case). By capturing the effect size of the Wilcoxon test, the r score reflects the magnitude of statistically significant distributional differences. Tables 5 and 6 show features with $|r| \geq 0.4$ and their mean and standard deviation for the Wikipedia and PaWac domains[8].

Quite interestingly, a subset of the reported features is shared across the two domains. This suggests that these features correspond to linguistic phenomena highly related to sentence complexity, regardless of the textual domain, and are typically modified to improve sentence readability. As expected, among these features we find sentence length (*sent_len*), which displays the highest r score in Wikipedia and the second highest in PaWaC. However, by inspecting the differences across domains, we observe that administrative sentences are particularly shortened compared to their originals. Since the majority of the features considered are closely tied to sentence length, this outcome may impact the distribution of the other most varying features.

Nevertheless, we can see that several features modeling different syntactic properties of sentences are highly ranked in terms of r score for both domains. One such feature is the distribution of verbal heads (*verbal_head*), i.e. tokens POS-tagged as verbs that function as the syntactic head in dependency relations, which is notably reduced in the simplified sentences. This reduction is closely linked to the decreased use of subordination, as indicated by lower values of a set of related features capturing this phenomenon. The set includes: the overall distribution of subordinate clauses (*subord_prop*), their position relative to the principal clause (*subord_post*), and their organization into sequences of embedded subordinate clauses (*avg_Schain_len*). Among these, we can also include a feature from the verb inflectional morphology group that is closely related to reduced subordination: the lower distribution of subjunctives (*aux_Sub*). Additionally, features modeling both global and local aspects of syntactic tree structure vary significantly in both domains. These include syntactic tree depth (*tree_depth*), indicative of sentence complexity [36], as well as two features associated with long-distance dependencies, well-known sources of cognitive load [37, 38]: the length of the longest dependency link (*links_len_max*) and the number of embedded sequences of prepositional complements (*n_prep_chains*). A similar pattern is observed in the lower frequency of subjects and objects in non-canonical position occurring in simplified sentences, specifically pre-verbal objects (*obj_pre*) and post-verbal subjects (*subj_post*), both known to be harder to process. On the lexical side, simplified sentences in both domains exhibit a reduced proportion of lemmas from the highest frequency class (*highest_class*). Interestingly, both domains display negative r scores for the distribution of auxiliary verbs (*upos_AUX* and *dep_aux*), indicating an increase in auxiliary usage in simplified versions. An in-depth analysis of verb forms reveals that this may reflect a higher prevalence of 'passato prossimo' tenses (roughly present perfect tenses) and a corresponding reduction of 'passato remoto' (roughly simple pasts), particularly in Wikipedia.

When focusing on features that vary significantly and with $|r| \geq 0.4$ in only one domain, we find that they capture finer-grained phenomena. They predominantly involve the distribution of specific verb tenses, such as

[8] The full list of features is reported in Appendix C.

present tense forms (**_Pres*) in Wikipedia (whereas in PaWaC they show only $|r| = 0.15$), and future (**_Fut*) and imperfect (**_Imp*) tenses in PaWaC (but not significantly varying in Wikipedia). A similar trend is observed for specific verb moods such as particles (**_Part*), which vary above our threshold only in Wikipedia, and conditionals (**_Cond*), varying significantly in PaWaC.

4.3. Linguistic Features and Readability

As a third level of analysis, we investigated which linguistic phenomena characterize automatically simplified sentences in relation to the differences in readability between the original and simplified versions. To this end, considering the *Randomly-selected simplification*, we computed Spearman correlations between the differences in the distribution of the linguistic features, extracted using Profiling-UD, and the corresponding differences in their READ-IT scores. The results are reported in Appendix B, where we compare the correlation scores for the Wikipedia and PaWac domains. We focus on the set of linguistic features that show statistically significant correlations (i.e. $p < 0.05$).

As can be seen, most of the correlation scores are positive. This suggests that an increase in the difference of specific linguistic features between original and simplified sentences is often directly proportional to the increase in their readability difference. This is the case, for example, for the distribution of subordinate clauses (*subordinate_proposition*) in both domains, which tend to be significantly reduced in the simplified sentences, leading to lower syntactic complexity and, consequently, a lower READ-IT score. By contrast, the difference in the distribution of auxiliary verbs (*upos_dist_AUX*) shows a negative correlation with the difference in READ-IT scores for both domains, as the distribution of auxiliaries increases in the simplified sentences.

Cross-Domain Correlation Patterns. When ranking the linguistic features in decreasing order of correlation, we observe that the most strongly correlated features are shared across both domains, despite differences in correlation scores. Notably, many of the top-ranked ones correspond to those discussed in the previous section. This seems to support the hypothesis that the linguistic phenomena mostly involved in the transformations of original sentences are also those that have the greatest impact on sentence readability.

As expected, the most strongly correlated feature is sentence length (*tokens_per_sent*), which is considerably reduced in the simplified sentences. Interestingly, even if this pattern holds across both domains, the correlation is stronger for Wikipedia ($r = 0.51$) than for PaWac ($r = 0.42$). This seems to align with and complement the intuition that simplifying administrative texts is particularly challenging, as many of the PaWac sentences tend to exhibit a relatively high level of linguistic complexity even after simplification (see Figure 1). It is therefore plausible that a surface-level transformation such as reducing sentence length is less predictive of changes in readability scores in this domain. This interpretation is also consistent with the MANOVA results, which indicate that simplified PaWaC sentences differ more substantially from their original versions across multiple linguistic features, suggesting a more articulated simplification process.

Among the top-ranked correlated features, we find several that, while sensitive to sentence length, also reflect deeper, linguistically motivated transformations involved in the simplification process. This is the case of the distribution of verbal heads (*verbal_head_per_sent*) and of a subset of related features modeling the subordination. These include: the overall distribution of subordinate clauses (*subordinate_proposition_dist*); their organization in recursively embedded subordinate clause chains within a top-level subordinate clause (*avg_subordinate_chain_len_diff*); their relative order with respect to the principal clause (*subordinate_post*), a characteristic associated with differences in cognitive processing difficulty [39]; and a specific type of subordinate clauses, i.e. relative clauses (*dep_dist_acl:relcl*), which are well-known sources of processing difficulty. In addition, we find two features related to long-distance constructions: the length of the longest dependency link in a sentence (*max_links_len*) and the number of embedded sequences of prepositional complements governed by a nominal head (*n_prepositional_chains*).

Focusing on lexical variation, the reduction in the proportion of lemmas belonging to the highest frequency class (*highest_class*) shows a positive correlation with readability improvement, particularly in PaWac ($r = 0.20$) compared to Wikipedia ($r = 0.16$). Conversely, a slight increase in the use of 'high availability words' (lower-frequency lemmas referring to everyday objects or actions and well known to speakers), as identified in the NBIV (*in_AD_types*), is negatively correlated in both domains.

4.4. Comparing Simplification Approaches

We complemented the linguistic profiling of the LLaMAntino-2–generated simplified sentences with a comparative analysis aimed at identifying whether certain linguistic phenomena are specific to the LLM-based approach to ATS resource construction or are shared across different simplification methodologies. To this end, we started from the findings of [15], who compared two Italian ATS resources created manually, "Teacher" and "Terence" [16], and one semi-automatically, PaCCSS-IT [7], focusing on the distribution of a set of linguistic features comparable to those used in the present study. Our main goal is to assess whether some linguistic features are characteristic of simplified sentences regardless of the simplification method adopted. While preliminary, our results provide initial insights into whether an LLM-based method yields simplified sentences with characteristics similar to those produced by human experts.

The first characteristic shared by sentences simplified by both human experts and automatically generated concerns their sentence length. Simplified sentences are always shorter than their original counterparts. This could be expected since sentence length has been considered as a shallow proxy of sentence complexity and is widely used by traditional readability assessment formulas. However, the different average length in original-simplified sentence pairs may differ according to textual genre, as shown in our analysis and discussed in [15].

A second group of features common to all ATS resources includes those modeling the morpho-syntactic profile of the simplified sentences[9]. Similarly to manually and semi-automatically built simplifications, the sentences automatically generated by LLaMAntino-2 tend to contain fewer pronouns, adverbs, and punctuation marks, and a higher proportion of determiners. However, in contrast to the findings reported in [15], which were also based on the Wilcoxon signed-rank test ($p < 0.05$), the LLM-generated simplified sentences exhibit a higher frequency of nouns, and the variation in the distribution of adjectives compared to the original sentences is not statistically significant. We leave to future work the investigation of whether this trend may be influenced by the textual genre of the original sentences.

Among the features common across approaches, we find those capturing global and local syntactic structure. As also observed in Section 4.2, simplified sentences tend to have shallower syntactic trees and shorter dependency links, suggesting that reducing syntactic depth and dependency length is a broadly adopted simplification strategy. However, when examining finer-grained syntactic properties, some differences emerge. A first example concerns the use of subordination. While previous studies suggest that subordinate clauses following the main clause are easier to process [39], only the "Terence" corpus and PaCCSS-IT show a higher percentage of post-verbal subordinates. By contrast, an opposite trend is observed in the sentences automatically generated by LLaMAntino-2 as well as in the manually built "Teacher" corpus, where post-verbal subordinates are less frequent. A second example is the distribution of subjects. All resources show an increased presence of overt subjects in simplified sentences, particularly in the "Teacher" corpus, representing an intuitive manual simplification in [15]. This aligns with observations about the insertion of explicit arguments to reduce the inference load associated with null-subject constructions [40]. Interestingly, however, the tendency to favor the canonical Italian argument order, with subjects preceding the verb and objects following it, is not consistently observed across resources. While unmarked word orders are generally preferred in simplification, as they are known to ease processing in free word-order languages [41], a higher proportion of pre-verbal subjects is found only in the PaWac LLaMAntino-2-generated simplifications and in the Teacher corpus. An even less consistent pattern emerges for post-verbal objects, whose distribution differs across original and simplified sentences without a systematic direction.

5. Conclusion

This study investigated the ability of small LLMs fine-tuned on the Italian language to generate sentence simplifications in a zero-shot setting, focusing on two linguistically distinct domains: Wikipedia and Public Administration. All tested models were able to produce simplified sentences that preserved the surface-level properties and semantic content of the original inputs while improving readability. Among them, LLaMAntino-2 consistently outperformed the other models across all evaluation metrics. Beyond single-sentence simplification, we also showed that prompting the model to generate multiple outputs for the same input sentence results in a meaningful gradient of linguistic complexity.

Domain-specific analyses revealed that, although simplified sentences in the administrative domain remain less accessible than their Wikipedia counterparts, simplifying administrative texts involves more substantial linguistic transformations, as suggested by MANOVA results, thus pointing to more complex simplification strategies in this domain. These findings highlight the potential of this approach to support the development of ATS resources tailored to specific reader profiles and domains. Despite a few cross-domain differences, our analysis of the linguistic features most affected by simplification shows that many transformations are shared across domains and closely align with known simplification patterns found in manually constructed ATS corpora.

These findings support two key directions for future work. First, the generation of synthetic simplifications using small, language-specific LLMs offers a promising method for building ATS resources in low-resource settings. Second, the linguistic properties characterizing LLM-generated simplifications can inform Controllable Text Generation approaches [42], enabling models to be guided toward specific simplification strategies aligned with the needs of different reader populations.

[9] The values of some linguistic features are not reported in Tables 6 and 5, as their rank-biserial correlation scores are $|r| \leq 0.4$.

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" (PNRR-MAD-2022-12376692_VADALA' – CUP F83C22002470001) and by the PRIN 2022 project TEAMING-UP - Teaming up with Social Artificial Agents (20177FX2A7) funded by the Italian Ministry of University and Research.

References

[1] F. Alva-Manchego, C. Scarton, L. Specia, Data-driven sentence simplification: Survey and benchmark, Computational Linguistics 46 (2020) 135–187. URL: https://aclanthology.org/2020.cl-1.4/. doi:10.1162/coli_a_00370.

[2] M. J. Ryan, T. Naous, W. Xu, Revisiting non-English text simplification: A unified multilingual benchmark, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 4898–4927. URL: https://aclanthology.org/2023.acl-long.269/. doi:10.18653/v1/2023.acl-long.269.

[3] D. Kauchak, Improving text simplification language modeling using unsimplified text data, in: H. Schuetze, P. Fung, M. Poesio (Eds.), Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 1537–1546. URL: https://aclanthology.org/P13-1151/.

[4] D. Pellow, M. Eskenazi, An open corpus of everyday documents for simplification tasks, in: S. Williams, A. Siddharthan, A. Nenkova (Eds.), Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR), Association for Computational Linguistics, Gothenburg, Sweden, 2014, pp. 84–93. URL: https://aclanthology.org/W14-1210/. doi:10.3115/v1/W14-1210.

[5] F. Alva-Manchego, L. Martin, A. Bordes, C. Scarton, B. Sagot, L. Specia, ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4668–4679. URL: https://aclanthology.org/2020.acl-main.424/. doi:10.18653/v1/2020.acl-main.424.

[6] W. Xu, C. Callison-Burch, C. Napoles, Problems in current text simplification research: New data can help, Transactions of the Association for Computational Linguistics 3 (2015) 283–297. URL: https://aclanthology.org/Q15-1021/. doi:10.1162/tacl_a_00139.

[7] D. Brunato, A. Cimino, F. Dell'Orletta, G. Venturi, PaCCSS-IT: A parallel corpus of complex-simple sentences for automatic text simplification, in: J. Su, K. Duh, X. Carreras (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 351–361. URL: https://aclanthology.org/D16-1034/. doi:10.18653/v1/D16-1034.

[8] L. Martin, A. Fan, É. de la Clergerie, A. Bordes, B. Sagot, MUSS: Multilingual unsupervised sentence simplification by mining paraphrases, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 1651–1664. URL: https://aclanthology.org/2022.lrec-1.176/.

[9] A. Palmero Aprosio, S. Tonelli, M. Turchi, M. Negri, M. A. Di Gangi, Neural text simplification in low-resource conditions using weak supervision, in: A. Bosselut, A. Celikyilmaz, M. Ghazvininejad, S. Iyer, U. Khandelwal, H. Rashkin, T. Wolf (Eds.), Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 37–44. URL: https://aclanthology.org/W19-2305/. doi:10.18653/v1/W19-2305.

[10] M. Miliani, F. Alva-Manchego, A. Lenci, Simplifying administrative texts for Italian L2 readers with controllable transformers models: A data-driven approach, in: F. Boschetti, G. E. Lebani, B. Magnini, N. Novielli (Eds.), Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023), CEUR Workshop Proceedings, Venice, Italy, 2023, pp. 303–315. URL: https://aclanthology.org/2023.clicit-1.37/.

[11] L. Long, R. Wang, R. Xiao, J. Zhao, X. Ding, G. Chen, H. Wang, On LLMs-driven synthetic data generation, curation, and evaluation: A survey, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 11065–11082. URL: https://aclanthology.org/2024.findings-acl.658/. doi:10.

18653/v1/2024.findings-acl.658.

[12] T. Kew, A. Chi, L. Vásquez-Rodríguez, S. Agrawal, D. Aumiller, F. Alva-Manchego, M. Shardlow, BLESS: Benchmarking large language models on sentence simplification, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 13291–13309. URL: https://aclanthology.org/2023.emnlp-main.821/. doi:10.18653/v1/2023.emnlp-main.821.

[13] D. Nozza, G. Attanasio, Is it really that simple? prompting large language models for automatic text simplification in Italian, in: F. Boschetti, G. E. Lebani, B. Magnini, N. Novielli (Eds.), Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023), CEUR Workshop Proceedings, Venice, Italy, 2023, pp. 322–333. URL: https://aclanthology.org/2023.clicit-1.39/.

[14] M. Russodivito, V. Ganfi, G. Fiorentino, R. Oliveto, AI vs. human: Effectiveness of LLMs in simplifying Italian administrative documents, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 842–853. URL: https://aclanthology.org/2024.clicit-1.91/.

[15] D. Brunato, F. Dell'Orletta, G. Venturi, Linguistically-based comparison of different approaches to building corpora for text simplification: A case study on italian, Frontiers in Psychology Volume 13 - 2022 (2022). URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.707630. doi:10.3389/fpsyg.2022.707630.

[16] D. Brunato, F. Dell'Orletta, G. Venturi, S. Montemagni, Design and annotation of the first Italian corpus for text simplification, in: A. Meyers, I. Rehbein, H. Zinsmeister (Eds.), Proceedings of the 9th Linguistic Annotation Workshop, Association for Computational Linguistics, Denver, Colorado, USA, 2015, pp. 31–41. URL: https://aclanthology.org/W15-1604/. doi:10.3115/v1/W15-1604.

[17] M. Polignano, P. Basile, G. Semeraro, Advanced natural-based interaction for the italian language: Llamantino-3-anita, 2024. arXiv:2405.07101.

[18] P. Basile, E. Musacchio, M. Polignano, L. Siciliani, G. Fiameni, G. Semeraro, Llamantino: Llama 2 models for effective text generation in italian language, 2023. arXiv:2312.09993.

[19] S. Tonelli, A. P. Aprosio, F. Saltori, Simpitiki: a simplification corpus for italian, Proceedings of CLiC-it (2016).

[20] M. Miliani, S. Auriemma, F. Alva-Manchego, A. Lenci, Neural readability pairwise ranking for sentences in Italian administrative language, in: Y. He, H. Ji, S. Li, Y. Liu, C.-H. Chang (Eds.), 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), Association for Computational Linguistics, Online only, 2022, pp. 849–866. URL: https://aclanthology.org/2022.aacl-main.63/. doi:10.18653/v1/2022.aacl-main.63.

[21] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040/. doi:10.3115/1073083.1073135.

[22] W. Xu, C. Napoles, E. Pavlick, Q. Chen, C. Callison-Burch, Optimizing statistical machine translation for text simplification, Transactions of the Association for Computational Linguistics 4 (2016) 401–415. URL: https://aclanthology.org/Q16-1029/. doi:10.1162/tacl_a_00107.

[23] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[24] N. Reimers, I. Gurevych, 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, 2019. URL: https://arxiv.org/abs/1908.10084.

[25] N. Reimers, I. Gurevych, 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, 2020. URL: https://arxiv.org/abs/2004.09813.

[26] F. Dell'Orletta, S. Montemagni, G. Venturi, READ–IT: Assessing readability of Italian texts with a view to text simplification, in: N. Alm (Ed.), Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies, Association for Computational Linguistics, Edinburgh, Scotland, UK, 2011, pp. 73–83. URL: https://aclanthology.org/W11-2308/.

[27] L. C. Passaro, A. Lenci, PaWaC - Public Administration Web as Corpus (Processed), http://data.europa.eu/88u/dataset/elrc_1282, 2019. [Data set].

[28] M. Cortelazzo, Il linguaggio amministrativo: principi e pratiche di modernizzazione, Carocci, 2021.

[29] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. J. Crandall, D. Batra, Diverse beam search: Decoding diverse solutions from neural sequence models, CoRR abs/1610.02424 (2016). URL: http://arxiv.org/abs/1610.02424. arXiv:1610.02424.

[30] D. Brunato, A. Cimino, F. Dell'Orletta, G. Venturi, S. Montemagni, Profiling-UD: a tool for linguistic profiling of texts, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 7145–7151. URL: https://aclanthology.org/2020.lrec-1.883/.

[31] M.-C. De Marneffe, C. D. Manning, J. Nivre, D. Zeman, Universal dependencies, Computational linguistics 47 (2021) 255–308.

[32] T. De Mauro, I. Chiari, Il nuovo vocabolario di base della lingua italiana, Internazionale [accessed on 03/03/2023] (2016). URL: https://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana.

[33] G. Sarti, M. Nissim, IT5: Text-to-text pretraining for Italian language understanding and generation, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italy, 2024, pp. 9422–9433. URL: https://aclanthology.org/2024.lrec-main.823.

[34] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, 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, Association for Computational Linguistics, Online, 2021, pp. 483–498. URL: https://aclanthology.org/2021.naacl-main.41. doi:10.18653/v1/2021.naacl-main.41.

[35] H. W. Wendt, Dealing with a common problem in social science: A simplified rank-biserial coefficient of correlation based on the statistic., European J. of Social Psychology (1972).

[36] L. Frazier, Syntactic complexity, in: D. Dowty, L. Karttunen, A. Zwicky (Eds.), Natural Language Parsing, Cambridge University Press, Cambridge, UK, 1985.

[37] E. Gibson, Linguistic complexity: Locality of syntactic dependencies, Cognition 24 (1998) 1–76.

[38] V. Demberg, F. Keller, Data from eye-tracking corpora as evidence for theories of syntactic processing complexity, Cognition 109 (2008) 193–210.

[39] J. Miller, R. Weinert, Spontaneous spoken language. Syntax and discourse, Oxford University Press, 1998.

[40] G. Barlacchi, S. Tonelli, Ernesta: A sentence simplification tool for children's stories in italian, in: Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Springer Berlin Heidelberg, 2013, pp. 476–487.

[41] M. HASPELMATH, Against markedness (and what to replace it with), Journal of Linguistics 42 (2006) 25–70. doi:10.1017/S0022226705003683.

[42] Z. Li, M. Shardlow, How do control tokens affect natural language generation tasks like text simplification, Natural Language Engineering 30 (2024) 915–942. doi:10.1017/S1351324923000566.

A. Prompt Template for Sentence Simplification

Each model was prompted using its respective system prompt provided in the Hugging Face documentation. We also provided a task-specific prompt to instruct the model to perform the Sentence Simplification task. The following prompt pattern was used:

```
### Istruzione: Semplifica la
    seguente frase mantenendo il
    più possibile intatto il
    significato.
### Input: {original_sentence}
### Output:
```

English translation: "`Instruction: Simplify the following sentence while keeping the meaning the same as much as possible.`".

B. Linguistic Features and Readability Correlation Heatmap

Figure 2 reports the full list of statistically significant Spearman correlations ($p < 0.05$) between the differences in linguistic feature distributions, automatically extracted using Profiling-UD, from the subset of 2,000 original/simplified sentence pairs, and the corresponding differences in their READ-IT scores.

C. Linguistic Features of Original and Simplified Sentences

Tables 7 and 8 complement the results discussed in Section 4.2 and focus on the differences in the distribution of linguistic features between the original and the corresponding *Randomly-selected* simplified sentences. They report the set of features that vary in a statistically significant way ($p < 0.05$) and have effect size scores from the Wilcoxon test where $|r| \leq 0.4$. Specifically, these results extend those in Tables 5 and 6, which highlight features with stronger effects ($|r| \geq 0.4$).

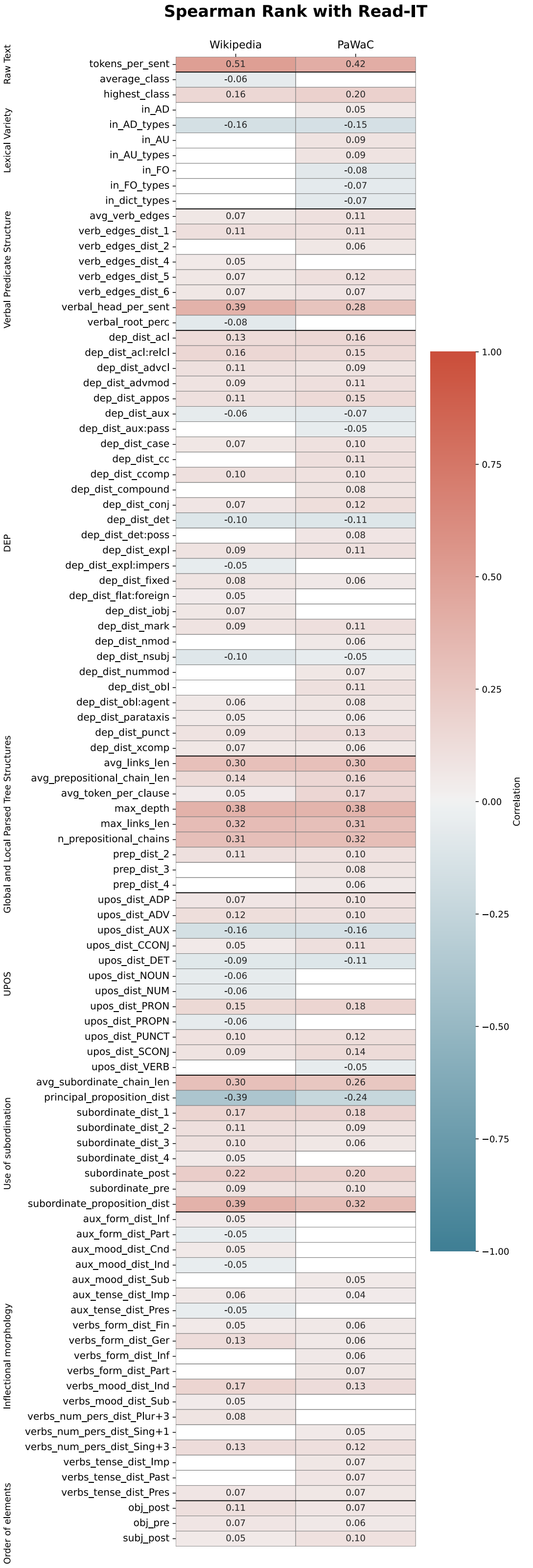

Figure 2: Correlation between linguistic feature differences (original vs. simplified) and READ-IT all scores. Each column refers to one domain (PaWaC or Wikipedia). White cells indicate non-significant correlations.

Feature	Original	Simplified	***r***
aux_form_dist_Inf	2.54 (±12.99)	1.88 (±10.75)	0.41
dep_dist_cop	1.04 (±2.00)	1.46 (±3.10)	-0.39
aux_tense_dist_Imp	9.36 (±27.71)	7.00 (±24.33)	0.38
verb_edges_dist_6	2.09 (±10.60)	1.32 (±9.30)	0.38
dep_dist_nsubj:pass	0.78 (±1.74)	1.09 (±2.46)	-0.37
upos_dist_PRON	2.60 (±3.38)	2.01 (±3.41)	0.36
subordinate_pre	9.48 (±25.68)	7.16 (±23.51)	0.36
dep_dist_acl:relcl	0.92 (±1.65)	0.66 (±1.63)	0.36
dep_dist_aux:pass	1.07 (±2.01)	1.38 (±2.67)	-0.35
dep_dist_flat:foreign	0.18 (±1.26)	0.21 (±3.05)	0.35
aux_mood_dist_Ind	58.20 (±48.66)	64.56 (±47.56)	-0.34
upos_dist_ADV	3.09 (±3.74)	2.50 (±3.99)	0.34
dep_dist_nsubj	3.57 (±3.11)	4.26 (±3.99)	-0.34
dep_dist_advmod	2.72 (±3.54)	2.19 (±3.81)	0.34
avg_verb_edges	2.73 (±1.14)	2.50 (±1.27)	0.34
prep_dist_1	66.10 (±41.15)	59.19 (±45.65)	0.33
upos_dist_X	0.26 (±1.95)	0.29 (±3.95)	0.32
in_AD_types	0.08 (±0.04)	0.08 (±0.05)	-0.32
dep_dist_acl	1.08 (±1.92)	0.81 (±1.87)	0.31
upos_dist_SCONJ	0.65 (±1.57)	0.50 (±1.59)	0.31
avg_token_per_clause	14.11 (±8.29)	12.70 (±7.55)	0.31
prep_dist_2	15.12 (±28.07)	12.25 (±27.86)	0.29
aux_num_pers_dist_Sing+3	49.20 (±48.95)	53.97 (±49.22)	-0.28
dep_dist_advcl	1.02 (±1.90)	0.81 (±1.95)	0.27
dep_dist_ccomp	0.42 (±1.31)	0.35 (±1.40)	0.27
dep_dist_fixed	0.40 (±1.22)	0.33 (±1.28)	0.27
dep_dist_punct	11.59 (±6.50)	10.17 (±6.09)	0.25
verbs_tense_dist_Imp	4.85 (±18.38)	4.10 (±17.67)	0.25
upos_dist_PUNCT	11.57 (±6.50)	10.25 (±6.71)	0.25
dep_dist_expl	0.85 (±1.81)	0.72 (±2.08)	0.25
aux_tense_dist_Past	14.40 (±31.69)	12.56 (±27.92)	0.25
verbs_form_dist_Part	39.29 (±39.79)	43.47 (±43.41)	-0.21
dep_dist_det	14.68 (±5.28)	15.29 (±6.25)	-0.20
in_dict_types	0.74 (±0.15)	0.75 (±0.16)	-0.20
upos_dist_DET	15.55 (±5.58)	16.26 (±6.93)	-0.19
dep_dist_compound	0.27 (±1.09)	0.26 (±1.95)	0.19
aux_form_dist_Fin	56.88 (±46.78)	59.99 (±45.23)	-0.18
upos_dist_PROPN	9.31 (±9.01)	9.95 (±10.85)	-0.17
char_per_tok	4.76 (±0.61)	4.70 (±0.69)	0.16
in_FO_types	0.55 (±0.14)	0.56 (±0.15)	-0.15
verb_edges_dist_4	19.46 (±30.55)	17.10 (±31.78)	0.15
upos_dist_NUM	3.02 (±4.36)	3.18 (±5.07)	-0.14
dep_dist_case	15.08 (±5.53)	14.42 (±6.37)	0.14
verbs_form_dist_Inf	10.52 (±20.64)	9.96 (±21.94)	0.14
in_FO	0.58 (±0.13)	0.59 (±0.15)	-0.14
upos_dist_ADP	15.99 (±5.41)	15.37 (±6.31)	0.13
dep_dist_mark	1.51 (±2.53)	1.37 (±2.76)	0.13
dep_dist_flat:name	2.91 (±4.83)	3.13 (±5.92)	-0.12
upos_dist_NOUN	17.75 (±6.95)	18.01 (±7.76)	-0.12
in_AU_types	0.11 (±0.07)	0.11 (±0.08)	0.12
in_AU	0.10 (±0.07)	0.09 (±0.08)	0.11
dep_dist_conj	3.33 (±4.07)	3.10 (±4.74)	0.11
dep_dist_cc	2.63 (±2.84)	2.40 (±3.23)	0.11
upos_dist_VERB	7.60 (±4.38)	7.81 (±5.16)	-0.11
upos_dist_CCONJ	2.62 (±2.85)	2.39 (±3.22)	0.11
average_class	7.63 (±1.22)	7.57 (±1.37)	0.11
verb_edges_dist_3	27.52 (±33.79)	30.08 (±38.78)	-0.10
verb_edges_dist_2	22.44 (±31.24)	24.35 (±35.54)	-0.09
dep_dist_obl	6.82 (±4.29)	6.50 (±5.24)	0.09
dep_dist_nmod	8.25 (±5.32)	7.92 (±5.97)	0.08
dep_dist_amod	5.84 (±4.91)	5.62 (±5.71)	0.06
lexical_density	0.50 (±0.09)	0.50 (±0.10)	0.06

Table 7
Mean (and standard deviation) of linguistic feature distribution in original and simplified Wikipedia sentences, ordered by decreasing $|r|$ value, with $|r| \leq 0.4$.

Feature	Original	Simplified	***r***
obj_post	50.37 (±49.69)	42.54 (±49.30)	0.39
verb_edges_dist_5	7.86 (±20.03)	4.83 (±17.28)	0.38
avg_token_per_clause	18.18 (±14.93)	14.54 (±12.07)	0.38
avg_prepositional_chain_len	1.33 (±0.61)	1.17 (±0.72)	0.38
dep_dist_aux:pass	0.92 (±1.63)	1.26 (±2.44)	-0.36
dep_dist_nsubj	1.73 (±2.17)	2.43 (±3.39)	-0.36
in_AD_types	0.07 (±0.04)	0.08 (±0.05)	-0.35
prep_dist_4	1.43 (±7.64)	0.99 (±7.10)	0.35
dep_dist_ccomp	0.39 (±1.13)	0.30 (±1.28)	0.34
subordinate_dist_2	6.07 (±19.20)	4.37 (±18.02)	0.34
verbs_tense_dist_Past	56.78 (±41.57)	50.16 (±45.24)	0.33
verbs_num_pers_dist_Plur+3	10.28 (±28.04)	8.48 (±26.64)	0.33
verbs_form_dist_Ger	2.37 (±9.24)	1.94 (±9.34)	0.33
avg_verb_edges	2.39 (±1.25)	2.12 (±1.35)	0.33
in_dict_types	0.73 (±0.17)	0.74 (±0.20)	-0.32
verb_edges_dist_1	21.80 (±29.30)	17.62 (±30.01)	0.32
dep_dist_acl	1.78 (±2.15)	1.41 (±2.37)	0.32
dep_dist_punct	10.89 (±8.39)	9.14 (±6.93)	0.30
verbs_form_dist_Part	51.67 (±39.55)	45.58 (±42.86)	0.30
upos_dist_PUNCT	11.22 (±9.91)	9.51 (±8.96)	0.30
aux_tense_dist_Pres	39.96 (±46.56)	45.47 (±47.59)	-0.30
in_FO	0.52 (±0.15)	0.54 (±0.18)	-0.29
dep_dist_compound	0.49 (±1.32)	0.39 (±1.37)	0.29
dep_dist_expl	0.54 (±1.33)	0.45 (±1.51)	0.29
dep_dist_appos	0.74 (±1.63)	0.58 (±1.73)	0.28
in_FO_types	0.50 (±0.15)	0.52 (±0.18)	-0.27
dep_dist_obl	5.36 (±3.85)	4.61 (±4.22)	0.27
upos_dist_DET	14.19 (±6.08)	15.08 (±7.38)	-0.26
dep_dist_det	13.88 (±5.93)	14.71 (±7.15)	-0.26
dep_dist_flat	0.58 (±1.92)	0.92 (±3.74)	-0.26
dep_dist_xcomp	0.38 (±1.13)	0.31 (±1.26)	0.24
upos_dist_PRON	1.78 (±2.50)	1.58 (±3.12)	0.24
dep_dist_nsubj:pass	1.03 (±1.82)	1.30 (±2.53)	-0.23
prep_dist_1	62.97 (±35.54)	56.87 (±40.85)	0.23
dep_dist_advmod	2.01 (±3.18)	1.77 (±3.28)	0.22
average_class	7.51 (±1.56)	7.34 (±1.62)	0.22
prep_dist_3	6.08 (±15.81)	5.21 (±16.30)	0.21
aux_num_pers_dist_Sing+3	33.23 (±45.29)	36.06 (±46.91)	-0.21
in_AU	0.13 (±0.07)	0.12 (±0.09)	0.20
upos_dist_ADV	2.22 (±3.37)	1.98 (±3.54)	0.20
dep_dist_case	16.44 (±5.96)	15.28 (±7.17)	0.20
dep_dist_det:poss	0.17 (±0.70)	0.16 (±0.85)	0.20
subj_pre	54.11 (±46.58)	57.81 (±47.49)	-0.19
in_AU_types	0.16 (±0.08)	0.15 (±0.10)	0.19
upos_dist_ADP	17.14 (±6.11)	16.12 (±7.68)	0.19
upos_dist_SCONJ	0.56 (±1.35)	0.50 (±1.49)	0.19
dep_dist_nummod	3.06 (±4.69)	2.70 (±5.28)	0.19
in_dict	0.78 (±0.16)	0.79 (±0.20)	-0.18
dep_dist_nmod	13.06 (±6.84)	12.01 (±8.09)	0.18
prep_dist_2	21.85 (±28.36)	19.33 (±29.69)	0.18
dep_dist_conj	4.04 (±4.39)	3.67 (±5.27)	0.17
dep_dist_mark	1.29 (±2.06)	1.18 (±2.38)	0.17
verbs_form_dist_Inf	15.53 (±26.94)	14.31 (±28.05)	0.16
aux_num_pers_dist_Plur+3	18.91 (±37.08)	17.47 (±36.64)	0.16
verbs_tense_dist_Pres	25.19 (±34.69)	22.91 (±36.64)	0.15
dep_dist_obj	1.87 (±2.39)	2.33 (±3.82)	-0.14
upos_dist_CCONJ	2.71 (±2.77)	2.51 (±3.29)	0.12
dep_dist_cc	2.69 (±2.77)	2.49 (±3.28)	0.12
in_AD	0.13 (±0.06)	0.12 (±0.07)	0.12
aux_form_dist_Fin	44.86 (±45.28)	46.77 (±45.74)	-0.12
aux_mood_dist_Ind	48.83 (±48.97)	50.95 (±49.40)	-0.11
dep_dist_acl:relcl	0.63 (±1.25)	0.63 (±1.70)	0.10
verbs_form_dist_Fin	20.98 (±29.55)	20.02 (±32.91)	0.10
upos_dist_NOUN	23.82 (±6.86)	24.10 (±8.98)	-0.09
upos_dist_VERB	6.14 (±4.48)	6.55 (±6.06)	-0.09
lexical_density	0.51 (±0.11)	0.51 (±0.13)	0.08

Table 8
Mean (and standard deviation) of linguistic feature distribution in original and simplified PaWac sentences, ordered by decreasing |*r*| value, with $|r| \leq 0.4$.