AI vs. Human
Effectiveness of LLMs in Simplifying Italian Administrative Documents

Marco Russodivito, Vittorio Ganfi, Giuliana Fiorentino and Rocco Oliveto
University of Molise, Italy

Introduction

Due to the increasing popularity of **Generative Artificial Intelligence** (AI) language tools [1, 2], significant attention has been devoted to the use of LLMs for text simplification [3]. Several studies have addressed the application of LLMs to simplify texts, particularly focusing on administrative documents, including those in Italian [4, 5, 6]. **Italian administrative texts** are often notably **complex** and **obscure** [7, 8, 9], which restricts a large segment of the popultion from fully accessing the content produced by the Italian public administration [10, 11]. This work aims to (a) **evaluate** the quality of **automatic text simplification** performed by several well-known **LLMs**, and (b) **compare** LLM-based simplification **with human-based simplification**.

Study Design

How effective are AI systems at simplifying administrative texts compared to humans?

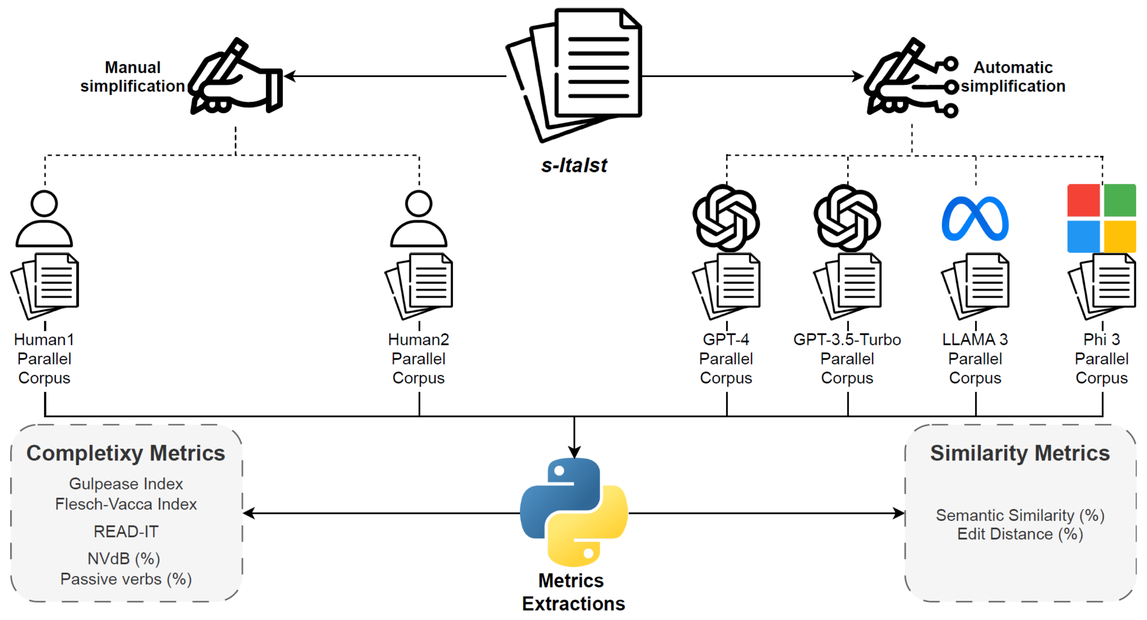
Experimental Procedure

Our empirical study can be summarized in three main steps:

i) constructing a **corpus of administrative documents** (i.e., s-ItaIst);

ii) **simplifying** this corpus using four **LLMs** and two **human** annotators;

iii) **comparing** the LLM-simplified corpora with the human-simplified **corpora**.



ItaIst Corpus

The **ItaIst corpus** (2,000,000 tokens and 45,000 types), created under the **VerbACxSS research project**, was composed by **linguists** and **jurists** to create a representative linguistic resource for **administrative Italian** [12, 13]. ItaIst was assembled by collecting recent official documents from local and regional public administration websites of **eight Italian regions** (Basilicata, Calabria, Campania, Lazio, Lombardy, Molise, Tuscany, and Veneto) covering topics such as **garbage, healthcare, and public services**. The corpus includes a variety of text types, such as **Tenders Notices, Planning Acts, Services Charters**.

To make a fair comparison between humans and AI, we extracted the **s-ItaIst** sub-corpus, composed by:

Documents	Sentences	Tokens	Types
8	1315	33295	5622

LLMs

To investigate both **open-source** and **commercial** models, the s-ItaIst corpus was simplified using four distinct commercial LLMs, namely **GPT-3.5-Turbo** and **GPT-4** by OpenAI, **LLaMA 3** by Meta, and **Phi 3** by Microsoft. A **detailed prompt** was formulated to instruct each model to perform the simplification task properly, avoiding summary:

*Sei un dipendente pubblico che deve scrivere dei documenti istituzionali italiani per renderli semplici e comprensibili per i cittadini. Ti verrà fornito un documento pubblico e il tuo compito sarà quello di riscriverlo applicando regole di semplificazione senza però modificare il significato del documento originale. Ad esempio potresti **rendere le frasi più brevi**, eliminare le **perifrasi**, **esplicitare** sempre il **soggetto**, utilizzare **parole** più **semplici**, **trasformare** i verbi **passivi** in verbi di forma **attiva**, spostare le **frasi parentetiche** alla **fine** del **periodo**.*

Results and Discussion

A preliminary analysis of our results reveals several significant similarities and differences between the human and LLM datasets. For instance, the **variation in the number of tokens** is **similar** across both **human** and **LLM** corpora, although **LLMs** generally **increase the number of sentences** more prominently than human annotators. Regarding complexity metrics, all the parallel corpora (both human and LLM) exhibit a **general increase in readability**, e.g., **Gulpease Index** [14], compared to the original texts.

The analysis of semantic and structural distance metrics from the original s-ItaIst shows more pronounced differences between human and LLM datasets. In terms of **Semantic Similarity**, the **Human1** and **Human2** corpora **are closer** to the **original meaning** than the LLM-simplified corpora. These differences are even more pronounced when considering Edit Distance. The percentage of **Edit Distance is higher in the LLM group**, with each LLM corpus exceeding the human ones by at least 10%.

GPT-4 achieved the **best results** across the majority of metrics.

	Original	*Human1*	*Human2*	*GPT-3.5-Turbo*	*GPT-4*	*LLaMA 3*	*Phi 3*
Tokens	33,295	34,135	29,755	30,032	31,722	36,035	36,056
Sentences	1,314	1,506	1,744	1,515	1,840	1,944	1,900
Tokens per Sentences	25.33	22.66	17.06	19.53	17.24	18.53	18.97
Sentences per Documents	164.25	188.25	218.00	189.37	230.00	243.00	237.50
Gulpease Index	44.31	49.72	50.64	48.49	**51.34**	50.26	50.16
Flesch Vacca Index	19.97	34.23	33.63	30.33	**36.75**	34.09	33.75
NVdB (%)	73.28	80.44	76.89	78.28	**81.07**	80.18	80.16
Passive (%)	20.87	15.78	17.71	13.99	**12.00**	15.81	15.72
READ-IT BASE (%)	75.91	68.62	51.00	66.61	**55.00**	58.37	57.69
READ-IT LEXICAL (%)	93.64	85.37	89.71	91.96	90.29	77.13	**75.74**
READ-IT SYNTACTIC (%)	63.72	53.14	40.09	38.42	**29.92**	40.97	41.24
READ-IT GLOBAL (%)	86.48	69.24	61.34	68.69	**54.60**	59.26	58.37
Semantic Similarity (%)	-	96.52	**97.26**	96.06	95.80	94.96	94.96
Edit distance (%)	-	35.84	29.20	49.21	52.14	55.48	55.44

GPT-4 simplifications can be comparable to human simplifications. GPT-4 simplifications are negligibly better for complexity metrics, moderately worse for similarity, and largely rephrased compared to human simplifications.

Examples

Original

L'operatore di Polizia Locale, quindi, rappresenta un importante punto di riferimento per la collettività. Nell'ambito delle sue funzioni, esso svolge i propri compiti in maniera autorevole, dando prova di preparazione professionale e sensibilità nel contatto relazionale. La sua attività, inoltre, è caratterizzata dal costante sforzo teso alla migliore interpretazione delle situazioni e delle problematiche incontrate, applicando un approccio educativo e orientato alla più adeguata risposta ai bisogni della cittadinanza.

Gulpease: 38 **Flesch-Vacca**: 12
NVdB: 77 % **Passive**: 28 %
Semantic Similarity: - %
Edit distance: - %

Human 1

La Polizia Locale è un punto di riferimento per i cittadini. La Polizia Locale ha autorevolezza, professionalità e sensibilità nel contatto con i cittadini. La Polizia Locale cerca sempre di interpretare al meglio situazioni e problematiche incontrate. La Polizia Locale si comporta in modo da educare e rispondere adeguatamente ai bisogni dei cittadini.

Gulpease: 55 **Flesch-Vacca**: 33
NVdB: 67 % **Passive**: 0 %
Semantic Similarity: 93 %
Edit distance: 56 %

Human 2

L'operatore di Polizia Locale, quindi, è un importante punto di riferimento per la collettività. Quando è in servizio, esso svolge i propri compiti in maniera autorevole. Dimostra preparazione professionale e sensibilità nel contatto con gli utenti. Cerca sempre di interpretare le situazioni e i problemi nel migliore dei modi. Applica un approccio educativo e cerca di adeguarsi ai bisogni della cittadinanza.

Gulpease: 58 **Flesch-Vacca**: 42
NVdB: 83 % **Passive**: 0 %
Semantic Similarity: 98 %
Edit distance: 35 %

GPT-4

L'operatore di Polizia Locale è un punto di riferimento importante per la comunità. Svolge i suoi compiti con autorità, mostrando competenza professionale e sensibilità nelle relazioni. Inoltre, lavora sempre per capire al meglio le situazioni e i problemi che incontra, usando un approccio educativo per rispondere adeguatamente ai bisogni dei cittadini.

Gulpease: 48 **Flesch-Vacca**: 32
NVdB: 84 % **Passive**: 0 %
Semantic Similarity: 97 %
Edit distance: 48 %

GPT-3.5-Turbo

L'operatore di Polizia Locale è un punto di riferimento importante per la comunità. Nell'esercizio delle sue funzioni, svolge i compiti con autorevolezza, dimostrando professionalità e sensibilità nei rapporti con le persone. Si impegna costantemente a comprendere al meglio le situazioni e le problematiche affrontate, adottando un approccio educativo e orientato a rispondere in modo adeguato ai bisogni dei cittadini.

Gulpease: 45 **Flesch-Vacca**: 27
NVdB: 78 % **Passive**: 0 %
Semantic Similarity: 98 %
Edit distance: 45 %

LLaMA 3

L'operatore di Polizia Locale è un punto di riferimento importante per la comunità. Esegue i suoi compiti con autorità, professionalità e sensibilità nel rapporto con la gente.
La sua attività è caratterizzata dal costante impegno per comprendere meglio le situazioni e i problemi, e per rispondere in modo educativo ai bisogni dei cittadini, con un approccio orientato alla loro assistenza.

Gulpease: 50 **Flesch-Vacca**: 37
NVdB: 85 % **Passive**: 28 %
Semantic Similarity: 96 %
Edit distance: 54 %

Phi 3

L'operatore di Polizia Locale è un punto di riferimento importante per la comunità. Esegue i suoi compiti con autorità, professionalità e sensibilità nel trattare le persone. La sua attività è guidata dal desiderio di capire meglio le situazioni e le problematiche, e di rispondere in modo appropriato ai bisogni dei cittadini, con un approccio educativo.

Gulpease: 52 **Flesch-Vacca**: 38
NVdB: 82 % **Passive**: 28 %
Semantic Similarity: 96 %
Edit distance: 56 %

References

1] Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems (NIPS), volume 30, 2017.
2] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State of the art natural language processing, in: Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP), 2020, pp. 38–45.
3] M. J. Ryan, T. Naous,W. Xu, Revisiting non-English text simplification: A unified multilingual benchmark, Association for Computational Linguistics (ACL) (2023).
4] D. Brunato, F. Dell'Orletta, G. Venturi, S. Montemagni, Design and Annotation of the First Italian Corpus for Text Simplification, in: Linguistic Annotation Workshop (LAW), 2015, pp. 31–41.
5] M. Miliani, S. Auriemma, F. Alva-Manchego, A. Lenci, Neural readability pairwise ranking for sentences in Italian administrative language, in: Asia-Pacific Chapter of the Association for Computational Linguistics (AACL) and International Joint Conference on Natural Language Processing (IJCNLP), 2022, pp. 849–866.
6] M. Miliani, M. S. Senaldi, G. Lebani, A. Lenci, Understanding Italian Administrative Texts: A Reader-Oriented Study for Readability Assessment and Text Simplification, in: Workshop on AI for Public Administration (AIxPA), 2022, pp. 71–87.
7] S. Lubello, La lingua del diritto e dell'amministrazione, Il mulino, Bologna, 2017.
8] M. Cortelazzo, Il linguaggio amministrativo. Principi e pratiche di modernizzazione, Carocci, Roma, 2021.
9] G. Fiorentino, V. Ganfi, Parametri per semplificare l'italiano istituzionale: Revisione della letteratura, Italiano LinguaDue 16 (2024) 220–237.
10] E. Piemontese (Ed.), Il dovere costituzionale di farsi capire. A trent'anni dal Codice di stile, Carocci, Roma, 2023.
11] S. Lubello, Da dembsher al codice di stile e oltre: un bilancio sul linguaggio burocratico, in: E. Piemontese (Ed.), Il dovere costituzionale di farsi capire A trent'anni dal Codice di stile, Carocci, Roma, 2023, pp. 54-70
12] D. Vellutino, et al., L'italiano istituzionale per la comunicazione pubblica, Il mulino, Bologna, 2018.
13] D. Vellutino, N. Cirillo, Corpus «itaist»: Note per lo sviluppo di una risorsa linguistica per lo studio dell'italiano istituzionale per il diritto di accesso civico, Italiano LinguaDue 16 (2024) 238–250.
14] P. Lucisano, M. E. Piemontese, Gulpease: una formula per la predizione della leggibilita di testi in lingua italiana, Scuola e città (1988) 110–124.

SCAN ME