You write like a GPT

Andrea Esuli[1], Fabrizio Falchi[1], Marco Malvaldi[1,2] and Giovanni Puccetti[1,†]

[1]Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"- Consiglio Nazionale delle Ricerche
[2]Professional writer

Abstract

We investigate how Raymond Queneau's *Exercises in Style* are evaluated by automatic methods for detection of artificially-generated text. We work with the Queneau's original French version, and the Italian translation by Umberto Eco.

We start by comparing how various methods for the detection of automatically generated text, also using different large language models, evaluate the different styles in the opera. We then link this automatic evaluation to distinct characteristic related to content and structure of the various styles.

This work is an initial attempt at exploring how methods for the detection of artificially-generated text can find application as tools to evaluate the qualities and characteristics of human writing, to support better writing in terms of originality, informativeness, clarity.

Keywords

GPT, style, generated text, human writing

1. Introduction

The extraordinary writing ability of the latest chatbots and virtual assistants based on Large Language Models (LLMs) poses a significant question for anyone who attempts to write today —- be they a scientist, a writer, or a lover: is it worth the effort to engage in the act of writing?

For those not hindered by excessive laziness and who, with courage, still tackle writing with determination and passion, this question implies a more specific one: am I writing a text that an artificial intelligence could not have produced?

We believe that the answer to this question may, in the future, come from the LLMs themselves given that they are designed to assess the probability of the occurrence of the next word in a text. We envision a future where LLMs, although widely used to produce essentially obvious texts, will assist those who still engage in writing to create texts worth reading, if only because the artificial intelligence, having read and statistically evaluated almost everything ever written, considers them non-obvious and distinct from what it would have produced itself.

The ability of LLMs to evaluate the probability of the next word in a text stems from the extensive corpus of writing they are trained on. Consequently, their evaluation of a piece of writing is ultimately based on an indirect comparison between the given text and the entire body of literature they have been exposed to. Using LLMs to assess how much a text differs from the production capabilities of LLMs inherently implies an evaluation of the novelty it represents compared to known literature.

Starting to move in this direction, this article explores whether an LLM can be used to help humans answer this question. In this first attempt we do this not based on the content intended for communication but on the style. We have conducted a preliminary study on the possibility of using LLMs to evaluate how and to what extent a certain writing style and/or a specific text differs from what a machine can achieve.

We took as a reference Raymond Queneau's "Exercises in Style" [1], which draws from Erasmus of Rotterdam's "De Utraque Verborum ac Rerum Copia" [2] a bestseller widely used for teaching how to rewrite pre-existing texts and how to incorporate them into a new composition. In Queneau's work, the same simple story is revisited each time in a different literary style. We asked ourselves and conducted experiments on how much the texts in various styles used by Queneau differ from the writing abilities of LLMs, which have acquired their skills by learning statistical relationships from vast amounts of text.

Calvino had already attempted to answer this question: "What would be the style of a literary automaton?" He replied, "The test for a poetic-electronic machine will be the production of traditional works, of poems with closed metric forms, of novels with all the rules". We believe it has indeed happened this way, as today's chatbots and virtual assistants are built from a language model.

In this work, we provide initial evidence that language models recognize those texts that are more traditional, particularly used in spoken language or by classical characters as more probable while they deem more unlikely experimental and innovative texts. However, we find

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

† All the authors contributed equally.

andrea.esuli@isti.cnr.it (A. Esuli); fabrizio.falchi@isti.cnr.it (F. Falchi); giovanni.puccetti@isti.cnr.it (G. Puccetti)

0000-0002-5725-4322 (A. Esuli); 0000-0001-6258-5313 (F. Falchi); 0000-0003-1866-5951 (G. Puccetti)

evidence that even for powerful LLMs it remains difficult to cut a clear line between experimental texts and those that instead incur the risk of becoming unreadable.

2. Related Work

The evaluation of text readability may dated back at least to the work of Flesch in 1948 [3]. Flesch's method was based on simple surface properties of text (i.e., words per sentence and syllables per word). Since then a steady evolution of methods involved more complex NLP and ML as new tools were developed (see the surveys [4, 5]).

An example of the use of LLMs on this topic is the work Miaschi et al. [6], which investigated the correlation between a readability score measured by an automatic readability tool (READ-IT [7]) and the perplexity measured by an LLM, yet they found no significant correlation between the two dimensions.

Hayati et al. [8] compared human and BERT-based relevance scoring of words in a sentence to determine its style, polite or offensive, as well as the expression of sentiment and emotions. They found a loose correlation in the way words are identified as relevant by humans and BERT, with BERT giving more relevance to context word (e.g. "baseball" for the emotion of joy), while human are more focused on words perceived as "typical" of the style. (e.g., "smile" for joy).

The style transfer process is the task of rewriting a passage of text changing the set of lexical choices and syntactic structures, yet not substantially changing the actual content of the text. Krishna et al. [9] surveys the style transfer literature and proposed a style transfer method trained on reconstructing a style-specific text (inverse paraphrase) on pseudo-parallel data generated using a diverse paraphrase model.

Qi et al. [10] proved that a change of the writing style, made using a trained model, can be an effective means of attack to BERT-based classifiers, e.g., letting an offensive text be classified as non-offensive just by rewriting it using a Bible-like style. Similarly Krishna et al. [11] have shown that automatic paraphrasing can be extremely effective at breaking the ability of detection method to recognize artificially generated text.

3. Writing with style

Queneau's original work in French of 1947 [1] tackles on telling the same short story using 99 different styles. The first style, Notations, is a clear report of a sequence of events, each with details that together define the actual content of the story that is reported in all of the other 98 versions. Each version has a defining title that denotes its style. Styles can be grouped by similarity; Barbara

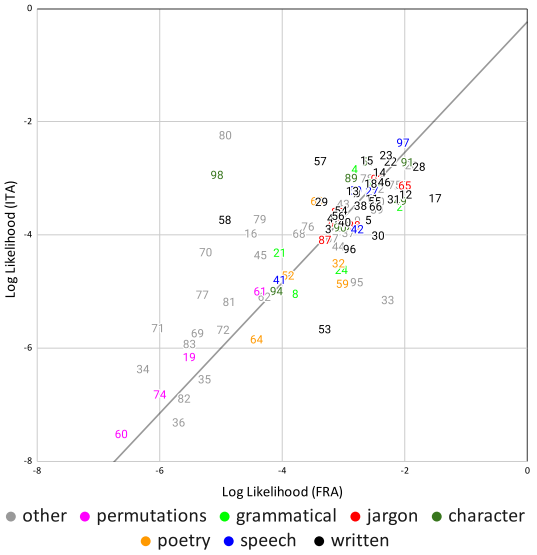

Figure 1: Log Likelihood for both the Italian and French versions of "Exercises in Style". The numbers provided correspond to the IDs in Table 1. The colors indicate the exercise group. The line show the correlation ($R^2 = 0.805$).

Wright, who made the English translation in 1958 [12], reports to have roughly identified seven groups[1]:

- different types of speech;
- different types of written prose, e.g., Official Letter, Philosophic;
- five poetry styles, e.g., Haiku, Ode;
- eight language-based character sketches, e.g., Reactionary, Biased, Abusive;
- grammatical and rhetorical forms, e.g., Litotes, Synchesis, Parts of speech;
- jargon, e.g., mathematical, botanical;
- and the very specific group of Permutations, by groups of letters or words.

Along time, new editions presented variations in the list of styles. For example, five styles in the original edition[2], were replaced by other five in the edition of 1969[3], the one we used in our experiments.

Queneau's opera has been translated in more than 30 languages. The Italian translation was made by Umberto Eco [13], in 1983. Similar to other translations, the Italian translation reports almost all the original styles, but some are considered untranslatable and replaced with variants

[1] In the preface of the book where the groups are listed, Wright did not report a complete assignment of all styles to these groups, only hinting a few cases for some of them.
[2] Réactionnaire, Feminine, Hai-Kai, Permutations de 2 á 5 lettres, Permutations de 9 á 12 lettres.
[3] Ensembliste, Définitionnel, Tanka, Translation, Lipogramme.

			Italian (Eco)				French (Queneau)			
			DetectGPT		l. likelihood		DetectGPT		l. likelihood	
ID	Title	gr.	value	rank	value	rank	value	rank	value	rank
64	Tanka	P	-.120	1	-5.85	9	.100	25	-5.85	19
35	Aferesi	O	-.056	2	-6.56	5	.077	19	-6.56	11
82	Perlee Englaysee	O	-.037	3	-6.90	3	.091	24	-6.90	6
36	Sincopi	O	.026	4	-7.32	2	-.102	2	-7.32	5
71	Epentesi	O	.033	5	-5.65	13	.148	40	-5.65	3
74	Metatesi	Π	.035	6	-6.83	4	.068	17	-6.83	4
60	Perm. ... lettere	Π	.037	7	-7.53	1	.012	9	-7.53	1
61	Perm. ... parole	Π	.057	8	-5.00	19	.048	15	-5.00	21
95	Interiezioni	O	.061	9	-4.84	22	.217	63	-4.84	61
19	Anagrammi	Π	.063	10	-6.17	7	-.090	3	-6.17	7
25	Analisi logica	O	.074	11	-3.13	78	.135	36	-3.13	92
58	Telegrafico	W	.077	12	-3.74	49	.025	11	-3.74	15
62	Ellenismi	O	.079	13	-5.10	16	.066	16	-5.10	23
81	Francesismi	O	.090	14	-5.19	14	.244	71	-5.19	17
83	Contre pèteries	O	.116	15	-5.94	8	-.042	5	-5.94	8
73	Parti del discorso	G	.144	16	-3.10	81	.084	20	-3.10	74
16	Parole composte	O	.154	17	-3.98	36	.084	21	-3.98	18
77	Giavanese	O	.170	18	-5.06	17	.028	12	-5.06	10
63	Versi liberi	P	.173	19	-3.41	65	.107	26	-3.41	31
94	Contadino	C	.175	20	-5.00	20	.120	30	-5.00	24
69	Anglicismi	O	.191	21	-5.74	10	-.029	7	-5.74	9
34	Apocopi	O	.194	22	-6.38	6	-.105	1	-6.38	2
93	Geometrico	J	.214	23	-3.02	84	.202	60	-3.02	78
65	Insiemista	J	.216	24	-3.13	77	.326	95	-3.13	94
53	Olfattivo	W	.219	25	-5.67	12	.042	14	-5.67	34
87	Gastronomico	J	.223	26	-4.09	32	.251	75	-4.09	35
32	Canzone	P	.224	27	-4.51	26	.316	92	-4.51	43
47	Filosofico	W	.230	28	-3.71	51	.128	34	-3.71	38
24	Onomatopee	G	.232	29	-4.62	25	.166	47	-4.62	47
52	Sonetto	P	.236	30	-4.72	24	.016	10	-4.72	27
8	Sinchisi	G	.268	31	-5.05	18	.072	18	-5.05	28
39	Dunque, cioè	O	.273	32	-3.56	56	.201	59	-3.56	77
59	Ode	P	.280	33	-4.87	21	.199	57	-4.87	48
72	Paragoge	O	.283	34	-5.68	11	.132	35	-5.68	14
41	Volgare	S	.286	35	-4.80	23	.159	46	-4.80	25
67	Lipogrammi	O	.291	36	-4.07	33	.148	41	-4.07	37
2	Litoti	G	.304	37	-3.51	57	.270	87	-3.51	90
76	Nomi propri	O	.309	38	-3.86	42	.127	33	-3.86	30
17	Negatività	W	.311	39	-3.36	69	.127	32	-3.36	99
21	Omoteleuti	G	.315	40	-4.32	28	.168	49	-4.32	26
43	Commedia	O	.316	41	-3.46	61	.245	72	-3.46	50
37	Me, guarda...	O	.320	42	-3.97	37	.119	29	-3.97	53
45	Parechesi	O	.322	43	-4.36	27	.034	13	-4.36	22
9	Arcobaleno	O	.324	44	-3.75	47	.144	39	-3.75	65
38	Esclamazioni	W	.325	45	-3.49	59	.240	69	-3.49	66
88	Zoologico	J	.328	46	-3.84	44	.111	28	-3.84	57
96	Prezioso	W	.344	47	-4.26	30	.250	74	-4.26	54
40	Ampolloso	W	.344	48	-3.78	46	.264	86	-3.78	52
50	Disinvolto	S	.348	49	-3.22	75	.182	51	-3.22	59
12	Precisazioni	W	.364	50	-3.29	70	.159	45	-3.29	95

			Italian (Eco)				French (Queneau)			
			DetectGPT		l. likelihood		DetectGPT		l. likelihood	
ID	Title	gr.	value	rank	value	rank	value	rank	value	rank
4	Retrogrado	G	.366	51	-2.85	89	.143	38	-2.85	58
20	Distinguo	O	.371	52	-3.88	40	.259	82	-3.88	49
57	Auditivo	W	.377	53	-2.70	95	.153	43	-2.70	32
79	Latino maccher.	O	.384	54	-3.73	50	.344	97	-3.73	20
56	Visivo	W	.387	55	-3.67	53	.123	31	-3.67	41
5	Sorprese	W	.390	56	-3.74	48	.319	94	-3.74	70
68	Sostituzioni	O	.393	57	-3.98	35	-.009	8	-3.98	29
48	Apostrofe	G	.394	58	-3.68	52	.214	62	-3.68	44
26	Insistenza	O	.403	59	-2.77	91	.084	22	-2.77	97
30	Passato remoto	W	.406	60	-4.02	34	.254	79	-4.02	80
75	Davanti e di dietro	O	.408	61	-3.12	79	.271	88	-3.12	89
29	Presente	W	.411	62	-3.42	62	.206	61	-3.42	33
54	Gustativo	W	.414	63	-3.57	55	.087	23	-3.57	46
80	Vero?	O	.422	64	-2.24	99	-.030	6	-2.24	16
86	Ingiurioso	C	.424	65	-3.85	43	.252	76	-3.85	51
89	Impotente	C	.426	66	-3.00	86	.195	56	-3.00	55
46	Fantomatico	W	.446	67	-3.07	82	.318	93	-3.07	84
70	Protesi	O	.454	68	-4.31	29	.109	27	-4.31	12
85	Medico	J	.462	69	-3.60	54	.138	37	-3.60	40
33	Poliptoti	O	.468	70	-5.16	15	.261	83	-5.16	86
10	Logo-rallye	O	.474	71	-3.42	63	.352	98	-3.42	81
3	Metaforicamente	W	.474	72	-3.90	39	.256	80	-3.90	36
44	A parte	O	.475	73	-4.21	31	.191	55	-4.21	42
78	Controverità	O	.478	74	-3.01	85	.191	54	-3.01	68
1	Partita doppia	O	.479	75	-3.48	60	.243	70	-3.48	62
18	Animismo	W	.479	76	-3.10	80	.222	66	-3.10	71
11	Esitazioni	W	.491	77	-3.37	68	.200	58	-3.37	73
22	Lettera ufficiale	W	.492	78	-2.71	93	.264	85	-2.71	87
55	Tattile	W	.500	79	-3.41	64	.154	44	-3.41	75
90	Modern style	C	.510	80	-3.87	41	.343	96	-3.87	45
92	Ritratto	O	.517	81	-3.19	76	.384	99	-3.19	79
13	Asp. soggettivo l	W	.521	82	-3.23	73	.185	53	-3.23	56
84	Botanico	J	.522	83	-3.79	45	.167	48	-3.79	39
27	Ignoranza	S	.522	84	-3.23	74	.221	65	-3.23	72
49	Maldestro	C	.524	85	-3.41	66	.168	50	-3.41	91
15	Svolgimento	W	.526	86	-2.69	96	.234	67	-2.69	67
7	Pronostici	W	.534	87	-3.27	71	.219	64	-3.27	60
0	Notazioni	W	.548	88	-3.27	72	.263	84	-3.27	63
6	Sogno	W	.548	89	-3.04	83	.257	81	-3.04	83
28	Passato prossimo	W	.555	90	-2.80	90	.288	91	-2.80	98
97	Inatteso	S	.581	91	-2.38	98	.284	90	-2.38	93
23	Com. stampa	W	.584	92	-2.60	97	.239	68	-2.60	85
66	Definizioni	W	.601	93	-3.50	58	.252	77	-3.50	76
42	Interrogatorio	S	.613	94	-3.91	38	.150	42	-3.91	64
14	Altro asp. sogg.	W	.618	95	-2.90	88	.248	73	-2.90	82
31	Imperfetto	W	.630	96	-3.38	67	.253	78	-3.38	88
91	Probabilista	C	.633	97	-2.72	92	.281	89	-2.72	96
51	Pregiudizi	C	.654	98	-2.71	94	.184	52	-2.71	69
98	Reazionario	C	.704	99	-2.95	87	-	-	-	-
98	Loucherbem	C	-	-	-	-	-.054	4	-2.95	13

C character G grammatical J jargon O other P poetry Π permutations S speech W written

Table 1
Scores and ranks of the various styles with respect to various detection methods. Styles are ranked by the DetectGPT score on Italian. Groups are indicated by their initials (Π is used for *permutations*) and are color-coded consistently with the previous figures.

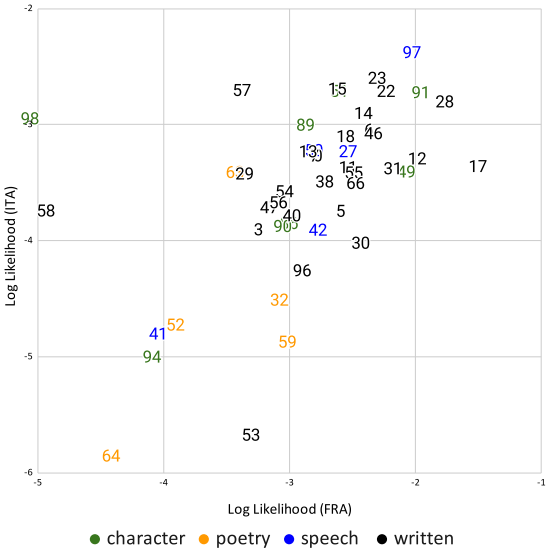

Figure 2: Log Likelihood for the main groups, presented in a zoomed-in view.

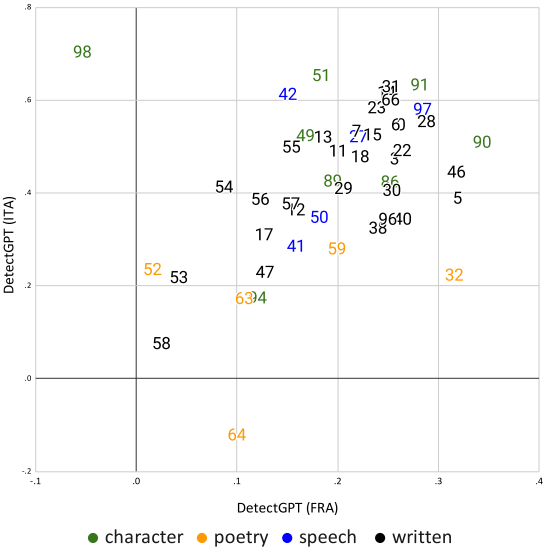

Figure 3: DetectGPT scores for the main groups.

semantically similar to the original ones, or relevant for other reasons. For example the style Homophonique was replaced by Eco with a style named Vero? (True?), because French has many homophones while Italian has not. The Vero? style links to the repeated use of intercalation and links to the Alors style of the French edition. Eco also decided to not translate the Loucherbem style, based on the slang spoke by Parisian and Lyonnaise butchers, considering not interesting to link it to an Italian slang or dialect, whereas dialect-based styles already were included in the opera. Eco replaced it with its own version of the Réactionnaire style from the first edition, which he liked more, as he detailed in the preface of his translation.

4. Style and detection, is there a relation?

The Research Question (RQ) we wish to answer is the following: **Can we use Machine Generated Text (MGT) detection methodologies to measure some qualities and characteristics of the style used in writing a piece of text?**

Our assumption supporting the relevance of this RQ is that LLMs, trained on trillions of tokens, naturally approximate an *average writing style* that is necessarily "average" and thus not original or unique. On the other hand, original and surprising writing styles, which by definition will come in many very different forms, will be less frequent, and sparse across the long tail in the distribution of training data, and thus modeled as less likely according to the LLMs.

We use two metrics to measure the style of texts according to language models, *Log Likelihood* (LL) and *DetectGPT* [14], these metrics are used to detect text generated by a given language model since on average they will be higher for text that a language model has generated, when compared to text written by a human.

We focus on Eco's Italian and Queneau's original French versions of the style exercises. To measure the scores, we use LLMs tuned for these languages. For Italian we use Anita [15] while for French Mistral [16].

As a first validation of our assumption, Figure 1 shows the correlation between the *Log Likelihood* each writing style passage is assigned in Italian (y-axis) and in French (x-axis). The Figure shows significant correlation and zooming in on the higher *Log Likelihood* texts, Figure 2, we see that the correlation persists.

Similar results hold for *DetectGPT*, Figure 3, shows the correlation between this score for the Italian texts and for the French ones, and the correlation is close to the one for *Log Likelihood* shown in Figure 2.

Both Figures 2 and 3 show style number 98 as a kind of outlier. This is a correct measurement as style 98 is actual two different styles between the two versions, Loucherbem in French, and Reazionario in Italian, as reported in Section 3.

Both *Log Likelihood* and *DetectGPT* appear to behave consistently across languages and styles, supporting our hypothesis that some characteristics of the writing styles are captured by these scores.

4.1. Analysis of Detection Scores of Styles

Table 1 shows the actual value of *Log Likelihood* and *DetectGPT* for each passage in both Italian and French as well as their ranking among all style exercises, ranked based on the *DetectGPT* score in Italian. We adopted Wright's grouping of styles, assigning each style to one of the seven groups listed in Section 3, and also adding an "other" group for styles for which we could not find a clear positioning in Wright's groups (typically the styles based on almost obsessive repeated use of some kind of expression). The (colored) *gr.* column reports the style group that is assigned to each style exercise and we can observe that ranking the styles based on the *DetectGPT* scores in Italian (as they are reported in the table) highlights a few prominent patterns which we now describe.

The ***permutation*** class is present only in the lower ranks, and indeed the texts belonging to this group are hard to read and don't show any recognizable stylistic pattern, they are more akin to games that makes sense only within the context of Queneau's book.

The texts belonging to the ***jargon*** are mostly in the lower end of the tail. The styles that are in higher ranks are likely to be present in higher quantity in LLMs training data justifying the ranking shift.

The ***poetic*** class is the next one in average rank, just higher than the *permutation* one, with the exception of the "Tanka" style, which is indeed a very short text, with almost no syntax connecting minimal sentences.

Interestingly, right above the *poetic* group stands the ***grammatical and rhetorical*** group; indeed rhetorical figures are a key component of poem writing. This group is evenly spread among the middle ranks, with the exception of "Parti del discorso" (Part of speech), which is in a lower position, and which also the one with more loose relation with *grammatical and rhetorical* group.

The ***writing*** group, contains a large number of styles and is spread across several ranks, however it is heavily skewed towards the higher ranks.

The ***speech*** group is entirely in the higher ranks and as its spoken source suggests it has a strong character-rooted component.

Accordingly, the only group that ranks higher than *speech* is ***character***[4] which, with only two exceptions, "Ingiurioso" (Offensive) and "Impotente" (Powerless), always ranks in the top quarter, takes all 3 top ranks and is the highest ranking one. The last line of Table 1 reports the ranks and scores for the Loucherbem style, which exists only in the French version. The ranks are very low as this style uses almost made up words to replicate the phonetics of the jargon.

The ***other*** group which contains all those styles which are harder to assign to a specific group is evenly spread across the lower ranks with few exceptions indicating that the texts that compose it are indeed quite varying and hard to group together.

An overall look at the ranking without considering the groups suggests a relation between the scores of detection methods and some characteristics of the styles. Styles that make use of unusual, or just made up words, or do not use a correct syntax, get low detection scores. Styles that are based on a clean, modern prose, with a simple syntax, get high detection scores. The middle ranks show a smooth transition among the two extremes, in which the use of unusual terms or syntax is more frequent as the detection scores get lower.

5. Conclusions

This work is a first exploration of the idea of designing tools that evaluate how and to what extent a writing style and/or a specific text differs from what a machine can achieve. We tested for this task the use machine generated text detection tools, under the hypothesis of a correlation between their detection scores and our goal of discovering the many facets that build an original human written text. We applied them to Queneau's exercises in style, in which the same story is written using a rich and varied set of writing styles. We have found a consistent correlation between the scores assigned by detection methods, across detection methods and across languages.

The comparison of the styles with their detection scores indicates that lower scores from detection methods are correlated with the use of unusual terms or syntax, while higher scores are more related to styles that are based on a clean and more prose, with a smooth transition among this two extremes. The ranks thus do not indicate a "better" or a more "interesting" style, yet they confirm Calvino's statement we reported in the introduction: content that is akin to a machine-generated one is the one that produce "traditional" content, following the main rules of writing.

Writers willing to depart from sounding "ordinary" could indeed use detection methods to estimate these aspects on their content, with the caveat that while a mid-level detection score may suggest some original traits in text, low scores may not indicate a more original or interesting text, but they may likely derive from an obscure or plainly unreadable text.

Given the positive results of this first investigation, future developments will be based on the use of texts specifically written for this activity. This will have the advantage of having full control over the contents and to have the guarantee that they have never been part of the LLMs training data.

[4]Character as in "the character of a play".

Acknowledgments

This work was partially supported by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by European Union - NextGenerationEU.

References

[1] R. Queneau, Exercises de style, Gallimard, 1947.

[2] D. Erasmus, De Utraque Verborum ac Rerum Copia, 1512.

[3] R. Flesch, A new readability yardstick., Journal of applied psychology 32 (1948) 221.

[4] K. Collins-Thompson, Computational assessment of text readability: A survey of current and future research, ITL-International Journal of Applied Linguistics 165 (2014) 97–135.

[5] S. Vajjala, Trends, limitations and open challenges in automatic readability assessment research, arXiv preprint arXiv:2105.00973 (2021).

[6] A. Miaschi, C. Alzetta, D. Brunato, F. Dell'Orletta, G. Venturi, Is neural language model perplexity related to readability?, in: J. Monti, F. Dell'Orletta, F. Tamburini (Eds.), Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021, volume 2769 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020.

[7] 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, SLPAT 2011, Edinburgh, Scotland, UK, July 30, 2011, Association for Computational Linguistics, 2011, pp. 73–83.

[8] S. A. Hayati, D. Kang, L. Ungar, Does bert learn as humans perceive? understanding linguistic styles through lexica, arXiv preprint arXiv:2109.02738 (2021).

[9] K. Krishna, J. Wieting, M. Iyyer, Reformulating unsupervised style transfer as paraphrase generation, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 737–762. URL: https://aclanthology.org/2020.emnlp-main.55. doi:10.18653/v1/2020.emnlp-main.55.

[10] F. Qi, Y. Chen, X. Zhang, M. Li, Z. Liu, M. Sun, Mind the style of text! adversarial and backdoor attacks based on text style transfer, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 4569–4580. URL: https://aclanthology.org/2021.emnlp-main.374. doi:10.18653/v1/2021.emnlp-main.374.

[11] K. Krishna, Y. Song, M. Karpinska, J. Wieting, M. Iyyer, Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 27469–27500.

[12] R. Queneau, B. Wright, Exercises in style, Gaberbocchus Press, 1958.

[13] R. Queneau, U. Eco, Esercizi di stile, Gli Struzzi, Einaudi, 1983.

[14] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, C. Finn, Detectgpt: zero-shot machine-generated text detection using probability curvature, in: Proceedings of the 40th International Conference on Machine Learning, ICML'23, JMLR.org, 2023.

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

[16] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. arXiv:2310.06825.