GPT-3 vs. Delta. Applying stylometry to large language models

*** ***[1]

[1] ***, *** _ ***

ABSTRACT

This paper tests the ability of large language models to deceive stylometric approaches in authorship attribution. A corpus of ten English authors is used as a reference point, while GPT-3 is asked to generate texts that imitate their style. After having defined a baseline for the efficiency of stylometric methods on human-generated texts, a series of analysis is performed on the artificially generated texts. Results show the inability of GPT-3 to deceive stylometry and allow a quantitative analysis of its distinctive linguistic features. Preliminary results are also presented for ChatGPT, indicating the efficiency of stylometry in detecting its authorial fingerprint.

KEYWORDS

Stylometry; Authorship attribution; Large language models; GPT-3; ChatGPT

1. INTRODUCTION

The growing interest towards large language models (LLMs) such as GPT-3, ChatGPT, and Bing, has been accompanied by the preoccupation if the text generated by these systems can be identified automatically, so to control its possible misuse. Works like the one by [5] have already aimed at identifying the features that allow such a recognition, while online services such as GPT-Zero offer the possibility to easily test its efficiency. Even if preliminary results are promising, [11] warns about the necessity of moving beyond traditional stylometric methods to identify artificially generated text. Apart from limited applications like [7], still, no extensive study has been dedicated to the automated recognition of LLMs in literary studies.

This paper tries to fill this gap by testing to what extent one language model (GPT-3) is able to deceive stylometric approaches in authorship attribution. In fact, while stylometry has clearly demonstrated its potential in recognizing the authorial fingerprint of literary authors [1], examples have already been shown where authors were able to deceive the algorithm, through the practices of pastiche and imitation (see for example Joyce in [8]).

2. DEFINING THE BASELINE

In order to perform such an analysis, a baseline needs to be defined, indicating the efficiency of stylometric methods in the attribution of text written by humans.

The ELTeC corpus of English literature (created in the context of the *Distant Reading for European Literary History* COST Action) was selected for this aim, because of its balanced structure and detailed documentation [10]. 30 texts were extracted from it, after having identified the authors that contributed more than two texts to the collection. Out of these, 20 texts (two per author) constituted the training set, while 10 (one per author) constituted the test set.

A series of preliminary analyses[1] was then performed on the training set to identify the best performing stylometric features. Based on the research by [4], the Cosine Delta distance measure (also known as Würzburg Delta) was kept constant, while most frequent words (MFWs) were varied. As Figure 1 shows, the dendrogram reaches stability at 150 MFWs, clearly distinguishing the different authors in separated clusters. While stability is also confirmed by higher MFW selections, the value of 150 was selected because it limits the analysis mainly to function words, already identified as the possible groundwork for the success of stylometry in authorship attribution [6].

After having identified the stylometric features, the baseline was defined by comparing the training set with the test set. The 10 texts in the test set were split into chunks of 5,000 words (identified as the minimum length for a reliable stylometric analysis [2]), thus producing a total of 345 text samples. These samples were compared one by one with the training set, verifying if the stylometric analysis were able to cluster them close to the right author. Overall, 83.5% of the analyses were successful, with a success rate always higher than 68.7% for single authors. These values were therefore taken as a baseline for the following experimentation with GPT-3.

[1] All analyses were performed using the *stylo* R package [3]. Scripts and datasets are available on GitHub. **[Repository redacted for blind review. Materials can be downloaded at the following link: https://bit.ly/GPT3_Stylometry]**

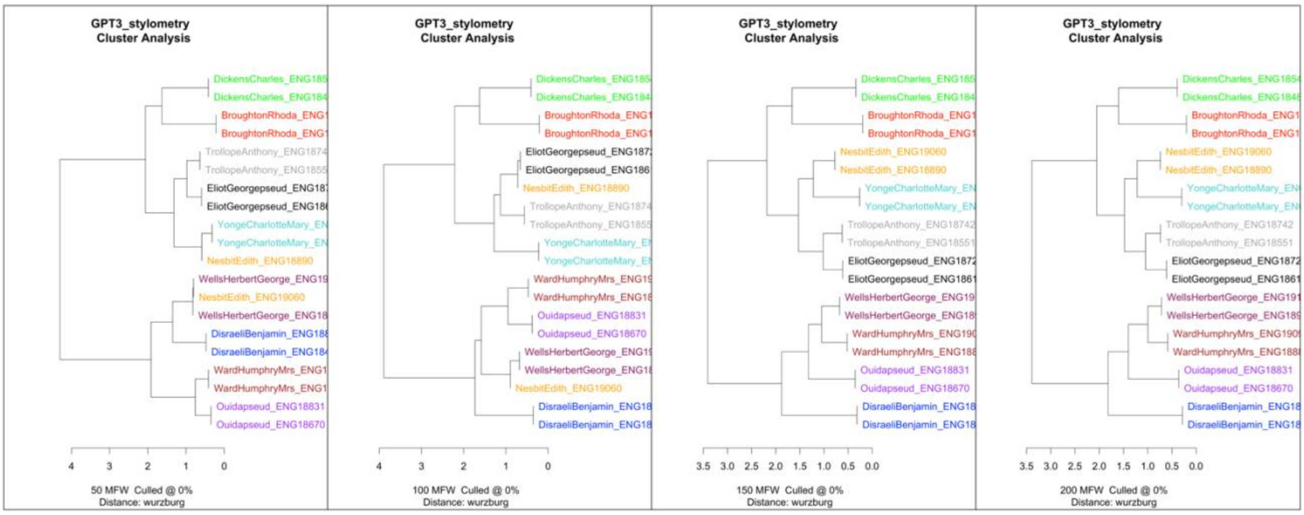

Figure 1. Cluster analyses on the training set

3. CREATING AN ARTIFICIAL TEST SET WITH GPT-3

To instruct GPT-3 into imitating the style of the 10 authors in the training set, a series of strategies were implemented via the OpenAI API. An overview is provided by Table 1, which shows the four features that were varied and combined:

- the name of the author to be imitated;
- a prompt giving instructions on how to generate the new text;
- the model to be used (being DaVinci the most advanced and Ada the most basic);
- the temperature of the model (intended as how much the model diverges from a deterministic behavior, thus becoming more and more "creative").

All combinations between these four features were tested, thus generating a total of 320 different configurations.

author_name	prompt	model	temperature
Rhoda Broughton	Write a chapter of a novel in the style of <author_name>	text-davinci-003	0.1
Charles Dickens			
Benjamin Disraeli	Write a novel with a complex structure. It should incude many events, dialogues, and descriptions. Write like <author_name>	text-curie-001	
George Eliot			
Edith Nesbit			
Ouida	Write a novel by imitating <author_name>	text-babbage-001	0.9
Anthony Trollope			
Mrs. Humphry Ward	Write a story as if it were written by <author_name>. Give it the structure typical of <author_name>'s narratives	text-ada-001	
Herbert George Wells			
Charlotte Mary Yonge			

Table 1. GPT-3 text generation strategy overview

One issue to be dealt with was that of text length. Independently from the prompt, in fact, the models always tended to generate short texts, with an average length of 330 tokens. To reach the minimum of 5,000 words requested for a reliable stylometric analysis, text generation was repeated multiple times for each configuration, until the total amount of generated text surpassed the limit of 6,000 tokens (roughly corresponding to 5,000 words). All the generated texts were then concatenated into single text samples.

A preliminary qualitative analysis of the corpus confirmed the impression that GPT-3 correctly interpreted the task. In the case of Charles Dickens, for example, texts were inhabited by the most famous characters of his novels, even with evident cases of imitation (like almost verbatim repetitions of the incipit of *A Tale of Two Cities*).

4. STYLOMETRIC ANALYSIS

The analysis of the artificially generated corpus was performed by repeating the procedure described in paragraph 2: each text sample was compared to the training set by using Cosine Delta distance and 150 MFW, verifying if it clustered together with the author it tried to imitate.

Overall, only 16.8% of the samples were attributed to the imitated author, thus indicating the substantial inability of GPT-3 to deceive the stylometric analysis. Some features showed higher efficiency than others (22.5% for the "Write a chapter of a novel in the style of <author_name>" prompt; 25% for the DaVinci model; 19.4% for the 0.9 temperature), but still much below the baseline.

Figure 2 shows the results for different imitated authors, indicating how the partial success for some of them might depend more on the intrinsic characteristics of GPT-3's "writing style" (more similar to the one of these authors—in particular Herbert George Wells), than on its efforts to imitate their style.

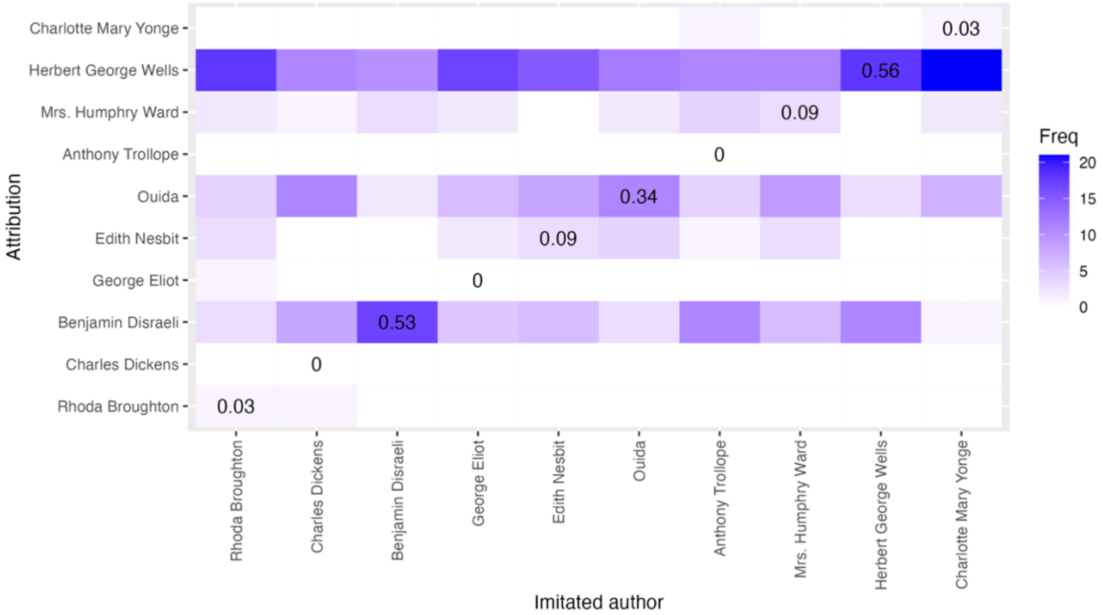

Figure 2. Attributions per imitated author (scores on the diagonal indicate the efficiency per author)

This result advises for a more detailed analysis of GPT-3's style. As the dendrogram in Figure 3a shows, in fact, all the texts created by GPT-3 tend to group in a separated cluster, independently from the imitated author. The effect of imitation emerges only marginally when reducing the analysis to the artificially generated texts: in this case, the strongest stylometric signal seems to be generated by the different models, with DaVinci distinguishing itself more clearly (see Figure 3b).

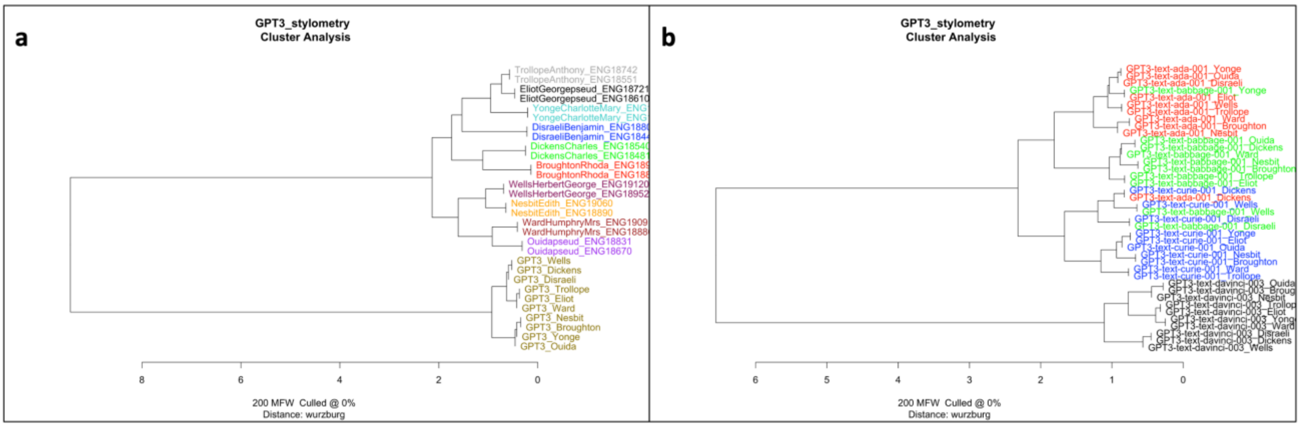

Figure 3. Cluster analysis of GPT-3 texts

To dig deeper into the distinctive characteristics of GPT-3's language, a procedure was developed by working on the Zeta scores, which constitute the groundwork for approaches like Cosine Delta. As shown by [9], in fact, the Delta method automatically looks for patterns in the Zeta scores (indicating how much the frequency of a word in a text deviates from its average frequency in a corpus). One possibility for (partially) reverse engineering the Delta procedure is that of looking for the words in a group of clustered texts that show the lowest variance in the Zeta scores (thus indicating an internally coherent behavior, which is at the same time distinct from the rest of the corpus).

Figure 4 shows the results of such a procedure when applied (a) to all the GPT-3 texts versus the training set and (b) to the GPT-3 DaVinci texts versus the training set. A few aspects need to be noticed here. First, GPT-3 voices distinguish themselves only by the underuse (and not the overuse) of words. It therefore seems that GPT-3 applies a form of self-restraint in text generation. In particular, DaVinci distinguishes itself through the underuse of negation (“don”, “not”) and of modal verbs such as “might” (generally underused by all GPT-3 models) and “should”. While further research is indeed required, this phenomenon seems to be caused by a tendency of GPT-3 towards a more assertive tone, which aims at synthesis and avoids complex constructions.

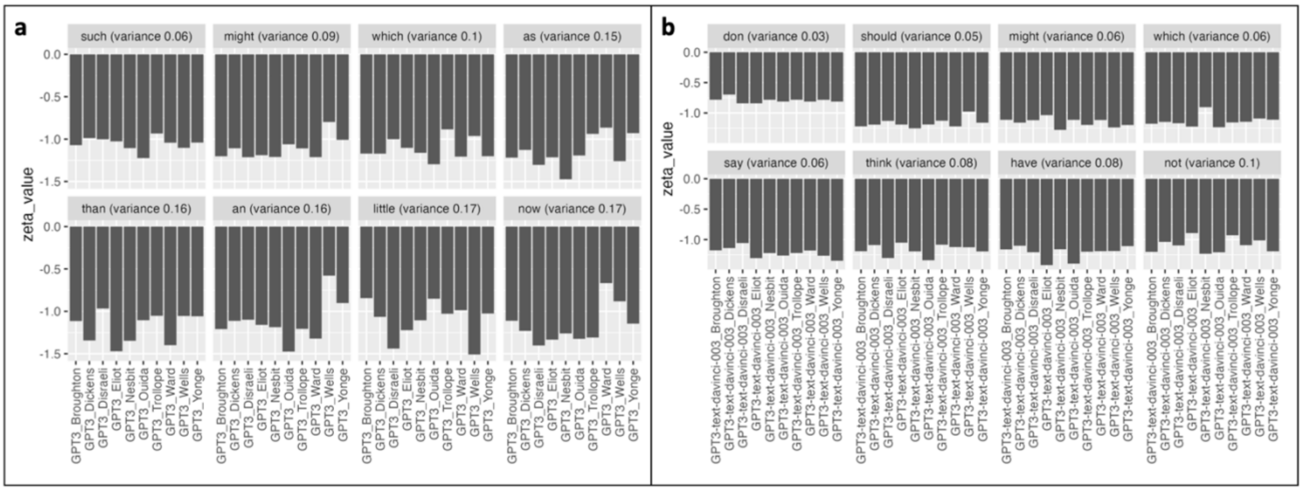

Figure 4. Most distinctive GPT-3 words

5. PRELIMINARY RESULTS WITH CHATGPT

GPT-3 was chosen for this study because, at the time of writing, it provided an API which allowed extensive testing. While a similar analysis is not (yet) possible with the more powerful ChatGPT, a preliminary experiment was attempted to verify if the above results could be confirmed.

Two strategies were applied with only one author. In the first attempt, the initial prompt: “Write the incipit of a novel in the style of Charles Dickens”; was followed by the repeated prompt: “Please go on in the style of Charles Dickens”. This approach was only marginally unsuccessful, because ChatGPT concluded the story already at the third interaction, continuously reshaping the ending in the following ones. For this reason, a second, more complex approach was adopted. A first prompt: “Set up a plan for writing the chapter of a novel in the style of Charles Dickens. It should contain actions, descriptions, and dialogues, in the proportion typical of Dicken’s narratives. It should be divided into 10 [or 15] parts”; was followed by the request to write the parts one by one. This approach was more successful, as it produced three coherent stories of a length between 1,300 and 3,300 words. In any case, more than 5,000 words were collected for both attempts.

Figure 5 confirms the inability of ChatGPT to deceive the stylometric approaches and it also indicates the possible distinctiveness of its voice, placing itself close to Ouida.

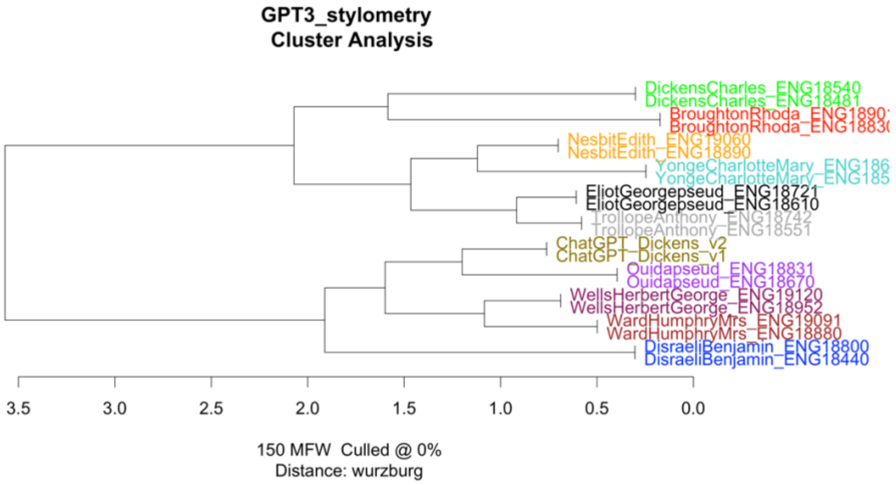

Figure 5. Cluster analysis of ChatGPT texts

6. CONCLUSIONS

One of the main issues in carrying out such a research is that LLM technology evolves at an increasing speed. What is not yet available at the time of writing (e.g., an API for ChatGPT, or the much advertised GPT-4) might be published in the very near future. For this reason, the results presented here should be considered as just a screenshot of a continuously evolving landscape. Still, the coherence between the results obtained with GPT-3 and with ChatGPT suggests how LLMs are still far from effectively imitating the style of literary authors.

A final note should be then added on the many alternative experimental designs that could have been applied for this study. First, different features (such as character n-grams or parts of speech) and approaches (such as machine learning) could have been used for the stylometric analysis: this study focused just on MFWs and distance measures as they allowed a higher explainability of the results, but further research with different approaches is indeed needed to test the full potential of stylometry. Another possible different design could have involved contemporary literary authors. Indeed, one might argue that LLMs have more difficulty in imitating the style of authors from the past, being mainly trained on contemporary text. While this is a sharable concern, copyright restrictions make such an experimentation much less feasible. The free availability of corpora like ELTeC is what made this study possible in the first place. Finally, another option would have been that of fine-tuning the models on the texts by the imitated authors: while this should probably improve the efficiency of the models, it might tell us less about their intrinsic characteristics. For this reason, this study avoided any fine-tuning, just testing the models in their "raw" potential.

Notwithstanding all these issues and limitations, the results presented here could still be used as the groundwork for further research and experimentation, profiting of the confrontation between stylometry and LLMs to better understand them both.

REFERENCES

[1] Burrows, John. "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing* 17, no. 3 (2002): 267–87. https://doi.org/10.1093/llc/17.3.267.

[2] Eder, Maciej. "Does Size Matter? Authorship Attribution, Small Samples, Big Problem." *Digital Scholarship in the Humanities* 30, no. 2 (2013): 167–82. https://doi.org/10.1093/llc/fqt066.

[3] Eder, Maciej, Jan Rybicki, and Mike Kestemont. "Stylometry with R: A Package for Computational Text Analysis." *The R Journal* 8, no. 1 (2016): 107–21.

[4] Evert, Stefan, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. "Understanding and Explaining Delta Measures for Authorship Attribution." *Digital Scholarship in the Humanities* 32, no. suppl_2 (2017): ii4–16. https://doi.org/10.1093/llc/fqx023.

[5] Fröhling, Leon, and Arkaitz Zubiaga. "Feature-Based Detection of Automated Language Models: Tackling GPT-2, GPT-3 and Grover." *PeerJ Computer Science* 7 (2021): e443. https://doi.org/10.7717/peerj-cs.443.

[6] Kestemont, Mike. "Function Words in Authorship Attribution. From Black Magic to Theory?" In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, 59–66. Gothenburg, Sweden: Association for Computational Linguistics, 2014. http://aclweb.org/anthology/W/W14/W14-0908.pdf.

[7] Lawson, Rebecca. "GPT-2: Girl Detective Analyzing AI-Generated Nancy Drew with Stylometry." *IPHS 200: Programming Humanity*, October 1, 2020. https://digital.kenyon.edu/dh_iphs_prog/32.

[8] Rebora, Simone. "Encyclopedic Novel Revisited. Joyce's Role in a Disputed Literary Genre." *Joyce Studies in Italy* 19 (2017): 147–68.

[9] ———. "Stylometry and Reader Response. An Experiment with Harry Potter Fanfiction." In *AIUCD 2022 - Proceedings*, edited by Fabio Ciracì, Giulia Miglietta, and Carola Gatto, 30–34. Bologna: AIUCD, 2022. http://amsacta.unibo.it/6848/.

[10] Schöch, Christof, Roxana Patras, Tomaž Erjavec, and Diana Santos. "Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives." *Modern Languages Open* 1 (2021): 25. https://doi.org/10.3828/mlo.v0i0.364.

[11] Schuster, Tal, Roei Schuster, Darsh J. Shah, and Regina Barzilay. "The Limitations of Stylometry for Detecting Machine-Generated Fake News." *Computational Linguistics* 46, no. 2 (2020): 499–510. https://doi.org/10.1162/coli_a_00380.