Could Style Help Plagiarism Detection? - A Sample-based Quantitative Study of Correlation between Style Specifics and Plagiarism

Adile Uka
Ruhr University Bochum
adile.uka@rub.de

Maria Berger
Ruhr University Bochum
maria.berger-a2l@rub.de

Abstract

The paper presents an empirical study based on samples of the correlation between an author's writing style and the plagiarism ratio in a text. Specifically, we investigate the research question whether a correlation in style can also hint to potential plagiarism, or at least some systematic copying in two texts. To gain an understanding of the characteristics a "copied" text might pertains, we collect different sample sets reaching from chapters by the Brontë sisters, over parallel samples of a plagiarism corpus up to the ChatGPT rephrased pendants of some essay pages by the main author. We also add sections by Matthew, Mark and Luke from a Brazilian Portuguese Bible translation to our samples. Results show that there exists a moderate positive correlation between style similarities and plagiarism overlap across four different genres.

1 Introduction

Plagiarism detection has always been an immensely important task in natural language processing. First, because it is an essential task in all-day publishing, second, because research in plagiarism detection also plays an elementary role in developing information retrieval-related algorithms and operations (c.f., Potthast et al., 2010; Foltỳnek et al., 2019; Alzahrani and Salim, 2008). However, highly-paraphrased plagiarism with many word substitutions and re-ordering still presents a challenge for many systems (c.f., Hunt et al., 2019; Alvi et al., 2021).

The humanities have always been an important driver for the development of technical and digital approaches to analyse and share information across time and space. This is why we also understand the historical use-cases of authorship attribution and user or author pseudonymization (also aliasing) as a means to investigate contemporary challenges of plagiarism and third-party authorship.

In this paper, we attempt to operationalize a procedure that helps us to understand whether style similarities among two texts can give us a hint on plagiarism. Strong overlaps in plagiarism between two texts usually mean that many areas in both texts (suspicious document and sample candidate) share very literal wording. This wording typically also shares style characteristics (Eissen and Stein, 2006). However, in style analysis, the focus is not on content words, but on the distribution of function words, which typically are the most frequent n words in a language. We also want to compare the text samples towards their ChatGPT-rephrased (OpenAI, 2024) versions to make a statement on how GPT's style might differ compared to manually written texts.

2 Related Work

Even though there exists lots of research on plagiarism detection and style analysis, the complementing of both tasks was not performed very often. However, searching for style deviations to find hints for third-party authorship is a common set-up. AlSallal et al. (2019) perform a study on intrinsic plagiarism detection by validating the inner parameters of an author's text. They use most-frequent-words features to represent an author's profile, and use a classifier to evaluate whether a text was actually written by that author. Another form of intrinsic plagiarism analysis is presented in Oberreuter and Velásquez (2013). The authors argue that it is not always possible to measure similarity at the document-level, because the referring documents are not necessarily available. Therefore, the authors use profiles based on linguistic features at the character and word level to compare segments of a text towards the whole text. Whether these paragraphs then are on par with the profiles of the entire document can be determined by measuring the significance. With the rise of LLMs-based

text generators, we have encountered another quality of "plagiarism". However, we hypothesize that we can still distinguish the writing style of artificially generated text from human-written text. We already found proof for this in Zwilling and Berger (2024). Gao et al. (2023) perform a study that shows that ChatGPT generated paper abstracts are unlikely to be plagiarized from the web (while their originals are moderately likely to contain plagiarism) and, further, these generated abstract have only moderately similar texts existing in the web while their originals have very similar versions existing online. This is due to the fact that plagiarism indicates at least area-wise very close stile similarities.

3 Method

3.1 Data Selected

We compile a diverse data set that comprises plagiarism (in the broadest sense) data samples from several sources.

1. Excerpts from British novels by the Brontë sisters as well as some texts under discussion of being authored by William Shakespeare versus Christopher Marlowe (henceforth, Novels samples).

2. Novels samples and their ChatGPT re-products (Novels & GPT).

3. Five pairs sampled form the Webis corpus for plagiarism detection[1] (Burrows et al., 2013) (Webis plagiarism samples).

4. Webis plagiarism samples and their ChatGPT re-products (Webis plagiarism & GPT).

5. Five pages from four different German term papers by one of our co-authors versus their ChatGPT re-product (Essays & GPT).

6. The books Mark & Luke and Mark & Matthew from the Almeida Revisada in Brazilian Portuguese language (Almeida Revisada).

Novels: We use text pairs that we assume suitable for our study because earlier work showed stylistic similarities of them. First, one chapter (Ch. 8) of Charlotte Brontë's "Jane Eyre" together with her sister's Anne Brontë's "The Tenant of Wildfell Hall" (Ch. 1). These works show an especially high stylistic overlap since the sisters' creative writing was exposed to a close exchange among each others from the beginning (c.f., Eder et al., 2016).[2] We downloaded these works from GitHub[3]. We further use texts by Shakespeare's "Henry VI" Part 1[4] (act 1, scene 2, featuring Joan Pucelle) versus "Henry VI" Part 2[5] (act 4, scene 2, featuring Cade at Blackheath). There is strong evidence that both scenes could be considered to be written by Christopher Marlowe, not Shakespeare himself or him alone (c.f., Craig and Kinney, 2009; Nance, 2017).[6]

Novels & GPT: We take each text solely from the Novels dataset and rephrase a ChatGPT version of it and compare it towards this version. Since ChatGPT usually shortens and summarizes such long texts, we opted for rephrasing the novels in chunks of about 200 words to ensure a proper rephrasing of the complete texts. We use the same ChatGPT prompt for the novels and the Webis data, which is "Rephrase this text as much as possible."

Webis Plagiarism samples: We use Random Lists[7], an online random number generator that takes a range and a list size as input. Hence, it returns 5 numeric values within a given range of numbers (the size of the Webis corpus). This way, we could safely select samples from the Webis-CPC-11 corpus for plagiarism detection to have a representative subset to investigate correlation between style and plagiarism. See Tab. 1 for an example of paraphrastic plagiarism.

Webis Plagiarism & GPT: We take each single text from the Webis Plagiarism dataset and have ChatGPT rephrase a version of it. Then, we compare the rephrased version towards its original text.

Essay & GPT: We take five pages from four different German essays by on of our co-authors and compare it towards their ChatGPT re-products. This is especially interesting, as these texts are very specific and we can assume that there is not too much redundancy of these texts around in the web—a fact that GPT would benefit from. The ChatGPT prompt for the rephrasing of the essays

[1] Acc. Feb. 2024 https://webis.de/data/webis-cpc-11.html

[2] Acc: Feb. 2024 https://alanabeeblog.wordpress.com/2013/12/19/the-bronte-sisters-a-stylometric-analysis/

[3] Acc. Feb 2024 https://github.com/computationalstylistics/A_Small_Collection_of_British_Fiction/tree/master/corpus

[4] Acc. Feb 2024 http://shakespeare.mit.edu/1henryvi/index.html

[5] Acc: Feb. 2024http://shakespeare.mit.edu/2henryvi/index.html

[6] The stylistic techniques that were applied to derive that hypothesis were later re-checked within a broader statistical test by OTA (2023).

[7] Acc. Feb. 2024 https://www.randomlists.com/random-numbers

original	paraphrastic plagiarism
The explanation of this estrangement given by my grandfather, was that there had been a disagreement about land; but perhaps he may have felt some delicacy about telling his children that his unambitious marriage had contributed to render the separation permanent.	Explanation of the disposition given by my grandfather, was that there was a disagreement on the land, but perhaps he may have felt some delicacy about his children modestly says his marriage helped make the separation permanent.

Table 1: Plagiarism example from Webis sample 4325

is: "Formuliere diesen Text soweit wie möglich um" ("Rephrase this text as much as possible").

Almeida Revisada: We also are interested in comparing Matthew and Mark, and Luke and Mark in a Brazilian Portuguese version of the Bible. The "Synoptic Gospels" are a strong use-case for plagiarism in the digital humanities. For us, these books show-case another form of similarity and thus are worth investigating. We compare Matthew and Mark, and Luke and Mark, because each of them show a strong overlap with Mark, which is historically acknowledged and computationally con fi rmed (c.f. Jänicke et al., 2014; Harder, 2022). We downloaded a Brazilian Portuguese version o f the Bible f rom the Mysword Bible repository f ootnoteAcc. F eb. 2024 https://www.mysword.info/download-mysword/bibles. Precisely the "Almeida Revisada de acordo com os Melhores Textos em Hebraico e Grego" f rom 1967, which is a rather modern version o f Almeida's translation. We use this version, because we think it is one o f the most common Brazilian Portuguese Bible translations. A version closer to Almeida's original version would also be interesting to study. However, it is beyond the scope o f this study to analyse historical spelling modi f ication and its e f f ect on plagiarism detection and style analysis tools. Matthew is a much longer work than Mark, hence, we cut the chapters 1, 2, 5, 6, 7, 11, and 25 in Matthew as these have no or almost no similarity edges to Mark (Harder, 2022). In Luke, we remove chapters 1, 2, and 12-17, because these also do not have many textual overlaps with Mark.

3.2 Tools Used

We use the Stylo R package[8] (Eder et al., 2016) for calculating stylistic relations between each pair of our sample set. We choose the cosine distance measure handing over our own code of cosine distance. Then, we subtract the resulting value from 1 (that represents the cosine similarity), and receive results that range between 0 and 1. Stylo's default cosine distance version scales these ranges back to the numeric origin scale, which allows also negative similarities scores.[9] A positive effect of cosine measure is that it is robust towards documents of different length (Evert et al., 2016). As we want to investigate the style of the texts—not necessarily the domain content—, we set features to most frequent words (MFW), which long proofed to be a good means to identify similar style (Damerau, 1975; Hoover, 2003).[10] For the books in the Brazilian Portuguese version of the Bible, we use the 200 MFW, because these documents are much longer than the others. For all the other documents, we use the 100 MFW. This setting is the most intuitive while simple and effective.

We further use WCopy fi nd (**?**) to calculate plagiarism overlaps among our sample pairs. Even though there are a lot of plagiarism tools available (also for free use), many of them are not very flexible and do not enable an extrinsic comparison to a local repository. We found WCopy fi nd a useful tool as it determines the similarity based on the partition of common sub-strings (of a given length), in a bi-directional manner (**L**eft, and **R**ight, c.f. Tab. 2), and it also highlights closed-reading overlaps so that we can easily fi nd very long string-overlaps when apparent. The parameters we use for "Shortest Phrase to Match" is 3. We require the system to match 100% of these words, which is a very strict setting for verbatim plagiarism.

4 Results & Discussion

Following, we describe how we investigate the correlation between both, overlapping style characteristics in two texts, and overlapping plagiarism.

4.1 Quantitative correlation measured

To calculate correlation between the style similarities and the plagiarism overlaps, we use the average

[8] Acc. Feb. 2024 https://github.com/computationalstylistics/stylo

[9] The cosine similarity can range between -1 and 1. Because the distance is 1-cosine-similarity, it can range from 0 to 2, see https://medium.com/@milana.shxanukova15/cosine-distance-and-cosine-similarity-a5da0e4d9ded, acc: Jan 2024

[10] Please note that the most frequent top n words of a text do not necessarily contain domain content words

Sample set	cos sim.	pla. %	example of longer overlaps
Novels samples			
Jane_8 & Tenant_1	.95	01L, 01R	"would not be"
HenryVI1_Pucelle & HenryVI2_Cade	.78	00L, 00R	-
Novels & GPT			
Jane_8 & Jane_8_gpt	.98	36L, 39R	"the swelling spring of pure, full, fervid eloquence? Such was the characteristic of Helen's discourse on that"
Tenant_1 & Tenant_1_gpt	.98	31L, 36R	"she is known to have entered the neighbourhood early last week, she did not make her appearance at church on Sunday; and she - Eliza, that is - will"
Webis Plagiarism samples			
3895-orig & 3895-para	.93	30L, 29R	"for me to impugn their honesty if"
4099-orig & 4099-para	.96	45L, 49R	"causes of discontent than they would naturally have independent of this circumstance"
4325-orig & 4325-para	.97	62L, 66R	"is difficult for me to realize the simple fact that she was niece to an uncle"
4475-orig & 4475-para	.82	02L, 04R	" frequently mistaken for"
7804-orig & 7804-para	.97	15L, 13R	"like the bottom of a well with"
Webis Plagiarism & GPT			
3895-original & 3895_gpt	.95	21L, 25R	"detected by one o f the three judges in the ring"
4099-original & 4099_gpt	.93	26L, 32R	"since the earliest settlement o f the country, would"
4325-original & 4325_gpt	.93	34L, 41R	"impression is that the child was asked to describe the vision more minutely"
4475-original & 4475_gpt	.97	38L, 46R	"This species is one of the most graceful birds"
7804-original & 7804_gpt	.94	25L, 34R	"from Shih-tien delivered his official dispatch at"
Essays & GPT (DE)			
text1 & text1_gpt	.97	62L, 67R	"umfasst jeden noch so kleinen intertextuellen Bezug im Text, was bedeuten würde, dass Intertextualität eine zentrale Eigenschaft von Texten ist."
text2.1 & text2.1_gpt	.93	46L, 44R	"indem sie zugibt, dass sie bei polnischen Gospelsongs eher an eine kulturelle Aneignung von schwarzer Kultur in den USA"
text2.2 & text2.2_gpt (5)	.97	41L, 47R	"die Ansichten zu diesem Thema, die Hand in Hand mit den Warnehmungen gehen, erarbeitet."
text3 & text3_gpt	.94	18L, 22R	"ein Ziellexikon mit manuell normalisierten Wort f ormen"
text4 & text4_gpt	.90	35L, 35R	"Ursprung, aber alles Neue in Natur und Kultur kann als Ergebnis der Schöpf ung durch Übernatürliches"
Almeida Revisada (PR)			
Matthew & Mark	.99	21L, 27R	"em verdade vos digo que de modo algum perderá a sua recompensa."
Luke & Mark	.99	16L, 18R	"Ora, para que saibais que o F ilho do homem tem sobre a terra autoridade para perdoar pecados (disse ao paralítico)"

Table 2: Overview o f style similarities and plagiarism "overlap" in our samples: style similarity represents the cosine similarities between the distribution o f the most f requent 100 words o f two texts; plagiarism (in %) represents the token-wise overlap with respect to all f ive-word-windows that overlap between two texts;

of the plagiarism percentage detected (towards the left and towards the right-hand sided texts), but keep them in the table separately to not lose any in formation (see Tab. 2).

We find a moderate positive correlation coefficient (c.f. Pearson, 1895) of 0.52 between style similarities and plagiarism overlap in our samples (excluding couples with ChatGPT-generated texts). For the texts coupled with ChatGPT texts only, this value is 0.32 (weakly positive). This is especially attributed to the fact that ChatGPT naturally does not use a specific style. Instead it might makes

heavily use of the texts input's style. The correlation of all texts amounts to a value of 0.5 indicating still a moderate positive correlation.

4.2 Qualitative correlation measured

Novels: The novels samples show very little to no plagiarism across both sets of texts, while both reveal similarities in the use of style with a maximum cosine similarity of up to .95. The novels especially show the case where the domain vocabulary obviously is very different in both texts. We find very strong style overlap, but only very little plagiarism.

Our novels samples, compared with their ChatGPT re-phrased version show very similar styles with a cosine similarity of 98%. This very identical style could hint to the fact that ChatGPT is not very creative in formulating its own wording and style.

Webis Plagiarism data: Plagiarism is defined by verbatim copying which also goes strongly together with a high stylistic similarity. The Webis data set is the most interesting one for us, because it ensures the domain overlaps and helps us to make predictions on the texts' style. Leaving aside text 4475 with a plagiarism percentage close to zero, the results show a plagiarism percentage ranging between 13% and 66%. In text 4325, we can find the highest result of detected plagiarism of 66% while simultaneously showing very similar style. Text 4475 does not show a meaningful plagiarism ratio, but it also ships with a significantly lower style similarity. The results generally show the correlation of the style similarity and the plagiarism detected: The more similar the writing style of two texts compared, the higher the percentage of plagiarism overlap can be. In comparison with the Webis plagiarism data, the GPT-rephrased versions show less plagiarism detected ranging relatively close between 21% and 46%. This is on par with the study by Gao et al. (2023) where the authors found that GPT-produced texts are less likely to be plagiarised. We still also observe a narrow range in the use of style, ranging between .93 and .97. Looking at samples, we find that GPT typically replaces content words with similar ones, but the overall sentence structure stays rather similar. We find that, although less obvious, the same style-plagiarism correlation is visible.

Essays & GPT: The essays also show a high ratio of plagiarism overlap and a very strong correlation with the relating style similarities. The observations are comparable with those from the Webis Plagiarism & GPT. Again, very similar style can be owed to the fact that ChatGPT does not do a good job in rephrasing sentences, it simply replaces words and phrases.

Almeida Gospels: The results show an identical cosine similarity of .99 while the detected plagiarism is between 16% and 27%, higher in the Matthew & Marc comparison than in the Luke & Marc comparison. These samples are possibly comparable with the Webis Plagiarism & GPT samples, which also show high stylistic similarity while also showing a meaningful plagiarism overlap.

5 Conclusion

We showed that there is a moderate correlation between the text samples coming from the English literature period, the Webis Plagiarism corpus together with their paraphrased versions and the books of the Brazilian Bible translation. We carefully selected the features that we utilize to measure style similarities considering function words distribution that do reliably represent style characteristics. In future work, we will fine-tune the procedure and have a closer look at how different register ranges affect our style similarities, and how pruning the most frequent n words (depending on the language) affects these correlation. We also found lower correlation between the style employed and the plagiarism detected in the texts re-phrased by ChatGPT. Which leads us to the conclusion that it copies the author's style, especially because it can sample style from pre-existing texts for German only with some effort.

References

Muna AlSallal, Rahat Iqbal, Vasile Palade, Saad Amin, and Victor Chang. 2019. An integrated approach for intrinsic plagiarism detection. *Future Generation Computer Systems*, 96:700–712.

Faisal Alvi, Mark Stevenson, and Paul Clough. 2021. Paraphrase type identification for plagiarism detection using contexts and word embeddings. *International Journal of Educational Technology in Higher Education*, 18(1):42.

Salha Mohammed Alzahrani and Naomie Salim. 2008. Plagiarism detection in arabic scripts using fuzzy information retrieval. In *Student Conf. Res. Develop., Johor Bahru, Malaysia*, pages 281–285.

Steven Burrows, Martin Potthast, and Benno Stein. 2013. Paraphrase Acquisition via Crowdsourcing and Machine Learning. *Transactions on Intelligent Systems and Technology (ACM TIST)*, 4(3):43:1–43:21.

Hugh Craig and Arthur F Kinney. 2009. *Shakespeare, computers, and the mystery of authorship*. Cambridge University Press.

Fred J Damerau. 1975. The use of function word frequencies as indicators of style. *Computers and the Humanities*, pages 271–280.

Maciej Eder, Jan Rybicki, and Mike Kestemont. 2016. Stylometry with R: A Package for Computational Text Analysis. *The R Journal*, 8(1):107–121.

Sven Meyer zu Eissen and Benno Stein. 2006. Intrinsic plagiarism detection. In *Advances in Information Retrieval: 28th European Conference on IR Research, ECIR 2006, London, UK, April 10-12, 2006. Proceedings 28*, pages 565–569. Springer.

Stefan Evert, Fotis Jannidis, Friedrich Michael Dimpel, Christof Schöch, Steffen Pielström, Thorsten Vitt, Isabella Reger, Andreas Büttner, and Thomas Proisl. 2016. " delta" in der stilometrischen autorschaftsattribution. In *DHd*.

Tomáš Foltỳnek, Norman Meuschke, and Bela Gipp. 2019. Academic plagiarism detection: a systematic literature review. *ACM Computing Surveys (CSUR)*, 52(6):1–42.

Catherine A Gao, Frederick M Howard, Nikolay S Markov, Emma C Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T Pearson. 2023. Comparing scientific abstracts generated by chatgpt to real abstracts with detectors and blinded human reviewers. *NPJ Digital Medicine*, 6(1):75.

Douglas Wilhelm Harder. 2022. Plagiarism in the gospels. Acc: Jan 2024 https://www.dwharder.org/plagiarism-in-the-gospels.

David L Hoover. 2003. Frequent collocations and authorial style. *Literary and Linguistic Computing*, 18(3):261–286.

Ethan Hunt, Ritvik Janamsetty, Chanana Kinares, Chanel Koh, Alexis Sanchez, Felix Zhan, Murat Ozdemir, Shabnam Waseem, Osman Yolcu, Binay Dahal, et al. 2019. Machine learning models for paraphrase identification and its applications on plagiarism detection. In *2019 IEEE International Conference on Big Knowledge (ICBK)*, pages 97–104. IEEE.

Stefan Jänicke, Annette Geßner, Marco Büchler, and Gerik Scheuermann. 2014. Visualizations for text re-use. In *2014 International Conference on Information Visualization Theory and Applications (IVAPP)*, pages 59–70. IEEE.

John V. Nance. 2017. "we, john cade": Shakespeare, marlowe, and the authorship of 4.2.33–189 2 henry vi. *Shakespeare*, 13(1):30–51.

Gabriel Oberreuter and Juan D Velásquez. 2013. Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style. *Expert Systems with Applications*, 40(9):3756–3763.

OpenAI. 2024. Chatgpt. Acc: Jan 2024 https://chat.openai.com.

Kazuaki OTA. 2023. Was marlowe shakespeare's collaborator?: Computational stylometry and the authorship of the three parts of henry vi.

Karl Pearson. 1895. Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London Series I*, 58:240–242.

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An evaluation framework for plagiarism detection. In *Coling 2010: Posters*, pages 997–1005.

Lukas Zwilling and Maria Berger. 2024. Chatgpt does not speak style!