Small-Scale Testing on Generative AI and Post-OCR Correction in Historical Datasets

Florentina Armaselu, University of Luxembourg
florentina.armaselu@uni.lu

Introduction

Recent developments in large language models (LLMs) and generative AI (GenAI) chatbots such as ChatGPT, Google Bard (now, Gemini) and YouChat (Brown et al., 2020; Manyika and Hsiao, 2023; Chaka, 2023) have fostered new types of interactions that can lower the barrier in human-machine communication through conversation in natural languages. We assume that such chatbots may be able to act as conversational assistants in tasks that otherwise require more complex processing, to improve the results produced by simpler or earlier, less performant techniques. This article proposes a set of small-scale tests with GenAI chatbots on post-OCR correction in historical datasets. It illustrates, through examples of responses obtained from GenAI agents integrated into post-OCR correction and assessment tasks, what types of challenges have to be addressed in this context when working with historical datasets.

Previous studies have shown that OCR errors in input data can have non-negligible impact on downstream language processing, such as sentence segmentation, named entity recognition (NER), topic modelling and word embedding (Strien et al., 2020). Therefore, various methods for tackling this problem have been envisaged. For instance, machine learning techniques for automatically estimating text quality and selecting candidates for OCR rerun have been examined within cultural institutions that deal with historical data of lower quality (Schneider and Maurer, 2022). On the other hand, studies on post-OCR correction have investigated the use of pretrained language models, such as GPT-2 family, in combining different OCR views with the goal of producing fewer errors (Gupta et al., 2021).

Our set of tests explore the potential of using GenAI agents in post-OCR correction. This type of enquiry is part of a larger project that uses word embedding results (neighbour lists) and citations extracted from a selection of historical French monographs (1690-1918)[1] intended to be converted and integrated into a multilingual diachronic collection of interconnected terms expressed in RDF-XML to be published in the linguistic linked open data (LLOD) cloud (Armaselu et al., 2024). In this context, the need for OCR corrections of neighbour lists and citations has been identified.

Methodology and results

The example presented below has been produced using three chatbots, ChatGPT-4, Google Bard and YouChat that were selected based on availability via subscription and free accounts and on reports considering them among the top AI chatbots (Davis, 2023). The conversations were intended to identify potential problems in the lists of neighbours of the word *révolution*, such as OCR errors (Table 1). More complex prompting included several exchanges with the GenAI agent (ChatGPT-4) when asking for error identification in citations extracted from the corpus corresponding to a given sense of the word *révolution*

---

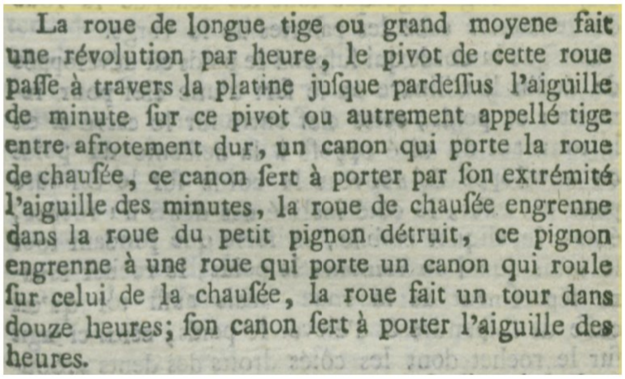
[1] The MONOGRAPH TEXT PACK from the BnL Open Data Collection. https://data.bnl.lu/data/historical-newspapers/.

(Figure 1). Information about the origin of the excerpts, i.e., French historical texts and their publication period, was also included in the prompt.

Table 1: Identifying OCR errors. Prompt: *Please try to identify the words that may be due to OCR errors in the following list: vandalisme, insurrection, insurgé, confération, lirnbourg, eévolution, vindépendance, holstein, usurper, exéquatur*.

Chatbot	**Interaction date**	**Answer excerpts**	**Chatbot confidence**
ChatGPT-4	15.12.2023	• confération - Possible Correction: "confédération" (confederation) • lirnbourg - Possible Correction: "Limbourg" (Limburg) • eévolution - Possible Correction: "révolution" (revolution) • vindépendance - Possible Correction: "indépendance" (independence)	85% 80% 95% 90%
Bard	18.12.2023	• confération - confédération • lirnbourg - Luxembourg • eévolution - évolution • vindépendance - indépendance	100% 95% 90% 80%
YouChat	16.12.2023	• "lirnbourg" to "Limbourg" • "eévolution" to "évolution" • "vindépendance" to "indépendance"	60% 80% 90%

Figure 1 (right) shows the OCR-ised text that was prompted to ChatGPT-4 for correction.



La roue de longue tige ou grand moyene fait une révolution par heure, le pivot de cette roue paſſe à travers la platine juſque pardeſſus l'aiguille de minute ſur ce pivot ou autrement appellé tige entre afrotement dur, un canon qui porte la roue de chauſée, ce canon ſert à porter par ſon extrémité l'aiguille des minutes, la roue de chauſée engrenne dans la roue du petit pignon détruit, ce pignon engrenne à une roue qui porte un canon qui roule ſur celui de la chauſée, la roue fait un tour dans douze heures; ſon canon ſert à porter l'aiguille des heures.

La roue de longue tige ou gand moyene fait une révolution par heure, le pivot de cette rous pale à travers la platune jufque pardeffus l'aiguille de minute 1ur ce pivot ou autrement appellé tige entre afrotement dur, un canon qu porte la roue de chaufée, ce canon fert à porter par fon extrémit l'aiguille des minuts, la roue de chaufée engrenne dans la roue du petit pignon détruit, ce pignon engrenne à uns roue qui porte un canon qui roule fur celui de la chaufée, la roue fäit un tour dans douze heures; fon canon fert à porer l'aiguille des neures.

Figure 1. Extract (Rosset, 1789: 13): a) image; b) OCR-ised text

Following this preliminary testing phase, a small-scale assessment of ChatGPT-4 was also performed using test and gold standard excerpts from the ICDAR 2017 competition on post-OCR text correction (Rigaud et al., 2019). The fragment contained 50 sentences from a text by Montesquieu, *Arsace et Isménie*, published in 1783, chosen to be close to the publication date of the text analysed above. Both the test fragment and the gold standard were extracted from the ICDAR2017 dataset, French monographs.[2]

ChatGPT-4 was asked to correct the first 50 sentences from the test fragment, that was supposed to contain OCR errors. Then, it was prompted to compare the initial and the corrected texts with the corresponding gold standard (GS), to compute the character error rate (CER), word error rate (WER) (Abadie et al., 2022) and corresponding accuracy values. Table 2 shows two iterations of this process, since it was observed that ChatGPT-4 had a tendency to correct the historical word forms in French, and

---

[2] https://sites.google.com/view/icdar2017-postcorrectionocr/dataset. Some preparations were necessary, for instance to remove the symbol '@' from the excerpts, which was used in the competition for alignment at the character level.

additional prompting was necessary to prevent this type of change. For comparison, the CER and WER values, indicated in brackets, were computed independently using the Python library jiwer.[3]

Table 2. ChatGPT-4 CER and WER values (with jiwer references) and examples (ICDAR 2017 excerpts).[4]

Text file	**Interaction date**	**CER**	**WER**	**Chatbot confidence**	**Differences in the analysed text / GS**
OCR_toInput_Arsace_et_Isménie_1783 (ICDAR test fragment)	03.05.2024	1.22% 1.38% (jiwer)	6.68% 7.55% (jiwer)	95%	“.on” / “son” “pres-que” / “presque” “Ambassa-deurs, ” / “Ambassadeurs, ”
ChatGPT-4_Corr1_Arsace_et_Isménie_1783	03.05.2024	5.44% 5.15% (jiwer)	18.22% 18.01% (jiwer)	95%	“règne” / “regne” “d'Artamène, ” / “d'Artamene, ” “laissant” / “laissa” “désirait” / “desiroit”
ChatGPT-4_Corr2_Arsace_et_Isménie_1783	03.05.2024	3.23% 3.00% (jiwer)	12.63% 12.31% (jiwer)	95%	“règne” / “regne” “d'Artamène,” / “d'Artamene,” “et” / “1” “âme” / “ame”

Discussion

A qualitative analysis of the chatbots' answers for the first task (Table 1) indicates that the three AI agents were quite successful in identifying possible errors. The word similarly identified by all the chatbots was “vindépendance”, while slight differences in agreement can be observed for the other words. ChatGPT-4 agreed with Bard on “confération” and with YouChat on "lirnbourg", while Bard and YouChat agreed on “eévolution”. Additional prompting was needed to remind Bard that the corrected words should be provided in French (table 3). When the prompt included more details about specific types of errors, such as the use of elongated 's' in older texts, often confused by the OCR software with the letter 'f', the chatbots were able to propose plausible corrections. For instance, YouChat after firstly proposing “fantastic” as a correction of “fuppofant”, reconsidered it in favour of the form “supposant” (confidence 70%) (table 4).

More complex prompting included several exchanges when asking for error identification in citations extracted from the corpus corresponding to a given sense of the word *révolution* (Figure 1). While errors such as “platune” instead of “platine”, “jufque pardeffus” instead of “jusque pardessus”, or “fert” instead of “sert” were easily identified (confidence 90%), the form “chaufée” was more difficult to analyse and required additional prompting. After asking ChatGPT-4 to reconsider the form to take into account the elongated ‘s’ issue, the context “roue de chaufée” and the possibility that it may refer to clock making in the 18-th century, the chatbot proposed the correction “roue de chaussée”, with “’chaussée’ potentially referring to a part or a process in the clock’s mechanism” (confidence 75%) (table 5). Other difficulties

---

[3] https://jitsi.github.io/jiwer/.

[4] ChatGPT-4 was selected for this type of task since it allowed downloading the files corresponding to the various versions to be compared. Due to instability in results and reported errors by the GenAI agent, the calculation of the CER and WER values was repeated several times (on 31.01.2024, 07.02.2024, 02.05.2024 and 03.05.2024). The last results that seemed more stable and closer to the jiwer-calculated values, were included in the table.

consisted in preventing corrections of possible historical word usages, which were not due to OCR errors, such as “moyene”, “engrenne”, “afrotement” and their transformation into modern forms (table 6).

The same issue related to historical forms was observed when comparing the corrected ChatGPT-4 version with the gold standard from ICDAR 2017 (Table 2). One can observe that the lowest error values corresponded to the ICDAR test fragment. For the two iterations using ChatGPT-4, the error rate decreased for correction 2 as compared with correction 1 after the use of specific prompts indicating that historical forms should be preserved.[5] An additional test was performed to compare the texts corrected by ChatGPT-4 with a more modern version (1876) (MD) of the fragment from *Arsace et Isménie*.[6] The computed CER and WER values using jiwer for the GPT-corrected texts using this text as a reference[7] were higher as compared with the gold standard-based values from table 2, which was a bit surprising. Some mismatches were due to the use of different types of punctuation marks (e.g., straight vs. curly apostrophes). A closer look at the differences also showed that ChatGPT-4 corrections targeted not only presumed OCR errors but sometimes involved changing the order of words or rephrasing.[8]

Conclusion and future work

This article proposes a small-scale investigation on the use of GenAI agents for post-OCR correction in historical datasets. While the preliminary results show a certain potential of this type of technology in solving tasks from this category, more testing is necessary to assess their capacity to respond to prompts specially conceived for historical text processing. In particular, it was shown that the agents have a tendency to replace historical forms with more modern ones, to reformulate whole phrases or change punctuation. Specific prompts should be devised to prevent these forms of modification. However, building modernised layers for documents with old spelling may be considered a potentially interesting application in tasks such as the transformation of older texts to be read by modern users or computer programs.

Some instability in computing character and word error rates (CER, WER) was also observed, which indicates that comparing these values with independently computed results should be envisaged. Moreover, the interaction with the agents involved the use of online platforms, while the integration of this type of technology into larger-scale pipelines would probably require more code-oriented solutions that need to be further examined.

References

Abadie, N., E. Carlinet, J. Chazalon, and B. Duménieu. 2022. “A Benchmark of Named Entity Recognition Approaches in Historical Documents Application to 19th Century French Directories.” In Document

---

[5] Example of such prompt: *Please try to correct only the clear OCR errors, such as misrecognized characters and incorrect word splits, while retaining the historical French forms, such as "reposoit", "faudroit".*

[6] https://fr.wikisource.org/wiki/Arsace_et_Ism%C3%A9nie.

[7] ChatGPT-4 correction 1 (5.52% CER and 20.36% WER); correction 2 (5.41% CER and 22.87% WER).

[8] For instance, “Il écrivait à la Reine les lettres les plus tendres du monde” (corr. 1) vs. “Il écrivoit à la Reine les lettres du monde les plus tendres” (GS, corr. 2) vs. “Il écrivait à la reine les lettres du monde les plus tendres” (MD); “Elle descendit de son char et entra dans le temple” (corr. 1) vs. “Elle descendit de son char, entra dans le temple” (GS, corr. 2) vs. “Elle descendit de son char, et entra dans le temple” (MD); “il cherchait l'étranger et le trouva plongé dans une affreuse tristesse” (corr. 1) vs. “il cherchoit l'étranger, il le trouva dans une affreuse tristesse” (GS) vs. “il cherchoit l'étranger, et le trouva dans une affreuse tristesse” (corr. 2) vs. “il cherchait l’étranger, et il le trouva dans une affreuse tristesse” (MD).

Analysis Systems, edited by Seiichi Uchida, Elisa Barney, and Véronique Eglin, 13237:445–60. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-06555-2_30.

Armaselu, Florentina, Chaya Liebeskind, Paola Marongiu, Barbara McGillivray, Giedrė Valūnaitė Oleškevičienė, Elena Simona Apostol, and Ciprian-Octavian Truică. "Linguistic Linked Open Data for Diachronic Analysis (LLODIA)." Dataset. *GitHub Nexus Linguarum*, April 3, 2024. https://doi.org/10.5281/zenodo.11065197.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models Are Few-Shot Learners." arXiv. http://arxiv.org/abs/2005.14165.

Chaka, Chaka. 2023. "Generative AI Chatbots – ChatGPT versus YouChat versus Chatsonic: Use Cases of Selected Areas of Applied English Language Studies." *International Journal of Learning, Teaching and Educational Research* 22 (6): 1–19. https://doi.org/10.26803/ijlter.22.6.1.

Davis, Krissy. 2023. *The Best AI chatbots: ChatGPT and Other Alternatives*. https://www.wearedevelopers.com/magazine/best-ai-chatbots-chatgpt-and-other-alternatives.

Gupta, Harsh, Luciano Del Corro, Samuel Broscheit, Johannes Hoffart, and Eliot Brenner. 2021. "Unsupervised Multi-View Post-OCR Error Correction With Language Models." In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 8647–52. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.680.

Manyika, James, and Sissie Hsiao. 2023. "An Overview of Bard: An Early Experiment with Generative AI." https://ai.google/static/documents/google-about-bard.pdf.

Rigaud, Christophe, Antoine Doucet, Mickael Coustaty, and Jean-Philippe Moreux. 2019. "ICDAR 2019 Competition on Post-OCR Text Correction." In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1588–93. Sydney, Australia: IEEE. https ://doi.org/10.1109/ICDAR.2019.00255.

Rosset, F. 1789. *L'art de conduire et regler les pendules et les montres.* Chez la Veuve de J. B. Kleber, Imp de Sa Majesté. Digitised by the National Library of Luxembourg, p. 13. https://viewer.eluxemburgensia.lu/ark:70795/dqgfr3/pages/17/articles/DTL612.

Schneider, Pit, and Yves Maurer. 2022. "Rerunning OCR: A Machine Learning Approach to Quality Assessment and Enhancement Prediction." Journal of Data Mining & Digital Humanities 2022, November 30, 2022. https://doi.org/10.46298/jdmdh.8561.

Strien, Daniel van, Kaspar Beelen, Mariona Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. 2020. "Assessing the Impact of OCR Quality on Downstream NLP Tasks." In *Proceedings of the [1]2th International Conference on Agents and Artificial Intelligence*, 484–96. Valletta, Malta: SCITEPRE–S - Science and Technology Publications. https://doi.org/10.5220/0009169004840496.

Appendix

Table 3. Prompt and response excerpts, reminding the language (Google Bard, 18.12.2023)

User	**GenAI agent (Bard)**
Please try to identify the words that may be due to OCR errors in the following list and assign a confidence level in percentage to each proposed correction. The words are in French and the corrections should also be in French. …	Sure, here are the words in the list that may be due to OCR errors, along with proposed corrections and confidence levels: … confération confédération 100% …

Table 4. Prompt and response excerpts, “supposant” (YouChat, 16.12.2023)

User	**GenAI agent (YouChat)**
Would your decision be different if you consider that usually elongated 's' in older texts is producing OCR errors, such as considering 'f' instead of 's'?	Considering the possibility of an OCR error substituting "f" for elongated "s" is indeed a relevant factor to consider. Given this information, the potential correct form for the word "fuppofant" might indeed be "supposant" if we consider the substitution of "s" with "f" due to OCR errors in older texts. Taking this into account, the corrected potential form for "fuppofant" could be "supposant" instead of "fantastic." …

Table 5. Prompt and response excerpts, "roue de chaussée" (ChatGPT-4, 15.12.2023)

User	**GenAI agent (ChatGPT-4)**
Can you please reanalyze the word "chaufée" given the context "roue de chaufée" and the fact that the text refer to clock making in the 18th century?	In the context of clockmaking during the 18th century, the term "chaufée" in "roue de chaufée" could be a historical or technical term related to the mechanism of a clock. Considering the OCR misinterpretation of long 's' (ſ) as 'f', and given the clockmaking context, the term could be "roue de chaussée", with "chaussée" potentially referring to a part or a process in the clock's mechanism. However, without specific knowledge of 18th-century clockmaking terminology, it's difficult to ascertain the exact term. "Roue de chauffée" could refer to a heated part or a wheel that is treated in a particular way, but this seems less likely in the context of mechanical clockwork where terms like "chaussée" do not clearly fit. Considering this, a more likely correction would be "roue de chaussée" if "chaussée" is a term used in the context of clock mechanics at the time. However, if "chauffée" is not due to an OCR error and is actually a correct transcription of the original text, it might require specialized historical knowledge to interpret accurately. Based on common French usage and the potential for an OCR error, "roue de chaussée" could be a plausible correction with a medium to high confidence level, but without specific expertise in historical clockmaking terminology, I cannot provide a definitive correction.

Table 6. Prompt and response excerpts, "afrotement", "engrenne" (ChatGPT-4, 15.12.2023)

User	**GenAI agent (ChatGPT-4)**
Do you have any information about the word "afrotement ", in the text "tige entre afrotement dur", and its possible usage in French in the 18th century? Same question for "engrenne ", in "ce pignon engrenne". It is possible that these are historical usages that should be kept in the transcription of the text. The goal is to correct just the errors that may be due to OCR misinterpretation.	In the 18th century, French language and spelling were less standardized than they are today, and many words could have various spellings or be spelled according to phonetics. With this in mind: • Afrotement: The correct modern French word is "frottement," which means "friction." It is plausible that "afrotement" is a historical spelling or a typographical error in the original text. If "afrotement" is not an OCR misinterpretation, it could be an archaic spelling of "frottement" that was used at the time. This would be especially true in the context of clock mechanisms where friction is an important concept. • Engrenne: The modern French word "engrène" comes from the verb "engrener," which means "to mesh" or "to engage," as gears do in a clock. "Engrenne" could be an old conjugation or spelling of the verb. If "engrenne" is correctly transcribed, it may not be an OCR error but rather a historical usage that should be preserved in the text. …