

TRANSCRIÇÃO AUTOMÁTICA DE MANUSCRITOS BRASILEIROS: ESTUDO PRELIMINAR

AUTOMATIC TRANSCRIPTION OF BRAZILIAN MANUSCRIPTS: A PRELIMINARY STUDY

XIV WICI 2026
WORKSHOP INTERNACIONAL
EM CIÊNCIA DA INFORMAÇÃO

MAURICIO MAIA VINHAS DE AZEVEDO

Universidade Federal Fluminense (UFF)

ORCID: <https://orcid.org/0000-0002-7396-1182>

ROR: <https://ror.org/02rjhbb08>

CARLOS HENRIQUE MARCONDES

Universidade Federal Fluminense (UFF)

ORCID: <https://orcid.org/0000-0003-0929-8475>

ROR: <https://ror.org/02rjhbb08>

MODALIDADE: RESUMO EXPANDIDO

Resumo: Há uma vasta quantidade de manuscritos históricos em acervos no Brasil e ao redor do mundo. Contudo, manuscritos podem ser de difícil leitura, prejudicando o acesso ao seu conteúdo, mesmo quando digitalizados. Pesquisas sobre Reconhecimento de Texto Manuscrito buscam soluções automatizadas para o problema. O objetivo deste trabalho é criar e avaliar modelos adaptados para o português do Brasil por meio do ajuste fino da arquitetura TrOCR. Experimento utilizando manuscrito produzido em 1874 demonstrou que o modelo ajustado obteve performance 69% melhor do que o modelo de referência, sugerindo a viabilidade da proposta e incentivando maior exploração.

Palavras-chave: reconhecimento de texto manuscrito; ajuste fino; TrOCR.



Abstract: Throughout most of human history, documents were written by hand. There are vast collections of manuscripts in Brazil and around the world. Manuscripts can be difficult to read, hindering access to their content, even when digitized. Research on Handwritten Text Recognition seeks automated solutions to this problem. The objective of this study is to create and evaluate preliminary models adapted for Brazilian Portuguese by fine-tuning the TrOCR architecture. Experimental results on an 1874 manuscript indicate that the fine-tuned model yields a 69% improvement over the baseline, demonstrating the feasibility of the proposal and warranting further investigation.

Keywords: handwritten text recognition; fine-tune; TrOCR.



1 INTRODUÇÃO

Há uma vasta quantidade de manuscritos históricos em acervos ao redor do mundo, no Brasil, só a Fundação Biblioteca Nacional têm mais de um milhão (Merege, 2019, p. 234). Os desafios da recuperação de informação em manuscritos vão desde a difícil leitura dos originais até o produto da digitalização, que normalmente é apenas uma imagem do texto, e não o texto digital (Toselli; Puigcerver; Vidal, 2024, p. 2). A falta de transcrição de manuscritos dificulta o trabalho dos pesquisadores (Terras, 2022, p. 182). Transcrevê-los **PODE FACILITAR A DESCOBERTA E RECUPERAÇÃO DE INFORMAÇÕES, POSSIBILITANDO NOVOS USOS E NOVOS CONHECIMENTOS**. Neste cenário, a transcrição paleográfica surgiria como uma poderosa aliada e a sua automação, tema dos estudos da área de Reconhecimento de Texto Manuscrito (HTR – *Handwritten Text Recognition*), proporcionaria uma escalabilidade inédita.

A transcrição paleográfica de manuscritos em larga escala ampliaria o acesso à informação e a disseminação do conteúdo dos documentos (Krüger, 2014, p. 221), contudo, é inviável fazê-la manualmente por conta da escalabilidade e custo (Toselli; Puigcerver; Vidal, 2024, p. 2). Pesquisas em HTR, atualmente, empregam inteligência artificial para a transcrição automatizada¹ de manuscritos digitalizados com o intuito de tornar seu conteúdo textual mais acessível (Paixão de Sousa *et al.*, 2024, p. 3).

Há poucos estudos sobre HTR para língua portuguesa. Destaca-se a pesquisa em que Matos *et al.* (2024) criaram e testaram modelos para português medieval. Lose *et al.* (2024) exploraram o “processo de constituição do modelo Padre Antônio Vieira” no Transkribus². Paixão de Sousa *et al.* (2024) construíram modelos a partir de documentos portugueses (1591-1595) e assinalaram a escassez de modelos de HTR específicos para língua portuguesa. Todos os estudos mencionados se baseiam em documentos portugueses.

Diante dos desafios da legibilidade por máquina do conteúdo textual de manuscritos digitalizados, a proposta deste trabalho é, em caráter preliminar, verificar o desempenho do TrOCR³ para a transcrição de documentos brasileiros produzidos no século XIX. Tem por objetivos verificar a viabilidade do ajuste fino⁴ somente do decodificador

1 Um exemplo de reconhecimento automatizado de texto é o OCR (optical character recognition). Diferente do OCR, o HTR processa as imperfeições da escrita manual e produz uma transcrição, isto é, uma nova escrita.

2 Ferramenta de HTR desenvolvida na Universidade de Innsbruck, Áustria: <https://www.transkribus.org/>

3 Arquitetura de Transformadores e aprendizado por máquina: <https://github.com/rsommerfeld/trocr>.

4 Ajuste fino (fine-tune) é o processo de adaptar um modelo já treinado para tarefas ou contextos específicos.



(*decoder*)⁵ do TrOCR adaptando-o para português brasileiro; testar ajuste fino com linhas completas e palavras individualizadas e, por fim, comparar o desempenho dos modelos. A viabilidade do HTR é condição para promoção de manuscritos enquanto dados para pesquisa e reúso.

O presente trabalho possui algum ineditismo, pois não foi identificado outro estudo abordando a criação de modelo público para língua portuguesa empregando documentos brasileiros. Sugere que os procedimentos efetuados poderão ser replicados. O potencial demonstrado incentiva a exploração de outras ferramentas de HTR e até mesmo o emprego de MLLMs (Multimodal Large Language Models) como Gemini ou ChatGPT.

⁵ O decodificador (*decoder*) escreve o que foi reconhecido pelo processador (*processor*).



2 MÉTODOS E MATERIAIS

O presente trabalho resulta de pesquisa em andamento, de natureza exploratória, quantitativa e experimental, cujo objetivo é identificar e avaliar ferramentas e métodos para o reconhecimento de textos manuscritos (HTR) que possam ser usados em acervos brasileiros. Como não há modelos públicos adaptados à língua portuguesa, foi feito ajuste fino do modelo de referência da aplicação TrOCR⁶. Todos os procedimentos foram efetuados por meio de *vibe coding*⁷ no ambiente do Google Colab⁸.

Os dados para ajuste fino foram construídos a partir da transcrição de uma página do livro de Atas da Congregação da Faculdade de Medicina do Rio de Janeiro em 1874⁹. O primeiro ajuste fino (ft1.0), com 105 linhas, é composto por variações das mesmas 35 linhas de texto¹⁰: cor original, escala de cinza e somente preto. O segundo (ft1.1) parte do "ft1.0", mas adicionando recortes das palavras que compõem cada linha, totalizando 354 linhas.

A avaliação da qualidade das transcrições foi feita por meio das métricas "índice de erros de caracteres" (CER – *Character Error Rate*) e índice de erros de palavras (WER – *Word Error Rate*)¹¹, amplamente utilizadas no contexto de HTR. O conjunto de avaliação é composto por quatro linhas de outra página do mesmo documento com versões pré-processadas totalizando nove itens¹².

6 Ferramenta admitida como estado da arte em reconhecimento de texto manuscrito (Meoded, 2025, p. 17).

7 "Vibe coding" é o desenvolvimento de software mediado por IA generativa.

8 Ambiente "Jupyter Notebooks" em: <https://colab.research.google.com/>

9 Diante do caráter preliminar do estudo, foi feita amostragem por conveniência. Documento compõe acervo em que o autor atua (UFRJ) e foi selecionado de forma espontânea ao ser requisitado por um pesquisador.

10 Tentativa de "Data augmentation", estratégia para lidar com poucos dados para treino (Meoded, 2025, p. 2).

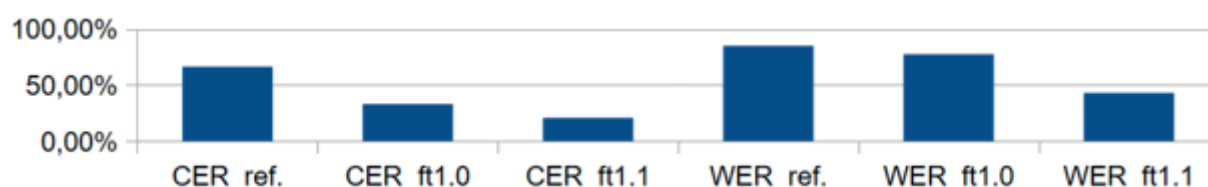
11 Valores menores representam melhores transcrições: um CER de 30% significa metade dos erros de um de 60%. Não distingue letras maiúsculas e minúsculas e baseia-se na posição dos caracteres ou palavras no texto.

12 Todos os modelos e códigos utilizados estão disponíveis em: <https://zenodo.org/records/20317590>

3 RESULTADOS INICIAIS

O modelo resultante da primeira rodada de ajuste fino (ft1.0) apresentou médias CER e WER de 32% e 77%. Portanto, reconheceu corretamente 50% mais caracteres e 9% mais palavras do que o modelo de referência que apresentou CER e WER médios de 62% e 85%. O segundo modelo (ft1.1) obteve médias CER e WER de 20% e 42%, melhora de 69% e 49% em comparação à referência¹³ - mais da metade das palavras foi corretamente reconhecida. Comparações no gráfico 1. O “ft1.1” demonstrou maior incremento em qualidade de identificação de palavras (40%) do que de letras (18%) em relação ao “ft1.0” sugerindo que a inclusão de palavras no conjunto de dados aprimora o reconhecimento de palavras.

Gráfico 1 – Índice de erros de caracteres (CER) e de palavras (WER)



Fonte: Autoria própria (2026).

Os modelos resultantes do processo de ajuste fino apresentaram crescente melhoria com redução nos índices de erros chegando a quase 70%. Contudo, mesmo os melhores índices alcançados no modelo “ft1.1”, ainda são altos, pois significam que 20% dos caracteres não foram identificados e que 42% das palavras foram transcritas com falha. Apesar disto, os resultados são animadores por conta da melhora a cada etapa.

¹³ Exemplo de transcrição (linha 3). Transcrição manual: “Medicina do Rio de Janeiro de 29 de Ja-”. Modelo de referência: “Middlesex vla Rio de Janeiro de Janeiro play”. Modelo ft1.0: “Mediando do Rio Janeiro da em dea Val”. Modelo ft1.1: “Medicina da Rio de Janeiro de 9osões”. Aparenta dificuldade com o início do texto.



4 CONSIDERAÇÕES FINAIS

Os resultados iniciais obtidos demonstram o potencial do TrOCR quando ajustado. Os objetivos foram alcançados e os dados indicam que o ajuste fino mediante *vibe coding* funcionou. Uma das etapas mais desafiadoras foi a da segmentação de linhas¹⁴, que exigiu ajustes manuais e seria um dos maiores desafios de HTR (Paixão de Sousa *et al.*, 2024).

A principal limitação do estudo é seu caráter preliminar. A avaliação com apenas quatro linhas de texto pareceu suficiente para o escopo definido, mas é reduzida demais para possibilitar qualquer generalização dos resultados.

Em cenários com alto índice de erros, as métricas CER e WER podem ser pouco esclarecedoras tanto para avaliar a legibilidade das transcrições como no contexto de extração de termos para indexação automática e recuperação da informação. Portanto, é necessário complementar a avaliação com outras medidas de análise.

¹⁴ Foi utilizada a ferramenta HTRFlow para segmentar o texto: <https://ai-riksarkivet.github.io/htrflow/latest/>



REFERÊNCIAS

- KRÜGER, A. C. O ensino de Paleografia no curso de graduação em Arquivologia da UFSC: um exercício com os documentos do Instituto Histórico e Geográfico de Santa Catarina. **Ágora**, v. 24, n. 48, p. 211-223, 2014. Disponível em: <https://agora.emnuvens.com.br/ra/article/view/483>. Acesso em: 15 mar. 2026.
- LOSE, A. D.; SANTOS, J. G. V. A. dos; JESUS, L. C. M. de; MAGALHÃES, L. B. S.; XAVIER, L. W. Transkribus: uma ferramenta de paleografia digital mediando pesquisas em fontes inquisitoriais. **LABORHISTÓRICO**, v. 10, n. 1, p. 1-22, 2024. Disponível em <https://doi.org/10.24206/lh.v10i1.63285>. Acesso em: 22 ago. 2025.
- MATOS, A.; ALMEIDA, P.; CORREIA, P. L.; PACHECO, O. iForal: Automated Handwritten Text Transcription for Historical Medieval Manuscripts. **Journal of Imaging**, v. 11, n. 36, 2025. DOI: <https://doi.org/10.3390/jimaging11020036>. Disponível em: <https://www.mdpi.com/2313-433X/11/2/36>. Acesso em: 15 jan. 2026.
- MEODED, E. Handwritten Text Recognition of Historical Manuscripts Using Transformer-Based Models. **ARXIV**, v. 1, ago. 2025. Disponível em: <https://arxiv.org/abs/2508.11499v1>. Acesso em: 19 nov. 2025.
- MEREGE, A. L. Desafios e perspectivas frente aos manuscritos da Biblioteca Nacional. **LABORHISTÓRICO**, v. 5, n. 2, p. 231-248, 2019. DOI: <https://doi.org/10.24206/lh.v5i2.24922>. Disponível em <https://revistas.ufrj.br/index.php/lh/article/view/24922>. Acesso em: 24 ago. 2025.
- PAIXÃO DE SOUSA, M. C.; MONTE, V. M.; NAMIUTI, C.; XAVIER, L. F. W.; MAGALHÃES, L. B. S.; SANTOS, J. V.; COSTA, B. S.; COSTA, A. S.; STURZENEKER, M. L. Digital technologies for early modern portuguese manuscripts: an experiment with two handwritten text recognition software applications. **Preprints.org**, v. 1, 2024. Disponível em: <https://www.preprints.org/manuscript/202412.0864>. Acesso em: 24 ago. 2025.
- TERRAS, M. Inviting AI into the archives: the reception of handwritten recognition technology into historical manuscript transcription. In: JAILLANT, L. (ed.). **Archives, access and artificial intelligence: working with born-digital and digitised archival collections**. Bielefeld: Transcript Verlag, 2022. p. 179-204. (Digital Humanities Research, v. 2).
- TOSELLI, A. H.; PUIGCERVER, J.; VIDAL, E. **Probabilistic indexing for information search and retrieval in large collections of handwritten text images**. Cham: Springer Nature Switzerland, 2024. (The Information Retrieval Series, 49).