

Transcritor-IA: utilizando inteligência artificial para transcrição de manuscritos históricos

Rodrigo C.¹, Alysson H.¹, Alessandro C.¹, Eduardo M.², Elizabeth K.¹, George Ney¹, Raimundo Valter¹

¹Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
Aracati, CE – Brasil

rodrigow1238@gmail.com, henriquealysson000@gmail.com

Alessandro.costasv@gmail.com, elizabethnicolau13@gmail.com

george.almeida@ifce.edu.br, valter.costa@ifce.edu.br

²Universidade de Fortaleza (UNIFOR) – Fortaleza, CE – Brasil

edmalfizeo@gmail.com

Abstract. *This paper presents Transcritor-IA, a tool for automated transcription of historical manuscripts using Handwritten Text Recognition (HTR) models trained with PyLaia. The platform performs hierarchical segmentation and iterative training based on user-verified transcriptions. We also compare Transcritor-IA with Transkribus[®], analyzing performance similarity and factors affecting transcription quality.*

Resumo. *Este artigo apresenta o Transcritor-IA, uma ferramenta para transcrição automatizada de manuscritos históricos com modelos de Handwritten Text Recognition (HTR) treinados via PyLaia. A plataforma realiza segmentação hierárquica e treinamento iterativo com apoio de transcrições validadas por usuários. Também comparamos o Transcritor-IA ao Transkribus[®], analisando a similaridade de desempenho e os fatores que impactam a qualidade das transcrições.*

1. Introdução

A Inteligência Artificial (IA) tem transformado significativamente diversas áreas, automatizando desde tarefas simples até processos complexos, como diagnósticos médicos e análise de grandes volumes de dados [Twinkle Sharma and Arora 2024, Lakshmi 2024]. No campo da transcrição de manuscritos, os avanços em aprendizado profundo têm impulsionado a área de reconhecimento de textos manuscritos — *Handwritten Text Recognition* (HTR) — permitindo converter escrita à mão em texto digital mesmo em contextos desafiadores [Rakesh et al. 2024].

O HTR é particularmente relevante em documentos históricos, como registros eclesiásticos e governamentais, que possuem valor para áreas como genealogia, história e linguística. A transcrição desses registros é fundamental para estudos da língua pretérita e preservação cultural [Mendes and Oliveira 2016, Moreira et al. 2007], sendo uma diretriz institucional em órgãos como a Biblioteca Nacional [Nacional 2022].

Entretanto, a digitalização enfrenta desafios como caligrafias variadas e material degradado. Apesar dos benefícios do acesso remoto e organização em bases estruturadas, a transcrição ainda demanda esforço humano, o que motivou o surgimento de soluções automatizadas como o *Transkribus*® [Humphries et al. 2024].

Neste contexto, o Transcritor-IA é apresentado como uma ferramenta integrada para transcrição automática de manuscritos, com funcionalidades como certificação, busca textual e retreinamento de modelos. Este artigo discute sua aplicação prática, os métodos empregados e os desafios enfrentados, utilizando modelos tradicionais de HTR, mais adequados para contextos com dados escassos e caligrafias complexas.

O trabalho está estruturado em seis seções: revisão de trabalhos relacionados, descrição do conjunto de dados, metodologia, resultados, discussões e considerações futuras.

2. Trabalhos Relacionados

A tarefa de HTR tem atraído atenção significativa nas últimas décadas, impulsionada pela necessidade de preservar e digitalizar acervos históricos. Nesse contexto, plataformas e *frameworks* têm desempenhado um papel essencial na viabilização de transcrições automáticas e no treinamento de modelos eficientes. Esta seção explora o estado da arte em HTR, destacando ferramentas relevantes e justificando a escolha do *framework* base para o Transcritor-IA.

Entre as plataformas mais reconhecidas para transcrição de manuscritos destaca-se o *Transkribus*® [Kahle et al. 2017], uma solução pioneira no campo de HTR. Inspirado por sua estrutura funcional e impacto na comunidade acadêmica, o Transcritor-IA busca oferecer funcionalidades semelhantes em termos de automação e organização de transcrições.

O *Transkribus*® tornou-se um marco ao possibilitar que pesquisadores e instituições transcrevam documentos históricos de forma automática, utilizando modelos avançados de HTR. A popularidade dessa plataforma ressalta a relevância de ferramentas que integrem tecnologias de reconhecimento de manuscritos com interfaces amigáveis e sistemas de treinamento dos modelos.

Para o desenvolvimento de modelos de HTR, diversos *frameworks* têm sido utilizados com sucesso, cada um oferecendo características distintas:

- **Kraken** [Kiessling et al. 2019]: Um *framework* de código aberto amplamente utilizado para reconhecimento óptico de caracteres - *Optical Character Recognition* (OCR) - e HTR, reconhecido por sua flexibilidade e capacidade de lidar com uma ampla variedade de manuscritos. É especialmente popular na digitalização de documentos históricos em diferentes idiomas.
- **Kaldi** [Arora et al. 2019]: Embora originalmente desenvolvido para reconhecimento automático de fala - *Automatic Speech Recognition* (ASR) -, o *Kaldi* também pode ser adaptado para tarefas de HTR, oferecendo recursos avançados para treinamento e decodificação de modelos.
- **HTR+** [Michael et al. 2018]: *Framework* proprietário desenvolvido pela empresa responsável pelo *Transkribus*®. Amplamente utilizado na plataforma, o *HTR+* oferece um desempenho robusto, mas é restrito aos usuários desta plataforma.

- **Pylaia** [Puigcerver and Mocholí 2018]: Um *framework* de código aberto projetado especificamente para tarefas de HTR. Reconhecido por sua versatilidade, o *Pylaia* permite o treinamento e a aplicação de modelos para reconhecimento de manuscritos com alto grau de personalização, sendo amplamente adotado por pesquisadores e desenvolvedores interessados em soluções abertas e flexíveis.

Estudos como “A Comprehensive Comparison of Open-Source Libraries for Handwritten Text Recognition in Norwegian” [Maarand et al. 2022] e “Transkribus and Beyond: Pioneering the Future of Transcription Technology” [Leifert et al. 2024] mostram que o *Pylaia* apresenta desempenho competitivo em comparação com outros *frameworks* de HTR, incluindo o *HTR+*, que é proprietário. Essas análises destacam que, em determinados cenários, o *Pylaia* alcança níveis de precisão comparáveis aos do *HTR+*, reforçando sua relevância como uma ferramenta eficiente e confiável para tarefas de reconhecimento de manuscritos.

Em setembro de 2020, o *Transkribus*® introduziu a possibilidade de seus usuários manipularem modelos de HTR utilizando o *Pylaia*, reforçando sua posição como uma opção de destaque [Readcoop 2020]. Essa integração evidenciou o *Pylaia* não apenas como um recurso viável para o *Transkribus*®, mas também como uma ferramenta fundamental no desenvolvimento do Transcritor-IA. Por ser um *framework* de código aberto, amplamente reconhecido por sua adaptabilidade e eficiência, o *Pylaia* foi selecionado como a base para o trabalho deste artigo.

Sua escolha se justifica por diversos fatores, entre eles seu desempenho sólido, muitas vezes equiparável a *frameworks* proprietários como o *HTR+*. Além disso, sua flexibilidade permite o treinamento em diferentes conjuntos de dados, tornando-o altamente adaptável a distintos tipos de manuscritos. A natureza de código aberto do *Pylaia* também favorece a colaboração, manutenibilidade e inovação dentro da comunidade, ao passo que sua crescente adoção reflete seu potencial e o suporte contínuo de desenvolvedores e pesquisadores. Com base nesses aspectos, o Transcritor-IA incorpora o *Pylaia* como núcleo de seu sistema de reconhecimento e treinamento de modelos, garantindo uma abordagem eficiente e alinhada com as melhores práticas no campo de HTR.

3. Conjunto de Dados

Este trabalho baseia-se em um livro eclesiástico manuscrito de 1890 do município de Crato (CE), que exemplifica os desafios típicos da transcrição de documentos históricos, como caligrafia difícil e linguagem arcaica. As entradas, dispostas em texto corrido sobre papel envelhecido e manuscritas com tinta, foram digitalizadas, apresentando manchas e desbotamentos característicos.

A segmentação automática das linhas é feita com a ferramenta Transcritor-IA, que emprega técnicas do *Open Computer Vision Library* (OpenCV) — como binarização, detecção de contornos e ajustes finos — para identificar e isolar linhas horizontais. Após essa etapa, aplica-se um refinamento de **limpeza de margens**, que remove ruídos e traços de linhas adjacentes por meio de operações morfológicas e análise de *pixels*. Em seguida, o **pré-processamento** realça contornos e normaliza o contraste, aumentando a legibilidade e a eficácia do modelo de transcrição. Esse fluxo visa preservar apenas

os caracteres da linha-alvo, reduzindo interferências e melhorando a qualidade da transcrição. Por fim, uma etapa opcional de correção manual pode ser usada para ajustar inconsistências, conforme ilustrado na Figura 1.



Figura 1. Exemplo de segmentação e transcrição de linha.

O algoritmo de divisão dos conjuntos de treino e validação foi projetado de forma que todas as palavras presentes no conjunto de validação também estejam no conjunto de treino. No entanto, cada imagem correspondente a uma ocorrência de palavra é única em cada conjunto — ou seja, não há imagens repetidas entre treino e validação. Isso garante que o modelo seja avaliado de forma consistente quanto à sua capacidade de generalização, ao mesmo tempo em que evita sobreposição direta de dados. As principais características do conjunto de dados estão relacionadas na Tabela 1. Este conjunto de transcrições certificadas está disponível no repositório *Hugging Face*¹ e na plataforma Transcritor-IA², onde pode ser acessado pelo livro *Eclesiástico 1890* (local: Crato, período: 1890–1900, escrevente: Antônio Alexandrino de Alencar) na seção “Coleções Públicas”.

Tabela 1. Sumarização do conjunto de dados

Característica	Quantidade
Número total de páginas	21
Número total de linhas	1759
Linhas para treinamento	1406 (80%)
Linhas para validação	353 (20%)
Número total de palavras (aprox.)	721

¹<https://huggingface.co/datasets/rodrig-crzz/transcritor-ia-eclesiastico-1890-htr>

²<https://transcritor-ia.com>

4. Metodologia

A plataforma Transcritor-IA permite gerar, validar e analisar modelos de HTR, utilizando o *Pylaia*. As etapas de treinamento, validação, transcrição manual e automática, bem como a análise de desempenho dos modelos estão permeadas nas funcionalidades da plataforma. A transcrição de manuscritos no Transcritor-IA, semelhante à plataforma *Transkribus*[©], ocorre em quatro etapas principais:

1. **Carregamento do livro na plataforma:** As imagens do livro são carregadas partindo de pastas contendo imagens de manuscritos.
2. **Transcrição Inicial:** As linhas de texto são processadas automaticamente pelo modelo de HTR pré-treinado com transcrições semelhantes, e a transcrição gerada é apresentada ao usuário como uma proposta inicial.
3. **Certificação Manual:** O usuário verifica a transcrição apresentada e realiza as correções necessárias. Linhas corrigidas são certificadas na plataforma, marcando-as como averiguadas e finalizadas.
4. **Treinamento/Retranscrição:** Após um certo número de linhas classificadas pelo usuário como certificadas, o modelo é retreinado e o manuscrito é retranscrito para nova análise.

Para avaliar a eficácia da plataforma, comparam-se os resultados de transcrição obtidos pelo Transcritor-IA e pelo *Transkribus*[©]. Para isso, o treinamento foi realizado com o conjunto de dados descrito na Seção 3, ajustando a divisão dos dados de modo que uma página seja reservada para validação dos modelos de HTR, enquanto as demais são utilizadas para o treinamento. Essa abordagem visa replicar as condições observadas na plataforma *Transkribus*[©].

Vale ressaltar que, na página reservada para validação, realiza-se uma análise prévia para selecionar as linhas transcritas que comporão o conjunto de validação. Apenas as linhas formadas exclusivamente por palavras já presentes no conjunto de treinamento são consideradas, garantindo que não existam palavras no conjunto de validação que não ocorram no conjunto de treinamento.

A validação foi realizada com base em métricas amplamente utilizadas no campo de HTR, como a Taxa de Erro por Caracter - *Character Error Rate* (CER) - e a Taxa de Erro por Palavra - *Word Error Rate* (WER). O CER quantifica a proporção de caracteres inseridos, deletados ou substituídos em relação ao total de caracteres da transcrição de referência. Por sua vez, o WER aplica essa mesma abordagem ao nível de palavras, avaliando a precisão da transcrição com base em palavras inteiras. Ambas as métricas são indicativas do desempenho do modelo, sendo que valores mais baixos refletem maior precisão. As fórmulas utilizadas para o cálculo dessas métricas são apresentadas a seguir:

$$\text{CER} = \frac{S_c + D_c + I_c}{\text{Total de Caracteres de Referência}}$$

- S_c : Substituições de caracteres;
- D_c : Deleções de caracteres;
- I_c : Inserções de caracteres;

$$\text{WER} = \frac{S_p + D_p + I_p}{\text{Total de Palavras de Referência}}$$

- S_p : Substituições de palavras;
- D_p : Deleções de palavras;
- I_p : Inserções de palavras.

5. Materiais e Métodos

Os experimentos realizados visam comparar a plataforma Transcritor-IA com a plataforma *Transkribus*[®]. O processo de utilização de ambas é semelhante. Considerando a plataforma Transcritor-IA, os usuários realizam *upload* de imagens correspondentes às páginas de documentos manuscritos, que são armazenadas em uma base de dados estruturada. Esse processo envolve diversas etapas automáticas. Inicialmente, ocorre a recepção e tratamento dos arquivos, em que documentos no formato *Portable Document Format* (PDF) são separados por página e submetidos a algoritmos de OCR para análise e correção da orientação da imagem, garantindo sua correta inclusão na base de dados. Em seguida, o sistema realiza a detecção automática das informações manuscritas, identificando regiões e linhas de texto relevantes dentro das páginas. Após essa etapa, as linhas detectadas são segmentadas em termos de altura (128 *pixels*) e passam por um pré-processamento que realça a qualidade visual e aprimora o contraste, facilitando a extração das características do texto. Com as imagens devidamente segmentadas, o modelo de HTR, desenvolvido com o *framework* PyLaia, realiza a transcrição automática, que são então apresentadas ao usuário para verificação e certificação.

A arquitetura do modelo de inteligência artificial utilizado pela plataforma Transcritor-IA está esquematizada na Figura 2. O modelo segue uma abordagem baseada em redes neurais profundas para realizar a transcrição automática de manuscritos históricos. Ele é estruturado em diferentes camadas, cada uma com uma função específica no processamento das imagens e na geração das transcrições.

O primeiro componente da arquitetura é a Entrada, que recebe as imagens contendo textos manuscritos. Cada imagem é processada em escala de cinza, possuindo um único canal e uma altura fixa de 128 *pixels*. Essa padronização facilita o treinamento e a inferência do modelo, garantindo uma entrada consistente para as camadas subsequentes.

A segunda parte é composta pelos Blocos Convolucionais, responsáveis pela extração de características da imagem. Essa etapa é formada por quatro blocos convolucionais, cada um contendo parâmetros específicos como a função de ativação *LeakyReLU*, diferentes tamanhos de *pooling*, número de filtros (neurônios por bloco) e normalização em lote (*BatchNorm*). Esses blocos reduzem a dimensionalidade da entrada e extraem padrões essenciais para o reconhecimento dos caracteres.

Após a extração de características, as informações são passadas para as Camadas Recorrentes. Essa etapa é composta por três camadas *Long Short-Term Memory* (LSTM) idênticas, cada uma com 256 células e um fator de *dropout* de 0.5 para evitar *overfitting*. As redes recorrentes capturam dependências temporais entre os caracteres e ajudam a modelar a sequência de escrita presente nos manuscritos.

A saída das camadas recorrentes é então processada por uma Camada Linear, que classifica os possíveis caracteres presentes na imagem e gera uma lista de símbolos. Essa representação é chamada de Transcrição Intermediária, pois ainda não passou por um refinamento linguístico. Essa transcrição intermediária foi utilizada para calcular a métrica CER no conjunto de validação e na comparação com o *Transkribus*[®], visto que o treinamento dos modelos de HTR não inclui a aplicação de um modelo de linguagem para melhorar a transcrição.

Para aprimorar a transcrição gerada pela rede neural, utilizamos uma camada de modelagem de linguagem - *Language Modeling Layer* (LML). Essa etapa aplica um modelo probabilístico baseado em *n-grams*, utilizando o modelo *ARPA Language Model* (ARPALM) treinado previamente. Esse modelo auxilia na correção de erros ortográficos e na reconstrução de palavras incompletas, refinando a transcrição intermediária antes de apresentá-la ao usuário.

Por fim, a Transcrição Final é gerada e disponibilizada ao usuário da plataforma Transcritor-IA. Essa transcrição representa o resultado final do processamento, já aprimorado pelo modelo de linguagem, tornando-se a versão pronta para análise e certificação.

Toda essa arquitetura foi configurada utilizando o *PyLaia*, que permite ajustes nos parâmetros das camadas convolucionais, recorrentes e lineares. O ARPALM também é compatível com o *PyLaia*, mas exige a criação prévia do modelo probabilístico através de ferramentas externas, como o *SRI Language Modeling Toolkit* (SRILM). Dessa forma, a plataforma Transcritor-IA combina o poder de redes neurais profundas com modelos probabilísticos para garantir um processo de transcrição robusto e eficiente.

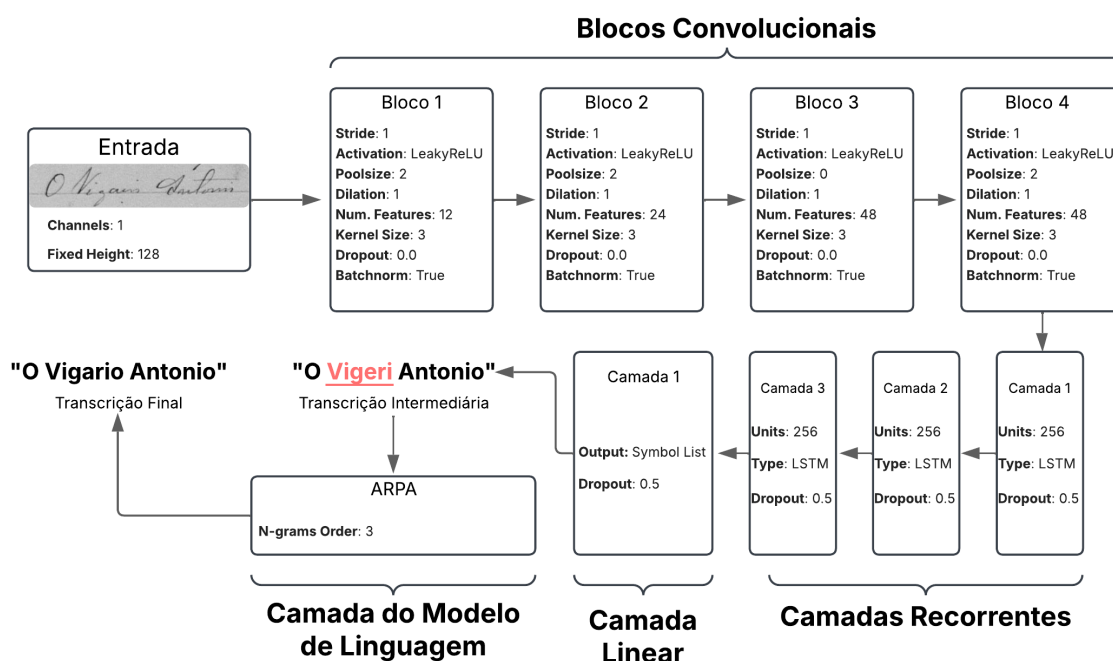


Figura 2. Arquitetura completa do modelo de transcrição utilizado no Transcritor-IA.

O treinamento do modelo de HTR segue um processo recorrente com base nas transcrições progressivamente certificadas pelos usuários dentro da plataforma. Para isso, as linhas transcritas são organizadas em dois subconjuntos: um destinado ao treinamento, utilizado para ajustar os parâmetros do modelo, e outro para validação, garantindo que todas as palavras presentes no conjunto completo apareçam no subconjunto de treino. Além disso, a plataforma implementa um treinamento recorrente, onde o modelo é atualizado automaticamente sempre que novas transcrições são certificadas. Após cada iteração de treinamento, o modelo retranscreve todo o livro, exceto as linhas já

certificadas. Para avaliar o desempenho do modelo, as métricas CER e WER são calculadas continuamente, permitindo a seleção da versão mais eficiente para futuras transcrições.

As linhas transcritas inicialmente pelo modelo são armazenadas como "não certificadas". O usuário tem a opção de:

- Verificar a exatidão da transcrição realizada pelo modelo atual;
- Corrigir eventuais erros;
- Classificar a linha segmentada como certificada.

Essas ações viabilizam o treinamento contínuo do modelo. A cada nova página transcrita — contendo todas as linhas validadas —, o re-treinamento é acionado automaticamente. Em média, cada página do conjunto contém **84 linhas transcritas**. Como o *Transkribus*® realiza a transcrição em nível de página, é necessário escolher uma página inteira para validação. Para isso, seleciona-se a página cuja **maior proporção de palavras transcritas já aparece nas demais páginas do conjunto de treinamento**. Isso garante que a validação ocorra com vocabulário conhecido, permitindo uma avaliação mais fiel da capacidade do modelo em reconhecer variações na escrita, sem introduzir novas palavras no processo.

O livro foi carregado em ambas as plataformas, página por página, com a certificação das linhas. A cada nova página certificada, o modelo foi treinado por **100 épocas**, seguido do cálculo do CER. Esse processo foi repetido para **16 das 21 páginas** do conjunto de dados, uma vez que essa quantidade já é suficiente para observar e comparar o desempenho das duas plataformas. Os valores de CER foram então tabulados para análise posterior.

6. Resultados e Discussão

Os resultados obtidos partindo do processo descrito na seção anterior são apresentados na Figura 3 e comparam o desempenho da ferramenta Transcritor-IA com a plataforma *Transkribus*®. A análise das taxas de erro indica que o Transcritor-IA atingiu um erro final de 3,4%, enquanto a *Transkribus*® obteve um valor ligeiramente superior, de 4,0%. Esse resultado demonstra que a Transcritor-IA não apenas se posiciona como uma solução competitiva, mas também apresenta um desempenho superior na transcrição automática de manuscritos dentro do conjunto testado.

Além disso, observa-se a presença de picos em ambas as curvas ao longo do treinamento. No caso do Transcritor-IA, um aumento da taxa de erro ocorre na etapa de 8 páginas, enquanto no *Transkribus*® um pico semelhante é identificado por volta da etapa de 9 páginas. Esse fenômeno pode ser explicado por diversos fatores. Um dos principais é a natureza dos métodos utilizados para manipulação e treinamento dos modelos de HTR, que frequentemente introduzem elementos de aleatoriedade nas etapas de aprendizado. A adição progressiva de novas páginas ao treinamento implicará a introdução de vocabulário inédito, o que pode, momentaneamente, impactar a taxa de erro até que o modelo se adapte a esses novos padrões.

Mesmo com essas variações, os resultados mostram que a Transcritor-IA mantém um desempenho estável ao longo do experimento, alcançando taxas de erro inferiores às do *Transkribus*® na maior parte do intervalo analisado. Isso evidencia o potencial

da abordagem utilizada na plataforma Transcritor-IA e sugere que novos ajustes e otimizações podem reduzir ainda mais a taxa de erro dos modelos.

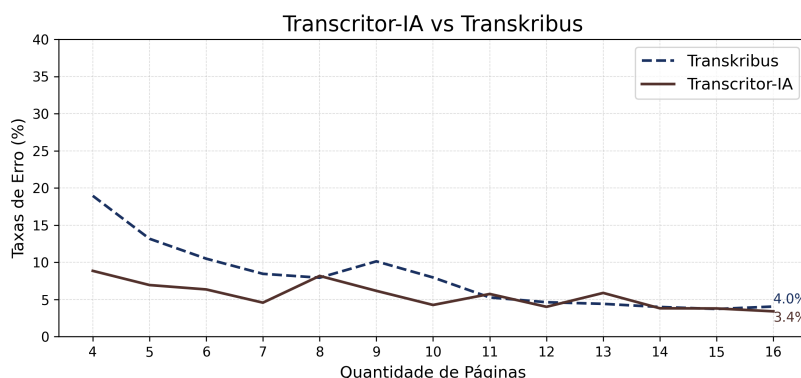


Figura 3. Gráfico da taxa de erro de carácter por página transcrita (conjunto de validação) utilizada no treinamento, comparativo entre as plataformas Transcritor-IA e Transkribus®.

Avaliando exclusivamente a plataforma Transcritor-IA em termos de CER e WER, obteve-se o gráfico apresentado na Figura 4. Esse resultado demonstra que, com o aumento do número de páginas transcritas, ambas taxas de erro diminuem progressivamente. Inicialmente, o WER supera 25%, mas estabiliza em torno de 11,7%, enquanto o CER reduz-se para 3,4%. Essa melhora indica que o modelo aprimora sua precisão à medida que recebe mais dados. Pequenas oscilações nas taxas podem ser atribuídas à variação na qualidade das imagens e na complexidade dos textos, mas a tendência geral confirma a evolução positiva do sistema.

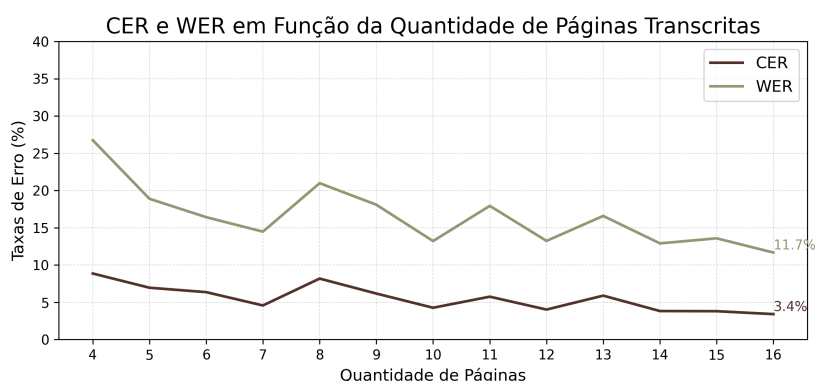


Figura 4. Gráfico da evolução das taxas de erro de carácter e palavra por página transcrita (conjunto de validação) no Transcritor-IA.

Ao avaliarmos os resultados obtidos na transcrição do conjunto de validação, as transcrições podem ser vistas na Tabela 2. Observa-se que, para a transcrição de referência 1, o modelo *Transkribus*® consegue capturar corretamente as construções com "rr" e "ll", como em "Ferreira" e "Mello". Em contrapartida, a plataforma Transcritor-IA transcreve corretamente "d'Oliveira", enquanto o modelo *Transkribus*® resulta em "do Oliveira".

Para a transcrição de referência 2, a plataforma Transcritor-IA apresenta um desempenho mais preciso já o *Transkribus*[®], no entanto, introduz erros significativos, como "doezecete" e "freguedo", demonstrando uma taxa de erro superior.

Na transcrição de referência 3, ambos os modelos obtêm uma transcrição exata, sem erros, com um CER de 0.0%. Observa-se que essa linha específica — 'O Vigário Antonio Alexandrino de Alencar' — ocorre com alta frequência no conjunto utilizado. Esse fator pode ter contribuído significativamente para que os modelos aprendessem essa estrutura com maior precisão, resultando em uma previsão exata. Esse caso evidencia que, em certas condições, os dois sistemas podem atingir níveis máximos de acurácia.

Já na transcrição de referência 4, ambas as plataformas apresentam um erro sutil, porém distinto. O modelo Transcritor-IA transcreve "Felicina" em vez de "Feliciana", enquanto o modelo *Transkribus*[®] reduz a palavra "Con" para "Co". Ambos os modelos obtêm um CER de 1.2%, indicando desempenho equivalente.

Por fim, na transcrição de referência 5, observa-se um desempenho inferior do modelo *Transkribus*[®], que introduz diversas modificações na grafia original, como "Juacinho" e "dos quais". Com isso, o CER atinge 8.7%. A plataforma Transcritor-IA também comete erros, como "Aelia Maria Dis" em vez de "Amelia Maria Dias", porém mantém um CER menor, de 5.8%.

A análise dos dados realizada na Tabela 2 revela que a plataforma Transcritor-IA apresentou um desempenho geral mais consistente em comparação com o *Transkribus*[®] ao processar linhas selecionadas aleatoriamente do conjunto de validação. Embora a diferença entre as duas seja reduzida, o Transcritor-IA registrou uma taxa de erro levemente inferior (3,4%) em relação ao *Transkribus*[®] (4,0%) para o conjunto de dados utilizado neste trabalho. Os resultados indicam, portanto, que ambas as ferramentas possuem eficiência semelhante.

Tabela 2. Comparação entre as transcrições da plataforma *Transcritor-IA* e *Transkribus* tomando como referência o conjunto de validação: erro de carácter no nível de linha.

Origem	Transcrição	CER
Linha de Validação 1	cisco Ferreira de Mello, e Maria Isabel d'Oliveira Mello, casados, os quais conheço	–
Transcritor-IA	cico Fereira de Melo, e Maria Isabel d'Oliveira Melo, casados, os quais conheço	4.8%
Transkribus	cisco Ferreira de Mello, e Maria Isabel do Oliveira Mello, casados, os quais conheço	2.4%
Linha de Validação 2	Aos dezecete de Junho de mil oitocentos noventa cinco, na capela do Juaseiro desta freguesia	–
Transcritor-IA	Aos dezecete de Junho de mil oitocentos noventa cinco, na capela do Juaseiro desta freguesi	1.1%
Transkribus	Aos doezecete de Junho de mil oitocentos noventa e cinco, na capela do Juaseiro desta freguedo	6.5%
Linha de Validação 3	O Vigário Antonio Alexandrino de Alencar	–
Transcritor-IA	O Vigário Antonio Alexandrino de Alencar	0.0%
Transkribus	O Vigário Antonio Alexandrino de Alencar	0.0%
Linha de Validação 4	de Abril do mesmo anno, filho legitimo de Joaquim Gomes da Silva e Feliciana da Con	–
Transcritor-IA	de Abril do mesmo anno, filho legitimo de Joaquim Gomes da Silva e Felicina da Con	1.2%
Transkribus	de Abril do mesmo anno, filho legitimo de Joaquim Gomes da Silva e Feliciana da Co	1.2%
Linha de Validação 5	Jacinto da Rocha Sobreira, e Amelia Maria Dias, casados, os quais co	–
Transcritor-IA	Jacinto da Rocha Sobreira e Aelia Maria Dis, casados, os quais co	5.8%
Transkribus	Juacinho da Rocha Sobreira, e Aelia Maria Dis, casados dos quais co	8.7%

7. Conclusão e Trabalhos Futuros

Os resultados obtidos indicam que ambas as plataformas apresentam desempenho semelhante na transcrição de documentos históricos do conjunto de dados avaliado. É

importante destacar que o modelo já emprega técnicas de pré-processamento de imagens, mas ainda há espaço para explorar abordagens mais avançadas com o intuito de torná-lo mais eficaz e potencialmente melhorar as métricas de transcrição.

Para reduzir a necessidade de transcrições e certificações manuais, técnicas como *data augmentation* e modelos gerativos de Difusão ou *Autoencoders* Variacionais podem ser explorados para aumentar a diversidade do conjunto de dados e aprimorar o treinamento do modelo. Além disso, a incorporação de Modelos de Linguagem Natural – *Natural Language Processing* (NLP), como *Bidirecional Encoder Representations from Transformers* (BERT) e *Generative pre-trained transformer* (GPT), pode refinar as transcrições, enquanto o uso de *transfer learning* com modelos pré-treinados facilitaria a adaptação do sistema a diferentes estilos de escrita.

Atualmente, a plataforma Transcritor-IA se restringe à transcrição de manuscritos. No entanto, sua ampliação para documentos datilografados exigiria aprimoramentos na detecção e segmentação de texto. Apesar dos desafios, os resultados demonstram que a plataforma possui potencial para evoluir, aumentando tanto a precisão quanto a aplicabilidade na transcrição de documentos históricos.

Agradecimentos

Agradecemos a Tânia Barroso, Marcos Barroso, Clarisier Azevedo, Carla Ribeiro, Herbert Duarte e Joatan Viana pelo apoio e colaboração ao longo deste trabalho. Estendemos nosso reconhecimento à Prefeitura de Cedro, na pessoa do prefeito Nilson Diniz, pelo suporte institucional, e ao Instituto Federal do Ceará (IFCE), na pessoa do reitor Wally Menezes, pelo incentivo e pela infraestrutura disponibilizados.

Referências

- Arora, A., Chang, C. C., Rekabdar, B., BabaAli, B., Povey, D., Etter, D., Raj, D., Hadian, H., Trmal, J., Garcia, P., Watanabe, S., Manohar, V., Shao, Y., and Khudanpur, S. (2019). Using asr methods for ocr. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 663–668.
- Humphries, M., Leddy, L. C., Downton, Q., Legace, M., McConnell, J., Murray, I., and Spence, E. (2024). Unlocking the archives: Large language models achieve state-of-the-art performance on the transcription of handwritten historical documents. *arXiv preprint arXiv:2411.03340*, 1.
- Kahle, P., Colutto, S., Hackl, G., and Mühlberger, G. (2017). Transkribus - a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 04, pages 19–24.
- Kiessling, B., Tissot, R., Stokes, P., and Stökl Ben Ezra, D. (2019). escriptorium: An open source platform for historical document analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 19–19.
- Lakshmi, D. V. (2024). Evolution of machine learning algorithms a comprehensive review. *AG Volumes*, pages 70–80.

- Leifert, G., Romein, C., Rabus, A., Ströbel, P. B., and Hodel, T. (2024). Transkribus and beyond: Pioneering the future of transcription technology. In *Proceedings of the Transkribus User Conference '24*, Innsbruck, Austria. Transkribus.
- Maarand, M., Beyer, Y., Kåsen, A., Fosseide, K. T., and Kermorvant, C. (2022). A comprehensive comparison of open-source libraries for handwritten text recognition in norwegian. In Uchida, S., Barney, E., and Eglin, V., editors, *Document Analysis Systems*, pages 399–413, Cham. Springer International Publishing.
- Mendes, S. T. d. P. and Oliveira, C. B. d. (2016). A transcrição de um manuscrito eclesiástico setecentista para a pesquisa na Área de linguística histórica. *e-hum Revista Científica das áreas de História, Letras, Educação e Serviço Social do Centro Universitário de Belo Horizonte*, 9(1):18–25.
- Michael, J., Weidemann, M., and Labahn, R. (2018). Htr engine based on nns p 3 optimizing speed and performance-htr+. Technical report, Technical report, READ-H2020 Project 674943.
- Moreira, A., de Paiva Oliveira, A., Mendes, F. F., de Queiroz, J. M., and Braga, V. (2007). Digitalização de manuscritos históricos: a experiência da casa setecentista de mariana. *Ci. Inf. (Brasília)*, 36(3):89–98.
- Nacional, B. (2022). Política de preservação digital da biblioteca nacional. Acessado em 19-05-2025.
- Puigcerver, J. and Mocholí, C. (2018). Pylaia. <https://github.com/jpuigcerver/PyLaia>. Acessado em 19-05-2025. Commit específico: 941.
- Rakesh, S., Reddy, P. K., Prashanth, V., and Reddy, K. S. (2024). Handwritten text recognition using deep learning techniques: a survey. *MATEC Web of Conferences*, 392.
- Readcoop (2020). + try out transkribus new recognition software pylaia! <https://readcoop.eu/try-out-transkribus-new-recognition-software-pylaia>. Acessado em 19-05-2025.
- Twinkle Sharma, P. and Arora, R. (2024). The evolution of artificial intelligence - a comprehensive review. *International Journal of Science, Engineering and Technology*, 12(3):1–7.