

O modelo de *OCR Early Portuguese Printing* no *Transkribus*: sobre as inovações na transcrição de impressos históricos e os desafios da infraestrutura de pesquisa em linguística histórica e paleografia digital no Brasil

The Early Portuguese Printing OCR Model in *Transkribus*: on innovations in the transcription of historical printed documents and the research infrastructure for historical linguistics and digital paleography in Brazil

Saulo Rogério Pacheco Rocha ^{*1}

¹Universidade Federal de Santa Catarina, Centro de Comunicação e Expressão, Programa de Pós-Graduação em Linguística, Florianópolis, SC, Brasil.

Resumo

Este trabalho apresenta o modelo de Reconhecimento Óptico de Caracteres (OCR) *Early Portuguese Printing* (EPP), desenvolvido na plataforma *Transkribus*, e discute o potencial, os desafios e a história dessa ferramenta para a pesquisa com documentos históricos. O *Transkribus*, mantido pela cooperativa europeia READ-COOP, permite que pesquisadores treinem modelos de IA especializados nas características de seus próprios *corpora*. O modelo EPP foi treinado especificamente para a transcrição de impressos em língua portuguesa dos séculos 16 ao 19, utilizando um corpus de gramáticas e obras linguísticas do período. Com um *training set* de 142.606 palavras (745 páginas), o EPP alcançou uma Taxa de Erro de Caracteres (CER) de apenas 2,58%. Este resultado representa um avanço significativo, pois demonstra a potencialidade de ferramentas do tipo para a formação de *corpora* quantitativos históricos em maior escala e em menos tempo, mantendo a precisão da transcrição de diacríticos, símbolos tipográficos e caracteres gregos, elementos que frequentemente limitam a eficácia de ferramentas de OCR generalistas. Ademais, o EPP está disponibilizado como um modelo de base (*base model*) que pode ser adaptado por outros pesquisadores para diferentes acervos impressos, mediante *fine-tuning*. Contudo, além de divulgar o potencial da ferramenta, este trabalho problematiza sua natureza. Por pertencer a uma entidade privada europeia e ser um produto *Software as a Service* (SaaS), o uso do *Transkribus* levanta questões sobre a centralização de dados e a sustentabilidade de sua aplicação em projetos de pesquisa brasileiros de grande escala, considerando o futuro e o volume de nossos acervos históricos.

Palavras-chave: Humanidades digitais. Linguística histórica. Paleografia digital. Soberania digital. Inteligência artificial.

Abstract

This paper presents the *Early Portuguese Printing* (EPP) Optical Character Recognition (OCR) model, developed on the *Transkribus* platform, and discusses the potential, the challenges and the history of such tools for research with historical documents. *Transkribus*, maintained by the European cooperative READ-COOP, allows researchers to train specialized AI models on the characteristics of their own corpora. The EPP model was specifically trained for the transcription of printed materials in the Portuguese language from the 16th to the 19th centuries, using a corpus of grammars and linguistic works from the period. With a training set of 142,606 words (745 pages), the EPP achieved a Character Error Rate (CER) of only 2.58%. This result represents a significant advancement, as it demonstrates the potential of such tools for creating historical quantitative corpora on large-scale and in less time, while maintaining accuracy in the transcription of diacritics, typographical symbols and Greek characters, elements that often limit the effectiveness of general-purpose OCR tools. Furthermore, the EPP is available as a base model that can be adapted by other researchers for different printed collections, through fine-tuning. However, in addition to publicizing the tool's potential, this paper also problematizes its nature. As it belongs to a private European entity and it is a Software as a Service product, the use of *Transkribus* raises questions about data centralization and the sustainability of its

Textolivre
Linguagem e Tecnologia

DOI: 10.1590/1983-
-3652.2026.64127

Seção:
Artigos

Autor Correspondente:
Saulo Rogério Pacheco Rocha

Editor de seção:
Daniervelin Pereira
Editor de layout:
Rayany Chaves

Recebido em:
29 de janeiro de 2026
Aceito em:
18 de março de 2026
Publicado em:
30 de setembro de 2026

Esta obra tem a licença
"CC BY 4.0".



*Email: saulorrp@gmail.com

application on large-scale Brazilian research projects, considering the future and the volume of our historical archives.

Keywords: Digital humanities. Historical linguistics. Digital paleography. Digital sovereignty. Artificial intelligence.

1 Introdução

O presente trabalho é um desdobramento metodológico de uma dissertação de mestrado que analisou a mudança nos padrões de colocação pronominal clítica em gramáticas portuguesas dos séculos 15 a 19, intitulada “O Português das ‘Obras Proueitasas’ e dos ‘Ilustres & Muito Reuerendos Senhores’: uma análise da mudança nos padrões de colocação pronominal clítica em gramáticas portuguesas do século 15 ao 19”. O objetivo dessa pesquisa era compreender como as transformações da língua se manifestavam em um gênero textual de natureza eminentemente conservadora e prescritiva. Para viabilizar essa análise, foi necessária a construção de um corpus linguístico específico. Embora as obras selecionadas estivessem publicamente disponíveis em acervos digitais, elas se encontravam em formato de fac-símile (imagem digital). Esse formato, por não ser textualmente pesquisável, impede a manipulação de dados em larga escala que as pesquisas sob a ótica da linguística de corpus exigem. Portanto, a etapa fundamental da metodologia consistiu na transcrição e edição integral dessas obras, a fim de constituir um corpus funcional para a análise.

A compilação documental desta pesquisa se fundamenta nos preceitos da Linguística de Corpus (LC), que demanda coleções textuais estruturadas e representativas (cf. McEnery e Hardie (2012)). Como a transcrição manual de fac-símiles históricos em larga escala é materialmente complexa, a inteligência artificial (IA) foi adotada para superar esse gargalo metodológico. Por meio de aprendizado de máquina supervisionado e aprendizado por transferência (Wang *et al.*, 2021), o treinamento do modelo *Early Portuguese Printing* (EPP) atuou como a ponte necessária para suprir a exigência da LC por dados quantitativos volumosos e confiáveis. Como resultado desse processo e em compromisso com a ciência aberta, o corpus integral transcrito foi disponibilizado no repositório *Zenodo*, garantindo a reprodutibilidade da pesquisa.

Para realizar essa tarefa, foi necessária a adoção de uma ferramenta de OCR. Dentre as opções avaliadas, como o *Tesseract OCR*, produtos como *Abby FineReader PDF* e alternativas de software livre, como o *eScriptorium*¹, o *Transkribus* foi a plataforma escolhida. As características da escrita histórica portuguesa, especialmente em um corpus diacrônico repleto de variações ortográficas, diacríticos e caracteres únicos, exigiam uma solução de transcrição automatizada que fosse personalizável. O *Transkribus* atende a essa necessidade sem precisar da operação técnica que o processamento local exigiria, além de permitir a criação e o treinamento de modelos de ATR e a edição dos documentos transcritos dentro de uma interface própria com acesso online aos dados. A adoção desse recurso de treinamento próprio tornou-se imperativa ao constataremos uma grave lacuna na biblioteca pública da plataforma: até o momento da pesquisa, havia apenas cinco modelos disponíveis para a língua portuguesa, sendo a maioria focada em manuscritos e inadequada para as especificidades tipográficas impressas do nosso corpus. O objetivo deste trabalho é, portanto, descrever e divulgar o processo de treinamento do EPP, tornado público numa primeira versão em outubro de 2024 e numa segunda em janeiro de 2025. A relevância dessa discussão reside na lacuna metodológica enfrentada por pesquisadores brasileiros que trabalham com documentação histórica, oferecendo um caminho prático e testado para a superação de desafios comuns de transcrição.

A pesquisa diacrônica brasileira enfrenta atualmente um gargalo metodológico significativo. A necessidade de transcrições manuais impede que pesquisadores individuais, em sua maioria alunos de iniciação científica e pós-graduandos, construam *corpora* volumosos, enquanto, até então, o uso de ferramentas de ATR generalistas compromete a precisão das análises históricas. Esse obstáculo se tornou cada vez menos relevante, uma vez que novas soluções baseadas em IA, como o *Transkribus*, permitiram avanços que antes eram pouco concebíveis. Ainda assim, além de apresentar a inovação

¹ As razões para a sua não adoção neste projeto estão fundamentadas na seção “Considerações Finais”.

técnica e o potencial do *Transkribus* e do modelo EPP, este artigo também busca problematizar a natureza da ferramenta. Embora o serviço ofereça benefícios claros, o fato de pertencer a uma cooperativa privada e centralizar os dados processados levanta questionamentos importantes. A sustentabilidade e a viabilidade de seu uso em iniciativas de grande escala, especialmente considerando o volume da documentação histórica brasileira e a necessidade de soberania digital, são temas que merecem um debate aprofundado.

Com o intuito de abordar essas questões de forma abrangente, o artigo está estruturado em quatro seções, além desta introdução. A primeira apresenta noções básicas sobre ferramentas de ATR, contextualizando o *Transkribus* nesse cenário. A segunda descreve a plataforma *Transkribus*, suas potencialidades e seu status atual. A terceira seção detalha o processo de treinamento do modelo *Early Portuguese Printing*, de modo a auxiliar outros pesquisadores em seus próprios projetos. Por fim, a quarta seção busca apresentar as considerações finais de maneira crítica, levantando os questionamentos necessários para a adoção e o uso consciente dessa e de outras ferramentas semelhantes.

2 Breve histórico das ferramentas de transcrição automatizada

A pesquisa histórica e linguística, em particular aquela que se debruça sobre a análise de documentos, demanda a manipulação de grandes acervos para a construção de corpora representativos das diferentes situações da língua. Tradicionalmente, esse trabalho estava circunscrito a métodos de transcrição manual, um processo que impõe um limite prático e metodológico ao volume de materiais que podem ser investigados num determinado espaço de tempo.

O advento do que se tem chamado de “Humanidades Digitais”, no entanto, introduziu o reconhecimento textual automatizado como uma alternativa capaz de otimizar e escalar o trabalho do pesquisador dessa área. Conforme apontam Baudry e Pedro (2023), esse campo se organiza em torno de duas tecnologias principais: o *Optical Character Recognition* (OCR), para textos impressos, e o *Handwritten Text Recognition* (HTR), para manuscritos, sendo ambos por vezes unificados sob o termo *Automatic Text Recognition* (ATR) ou unidas de forma indistinta, como, “OCR”, apenas. Apesar da significativa evolução na área, impulsionada por arquiteturas de aprendizado profundo (cf. Alkendi *et al.* (2024)), a aplicação de ferramentas de OCR de propósito geral a documentos históricos sempre revelou limitações técnicas, de forma a justificar a necessidade de soluções especializadas, papel que o *Transkribus* atualmente desempenha como nenhuma outra ferramenta na mesma escala.

A história do Reconhecimento Óptico de Caracteres (OCR) é, em essência, a história da luta contra a variabilidade textual. Embora protótipos conceituais existam desde o mosaico de retina de Carey, em 1870, as primeiras aplicações comerciais, que surgiram nas décadas de 1950 e 1960, só foram viáveis ao restringir drasticamente essa variabilidade (Wang, 2023; Narang; Jindal; Kumar, 2021). Lideradas por empresas como a IBM para atender às demandas de processamento de dados em massa (como cheques bancários e os antigos cartões de crédito, aqueles números quadrados em alto relevo), essas máquinas operavam pelo princípio de *pattern Recognition* através de *template matching*. Essa técnica comparava a imagem de um caractere a uma biblioteca de protótipos, exigindo uma correspondência quase perfeita. Para garantir o sucesso, foram criadas fontes específicas, como a OCR-A e a OCR-B, forçando a padronização do texto para se adequar às limitações da máquina (Narang; Jindal; Kumar, 2021, p. 4). A limitação inerente a essa abordagem, chamadas por Wang *et al.* (2021, p. 6) de “hand-crafted features”, é a sua rigidez: qualquer desvio do template predeterminado, mudança de fonte, tamanho, quantidade de tinta ou qualidade de impressão, resultava em falhas. Essa dificuldade fundamental em lidar com a variação, vista nos primórdios da tecnologia, espelha o desafio que ferramentas de OCR de propósito geral enfrentam hoje ao se depararem com a diversidade tipográfica e ortográfica de um corpus de características históricas, arcaicas. Elas falham porque, em sua essência, ainda esperam o grau de uniformidade dos documentos modernos, nas quais são treinadas, algo que documentação histórica raramente tem.

Para superar a rigidez do *template matching*, os avanços em OCR e HTR dependiam de uma nova abordagem. Conforme aponta Wang (2023), essa revolução veio com a incorporação de técnicas de inteligência artificial já no século 21, especificamente o *deep learning*. A arquitetura que se tornou padrão combina duas tecnologias-chave. Primeiramente, as Redes Neurais Convolucionais (CNNs)

abandonam a comparação de pixels e, em vez disso, aprendem a extrair hierarquias de características visuais, de linhas e curvas a formas mais complexas e/ou combináveis (Wang *et al.*, 2021, p. 7). Em seguida, as Redes Neurais Recorrentes (RNNs) analisam a sequência de características extraídas, incorporando o contexto linguístico para desambiguar caracteres visualmente semelhantes. Isso permite que o sistema, por exemplo, reconheça um “a” independentemente de sua fonte tipográfica ou de pequenas variações de impressão, pois, a partir do seu treinamento, ele seria capaz de extrair informações o bastante para chegar a uma categorização mais ou menos correta. Esse trabalho sequencial entre reconhecimento de características (CNN) e análise de sequência (RNN) define o que Narang, Jindal e Kumar (2021) classificam como a “Quarta Geração” de sistemas de OCR. Tais sistemas são capazes de lidar com documentos complexos e manuscritos antigos, como demonstram os autores com o sistema de escrita indiano Devanágari.

Para esses contextos, ferramentas modernas como o *Tesseract OCR*, mesmo que robustas, não foram projetadas para lidar com a ambiguidade visual de fontes históricas. A degradação do material, a variabilidade da impressão e os ruídos, como o espreadimento inevitável da tinta no papel, são um desafio significativo para modelos automatizados e genéricos, mesmo os mais avançados, quando a pesquisa exige alta fidelidade textual. A plataforma *Transkribus* se posiciona, portanto, como uma resposta direta a essa lacuna mercadológica, oferecendo uma solução tecnológica direcionada a um nicho de mercado específico: a comunidade de pesquisa com documentos históricos, que inclui desde universidades até arquivos públicos, museus etc. Ela operacionaliza o potencial dos sistemas de OCR de “Quarta Geração”, permitindo que especialistas criem modelos altamente personalizados. O diferencial reside na capacidade de treinar um modelo para um padrão gráfico específico, um processo cuja qualidade depende diretamente da expertise do filólogo na criação de transcrições manualmente verificadas e confiáveis para o treinamento (*Ground Truth* ou GT). O maior desafio dessa abordagem, no entanto, é a massiva necessidade de dados. Conforme Wang *et al.* (2021, p. 18), a obtenção de resultados de alta qualidade frequentemente exige um volume de treinamento que pode ultrapassar um milhão de palavras, um obstáculo significativo que será discutido em detalhe mais adiante.

2.1 Sobre o *Unicode*

A transcrição de documentos históricos para o meio digital confronta o pesquisador com a necessidade de uma codificação de caracteres consistente. Para um computador, o texto visível é apenas a representação de uma sequência numérica subjacente e a atribuição padronizada desses números a grafemas de séculos passados é crucial para a integridade de qualquer análise computacional. No início da computação, a ausência de um padrão universal resultou na proliferação de diferentes “páginas de código” (*code pages*), sistemas de caracteres locais que eram frequentemente incompatíveis entre si. Como explica Spolsky (2003) em seu texto clássico sobre o tema, “The Absolute Minimum Every Software Developer Absolutely, Positively Must Know About *Unicode* and Character Sets (No Excuses!)”², essa descentralização tornava um documento criado em um sistema frequentemente ilegível em outro, impedindo a agregação e a análise comparativa de textos. Foi para resolver esse caos que, no final da década de 1980, surgiu a iniciativa que levaria à criação do padrão *Unicode*. Formalizado em janeiro de 1991, seu propósito central é substituir a multiplicidade de padrões por um sistema universal, permitindo o uso de computadores em qualquer idioma. Para alcançar tal objetivo e facilitar sua adoção, o padrão se baseia em duas estratégias: primeiro, ele mantém compatibilidade retroativa com o *American Standard Code for Information Interchange* ou ASCII (o padrão mais difundido para a língua inglesa), mas, fundamentalmente, ele estabelece um princípio universal – fornecer um identificador numérico único e inequívoco para cada caractere, independentemente da plataforma, programa ou língua.

Tecnicamente, o *Unicode* atua como o alicerce para a internacionalização de softwares. O consórcio opera atribuindo a cada caractere um valor numérico único, conhecido como “ponto de código” (*code point*), que é expresso na notação hexadecimal (por exemplo, (U+0041) para a letra “A”) (Spolsky,

² Disponível em: SPOLSKY, Joel. The Absolute Minimum Every Software Developer Absolutely, Positively Must Know About *Unicode* and Character Sets (No Excuses!). Joel on Software. San Francisco: 8 oct. 2003.

2003). Segundo o site do próprio *Unicode Consortium*³, a versão 16.0.0 do padrão, lançada em setembro de 2024, já contém 154.998 caracteres, abrangendo quase todos os sistemas de escrita vivos do mundo, embora o espaço total de pontos de código permita uma expansão muito maior. É crucial entender que o Unicode é um repertório de caracteres abstratos, e não uma fonte, que é uma coleção de glifos (representações visuais dos caracteres). A abrangência do padrão é vasta, incluindo não apenas escritas modernas, mas também sistemas históricos, como hieróglifos egípcios, símbolos técnicos como os do Alfabeto Fonético Internacional (IPA), emojis; até tentativas de inclusão de sistemas de escrita ficcionais, como o *tengwar*⁴, foram feitas.

Mais do que um catálogo, o *Unicode* define um ecossistema de regras que garantem o funcionamento correto do texto nos mais diferentes sistemas digitais. Ele estabelece propriedades para cada caractere, como regras para o comportamento bidirecional (essencial para idiomas como o árabe e o hebraico) e padrões de normalização e ordenação (*collation*). Para o pesquisador de humanidades, de maneira geral, essas regras são fundamentais, pois asseguram que a busca, a manipulação e a comparação de dados textuais de fontes históricas sejam realizadas de forma consistente e que sejam confiáveis em qualquer ambiente computacional.

Ao transcrever textos antigos em um ambiente digital, especialmente para treinar modelos de reconhecimento baseados em IA, é fundamental compreender que o trabalho vai além da simples digitação ou escrita. Se trata, também, de um ato de codificação: cada símbolo do documento original é associado a um ponto de código numérico que a máquina pode processar. A interoperabilidade entre diferentes sistemas depende inteiramente do uso de um padrão universal, como o *Unicode*, para interpretar corretamente essa sequência de códigos. Para uma transcrição básica, o conhecimento do alfabeto padrão pode ser suficiente. Contudo, a pesquisa filológica e o treinamento de modelos de alta precisão exigem um refino muito maior. É necessário, por exemplo, diferenciar o “s” curto (“s”, U+0073) do “s” longo (“ſ”, U+017F) e de sua ligatura (“ß”, U+00DF), ou distinguir diferentes vogais modificadas, como o “e com til” (“ë”, U+1EBD) e o “e com cedilha” (“ẽ”, U+0119). Cada uma dessas distinções corresponde a um ponto de código único e impacta diretamente a qualidade das *ground truths* utilizadas para treinar o modelo.

O *Unicode*, no entanto, não é uma solução exaustiva e apresenta limites. Um exemplo notório na história da língua portuguesa é o “q com til” (“q̃”), um caractere extremamente comum em obras antigas que, diferentemente do “ẽ”, não possui um ponto de código pré-composto. A solução do padrão para esses casos é o uso de marcas diacríticas combinatórias: o grafema “q̃” é representado pela união de dois pontos de código, a letra base “q” (U+0071) seguida pelo til combinatório “̃” (U+0303). Essa dualidade entre representações pré-compostas e combinatórias exige do pesquisador a adoção de uma estratégia de normalização consistente⁵ para garantir a integridade dos dados do corpus. Ainda assim, cabe ressaltar que plataformas como o *Transkribus* muito provavelmente mitigam parte desse desafio ao, por padrão, salvar o texto inserido em seu editor em uma forma normalizada consistente.

Ainda sobre as limitações do *Unicode*, é importante destacar que, especialmente no que tange a grafemas e abreviações idiossincráticas de documentos históricos, há uma imensa lacuna no *Unicode Standard*. Para os casos em que um caractere simplesmente não existe no padrão, há a especificação reserva de uma banda de pontos de código conhecidos como Área de Uso Privado (*Private Use Area* ou PUA), como no caso do *tengwar*. Esse é um espaço intencionalmente vago para que projetos específicos, como a Medieval *Unicode* Font Initiative (MUFI⁶), definam seus próprios caracteres e criem fontes capazes de lê-los. O uso da PUA, contudo, implica num dilema fundamental: por

³ Disponível em: <https://unicode.org/versions/Unicode16.0.0/>

⁴ *Tengwar* é um dos sistemas de escrita élficos do universo de Senhor dos Anéis, de J. R. R. Tolkien. O sistema de escrita atualmente é apenas incluso numa *Private Use Area* (PUA), que é uma banda de pontos de códigos reservada para usos privados; para os caracteres estarem legíveis da forma desejada, é necessário instalar uma fonte dedicada à leitura desses pontos de código – caso contrário, o caractere será representado com um glifo padrão para caracteres não reconhecidos, normalmente uma caixa vazia (□) ou um losango com uma interrogação (◆). Desde a proposta de Michael Everson, em 1997, foram feitas tentativas para a inclusão do *tengwar* no padrão *Unicode*, mas, em 2025, o planejamento oficial do consórcio inclui a adoção do *tengwar* na banda (U+016080 a U+0160FF) do *Supplementary Multilingual Plane*. Disponível em: <https://www.unicode.org/roadmaps/smp/smp-16-0-2.html>.

⁵ Como descrito pelo Consórcio em <https://www.unicode.org/versions/Unicode16.0.0/core-spec/chapter-2/#G1708>

⁶ Disponível em: <https://mufi.info>.

definição, esses caracteres não são padronizados e, portanto, quebram a interoperabilidade universal, que é o próprio cerne da filosofia *Unicode*. Um documento que utiliza fontes que façam uso da PUA só pode ser lido corretamente em sistemas que possuam a fonte dedicada à convenção de mapeamento adotada. O pesquisador se vê, então, diante de uma escolha metodológica crucial: optar pela fidelidade máxima da transcrição, ao custo da acessibilidade e portabilidade universal dos dados.

3 O Transkribus

A plataforma *Transkribus* constitui, efetivamente, uma das ferramentas mais avançadas no campo das Humanidades Digitais para a transcrição de documentos históricos atualmente. Desenvolvida originalmente no âmbito do projeto europeu READ (*Recognition and Enrichment of Archival Documents*), financiado com cerca de 8 milhões de euros pela Comissão Europeia entre 2016 e 2019⁷, a plataforma foi concebida, em sua origem, com foco estrito no Reconhecimento de Texto Manuscrito (HTR) de acervos arquivísticos. Contudo, impulsionada pelo trabalho colaborativo de sua comunidade e pelas inovações nas arquiteturas de redes neurais, a ferramenta expandiu significativamente o seu escopo. O *Transkribus* evoluiu para abarcar também o Reconhecimento de Texto Impresso (PTR), processando obras que vão desde os primórdios da tipografia até o século 20, além de desenvolver modelos básicos (*base models*) cada vez mais generalistas e multilíngues.

Sua relevância global se deve, dentre outros motivos, à sua capacidade de processar um grande número de idiomas e sistemas de escrita, desde o latim e línguas europeias até alfabetos não latinos, se apoiando numa extensa biblioteca de modelos públicos pré-treinados e na liberdade de criação de modelos customizados pelos usuários do serviço. A plataforma é hoje gerida pela READ-COOP SCE, uma sociedade cooperativa europeia⁸ sediada na Universidade de Innsbruck, na Áustria. Seu modelo técnico e de negócios é o de Software como Serviço (SaaS). Diferentemente de ferramentas de OCR que operam localmente, o *Transkribus* é uma solução baseada em nuvem, acessada exclusivamente pela internet⁹. Isso implica que todo o processamento dos modelos de IA e o armazenamento dos dados são centralizados nos servidores da cooperativa, permitindo que pesquisadores utilizem a ferramenta, que exige um poder computacional robusto, sem a necessidade de hardware especializado. A monetização da plataforma opera por um sistema de créditos, que funcionam como uma unidade de medida para o processamento utilizado. Usuários do plano gratuito recebem uma cota mensal de créditos, porém suas tarefas são processadas com menor prioridade. Planos pagos, por sua vez, oferecem maior volume de créditos, prioridade na fila de processamento e acesso a recursos avançados, como os “super modelos” pré-treinados e *smartsearch*. Adicionalmente, um modelo de parceria institucional permite que membros da cooperativa, como universidades, arquivos e museus, contratem planos personalizados e de larga escala, “tailored for your needs”.

Conforme Terras *et al.* (2025) demonstra, a READ-COOP SCE não opera sob a lógica corporativa tradicional, o que implica num regime legal circunscrito às necessidades dos membros da cooperativa, que são universidades, fundos arquivísticos, etc. Além disso, o compromisso com o lucro, de forma semelhante às universidades comunitárias brasileiras, é restringido pela obrigação fiscal de reinvestir seus ganhos integralmente na melhoria da sua própria infraestrutura de operação. Esse regime legal também blinda a ferramenta contra a especulação financeira, visto que as ações são mantidas em seu valor nominal, num modelo que busca, segundo os autores, assegurar a governança democrática e a propriedade comunitária entre seus membros. Para evitar a monopolização das decisões, o poder de voto para cargos internos é limitado a um máximo de cinco votos para membros institucionais, independentemente do volume de ações adquiridas, o que mitiga votos para membros institucionais, independentemente do volume de ações adquiridas, em especial os riscos associados à centralização

⁷ Disponível em: <https://cordis.europa.eu/project/id/674943/reporting>.

⁸ Trata-se de uma cooperativa regulamentada em âmbitos específicos da União Europeia, a Sociedade Cooperativa Europeia (SCE), disponível em: <https://aurora-universities.eu/infrastructure/read-coop/>.

⁹ A READ-COOP SCE também oferece o *Transkribus* On-Prem, que permite *selfhost* do processamento e armazenamento dos dados em infraestrutura própria. O produto que rodaria localmente teria acesso à biblioteca de modelos e às capacidades de treinamento de novos modelos do *Transkribus* sem que a cooperativa armazenasse os dados ou o processamento, mas mais informações sobre o valor e a disponibilidade do produto são mantidas como parte da contratação do serviço. Disponível em: <https://transkribus.org/onprem>.

por uma instituição que invista mais e permite a participação ativa da comunidade científica nos rumos tecnológicos da ferramenta (cf. Terras *et al.* (2025, p. 4-5)).

Embora o *Transkribus* domine o cenário atual pela facilidade de uso de sua interface gráfica, o ecossistema de ATR conta com outras plataformas robustas que oferecem abordagens distintas. Ferramentas como o *eScriptorium*, o *transKriptorium* e as soluções da TEKLI (como a plataforma *Arkindex*) representam alternativas excelentes baseadas em software livre. A diferença fundamental reside no modelo de implementação: enquanto o *Transkribus* opera como um SaaS, o que facilita o acesso para pesquisadores sem formação em computação e centraliza a distribuição e o processamento dos modelos, plataformas como o *eScriptorium* e o *transKriptorium* exigem, via de regra, uma infraestrutura de servidor própria, muitas vezes o próprio computador do pesquisador, e conhecimentos em programação para sua instalação e manutenção. Essas ferramentas permitem um controle total sobre o ciclo de processamento e a soberania dos dados, mas impõem uma curva de aprendizado técnica e uma exigência de hardware, o que o modelo centralizado da READ-COOP consegue mitigar em troca de créditos de processamento, que são gratuitamente renováveis para pesquisadores individuais ou projetos de pequena escala.

Ao longo dos últimos anos, a plataforma *Transkribus* consolidou uma base crescente de modelos públicos e privados, voltados a diferentes idiomas, períodos históricos e sistemas tipográficos. A respeito da escala da plataforma, Baudry e Pedro (2023, p. 236) afirmam que, “nos finais de 2024, [contava com] mais de 300 000 utilizadores” e que “já existem milhares de modelos de reconhecimento de texto na plataforma [...], mas, até ao presente, só menos de três centenas foram tornados públicos pelos seus autores”. Esses números revelam uma dinâmica importante a ser levada em consideração no cenário atual das Humanidades Digitais: embora a produção de modelos seja vasta, a grande maioria permanece em uso privado. A esse cenário se soma o fato de que o *Transkribus*, por sua origem institucional em projetos da Europa Central, não teve a língua portuguesa como foco primário de seu desenvolvimento inicial, resultando em uma escassez natural de soluções otimizadas e prontas para o nosso idioma. Essa tendência à não publicação dos modelos, que pode ser ainda mais grave em projetos descentralizados (como o *eScriptorium*), embora compreensível no contexto de projetos individuais ou institucionais, gera uma duplicação de esforços na comunidade e retarda o avanço coletivo, o que cria um novo gargalo para a pesquisa em larga escala. É precisamente neste cenário que a criação e a divulgação de modelos públicos robustos e bem documentados, como o EPP descrito neste trabalho, são importantes, não apenas como divulgação para escrutínio científico de pares, mas como divulgação, contribuição técnica e convite para a construção de uma infraestrutura de pesquisa mais aberta e eficiente para os estudos históricos no Brasil e na lusofonia de maneira geral.

O treinamento dos modelos depende da construção das GT, o conjunto de dados de referência transcritos ou rigorosamente corrigidos por um filólogo. É a qualidade e o volume da GT que permite às redes neurais aprenderem a converter imagens de documentos em texto editável com precisão superior a de ferramentas generalistas. E a construção de modelos especializados a partir de outros modelos de base facilita o processo. Atualmente, o processamento desse aprendizado no *Transkribus* é operacionalizado pelo *PyLaia HTR*¹⁰. Conforme também comentam Baudry e Pedro (2023, p. 237), a especialização pode ser tão granular que os modelos são, efetivamente, “treinados para mãos específicas”, como nos casos dos manuscritos de Jeremy Bentham ou Michel Foucault, exemplificados pelos autores. Essa capacidade, no entanto, expõe o paradoxo central do HTR para documentos históricos: o problema da escassez de dados (ou *data scarcity*), amplamente reconhecido na literatura atual como um dos maiores entraves para o treinamento de modelos robustos (Alkendi *et al.*, 2024). Os modelos de *deep learning*, como os do *Transkribus*, exigem um volume massivo de texto para atingir alta precisão, mas são poucos os autores históricos cuja produção manuscrita sobrevive em quantidade suficiente para treinar um modelo escritor-específico eficiente do zero. A solução da

¹⁰ Um *framework* de código aberto que sucedeu o motor *CITlab HTR+* cf. (Terras *et al.*, 2025; Nockels *et al.*, 2022). Desenvolvido por pesquisadores do grupo Pattern Recognition and Human Language Technology (Tecnologia em Reconhecimento de Padrões e Linguagem Humana, tradução própria) da Universitat Politècnica de València, notadamente Joan Puigcerver cf. (Puigcerver, 2017, 2018), Carlos Mocholí, Daniel Martín-Albo e Mauricio Villegas. O *PyLaia* funciona como o motor de reconhecimento (*engine*) responsável por implementar a arquitetura de IA de “Quarta Geração”. Disponível em: <https://github.com/jpuigcerver/PyLaia/wiki>.

comunidade para esse impasse é o desenvolvimento de modelos “genéricos”, que são treinados em uma coleção de manuscritos de diferentes autores, mas com caligrafia, idioma e período histórico semelhantes. Esses modelos, chamados pelos manuais do *Transkribus* de “modelos de base” (*base models*), servem como ponto de partida para o aprendizado por transferência (*transfer learning*) de modelos mais especializados. Em vez de começar do zero, o pesquisador pode “refinar” (*fine-tuning*) um modelo genérico robusto com uma quantidade muito menor de seus próprios dados. De acordo com os manuais do *Transkribus*, enquanto um modelo novo exigiria um volume imenso, um *fine-tuning* eficaz pode ser alcançado com cerca de 5.000 a 15.000 palavras de GT, tornando a transcrição automatizada de alta precisão viável para a grande maioria dos projetos de pesquisa. Conforme afirmam as próprias sessões tutoriais do serviço:

Dependendo do tipo de material e do número de caligrafias (mãos), são necessárias entre 5.000 e 15.000 palavras (cerca de 25 a 75 páginas) de material transcrito para iniciar o processo. Em geral, as redes neurais do motor de Reconhecimento de Texto aprendem rapidamente: quanto mais dados de treinamento tiverem, melhores serão os resultados. Se o trabalho for com material impresso, 5.000 palavras devem ser suficientes para alcançar uma boa Taxa de Erro de Caracteres [CER]. No caso de documentos manuscritos, a recomendação é treinar o modelo com pelo menos 10.000 palavras para cada caligrafia. Modelos treinados com um grande volume de dados de treinamento (mais de 100.000 palavras), compreendendo várias caligrafias do mesmo período e região, devem ser capazes de reconhecer mãos não vistas em nenhum momento durante o treinamento: os resultados, no entanto, provavelmente serão um pouco inferiores à Taxa de Erro de Caracteres (que é medida nos Dados de Validação) (READ-COOP SCE, 2025, tradução própria)¹¹.

O processo de treinamento de um modelo no *Transkribus* se inicia com a preparação das GTs, o conjunto de dados que servirá de verdade absoluta para o treinamento da IA. Como comenta o manual, para criá-lo, o pesquisador pode seguir dois caminhos: a transcrição puramente manual dentro das ferramentas da plataforma ou um método híbrido. Nele, se utiliza um modelo público já existente, normalmente o *base model* do próprio modelo *fine-tuned* que se pretende criar¹², como ponto de partida para gerar uma primeira versão da transcrição (*pre-labeling*), que é, então, corrigida e revisada para garantir a qualidade da GT. Para isso, é fundamental compreender que a plataforma opera com uma arquitetura de dois modelos de IA distintos, mas interligados, a *pipeline* padrão da área (cf. Wang *et al.* (2021, p. 15)). O primeiro é o modelo de Análise de Layout (*Layout Analysis*), responsável pela segmentação da página, por detectar as regiões de texto e as linhas de base. O segundo é o modelo de Reconhecimento de Texto propriamente dito (o modelo de HTR/OCR), que efetivamente transcreve os pixels das linhas detectadas em caracteres. Embora ambos os modelos possam ser treinados em conjunto, a análise de layout é uma tarefa computacionalmente mais simples, e os modelos genéricos oferecidos pela plataforma frequentemente já produzem resultados excelentes. Por isso, é comum e eficiente executar apenas a segmentação automática para depois realizar a transcrição manual da GT, agilizando o trabalho e economizando créditos. Uma segmentação precisa é crucial, pois ela define não apenas a localização, mas também a ordem de leitura do texto na página, um desafio notório em documentos com layouts complexos, como tabelas ou notas de margem de página.

Outro fator importante para saber em relação aos modelos de reconhecimento do *Transkribus* é a forma de avaliação de um modelo de IA, que exige a separação dos dados de input em conjuntos, um de treinamento e outro de validação. O conjunto de treinamento é a GT utilizada justamente para alimentar o modelo, enquanto o conjunto de validação (*validation set*) consiste em uma porção da GT (geralmente 10%, conforme recomendado pela plataforma) que não é usada para o aprendizado do modelo, mas para testar o desempenho do modelo em dados não treinados e checar o seu desempenho com uma prova real. É a partir desse conjunto que o *Transkribus* calcula a sua principal métrica de

¹¹ Disponível em: help.transkribus.org/data-preparation

¹² Que, no caso da primeira versão do EPP, foi o *Portuguese Printed M1*.

performance: a Taxa de Erro de Caracteres (*Character Error Rate*¹³ ou CER). Segundo o manual da plataforma,

O desempenho de um modelo é determinado com base na 'distância' entre uma transcrição perfeita (o seu *Ground Truth*) e o texto reconhecido automaticamente, sendo medido pela Taxa de Erro de Caracteres (CER). O Erro de Caractere é a porcentagem de caracteres que foram transcritos incorretamente pelo modelo de Reconhecimento de Texto. Por exemplo, um CER de 5% significa que o modelo de texto transcreveu automaticamente de forma correta 95 de 100 caracteres, enquanto interpretou erroneamente apenas 5 caracteres (READ-COOP SCE, 2025, tradução própria)¹⁴.

Para dar alguma noção sobre o desempenho dos modelos, o manual do *Transkribus* estabelece *benchmarks* práticos para o CER dos modelos. Segundo a documentação do *Transkribus* citada acima, para modelos de HTR, um CER abaixo de 10% já é considerado eficiente, com o ideal sendo algo entre 2% e 8%. Para modelos de OCR treinados para textos impressos, a expectativa de precisão é maior, com um CER recomendado entre 0,5% e 2%. Contudo, esses valores se referem a textos similares aos do treinamento; textos impressos tendem a ser mais modernos, com fontes e qualidade de impressão padronizadas. Ao aplicar um modelo a dados não vistos e estilisticamente diferentes, como uma caligrafia de outra mão ou novos modelos de impressão/tipografia, o CER pode aumentar drasticamente, chegando a 20% ou 30%. Esse desafio de “desvio de domínio” foi o principal obstáculo no treinamento do modelo EPP. Embora se tratassem de textos impressos, a extensão diacrônica do corpus introduziu uma imensa variabilidade tipográfica e ortográfica. A diferença gráfica entre as gramáticas de Oliveira (1536) e de Barros (1540), por exemplo, separadas por quatro anos entre as publicações, já foi suficiente para tornar os primeiros passos de treinamento difíceis e degradar severamente o desempenho nas rodadas iniciais de treinamento. A solução metodológica foi uma forma de adaptação de domínio: o corpus foi ampliado com a transcrição de outros impressos do século 16. Embora não fossem o foco da análise linguística do trabalho final, esses materiais adicionais criaram um conjunto de treinamento mais diverso e robusto, de forma a permitir ao modelo aprender a generalizar para as múltiplas variações tipográficas do período.

A partir do CER calculado sobre os dados de treinamento e validação, a plataforma gera uma “curva de aprendizado”, um gráfico que serve para avaliar a performance do modelo ao longo das épocas de treinamento. Nesse gráfico, o eixo vertical representa o CER, enquanto o eixo horizontal marca o progresso do treinamento em épocas, cada época correspondendo a um ciclo completo de processamento de todo o *training set*. A principal utilidade da curva é diagnosticar possíveis problemas do treinamento ou vícios nas GTs, por meio da análise do comportamento das duas linhas: a do erro de treinamento e a do erro de validação. O cenário ideal é a convergência das duas para um valor de CER baixo, o que indica que o modelo está aprendendo e generalizando o que aprende para novos grupos de dados.

No entanto, a curva é crucial para identificar duas condições problemáticas comuns: quando o erro de treinamento continua a diminuir, mas o erro de validação estagna ou aumenta, o que significa que o modelo está apenas “decorando” os dados de treinamento e perdendo a capacidade de generalizar para documentos novos, um problema denominado *overfitting*; e quando ambas as linhas de erros permanecem altas, o que indica que o modelo é incapaz de aprender os padrões do corpus; nesse caso tem-se *underfitting* (Goodfellow; Bengio; Courville, 2016, p. 111). Verificar a curva de aprendizado de um modelo, portanto, é um passo metodológico necessário, não só para determinar a sua confiabilidade, mas também para decidir o momento ótimo de interromper o treinamento num tipo específico de GT e evitar *overfitting*.

A curva de aprendizado da última versão do modelo EPP, apresentada na Figura 1, ilustra a aplicação prática desses conceitos. No gráfico, a linha azul escura representa a Taxa de Erro de Caracteres nos dados de treinamento (*Train CER*), enquanto a linha verde clara indica o erro correspondente nos dados de validação (*Validation CER*).

¹³ Cabe ressaltar que essa métrica para a Taxa de Erro de Caracteres se fundamenta numa métrica conhecida como



Figura 1. Curva de aprendizado da última versão do EPP.

Fonte: Reprodução de metadados do modelo, disponível em: <https://app.transkribus.org/models/text/267229>.

A análise da curva de aprendizado da versão final do modelo EPP (Figura 1) demonstra um processo de treinamento bem-sucedido. Ainda nas primeiras épocas de treinamento, há uma fase de convergência rápida, na qual o *Train CER* cai de aproximadamente 85% e o *Validation CER* de aproximadamente 45% para um patamar inferior a 10% já nas primeiras épocas. A queda acentuada em ambas as curvas descarta um caso de *underfitting*. Mais importante, a distância mínima e estável entre as duas linhas demonstra que o modelo generaliza bem para dados não vistos, sinal de que também não há *overfitting*. Após a décima época, as curvas atingem um “platô” de convergência, se estabilizando num CER final de 2,58% para validação e 1,91% para treinamento. Mesmo que esses valores não atinjam o *benchmark* de 0,5% para modelos de PTR, eles representam um resultado excelente para um modelo generalista treinado num corpus diacrônico, de natureza gráfica desuniforme. A capacidade de manter um CER tão baixo num corpus que abrange a vasta diversidade tipográfica e ortográfica de três séculos (16 a 19) comprova a robustez e a alta capacidade de generalização dos modelos *PyLaia* treináveis no *Transkribus*.

4 O Early Portuguese Printing

O modelo de reconhecimento *Early Portuguese Printing* foi criado em abril de 2023 para auxiliar na transcrição das obras analisadas na dissertação “O Português das ‘Obras Proueitasas’ e dos ‘Ilustres & Muito Reuerendos Senhores’: uma análise da mudança nos padrões de colocação pronominal clítica em gramáticas portuguesas do século 15 ao 19”. O EPP teve diversas versões ao longo do seu processo de treinamento. A primeira a ser tornada pública foi a versão V-e (model ID 191805), em outubro de 2024, com 122.337 palavras treinadas e CER de 2,67%; a segunda versão (e final até o momento) é a VI-b, com 142.606 palavras e CER de 2,58%, publicada em janeiro de 2025.

O treinamento do modelo EPP, como dito, partiu da constatação de que faltavam opções especializadas na biblioteca pública do *Transkribus* para o tipo de obra impressa necessária na pesquisa. Para fins de documentação, os modelos voltados para a língua portuguesa disponíveis na plataforma até então eram apenas os disponíveis na Tabela 1.

Como se pode perceber pela Tabela 1, os modelos com maior treinamento em língua portuguesa eram os modelos voltados à transcrição de manuscritos. Os únicos modelos voltados à escrita impressa entre os séculos 16-19 (natureza das obras) são os modelos *SPJCL 17C 4.2* (ID: 52754), da Spanish & Portuguese Jews’ Congregation of London, que mistura o treinamento com textos em espanhol e em hebraico romanizado, o que torna a capacidade de transcrição em língua portuguesa muito difícil, e o *XXth century Typewritten Portuguese*, treinado em textos já muito mais modernos. O corpus utilizado para o treinamento do EPP possui diversas características únicas que tornavam impreciso o uso de outros modelos – por se tratarem de gramáticas e obras de cunho linguístico impressas

“distância de Levenshtein” (Levenshtein, 1966 apud Wang *et al.* (2021, p. 22)).

¹⁴ Disponível em <https://help.transkribus.org/character-error-rate-and-learning-curve>.

Tabela 1. Modelos de ATR do Transkribus para a língua portuguesa até julho de 2023.

Hand-written/ Print	Name e ID	Centuries	Author	Publication	Word count	CER
HTR	<i>General Portuguese</i> M1 (35459)	0-21	by luciafws@icloud.com	Sep. 27, 2022	64.847	3,80%
HTR	<i>Transkribus portuguese handwriting</i> M2 (45090)	–	by Transkribus Community	Oct. 4, 2022	707.803	8,00%
PTR	<i>Latin Portuguese Print 17th century</i> (41932)	17	by Hervé Baudry (hbaudry@fcsh.unl.pt)	Nov. 9, 2022	23.363	1,5%
PTR	<i>SPJCL 17C 4.2</i> (52754)	–	by oliver2@sampabym.com	Jan. 13, 2023	64.324	1,60%
HTR	<i>Portuguese Handwriting 16th-19th c.</i> (53270)	16-19	by TraPrInq Project	Jul. 5, 2023	1.158.586	5,20%
PTR	<i>XXth century Typewritten Portuguese</i> (56026)	20	by Natalia C. Salvador	Nov. 24, 2023	7.468	2,60%

Fonte: Elaborado pelo autor a partir dos dados disponíveis em <https://transkribus.org>.

em massa desde 1536 até 1894, diversos aspectos filológicos, tipográficos e linguísticos precisavam ser levados em consideração para a construção de um modelo suficientemente eficiente em obras de diferentes momentos históricos e tecnológicos.

As obras metalinguísticas, livros como gramáticas e tratados ortográficos, utilizadas até então para a construção do modelo final são as descritas na Tabela 2.

Tabela 2. Corpus de gramáticas utilizado para o treinamento do EPP.

Título	Autor	Publ.	N. de palavras
<i>Grammatica da lingua portuguesa</i>	Fernão de Oliveira (c. 1507-1581)	1536	22.974
<i>Grammatica da lingua portuguesa</i>	João de Barros (1496-1570)	1540	23.873
<i>Origem da Lingoa portuguesa</i>	Duarte Nunes de Leão (c. 1530-1608)	1606	21.263
<i>Orthographia da Lingua Portu-gueza</i>	Luís Caetano de Lima, C.R. (1671-1757)	1736	26.997
<i>Grammatica philosophica e ortho-graphia racional da lingua portu-guesa</i>	Bernardo de Lima e Melo Bacellar (1736-)	1783	25.219
<i>Grammatica nacional</i>	Francisco Júlio Caldas Aulete (1823-1878)	1864	21.353
			Total: 141.679 ¹⁵

Fonte: Pacheco Rocha (2025).

Guiado pelos preceitos teóricos do aprendizado de máquina para redes neurais, o modelo EPP final foi o resultado de um processo de treinamento “iterativo” conduzido ao longo de cerca de um ano, e não de um único evento de treinamento, partindo das obras mais antigas (séc. 16) às mais recentes (séc. 19). Essa forma pouco intuitiva de aprendizado “anti-curricular” (Bengio, 2009, p. 103) foi a estratégia de treinamento de IA adotada sob a premissa de que, ao investir tempo e esforço iniciais na disponibilização de GTs de melhor qualidade logo no início, o modelo desenvolveria uma base de reconhecimento mais sólida e aprenderia as exceções descartáveis, formas arcaicas que não encontraria no restante do corpus. Isso ocorreria ao aprender primeiro a vasta variabilidade dos impressos antigos, para, então, refinar esse conhecimento com textos mais padronizados. A metodologia consistiu em ciclos sequenciais de treinamento, nos quais novos documentos eram progressivamente transcritos e adicionados ao conjunto de dados para treinamento de novas versões do modelo, a partir das antigas

¹⁵ A diferença de 927 palavras para o total do modelo EPP se deve a obras que foram parcialmente transcritas e utilizadas no treinamento do modelo ainda no início do processo, em 2023, e que não foram incluídas no corpus da dissertação.

e de mais GTs. A cada ciclo, uma nova e mais robusta versão do modelo era gerada a partir da versão anterior, por meio do transfer learning já descrito. Essa abordagem permitiu que a precisão do modelo evoluísse gradualmente e o trabalho de transcrição de GTs fosse distribuído, já que se tratava de apenas um pesquisador. O treinamento das primeiras versões do modelo começou com as obras de Barros (1540) e Oliveira (1536), que já contam com versões transcritas disponíveis, especialmente as edições de Barros (1540) por Machado (1957) e as de Oliveira (1536) por Buescu (1975).

A etapa de transcrição para a geração das GT foi realizada diretamente na interface de edição do *Transkribus eXpert client* e, alternativamente, no webapp. Para o treinamento do modelo, a transcrição buscou ser conservadora, ou seja, caracterizou-se pela ausência de modernização ortográfica ou sintática, buscando a máxima fidelidade à materialidade do texto. No ambiente digital, essa postura diplomática consiste em, basicamente, manter os caracteres encontrados nos documentos originais por meio de seus correspondentes no padrão *Unicode*. Esse processo, no entanto, buscou também equilibrar a fidelidade ao corpus com a necessidade de estruturação dos dados para o treinamento do modelo, o que potencializaria a capacidade de transcrever mais obras; esse equilíbrio implica uma série de decisões que buscam tornar o treinamento mais eficiente em detrimento ao conservadorismo da transcrição. Essa seção se dedicará a explicitar esse equilíbrio e descrever as potencialidades do modelo.

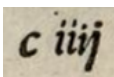
No que tange à segmentação de palavras, optou-se por uma intervenção padronizada: vocábulos que aparecem unidos no original foram separados segundo a norma ortográfica atual. Inversamente, nos casos em que erros de impressão separaram graficamente uma única palavra, o espaçamento anômalo foi mantido. Tal decisão visou preservar a previsibilidade do texto para o modelo de transcrição, evitando introduzir correções que não estivessem presentes no fac-símile. Seguindo a mesma lógica, as quebras de linha, hífen e outras marcações que indicam a interrupção de uma palavra para sua continuação na linha subsequente foram igualmente preservadas na transcrição final.

4.1 Descrição filológica do modelo

Ao apresentar a descrição filológica do modelo EPP, essa seção buscará detalhar as decisões tomadas ao longo do processo, de forma a exemplificar o tipo de trabalho necessário para o treinamento de modelos do tipo. Cada escolha filológica não apenas interfere diretamente na fidelidade do resultado final, mas também na fidelidade e capacidade de transcrição dos resultados futuros, pois o modelo é treinado com o próprio corpus que busca transcrever, nesse caso. A natureza diacrônica do corpus, nesse ínterim, é parte desse desafio, pois as soluções dadas a problemas de transcrição dos séculos anteriores devem ser compatíveis com as tradições gráficas dos séculos posteriores. Dessa forma, portanto, nesse trecho, será exposta uma série de exemplos de decisões e formas adotadas para a transcrição e treinamento do EPP.

Em relação aos algarismos, se optou pela manutenção das formas numéricas tal como se encontram no documento de base, tanto no caso dos algarismos arábicos quanto das formas românicas, se observando, por exemplo, o uso do “l” inicial como “i” com pingo e a representação do último “i” como “j”, conforme ilustrado na Tabela 3.

Tabela 3. Exemplos de algarismos.

	
¶ Capitolo xviiij.	ciiij

Fonte: Oliveira (1536) e Leão (1606), adaptado por Pacheco Rocha (2025).

O mesmo princípio de preservação foi aplicado às abreviaturas, cuja forma original foi integralmente mantida. Nesse sentido, decidiu-se unificar a representação da sobrescrição horizontal indicativa de abreviatura por meio do til (“~”), por compreender que esta seria a intenção dos autores das obras transcritas; como demonstram os primeiros exemplos da Tabela 4, o tipo-móvel utilizado para representar o (“~”) parecia ser um “r” pequeno na horizontal. Além disso, se respeitou o posicionamento

original do til, fosse ele colocado sobre vogais ou sobre consoantes, preservando assim as convenções gráficas observadas nos exemplares de referência, conforme demonstrado nos exemplos da Tabela 4.

Tabela 4. Exemplos de abreviaturas.

Forma original					
Transcrição adotada	Ntõ	nũero	por q̃	ẽ os	conuẽ
Legenda	(Nominativo)	(número)	(por que)	(em os)	(convem) ¹⁶
Forma original					
Transcrição adotada	aqlas	q̃lq̃r q̃	semp̃	q̃	gêtes
Legenda	(aquelas)	(qualquer que)	(sempre)	(que)	(gentes)

Fonte: Oliveira (1536), Barros (1539) e Leão (1606), adaptado por Pacheco Rocha (2025).

Opções como essas apresentam certos desafios, já que o número de exemplos para o treinamento precisa ser suficiente para as redes neurais aprenderem a generalizar esse tipo de ocorrência. Além disso, outros modelos do *Transkribus*, baseados noutras línguas e sistemas grafemáticos, podem influenciar ou causar vícios ao modelo treinado (normalmente a “criação” de um modelo se dá a partir da especialização de outro mais próximo); modelos treinados em castelhano, por exemplo, podem generalizar formas como “ão” em “ano”. Para tal, portanto, é importante que todos os exemplos do tipo sejam devidamente transcritos e que, nas ground truths, os diacríticos estejam devidamente unidos às letras que sobrescrevem.

Em complemento a essas escolhas que buscam otimizar o trabalho de treinamento do modelo, se optou por não registrar as formas de apontamento de abreviatura cuja recorrência é reduzida diacronicamente e cuja clareza gráfica é baixa, pois a baixa nitidez de tais sinais poderia ser interpretada pelo modelo de transcrição como ruído, comprometendo o processo de treinamento. Considerou-se, nesse caso, que o eventual ganho de fidelidade filológica não se justificaria, uma vez que tais elementos são isolados e pouco numerosos o bastante para serem desconsiderados. Entre os exemplos de sinais desconsiderados, estão o uso do “r” rotundo (“ʀ”), empregado por Oliveira (1536) em todos os dígrafos com consoante complexa de tepe, como no primeiro exemplo, bem como as formas diferenciadas de abreviatura para a preposição “para” ou “pr”, observadas nos exemplos a seguir e em outras ocorrências análogas (Tabela 5).

Tabela 5. Exemplos de formas diferenciadas de abreviatura.

Ponto de código	U+A75B	U+A751	U+A753	U+0223	U+0223
Forma original					
Transcrição adotada	proprio	p	p	d	ds
Transcrição diplomática	pʀopʀio	pʀ	pʀ	ʀ ¹⁷	ʀs

Fonte: Oliveira (1536) e Barros (1540), adaptado por Pacheco Rocha (2025).

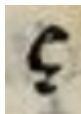
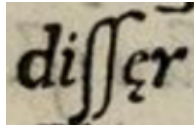
Seguindo o mesmo princípio de fidelidade, todos os diacríticos, como o acento agudo (“´”), o

¹⁶ Fernão de Oliveira não diferenciava singular e plural utilizando acentos gráficos como a convenção ortográfica moderna exige, portanto, a legenda acentuada para oxítone terminada em -em “convém” implicaria um sujeito singular (em oposição ao plural “convêm”), sentido que inexistia no documento original.

¹⁷ Embora a transcrição adotada normalize a abreviatura para “d”, adoto neste trabalho a hipótese de que o glifo original seja “ʀ” (ligatura de “ou”, U+0223), provável reaproveitamento de tipos móveis destinados à impressão do grego. O uso como uma abreviação logográfica de palavras iniciadas com “d” é corroborado pela clássica transcrição de Buescu (1975, p. 38, p. 65), que transcreve a forma “ʀs” como “Deus”.

grave (“`”), o circunflexo (“^”) e o til (“~”), foram preservados em suas posições originais. Essa manutenção é importante, pois tais marcas refletem as convenções ortográficas do corpus e possuem função distintiva na fonética e morfologia das palavras, que não pode ser perdida. Além disso, omiti-las poderia levar o modelo de transcrição a generalizar incorretamente a exclusão de outros sinais gráficos importantes. Um caso específico de preservação foi o do e-caudata (“ę”), também chamado de “e-cedilha”, utilizado por Barros (1540). Este caractere, nomeado no padrão *Unicode* como “Latin Small Letter E with Ogonek¹⁸”, apresenta um diacrítico subscrito em forma de gancho. Sua inclusão no treinamento foi considerada essencial, pois, devido à sua proeminência visual e posicionamento em relação à letra, ignorá-lo criaria o risco de ensinar o modelo a descartar outros diacríticos subscritos, especialmente no que diz respeito à semelhança gráfica entre “c”, “e”, “ç” e “ę”, o que teria um resultado prejudicial para a precisão de qualquer transcrição em língua portuguesa (Tabela 6).

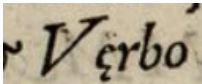
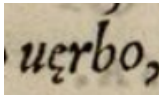
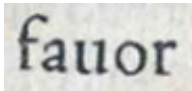
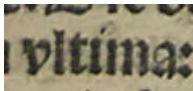
Tabela 6. Exemplos de diacrítico subscrito.

Ponto de código	U+0119	U+0119
		
Forma original		
Transcrição adotada	ę	disser

Fonte: Barros (1540), adaptado por Pacheco Rocha (2025).

Quanto à pontuação, todos os sinais foram preservados conforme aparecem no documento original, incluindo ponto (”), vírgula (”), ponto e vírgula (“;”), dois-pontos (“:”), barra (“/”), barra vertical (“|”), ponto de interrogação (“?”), ponto de exclamação (“!”), hífen (“-”), sinal de negação (“¬”) e apóstrofo (“’”). Essa preservação buscou assegurar a fidelidade tipográfica e a interpretação correta das estruturas sintáticas, uma vez que a pontuação nos impressos históricos frequentemente apresenta variações que podem impactar a análise linguística. Além disso, se manteve a diferenciação funcional entre “u” e “v”, herança latina ainda observada pelos autores das obras transcritas. Essa distinção, importante para a representação gráfica e fonética das palavras, foi respeitada em todos os casos do corpus, conforme ilustrado nos exemplos a seguir. No entanto, a sutileza dessa distinção pode representar um desafio para o modelo computacional. Permanece incerto se a capacidade estatística do *Transkribus* é suficiente para generalizar corretamente o uso de “u” e “v” sem produzir erros, uma limitação que talvez pudesse ser superada com um volume de dados de treinamento maior ou maior dispêndio de processamento estatístico das transcrições, como o uso de modelos de linguagem ou o serviço de smart search, oferecido para os usuários pagantes (Tabela 7).

Tabela 7. Exemplos de distinção funcional “u”/“v”.

				
Forma original				
Transcrição adotada	Verbo	uërbo	fauor	vltima:

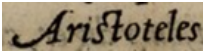
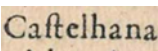
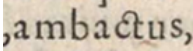
Fonte: Oliveira (1536), Barros (1540) e Leão (1606), adaptado por Pacheco Rocha (2025).

O tratamento das ligaturas (elementos tipográficos que unem dois ou mais caracteres em um único tipo) também exigiu a definição de um critério claro. Nas gramáticas transcritas, as ligaturas aparecem em duas formas básicas: a consonantal, que cumpria uma função pragmática de facilitar o trabalho do tipógrafo (por exemplo, criando um tipo único “st” para a sequência “st”, que exigiria dois tipos); e a vocálica, cujo uso aparenta ter também uma motivação ortográfica, especialmente em passagens em latim. Contudo, a prática de transcrição revelou que o emprego de ambas as formas

¹⁸ Se trata de um glifo (Latin Small Letter E with Ogonek) atualmente utilizado na escrita de línguas como o polonês e o lituano. A adoção deste code point (U+0119) para representar o e-caudata segue as recomendações técnicas da Medieval Unicode Font Initiative (Haugen, 2015).

é bastante idiossincrático e pouco coerente ao longo do corpus. Diante dessa inconsistência e de desafios técnicos adicionais, a norma adotada foi, via de regra, ignorar as ligaturas, desfazendo-as em seus caracteres componentes. Esta decisão foi fundamentada por três fatores principais: primeiro, a baixa representatividade da maioria das ligaturas não justificava o complexo esforço de treinamento; segundo, o risco de que formas raras fossem confundidas com ruídos do fac-símile pelo modelo, comprometendo a transcrição; e, por fim, a inexistência de certas ligaturas no padrão Unicode, o que impedia a correta representação digital das ligaturas por completo (Tabela 8).

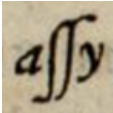
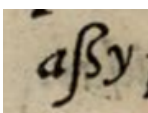
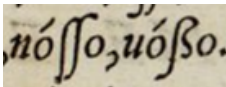
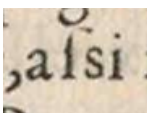
Tabela 8. Exemplos de ligaturas tipográficas.

Ponto de código	U+FB06	U+FB05	–
Glifo original			
Transcrição adotada	Aristoteles	Castelhana	ambactus
Transcrição diplomática	Aristoteles	Castelhana	Sem correspondência da ligatura no Unicode

Fonte: Barros (1540) e Leão (1606), adaptado por Pacheco Rocha (2025).

À norma geral de desfazer as ligaturas foi estabelecida uma única exceção: as formadas pela combinação das variantes da letra “s”. Nos textos, se observam tanto a forma “s-longo” (“f”) quanto a “s-curta” (“s”), cuja distinção era meramente funcional, pouco sistemática e foi, via de regra, ignorada para transcrição e treinamento. A junção das duas, no entanto, formava um caractere específico, a ligatura conhecida atualmente como eszett (“ß”). A decisão de preservar esse caractere se fundamenta basicamente na sua saliência gráfica, que é proeminente o suficiente para comprometer o treinamento do modelo caso fosse ignorada. Portanto, para garantir a estabilidade e a precisão do reconhecimento, esta ligatura em particular foi mantida na transcrição final (Tabela 9).

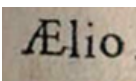
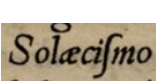
Tabela 9. Exemplos de distinção “s-curto”/“s-longo” e sua ligatura.

Ponto de código	U+017F	U+00DF	U+00DF	U+0073
Forma original				
Transcrição adotada	assy	aßy	nósso,uóßo.	,aßsi
Transcrição diplomática	affy	aßy	nóßfo,uóßo.	,aßsi

Fonte: Barros (1540) e Leão (1606), adaptado por Pacheco Rocha (2025).

O tratamento de outros casos específicos incluiu a manutenção de nexos vocálicos, como a ligatura “æ”. De modo semelhante, se optou por preservar o “e comercial” (“&”), a forma ligada da conjunção latina “et”, de modo a uniformizar a representação da conjunção coordenativa numa única forma sempre que esta aparecia como uma ligatura ou na notação tironiana adotada por Oliveira (1536) (Tabela 10).

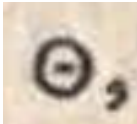
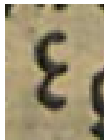
Tabela 10. Exemplos de outras ligaturas.

Ponto de código	U+00C6	U+00E6	U+0026	U+0026	U+204A
Forma original					
Transcrição adotada	Ælio	Solæcismo	&	&c	&
Transcrição diplomática	Ælio	Solæcismo	&	&c	j

Fonte: Oliveira (1536), Barros (1540) e Leão (1606), adaptado por Pacheco Rocha (2025).

O tratamento de caracteres não latinos e formas tipográficas especiais também seguiu critérios específicos. Um caso notável é a presença de letras do alfabeto grego. Devido à forte ligação do conhecimento letrado da época com a cultura greco-latina, a ocorrência desses caracteres é relativamente frequente nas gramáticas transcritas. Diante de sua relevância contextual, se optou por manter todas as letras gregas na transcrição, garantindo assim a máxima fidelidade ao texto original (Tabela 11).

Tabela 11. Exemplos de caracteres gregos.

Ponto de código	U+0398	U+03A6	U+03B5	U+03C9
Forma original				
Transcrição adotada	Θ	Φ	Ε	ω

Fonte: Oliveira (1536) e Barros (1540), adaptado por Pacheco Rocha (2025).

Contudo, a essa norma de preservação aplicou-se um critério de otimização para letras gregas com formas visualmente análogas às latinas, como as maiúsculas Chi ("X", U+03A7) e Zeta ("Z", U+0396). Nesses casos, optou-se por transcrevê-las utilizando seus equivalentes latinos ("X", U+0058 e "Z", U+005A), a fim de reduzir o conjunto de caracteres a ser aprendido pelo modelo. Uma exceção crucial a este critério de simplificação foi a letra alfa latina pequena ("α", U+0251). Esse caractere, empregado por Oliveira (1536) para marcar uma distinção explícita com a letra "a" em sua proposta ortográfica, foi mantido em sua forma original, pois sua normalização tornaria o argumento do autor incompreensível. Embora a raridade do "α" torne seu reconhecimento pelo modelo pouco confiável, o impacto dessa decisão na performance geral do sistema foi considerado irrisório, dada sua frequência de uso extremamente baixa em comparação com a vogal "a" (Tabela 12).

Tabela 12. Exemplo de letra latina diferencial.

Ponto de código	U+0251
Forma original	
Transcrição adotada	α

Fonte: Oliveira (1536), adaptado por Pacheco Rocha (2025).

O tratamento de elementos tipográficos e ornamentais seguiu um critério de frequência e relevância no corpus. Símbolos de presença constante, como o *pilcrow* ou caldeirão ("¶") e os diferentes tipos de fleurão ("§"), foram mantidos na transcrição para preservar a estrutura visual e editorial das obras. No entanto, se optou por ignorar distinções sutis dentro de uma mesma categoria de símbolo quando a representatividade de uma das variantes era baixa. Foi o caso do caldeirão de capítulo, conhecido como *capitulum* ("‡"), que foi representado de forma unificada como um caldeirão simples ("¶"), dada a sua rara ocorrência (Tabela 13).

Da mesma forma que o caldeirão, outros elementos ornamentais foram preservados com base em sua frequência e clareza. É o caso do fleurão, também chamado de "hera", um elemento decorativo de uso recorrente, especialmente por Barros (1540). Sua forma graficamente distinta e seu uso frequente justificaram a decisão não apenas de mantê-lo na transcrição, mas também de incluir seu reconhecimento como uma tarefa específica no treinamento do modelo, garantindo que não fosse interpretado como ruído (Tabela 14).

Tabela 13. Exemplos de elementos tipográficos funcionais.

Ponto de código	U+00B6	U+2E3F
Forma original		
Transcrição adotada	¶	¶
Transcrição diplomática	¶	¶

Fonte: Oliveira (1536) e Barros (1540), adaptado por Pacheco Rocha (2025).

Tabela 14. Exemplos de elementos tipográficos funcionais/ornamentais.

Ponto de código	U+2767	U+2619	U+2766
Forma original			
Transcrição adotada	☹	☹	☹

Fonte: Barros (1540) e Leão (1606), adaptado por Pacheco Rocha (2025).

4.2 Potencialidades e Aplicações do Modelo EPP

Embora o modelo EPP tenha sido treinado primariamente com obras de natureza metalinguística impressas em Portugal, suas potencialidades se estendem muito além desse gênero textual e de suas fronteiras geográficas iniciais. Na prática, a eficácia da ferramenta está circunscrita ao idioma português e à tipografia da época, independentemente do local de impressão do documento. Durante os períodos colonial e imperial, a circulação de impressos europeus e de impressão de textos em português, como jornais, decretos, alvarás, regimentos administrativos, além das próprias gramáticas, foi fundamental na formatação da ecologia linguística do Brasil, desde os centros administrativos até as zonas de fronteira em expansão no território.

Pesquisadores brasileiros que se debruçam sobre esses acervos impressos de circulação colonial podem usufruir do EPP de duas maneiras: a primeira é a aplicação direta do modelo sobre documentos impressos em língua portuguesa entre os séculos 16 e 19, a partir de sua versão disponibilizada publicamente na biblioteca do *Transkribus*. Isso permite a extração automatizada de textos em larga escala e com alta confiabilidade, de maneira a viabilizar análises quantitativas que levariam anos caso dependessem de transcrição manual. A segunda, e talvez mais promissora, é o uso do EPP como um “modelo de base”. Para pesquisadores que trabalham com corpora em língua portuguesa que apresentam variações tipográficas muito específicas, não é necessário iniciar um treinamento do zero. Equipes de pesquisa podem aplicar o EPP como base e realizar o fine-tuning fornecendo apenas algumas dezenas de páginas do seu corpus específico. Essa dinâmica de transferência de aprendizado reduz drasticamente a barreira técnica e a necessidade de transcrição de corpora massivos para o treinamento a partir do zero, democratizando o alcance de um baixo CER independentemente da especificidade do material analisado.

5 Considerações Finais

O presente trabalho demonstrou a viabilidade e a eficácia da criação de modelos de reconhecimento de texto especializados para corpora históricos. Por meio de um processo de treinamento iterativo na plataforma *Transkribus*, o modelo EPP alcançou uma Taxa de Erro de Caracteres (CER) de 2,58% num corpus diacrônico de alta complexidade. Esse resultado comprova que a “Quarta Geração” de tecnologias de ATR, baseada em redes neurais profundas, permite que pesquisadores individuais ou pequenas equipes superem o tradicional gargalo da transcrição manual, viabilizando as análises quantitativas em larga escala exigidas pela Linguística de Corpus, que antes eram pouco concebíveis.

Contudo, a adoção de plataformas como o *Transkribus*, apesar de seus inegáveis benefícios técni-

cos, suscita importantes questionamentos sobre a infraestrutura da pesquisa digital. É natural que a dependência de uma solução centralizada e gerida por uma entidade europeia, na qual dados e modelos que representam o patrimônio documental e científico brasileiro são processados e armazenados, tensiona o debate sobre a soberania digital. Por outro lado, para que essa crítica não se encerre no mero ceticismo tecnológico, é fundamental resgatar a natureza cooperativa da READ-COOP SCE. Conforme demonstrado por Terras *et al.* (2025, p.4 5) e discutido ao longo deste trabalho, a plataforma é gerida sob um regime legal que foge à lógica corporativa tradicional de maximização de lucros e obriga uma gestão controlada pelos membros da cooperativa. Como não-membros da cooperativa, os usuários gratuitos do serviço não estão desprovidos de alternativas.

O desafio atual não reside na ausência de ferramentas alinhadas a uma filosofia de ciência aberta, mas sim nas condições estruturais para a sua adoção. Plataformas de código aberto e acesso livre, como o *eScriptorium* (baseado no motor Kraken), já permitem o processamento de documentos em servidores institucionais próprios. Adicionalmente, também nada impede que pesquisadores operem motores como o próprio PyLaia de forma autônoma, por meio de scripts locais, dispensando interfaces comerciais. Entretanto, a adoção prática dessas alternativas esbarra numa limitação de formação tecnológica, ecossistema e usabilidade. Diferentemente do Transkribus, que atua como um agregador com biblioteca de modelos pública nativa, as soluções abertas possuem uma dinâmica fragmentária. Embora existam louváveis catálogos independentes mantidos pela comunidade, como o diretório HTR-United, eles operam de forma externa à ferramenta de transcrição. Essa necessidade de buscar, baixar arquivos em repositórios de dados (como o Zenodo) e importá-los manualmente para, assim, testar sua aplicabilidade ao corpus desejado eleva a curva de adoção técnica e dificulta o reuso de modelos pré-treinados por pesquisadores cuja especialidade não é tecnologia, mas paleografia, linguística etc. Na prática, essa fricção no compartilhamento leva equipes a iniciarem seus treinamentos a partir do zero.

Além disso, a transição para um modelo puramente descentralizado de inteligência artificial enfrenta grandes obstáculos, especialmente no contexto do financiamento da ciência pública brasileira. Primeiramente, o processamento e o treinamento de redes neurais exigem um hardware custoso, cuja aquisição e manutenção pulverizada por diferentes universidades se mostra financeiramente difícil no cenário atual. Em segundo lugar, a descentralização excessiva do processamento fragmenta os corpora; ao treinar modelos isoladamente em computadores locais, a comunidade perde a capacidade de escalonamento colaborativo. A centralização de dados é uma condição para pesquisa com IA e o treinamento de modelos generalistas, como os do *Transkribus*, treinados com milhões de palavras, dependem fundamentalmente do acúmulo e da centralização de dados e de trabalho humano. Embora plataformas de código aberto como o *eScriptorium* permitam o trabalho colaborativo e o compartilhamento interno de modelos entre usuários de um mesmo projeto, essa funcionalidade é restrita ao ambiente em que o sistema está hospedado. Sua implementação exige, invariavelmente, a manutenção de um servidor institucional dedicado. Sem um projeto interinstitucional que centralize os custos de processamento e o armazenamento de uma biblioteca de modelos à semelhança do modelo cooperativo europeu, a alternativa ao *Transkribus* no Brasil seria esperar que os pesquisadores brasileiros tenham o conhecimento técnico necessário e poder de processamento em suas máquinas pessoais, cenário que ainda inviabilizaria a organização do trabalho colaborativo.

Dessa forma, a sustentabilidade e a autonomia da pesquisa nacional no treinamento de modelos de HTR para transcrição de corpora históricos exigem, mais do que a mera adoção de tecnologias livres, a consolidação de uma estrutura de pesquisa capaz de planificar investimentos em infraestrutura tecnológica e em bases de dados integradas. Mediante uma organização do tipo, seria possível desenvolver alternativas reais, públicas e nacionais ao modelo da sociedade cooperativa europeia. Em suma, a despeito das críticas legítimas, a iniciativa europeia conseguiu alcançar, dentro de suas limitações, um ecossistema colaborativo e, de certa forma, gratuito, erigindo, nas palavras de Terras *et al.* (2025), uma verdadeira “infraestrutura comunitária compartilhada”.

Ainda assim, o trabalho metodológico descrito em torno do modelo EPP aponta para um futuro promissor e alinhado aos princípios da Ciência Aberta. Para superar o entrave da centralização corporativa e garantir que o modelo se consolide como uma ferramenta de acesso livre e amplo,

o compromisso deste trabalho se estende à disponibilização pública de seus dados formadores. Ao tornar o corpus e os parâmetros do modelo publicamente acessíveis sob licenças abertas, garante-se a interoperabilidade dos dados. Na prática, isso permite que a comunidade acadêmica exporte esse conhecimento e o aplique diretamente em ecossistemas fundamentados em software livre, como o próprio eScriptorium, ou em instâncias locais do *PyLaia*. Mais do que uma ferramenta de transcrição restrita a um único ambiente, o EPP representa um ativo de pesquisa e dados abertos que pode ser continuamente aprimorado, auditado e aplicado a novos acervos. Dessa forma, acelera-se o trabalho de filólogos, historiadores e linguistas, buscando promover a colaboratividade e garantir que o avanço das Humanidades Digitais sobre os vestígios escritos do passado seja, em sua essência, um bem comum, auditável e inalienável.

Referências

ALKENDI, Wissam; GECHTER, Franck; HEYBERGER, Laurent; GUYEUX, Christophe. Advancements and Challenges in Handwritten Text Recognition: A Comprehensive Survey. *Journal of Imaging*, v. 10, n. 1, p. 1–18, 2024. DOI: 10.3390/jimaging10010018.

BARROS, João de. *Grammatica da língua portuguesa com os mandamentos da santa madre igreja*. Lisboa: Luís Rodrigues, 1539. Autor: João de Barros (1496–1570). Disponível em: http://acervo.bndigital.bn.br/sophia/index.asp?codigo_sophia=504. Acesso em: 4 set. 2025.

BARROS, João de. *Grammatica da lingua portuguesa*. Olyssipone: apud Lodouicum Rotorigiu[m], Typographum, 1540. Autor: João de Barros (1496–1570). Disponível em: <https://purl.pt/12148>. Acesso em: 4 set. 2025.

BAUDRY, Hervé; PEDRO, Susana Tavares. Humanidades digitais e documentos históricos: transcrever, catalogar. *Cultura. Revista de História e Teoria das Ideias*, Lisboa, n. 41–42, p. 235–246, 2023. Disponível em: <https://revistas.fcsh.unl.pt/cultura/article/view/1223>. Acesso em: 4 set. 2025.

BENGIO, Yoshua. Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning*, Hanover, v. 2, n. 1, p. 1–127, 2009. DOI: 10.1561/22000000006.

BUESCU, Maria Leonor Carvalhão. *A Gramática da Linguagem Portuguesa (Fernão de Oliveira, 1536): Introdução, leitura actualizada e notas*. Lisboa: Imprensa Nacional-Casa da Moeda, 1975.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. *Deep Learning*. Cambridge: MIT Press, 2016.

HAUGEN, Odd Einar (ed.). *MUFI Character Recommendation*. 4.0. ed. Bergen: Medieval Unicode Font Initiative, 2015. Disponível em: <https://nva.sikt.no/registration/0198f0c7becc-8aff5ed6-4690-4727-96bf-69e5d752c276>. Acesso em: 2 jan. 2026.

LEÃO, Duarte Nunes de (fl. 1530–1608). *Origem da Lingoa portuguesa*. Lisboa: Pedro Crasbeeck, 1606. Disponível em: <https://purl.pt/50>. Acesso em: 4 set. 2025.

MACHADO, José Pedro (ed.). *João de Barros: Gramática da Língua Portuguesa*. 3. ed. Lisboa: Sociedade Astória, 1957.

MCENERY, Tony; HARDIE, Andrew. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press, 2012.

NARANG, Sonika Rani; JINDAL, Munish Kumar; KUMAR, Munish. Ancient Text Recognition: A Review. *Artificial Intelligence Review*, v. 54, p. 1–28, 2021. DOI: 10.1007/s10462-020-09827-4.

NOCKELS, J.; GOODING, P.; TERRAS, M.; AMES, S. Understanding the Application of Handwritten Text Recognition Technology in Heritage Contexts: A Systematic Review of Transkribus in Published Research. *Archival Science*, v. 22, n. 3, p. 367–392, 2022. DOI: 10.1007/s10502-022-09397-0.

OLIVEIRA, Fernão de. *Grammatica da lingoagem portuguesa*. Lisboa: Germão Galharde, 1536. Autor: Fernão de Oliveira (c. 1507–1581). Disponível em: <https://purl.pt/120>. Acesso em: 4 set. 2025.

PACHECO ROCHA, Saulo Rogério. *O Português das “Obras Proueitasas” e dos “Ilustres & Muito Reuerendos Senhores”: uma análise da mudança nos padrões de colocação pronominal clítica em gramáticas portuguesas do século 15 ao 19*. 2025. f. 202. Dissertação (Mestrado em Linguística) – Universidade Federal de Santa Catarina, Florianópolis.

PUIGCERVER, Joan. *PyLaia: Deep Learning-based Handwriting Recognition Toolkit*. [S. l.: s. n.], 2017. Disponível em: <https://github.com/jpuigcerver/PyLaia>. Acesso em: 4 set. 2025.

PUIGCERVER, Joan. *A Probabilistic Formulation of Keyword Spotting*. 2018. Tese (Doutorado) – Universitat Politècnica de València, València.

READ-COOP SCE. *Character Error Rate and Learning Curve*. Innsbruck: [s. n.], 2025. Disponível em: <https://help.transkribus.org/character-error-rate-and-learning-curve>. Acesso em: 28 mar. 2026.

SPOLSKY, Joel. *The Absolute Minimum Every Software Developer Absolutely, Positively Must Know About Unicode and Character Sets (No Excuses!)* [S. l.: s. n.], out. 2003. Disponível em: <https://www.joelonsoftware.com/2003/10/08/the-absolute-minimum-every-software-developer-absolutely-positively-must-know-about-unicode-and-character-sets-no-excuses/>. Acesso em: 28 mar. 2026.

TERRAS, M. *et al.* The Artificial Intelligence Cooperative: READ-COOP, Transkribus, and the Benefits of Shared Community Infrastructure for Automated Text Recognition. *Open Research Europe*, v. 5, n. 16, 2025. DOI: 10.12688/openreseurope.18747.2.

WANG, Haifeng; PAN, Changzai; GUO, Xiao; JI, Chunlin; DENG, Ke. From Object Detection to Text Detection and Recognition: A Brief Evolution History of Optical Character Recognition. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2021. DOI: 10.1002/wics.1547. Acesso em: 4 set. 2025.

WANG, Junmiao. A Study of the OCR Development History and Directions of Development. *In: HIGHLIGHTS in Science, Engineering and Technology*. [S. l.: s. n.], 2023. v. 72, p. 409. DOI: 10.54097/bm665j77. Acesso em: 4 set. 2025.

Disponibilização de dados

Os dados de pesquisa estão disponíveis em repositório.

Financiamento

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Declaração de disponibilidade de dados de pesquisa

Em conformidade com as práticas de Ciência Aberta, o *dataset* de gramáticas e obras metalinguísticas utilizado para o treinamento e validação do modelo *Early Portuguese Printing* (EPP) está depositado publicamente no repositório Zenodo. O conjunto de dados inclui os fac-símiles das obras em PDF com camada de texto (ATR), o banco de dados relacional em formato SQLite (.db), que integra o texto integral a metadados estratificados (autor, nascimento, publicação e título), e as transcrições em arquivos de texto plano (.txt) para máxima interoperabilidade. Os dados podem ser acessados livremente e citados por meio do identificador de objeto digital: <https://doi.org/10.5281/zenodo.18354169>.

Declaração de conflitos de interesses

O autor declara que a pesquisa foi conduzida na ausência de quaisquer relações comerciais ou financeiras que possam ser interpretadas como um potencial conflito de interesses.