# Transkribus e o Modelo de OCR Early Portuguese Printing (EPP): Inovações na Transcrição de Documentos Históricos e suas Potencialidades para as Humanidades Digitais

Saulo Rogério Pacheco Rocha

https://doi.org/10.1590/SciELOPreprints.13650

Submetido em: 2025-10-03
Postado em: 2025-12-01 (versão 1)
(AAAA-MM-DD)

# Transkribus e o Modelo de OCR Early Portuguese Printing: Inovações na Transcrição de Documentos Históricos e suas Potencialidades para as Humanidades Digitais

Transkribus and the Early Portuguese Printing OCR Model: Innovations in the Transcription of Historical Documents and their Potential for the Digital Humanities

Saulo Rogério Pacheco Rocha (UFSC)[1]
https://orcid.org/0000-0003-3715-6706

**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 dessas ferramenta para a pesquisa com documentos históricos brasileiros. 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 de 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. Contudo, além de divulgar o potencial da ferramenta, este trabalho problematiza sua natureza. Por pertencer a uma entidade privada europeia e ser um produto 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; Transcrição OCR; Filologia.

**Abstract:** This paper presents the "Early Portuguese Printing" (EPP) Optical Character Recognition (OCR) model, developed on the Transkribus platform, and discusses the potential, challenges, and history of such tools for research with Brazilian 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 just 2.58%. This result represents a significant advancement, as it demonstrates the potential of such tools for creating large-scale historical quantitative corpora 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. However, in addition to publicizing the tool's potential, this paper also problematizes its nature. As it belongs to a private European entity and is a SaaS product, the use of Transkribus raises questions about data centralization and the sustainability of its application in large-scale Brazilian research projects, considering the future and the volume of our historical archives.

[1] Doutorando em Linguística, Universidade Federal de Santa Catarina (UFSC), Programa de Pós-graduação em Linguística (PPGL). Florianópolis/SC. eu@saulo.ru, saulorrp@gmail.com.

**Keywords:** Digital Humanities; Historical Linguistics; OCR Transcription; Philology.

## 0. 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 Proueitosas' 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 daquela 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 a pesquisa em linguística de *corpus* exige. Portanto, a etapa fundamental da metodologia consistiu na transcrição e edição integral dessas obras, a fim de constituir um corpus funcional para a análise.

Para realizar a tarefa, foi necessária a adoção de uma ferramenta de Reconhecimento Óptico de Caracteres (OCR). Dentre as opções avaliadas, como Tesseract ou produtos como Abbyy FineReader, 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 de um *fine-tuning* mais técnico, ao permitir a criação e o treinamento de modelos de OCR, além da edição dos documentos transcritos, dentro de uma interface gráfica própria, especializados no corpus desejado. O objetivo deste trabalho é, portanto, descrever e divulgar o processo de treinamento do modelo Early Portuguese Printing (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 ICs e pós-graduandos, construam *corpora* volumosos, enquanto, até então, o

uso de ferramentas de OCR generalistas compromete a precisão das análises históricas. Este 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 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 OCR, 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 buscará apresentar as considerações finais de maneira crítica, levantando os questionamentos necessários para a adoção e o uso consciente dessa e de ferramentas semelhantes.

## 1. 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 de 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), este 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 apenas “OCR”. Apesar da significativa evolução na área, 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 "*scanner* 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 et al., 2020). Lideradas por empresas como a IBM para atender às demandas de processamento de dados em massa (como cheques bancários ou 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, se criaram fontes específicas como a OCR-A e a OCR-B, forçando a padronização do texto para se adequar às limitações da máquina (Narang et al., 2020, p. 4). A limitação inerente a essa abordagem, chamadas por Wang et al. (2021, p.6) de *"hand-crafted features",* é 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 reconheça um 'a' independentemente de sua fonte tipográfica ou pequenas variações de impressão, pois a patir 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 et al. (2020) 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 Devanagari.

Para esses contextos, ferramentas modernas como o Tesseract, 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 espraiamento 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, 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 *ground truths* (transcrições de treinamento) de alta fidelidade. 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.

## 1.1. 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 este 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 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, 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 (as 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 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 árabe e hebraico) e padrões de normalização e ordenação

[2] Disponível em <joelonsoftware.com/2003/10/08/the-absolute-minimum-every-software-developer-absolutely-positively-must-know-about-unicode-and-character-sets-no-excuses/>

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

[4] Tengwar é um dos sistemas de escrita élficos do universo de Senhor dos Aneis, de J. R. R. Tolkien. O sistema de escrita atualmente é apenas incluso numa private use area, que é uma banda de pontos de códigos reservada para usos privados; ou seja, para os caracteres estarem legíveis da forma desejada, é necessário instalar uma fonte dedicada a leitura desses pontos de código, caso contrário aparecerão apenas "caixas vazias" no lugar das letras. Desde a poposta de Michael Everson de 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 to U+0160FF do Supplementary Multilingual Plane, disponível em <https://www.unicode.org/roadmaps/smp/smp-16-0-2.html>

(*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 confiável 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 entre diferentes vogais modificadas, como o “e com til” (ẽ, U+1EBD) e o “e com cedilha” (ę, 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 consistente[5] 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, se tem a especificação reserva de uma banda de pontos de código conhecidos como Área de Uso Privado (*Private Use Area, PUA*), como no caso do *tengwar*.

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

Esee é um espaço intencionalmente vago para que projetos específicos, como a *Medieval Unicode Font Initiative* (MUFI[6]), definam seus próprios caracteres e criem fontes capazes de lê-los. O uso da PUA, contudo, implica um dilema fundamental: por 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 adotadas. 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.

## 2. 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*), o projeto recebeu cerca de 8 milhões de euros de fundos da Comissão Europeia entre 2016 e 2019[7]. A plataforma é hoje gerida pela Read-Coop SCE, uma sociedade cooperativa europeia[8] sediada na Universidade de Innsbruck (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 internet[9]. 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 no plano gratuito recebem uma cota mensal de créditos, porém suas tarefas são processadas com menor prioridade. Planos pagos, por sua vez,

---

[6] Disponível em <mufi.info>

[7] Disponível em <cordis.europa.eu/project/id/674943/reporting>

[8] Se trata de uma corporação regulamentada em âmbito específicos da União Europeia, uma *Societas cooperativa Europaea (*SCE*)*, disponível em <

[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 as 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 <transkribus.org/onprem>

oferecem maior volume de créditos, prioridade na fila de processamento e acesso a recursos avançados, como os "super modelos" pré-treinados e *smartseach*. 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".

A principal função do Transkribus é viabilizar o reconhecimento automático de textos, tanto manuscritos quanto impressos, por meio de modelos de inteligência artificial. O treinamento desses modelos depende, como dito, dos dados transcritos corretamente, normalmente feitos manualmente ou corrigidos e analisados por um filólogo; esse conjunto de dados é o que se chama de *Ground Truth* (GT). É a qualidade e o volume deste GT, criado e editado pelo pesquisador, que permite às redes neurais aprenderem a converter imagens de documentos em texto editável com um nível de precisão dificilmente alcançado por ferramentas generalistas. Essa capacidade de treinar modelos personalizados é impulsionada pelo "motor" da plataforma, o PyLaia, um *toolkit* de código aberto que recentemente substituiu a tecnologia anterior, HTR+ (agora descontinuada). Desenvolvido por pesquisadores da Universitat Politècnica de València (PRHLT): Joan Puigcerver (2018) conjuntamente com Carlos Mocholí; e Daniel Martin-Albo e Mauricio Villegas para a tecnologia sua antecessora, o Laia. (Cf. Puigcerver e Mocholí, 2020[10])

O PyLaia a ferramenta responsável é a implementação da arquitetura de IA discutida na seção anterior, que se tornou o padrão para reconhecimento de texto atualmente. Conforme descrito por Wang et al. (2021) e Narang et al. (2020), essa arquitetura de "Quarta Geração" combina as Redes Neurais Convolucionais (CNNs) para a extração de características visuais da imagem, com Redes Neurais Recorrentes (RNNs) para a análise do contexto sequencial do texto. Essa sequência de soluções tecnológicas, implementadas hierarquicamente, permite ao modelo não apenas "ver" um caractere, mas "ler" a palavra, resultando na capacidade de reconhecimento de padrões, mesmo em documentos históricos manuscritos, pela qual o Transkribus tem sido reconhecido.

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

[10] Disponível em <github.com/jpuigcerver/PyLaia/wiki>

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 de uso privado. Essa tendência à não-publicação, 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 uma 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.

Uma das capacidades mais relevantes de plataformas como o Transkribus, como discutido até então, é o treinamento de modelos de HTR altamente especializados. 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*). 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 comunidade para este 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 Trankribus de "modelos de base" (*base models)* servem como ponto de partida para o aprendizado por transferência (t*ransfer 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. Conforme 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 *ground truth*, tornando a transcrição automatizada de alta precisão viável para a grande maioria dos projetos de pesquisa. Conforme afirmam as próprias sessões tutoriais do serviço:

> Depending on the type of material and the number of hands, between 5,000 and 15,000 words (around 25-75 pages) of transcribed material are required to start. In general, the neural networks of the Text Recognition engine learn quickly: the more training data they have, the better the results will be. If you are working on printed material, 5,000 words should be sufficient to achieve a good Character Error Rate

> [CER]. In the case of handwritten documents, our advice is to train the model on at least 10,000 words for each hand. Models trained on a large training data (more than 100,000 words) comprising many hands from the same period and region should be capable of recognising hands not seen in any way during the training: the results, however, will probably be somewhat worse than the Character Error Rate (which is measured on the Validation Data). (Transkribus by READ-COOP SCE [11], 2025)

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, normalemte o *base model* do próprio modelo *fine-tuned* que se pretende criar[12], como ponto de partida para gerar uma primeira versão da transcrição (*pre-labeling*), que é então 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

[11] Disponível em <help.transkribus.org/data-preparation>

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

deste conjunto que o Transkribus calcula a sua principal métrica de performance: a Taxa de Erro de Caracteres (*Character Error Rate*[13], CER). Segundo o manual da plataforma,

> The performance of a model is determined based on the "distance" between a perfect transcription (your Ground Truth) and the automatically recognised text and is measured by the Character Error Rate (CER). The Character Error is the percentage of characters that have been transcribed incorrectly by the Text Recognition model. For example, a 5% CER means that the text model has automatically transcribed correctly 95 characters out of 100, while it has misread only 5 characters.(Transkribus by READ-COOP SCE, 2025)

Para dar alguma noção sobre o desempenho dos modelos, manual do Trankribus estabelece *benchmarks* práticos para o CER dos modelos. Segundo a documentação do Transkribus[14], 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 Early Portuguese Printing (EPP). Embora se tratassem de textos impressos, a vasta extensão diacrônica do *corpus* introduziu uma imensa variabilidade tipográfica e ortográfica. A diferença gráfica entre as gramáticas de Fernão de Oliveira (1536) e de João de Barros (1540), por exemplo, separadas por 4 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, permitindo 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,

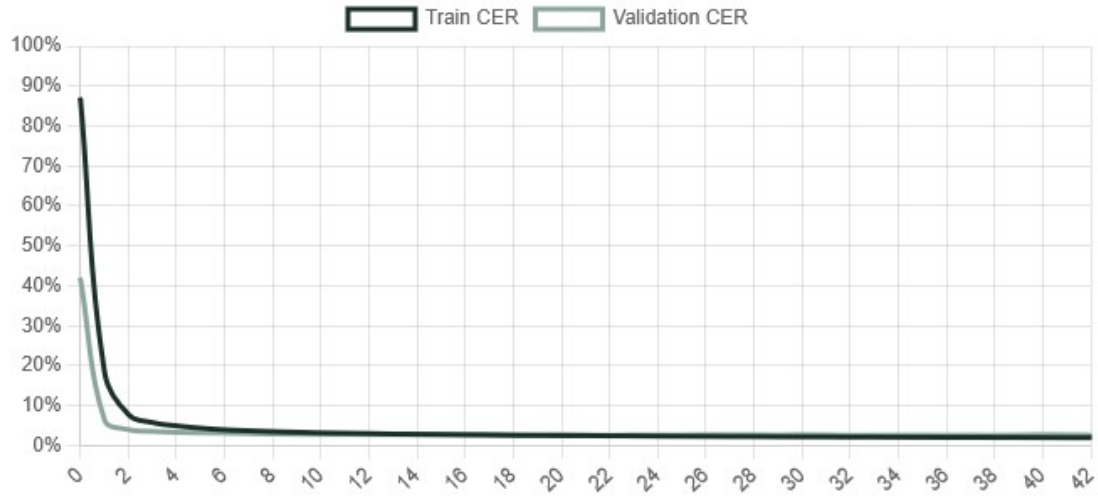
[13] Cabe ressaltar que essa métrica para a taxa de erro de caracteres é conhecida academicamente como "distância de Levenshtein" (Levenshtein, 1966 *apud* Wang *et al.* 2021, p. 22)

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

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, oq ue 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, isso significa que o modelo está apenas "decorando" os dados de treinamento e perdendo a capacidade de generalizar para documentos novos, esse problema é denominado *overfitting*. Quando ambas as linhas de erros permanecem altas, o que indica que o modelo é incapaz de aprender os padrões do corpus, o chamado *underfitting (*Goodfellow *et al,* 2016 *p. 111*.) Verificar a curva de aprendizado de um modelo, portanto, é um passo metodológico necessário, não só para determinar a confiabilidade de um modelo, mas também 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 nos dados de treinamento (*Train CER*), enquanto a linha verde claro indica o erro correspondente nos dados de validação (*Validation CER*):



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 OCR, ele representa um resultado excelente para um modelo generalista treinado num *corpus* diacrônico. 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 treináveis do Transkribus.

## 3. 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 Proueitosas' e dos 'Ilustres & Muito Reuerendos Senhores': uma análise da mudança nos padrões de colocação pronominal clítica em gramáticas portuguesas do século 15 ao 19", defendida em Maio de 2025; o modelo 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% foi publicada em janeiro de 2025.

O treinamento desses modelos se deu a partir da constatação de que faltavam modelos especializados na biblioteca pública de modelos do Transkribus para o tipo de obra necessária na pesquisa. Os modelos voltados para a língua portuguesa, até então, eram os seguintes:

| Nome e ID | Séculos | Autor | Publicação | Num. Palav. | CER | Hand/Print |
|---|---|---|---|---|---|---|
| General Portuguese M1 (44949) | 0-21 | by luciafwx@icloud.com | Sep 27, 2022 | 64.842 | 3,80% | HTR |

| | | | | | | |
|---|---|---|---|---|---|---|
| Transkribus portuguese handwriting M2 (45090) | - | by Transkribus Community | Oct 4, 2022 | 707.803 | 8,00% | HTR |
| SPJCL 17C 4.2 (52754) | - | by dsilver2@tampabay.rr.com | Jun 11, 2023 | 64.324 | 5,60% | PTR |
| Portuguese Handwriting 16th-19th c. (53270 ) | 16-19 | by TraPrInq Project | Jul 5, 2023 | 1.159.586 | 5,20% | HTR |
| XXth century Typewritten Portuguese (56926 ) | 20 | by Natalia C. Salvador | Nov 24, 2023 | 7.468 | 2,60% | PTR |

Como se pode perceber pela tabela acima, 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 (52754) (Spanish & Portuguese Jews Congregation 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 tornava impreciso o uso de outros modelos, por se tratarem de gramáticas e obras de cunho linguístico impressas 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 utilizadas, até então, para a construção do modelo são as seguintes:

1. Grammatica da lingoagem portuguesa, Fernão de Oliveira (c. 1507-1581), pub. 1536
2. Grammatica da língua portuguesa com os mandamentos da santa madre igreja, João de Barros (1496-1570), pub. 1539
3. Grammatica da lingua portuguesa, João de Barros (1496-1570), pub. 1540
4. Regras que ensinam a maneira de escreuer a orthographia da lingua portuguesa: com hum Dialogo que adiante se segue em defensam da mesma lingua, Pêro de Magalhães Gandavo (c. 1540-1579), pub. 1574
5. Orthographia da lingoa portuguesa: obra vtil & necessaria assi pera bem screuer a lingoa Hespanhol como a Latina & quaesquer outras que da Latina teem origem; Item hum tractado dos pontos das clausulas, Duarte Nunes de Leão (c. 1530-1608), pub. 1576
6. Origem da Lingoa portuguesa, Duarte Nunes de Leão (c. 530-1608), pub. 1606
7. Orthographia, ou modo para escrever certo na lingua portuguesa: com hum Trattado de memoria artificial: outro da muita semelhança, que tem a lingua portuguesa com a latina, Álvaro Ferreira de Vera (c. 1590- c. 1645), pub. 1631
8. Promptuario de syntaxe: dividido em duas partes, António Franco, S.J. (1662-1732), pub. 1699

9. Orthographia da Lingua Portugueza, Luís Caetano de Lima, C.R. (1671-1757), pub. 1736
10. Grammatica philosophica e orthographia racional da lingua portugueza, Bernardo de Lima e Melo Bacellar (1736-), pub. 1783
11. Breve tratado de orthographia, João Pinheiro Freire da Cunha (1738-1811), pub. 1792
12. Grammatica nacional, Francisco Júlio Caldas Aulete (1823-1878), pub. 1864
13. Grammatica portugueza elementar, Augusto Epifânio da Silva Dias (1841-1916), pub. 1894

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) parte da 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, 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 aprendizado por transferência (*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 João de Barros (1540) e Fernão de Oliveira (1536), que já contam com versões transcritas disponíveis, especialmente edições de Barros (1540) por José Pedro Machado (1957) e de Oliveira (1536) por Buescu (1975).

A metodologia de transcrição adotada para o treinamento do modelo buscou ser conservadora, tentando manter, na medida do possível, 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, que potencializaria a capacidade de transcrever mais obras; esse equilíbrio implica numa 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.

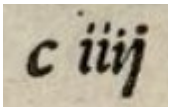
No que tange à segmentação de palavras, se optou por uma intervenção padronizada: vocábulos que aparecem unidos no original foram separados segundo a norma ortográfica atual. Inversamente, nos casos em que erros de impressão separaram graficamente uma única palavra, o espaçamento anômalo foi mantido. 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, hifens e outras marcações que indicam a interrupção de uma palavra para sua continuação na linha subsequente foram igualmente preservadas na transcrição final.

### 3.1 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, serão expostos 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 “I” inicial como “i” com pingo e a representação do último “i” como “j”, conforme ilustrado abaixo:

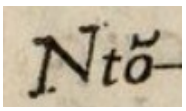
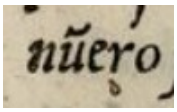
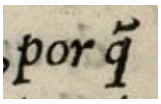
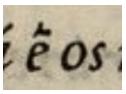
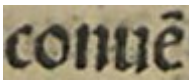
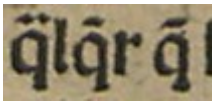
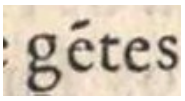
**¶ Capitolo xviij.**



**Ciiij**

O mesmo princípio de preservação foi aplicado às abreviaturas, cuja forma original foi integralmente mantida. Nesse sentido, se decidiu 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

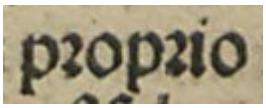
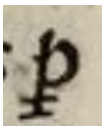
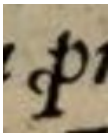
exemplos da tabela abaixo, o tipo-móvel utilizado para representar o "~" provavelmente era um "r" pequeno. Além disso, se respeitou o posicionamento original do til, fosse ele colocado sobre vogais ou sobre consoantes, preservando assim as convenções gráficas observadas nos exemplares de referência, conforme demonstrado nos exemplos a seguir.

| | | | | |
|---|---|---|---|---|
|  |  |  |  |  |
| Ntõ | nũero | por q̃ | ẽ os | conuẽ |
| Nominativo | numero | por que | em os | convem |
|  |  |  |  |  |
| aq̃las | q̃lq̃r q̃ | semp̃ | q̃ | gẽtes |
| aquelas | qualquer que | sempre | que | gentes |

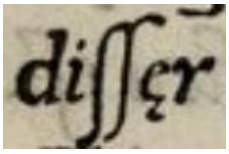
Opções como essas apresentam certos desafios, já que o número o de exemplos para o treinamento precisa ser suficiente para as redes neurais aprenderem a generalizar esse tipo de ocorrência. Além disso, outros modelos do Transkribus, baseados noutras línguas e sistemas grafemáticos, podem influenciar ou causar vícios ao modelo treinado (normalmente a "criação" de um modelo se dá a partir da especialização dum 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 *groud 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 a clareza gráfica, pois a baixa nitidez de tais sinais poderia ser interpretada pelo modelo de transcrição como ruído, comprometendo o processo de treinamento. Se considerou, 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 (ꝛ) 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.

| Cód. Unicode | U+A75B | U+A751 | U+A753 | U+0223 | U+0223 |
|---|---|---|---|---|---|
| Forma Original |  |  |  |  |  |
| Transcr. Adotada | proprio | p | p | d’ | d’s |
| Transcr. Verossímil | pɾopɾio | ꝑ | ꝑ | ȣ[15] | ȣs |

Seguindo o mesmo princípio de fidelidade, todos os diacríticos, como o acento agudo (´), o 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 Ogonek[16]”, 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, o que seria um resultado prejudicial para a precisão de qualquer transcrição em língua portuguesa.

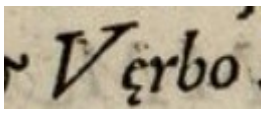
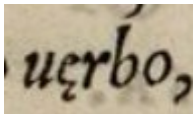
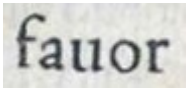
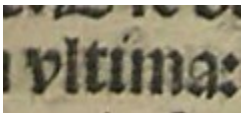
| Cód. Unicode | U+0119 | U+0119 |
|---|---|---|
| Forma Original |  |  |

[15] A análise tipográfica do caractere representado na imagem, presente na obra de Oliveira (1536), permite levantar a hipótese, que aqui adoto, de que se trata de uma adaptação da ligatura grega “ȣ” (ou). A aplicação deste caractere para abreviar palavras com a inicial ‘d’ seria, então, motivada pela provável semelhança gráfica entre a ligatura e o tipo gótico de ‘d’ da prensa de Germão Galharde, impressor e tipógrafo da obra. Uma interpretação alternativa, de que se trataria de uma letra ‘d’ com um til sobrescrito, mostra-se pouco provável, uma vez que a análise do tipo gráfico revela que os elementos acima do círculo não correspondem a uma marca de abreviação, mas sim a um caractere único, provavelmente importado junto com a prensa. Sem outra função evidente na língua portuguesa, este tipo teria sido reaproveitado para abreviar a preposição “de” (com ou sem artigo) e outros vocábulos comuns iniciados por 'd'. Este uso é corroborado por Buescu (1975, p. 38, p. 65), que transcreve a forma “ȣs” como “Deus”..

[16] Se trata de uma letra atualmente utilizada em polonês e lituano.

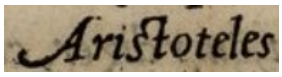
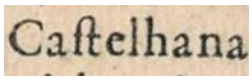
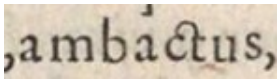
| | | |
|---|---|---|
| Transcr. Adotada | ę | disseę |
| Transcr. Verossímil | ę | diſſęr |

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.

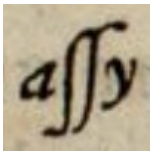
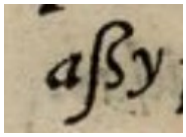
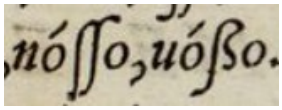
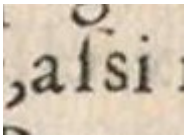
| | | | | |
|---|---|---|---|---|
| Forma Original |  |  |  |  |
| Transcr. Adotada | Vęrbo | uęrbo | fauor | vltima: |
| Transcr. Diplomática | Verbo | verbo | favor | ultima |

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: as consonantais, que cumpriam 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 as vocálicas, 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 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 elas 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 sua correta representação digital das ligaturas por completo.

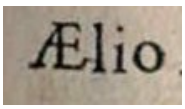
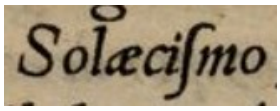
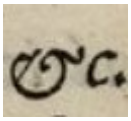
| Cód. Unicode | U+FB06 | U+FB05 | - |
|---|---|---|---|
| Forma Original |  |  |  |
| Transcr. Adotada | Aristoteles | Castelhana | ambactus |
| Transcr. Verossímil | Ariſtoteles | Caſtelhana | Sem correspondência da ligatura no Unicode |

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’ (ſ) 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.

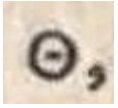
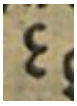
| Cód. Unicode | U+017F | U+00DF | U+00DF | U+0073 |
|---|---|---|---|---|
| Forma Original |  |  |  |  |
| Transcr. Adotada | assy | aßy | nósso,uóßo. | ,assi |

| | | | | |
|---|---|---|---|---|
| Transcr. Verossímil | aſſy | aßy | nóſſo,uóßo. | ,aſsi |

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 em uma única forma sempre que esta aparecia como uma ligatura ou na notação tironiana adotada por Oliveira (1536).

| Cód.Unicode | U+00C6 | U+00E6 | U+0026 | U+0026 | U+204A |
|---|---|---|---|---|---|
| Forma Original |  |  |  |  |  |
| Transcr. Adotada | Ælio | Solæcismo | & | &c | & |
| Transcr. Verossímil | Ælio | Solæcismo | & | &c | 7 |

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.

| Cód.Unicode | U+0398 | U+03A6 | U+025B | U+03C9 |
|---|---|---|---|---|
| Forma Original |  |  |  |  |
| Transcr. Adotada | Θ, | Φ, | ε | ω |

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 (ɑ). Este 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 ɑ 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’.

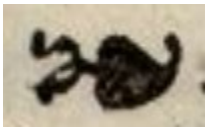
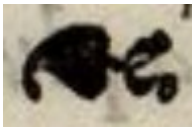
| Cód.Unicode | U+0251 |
|---|---|
| Forma Original |  |
| Transcr. Adotada | ɑ |

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 ocorrência rara.

| Cód.Unicode | U+00B6 | U+2E3F |
|---|---|---|
| Forma Original |  |  |
| Transcr. Adotada | ¶ | ¶ |
| Transcr. Verossímil | ¶ | ⁋ |

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.

| Cód.Unicode | U+2767 | U+2619 | U+2766 |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Forma Original |  |  |  |
| Transcr. Adotada | ꝺ | ꝺ | ꝺ |

## 4. Considerações Finais

O presente trabalho demonstrou a viabilidade e a eficácia da criação de modelos de reconhecimento de texto especializados para *corpora* históricos. Por mieo de um processo de treinamento iterativo na plataforma Transkribus, o modelo Early Portuguese Printing (EPP) alcançou uma taxa de erro de caracteres (CER) de 2,58% em um *corpus* diacrônico de alta complexidade. Esse resultado comprova que a "Quarta Geração" de tecnologias de OCR, baseada em *redes neurais profundas*, permite que pesquisadores individuais ou pequenas equipes superem o tradicional gargalo da transcrição manual, viabilizando análises quantitativas em larga escala que antes eram pouco concebíveis.

Contudo, a adoção de plataformas como o Transkribus, apesar de seus inegáveis benefícios técnicos, suscita importantes questionamentos sobre a infraestrutura da pesquisa digital. A dependência de uma solução centralizada e gerida por uma cooperativa europeia privada, na qual os dados e modelos que representam o patrimônio documental e científico brasileiro são processados e armazenados, levanta o debate sobre a soberania digital. A sustentabilidade a longo prazo e a autonomia da pesquisa nacional podem depender da criação de políticas e infraestruturas locais que incentivem o desenvolvimento de tecnologias de código aberto, como o motor PyLaia, e promovam o treinamento de modelos em repositórios nacionais, com tecnologias abertas e gratuitas, para fomento da ciência pública e gratuita.

Ainda assim, o trabalho descrito sobre o modelo EPP aponta para um futuro promissor. Mais do que uma ferramenta de transcrição, ele representa um ativo de pesquisa que pode ser continuamente aprimorado e aplicado a novos acervos, acelerando o trabalho de filólogos, historiadores e linguistas e permitindo que mais e novos modelos sejam criados e permitam o avanço das humanidades digitais aos vestígios escritos do passado.

## 5. Referências Bibliográficas

AL KENDI, W.; GECHTER, F.; HEYBERGER, L.; GUYEUX, C. Advancements and Challenges in Handwritten Text Recognition: A Comprehensive Survey. **Journal of Imaging**, [S. l.], v. 10, n. 1, p. 1-18, 2024. DOI: 10.3390/jimaging10010018.

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

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

BAUDRY, H.; PEDRO, S. T. Humanidades digitais e documentos históricos: transcrever, catalogar,ar. **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, Y. Learning deep architectures for AI. **Foundations and Trends in Machine Learning**, Hanover, v. 2, n. 1, p. 1-127, 2009. DOI: 10.1561/2200000006.

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

GOODFELLOW, I. et al. **Deep learning**. Cambridge: MIT press, 2016.

LEÃO, D. N. de (fl. 1530-1608). **Origem da Lingoa portuguesa**. Lisboa: Pedro Crasbeeck, 1606. Disponível em: https://purl.pt/50. Acesso em: 4 set. 2025.

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

NARANG, S. R.; JINDAL, M. K.; KUMAR, M. Ancient text recognition: a review. **Artificial Intelligence Review**, [S. l.], v. 54, p. 1-28, 2021. DOI: 10.1007/s10462-020-09827-4.

OLIVEIRA, F. de (c. 1507-1581). **Grammatica da lingoagem portuguesa**. Lisboa: Germão Galharde, 1536. Disponível em: https://purl.pt/120. Acesso em: 4 set. 2025.

PUIGCERVER, J. **A probabilistic formulation of keyword spotting**. 2018. Tese (Doutorado) - Universitat Politècnica de València, València, 2018.

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

SPOLSKY, J. **The Absolute Minimum Every Software Developer Absolutely, Positively Must Know About Unicode and Character Sets (No Excuses!)**. [S. l.], 8 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: 4 set. 2025.

WANG, H.; PAN, C.; GUO, X.; JI, C.; DENG, K. From object detection to text detection and recognition: a brief evolution history of optical character recognition. **Wiley Interdisciplinary Reviews: Computational Statistics**, [S. l.], 2021. DOI: 10.1002/wics.1547.

WANG, J. A study of the OCR development history and directions of development. In: **Highlights in Science, Engineering and Technology**. [S. l.], v. 72, p. 409, 2023. DOI: 10.54097/bm665j77.

## 6. 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.

## 7. Declaração de Disponibilidade de Dados de Pesquisa

Todos os dados relevantes gerados ou analisados durante este estudo estão incluídos neste artigo pré-publicado.

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença Creative Commons CC-BY.
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.