15º Congresso BAD

PAPER

Acelerar a História: a Inteligência Artificial na transcrição de manuscritos

José Marques[a], Catarina Serafim[a]

[a]Arquivo Histórico Parlamentar Expediente e Gestão Documental, Assembleia da República, Portugal, jose.carlosmarques@ar.parlamento.pt, catarina.serafim@ar.parlamento.pt

Resumo

O Arquivo Histórico Parlamentar, Expediente e Gestão Documental (AHPEGD) iniciou em 2024 um projeto de transcrição automática de documentos históricos com recurso à plataforma Transkribus. Trata-se de um *software* que recorre a tecnologias de Inteligência Artificial para reconhecer caracteres manuscritos e proceder à sua transcrição para ficheiros de texto. No decurso deste projeto, foi criado um modelo próprio de transcrição automática do AHPEGD, que resultou de várias sessões de treino do algoritmo. Esta valência está agora a ser incorporada na base de dados do arquivo, com o objetivo de tornar os documentos pesquisáveis e auxiliar na criação de descrições detalhadas dos documentos. O âmbito deste projeto é um lote de cerca de 30 mil documentos manuscritos dos séculos XIX e início do séc. XX. Trata-se de cartas, requerimentos, representações e petições enviados à Câmara dos Deputados durante o período da Monarquia Constitucional (1821-1910).

Palavras-chave: Inteligência Artificial, Transkribus, transcrição automática, documentos históricos, HTR

Introdução

A paleografia, entendida como a ciência que se ocupa do "estudo e conhecimento das escritas antigas" (Gomes, 2018), conheceu nos últimos anos uma evolução muito acentuada. Desde logo por uma extensão do próprio campo de estudo. Se, "na segunda metade do século XX, essas escritas antigas, para quem estudava Paleografia, terminavam nos finais do século XVIII, (...) naturalmente que as escritas oitocentistas e novecentistas se têm de integrar na longa evolução milenar da transmissão da escrita", completa o autor. A par deste alargamento, o advento de novas tecnologias no final do século XX permitiu levar a disciplina para um novo patamar, há muito desejado por arquivistas, bibliotecários e leitores. A análise computacional de textos impressos começou por se focar no reconhecimento de caracteres em letra de imprensa através da imagem (*Optical Character Recognition*). A pré-existência de padrões constantes nesta forma de escrita permitiu atingir rapidamente um grau de fiabilidade muito elevado no reconhecimento de texto, sobretudo a partir da década de 1990, quando surgiram no mercado diversos produtos comerciais.

Já o reconhecimento de textos manuscritos teve uma evolução muito mais lenta, dadas as dificuldades da tarefa. As primeiras tentativas procuraram emular os avanços já conseguidos com a tecnologia OCR, com o desenvolvimento de modelos de deteção de padrões (primeiro de caracteres individuais, depois de palavras). Os resultados não impressionaram. O grande salto dá-se com a explosão das tecnologias de Inteligência Artificial (IA) que aconteceu já no século XXI, sobretudo na última década. O desenvolvimento

https://doi.org/10.48798/congressobad.3195

do conceito de *Deep Learning* (que utiliza redes neurais artificiais multicamadas para imitar a inteligência humana, aprendendo representações automaticamente a partir de enormes conjuntos de dados não estruturados) potenciou o desenvolvimento das tecnologias HTR (*Handwritten Text Recognition – reconhecimento de texto manuscrito*). Em termos conceptuais, o reconhecimento de escrita manual consiste na tarefa de transformar uma língua representada na sua forma espacial de sinais gráficos na sua representação simbólica. Nas línguas baseadas no alfabeto latino, esta representação simbólica é normalmente a representação ASCII de 8 bits dos caracteres (Plamondon & Srihari, 2000).

O aparecimento de modelos que se alimentam de milhões de documentos permite à "máquina" identificar padrões de escrita a uma escala até aqui impossível. É a aprendizagem destes padrões que permite, com um grau de certeza incomparavelmente superior ao de um passado recente, inferir qual a palavra manuscrita presente num documento. Os resultados impressionam: "enquanto há algumas décadas o sonho do HTR era ainda, em grande medida, uma fantasia promissora (mas uma área de investigação significativa), tem-se desenvolvido rapidamente. Evoluiu de uma capacidade limitada de reconhecer apenas caligrafias extremamente claras com um corpus de treino muito extenso, passando por sucessivas fases de aperfeiçoamento, e consegue agora reconhecer textos manuscritos bastante complexos com quantidades significativamente menores de dados de treino. Estas melhorias são evidentes no aumento da precisão, velocidade e diversidade dos textos manuscritos que os sistemas HTR podem agora processar, frequentemente com taxas de erro de caracteres (CER) muito baixas" (Cummings, 2025).

Este desenvolvimento tecnológico teve também como consequência uma redução exponencial dos custos do processo de transcrição. Os novos produtos comerciais disponíveis no mercado, muitos deles em formato de *open source*, têm facilitado o acesso a estas tecnologias, que permitem a utilizadores individuais e institucionais lançarem-se na tarefa de transcrever documentação histórica, dando-lhes o que se pode qualificar como uma nova vida. De facto, "a democratização do acesso à informação tem sido conseguida através da disponibilização de conteúdos on-line, acessíveis 24 horas por dia, todos os dias, a quem tenha acesso a um computador com Internet. A crescente utilização de plataformas colaborativas da Web 2.0 ou de IA para a transcrição massiva de documentos têm dado um novo uso à paleografia, cujo conhecimento é essencial para a leitura de manuscritos" (Silva, 2025).

O que resulta da transição de um documento histórico digitalizado para um documento histórico digitalizado e transcrito é um novo leque de possibilidades no que diz respeito ao conhecimento dos acervos documentais. A tecnologia HTR permite que um leque mais vasto de dados se torne estruturado, localizável, pesquisável e legível, facilitando a sua utilização tanto por pessoas como por algoritmos (Nockels et al., 2024). As vantagens aplicam-se tanto do lado do serviço detentor da informação, a quem são dadas novas ferramentas de recuperação e descrição do acervo histórico à sua guarda, como, sobretudo, para o cliente, que vê imensamente facilitadas as tarefas de localizar, selecionar e analisar a informação relevante para os fins pretendidos. O desenvolvimento tecnológico permite às instituições uma maior eficiência no cumprimento da sua missão. Isto porque "as designadas instituições de memória (arquivos, bibliotecas e museus) fornecem, para além das funções de salvaguarda, organização e descrição, acesso à informação. Esta última será, porventura, a sua função mais importante visto que permite a recuperação de informação relevante quer para a própria organização quer para os investigadores externos" (Borges & Silva, 2018).

Estudos de caso

A ideia de aplicar no AHPEGD a transcrição automática de documentos a partir de tecnologia HTR surgiu do conhecimento que tivemos de outros projetos que utilizam este recurso. Em novembro de 2023, o Parlamento Europeu organizou um seminário sobre IA na sua sede no Luxemburgo. Um dos projetos apresentados foi o que está a ser desenvolvido desde 2019 pelo Arquivo Nacional da Hungria, sobre prisioneiros de guerra na II Guerra Mundial. Mais concretamente, o projeto pretende ajudar a esclarecer o que aconteceu às centenas de milhares de pessoas que foram detidas pelos exércitos russos, sendo que muitas delas foram deportadas para a União Soviética, onde permaneceram detidas e submetidas a trabalho forçado.

Um acordo estabelecido com o estado russo permitiu ao arquivo húngaro ter acesso a 682 mil fichas de prisioneiros de guerra, capturados pelas tropas russas durante a II Guerra Mundial. De notar que estas fichas foram escritas à mão, em língua russa, usando o alfabeto cirílico, com a dificuldade extra de muitos dos nomes húngaros terem sido transcritos para o russo por soldados soviéticos, havendo por isso muitos erros de grafia. Podemos facilmente imaginar as dificuldades técnicas e humanas que o tratamento desta massa de informação traria a qualquer arquivo. A resposta a este desafio envolveu o Centro Húngaro de Pesquisa de Linguística (NYTK) e o contributo inestimável da IA, nomeadamente com o recurso ao software Transkribus. Através do treino intensivo de um novo modelo de transcrição, alimentado com milhares de documentos, foi possível criar um algoritmo capaz de traduzir e corrigir os nomes do russo original para o húngaro, assim como a informação de contexto contida nas fichas individuais dos antigos prisioneiros de guerra.

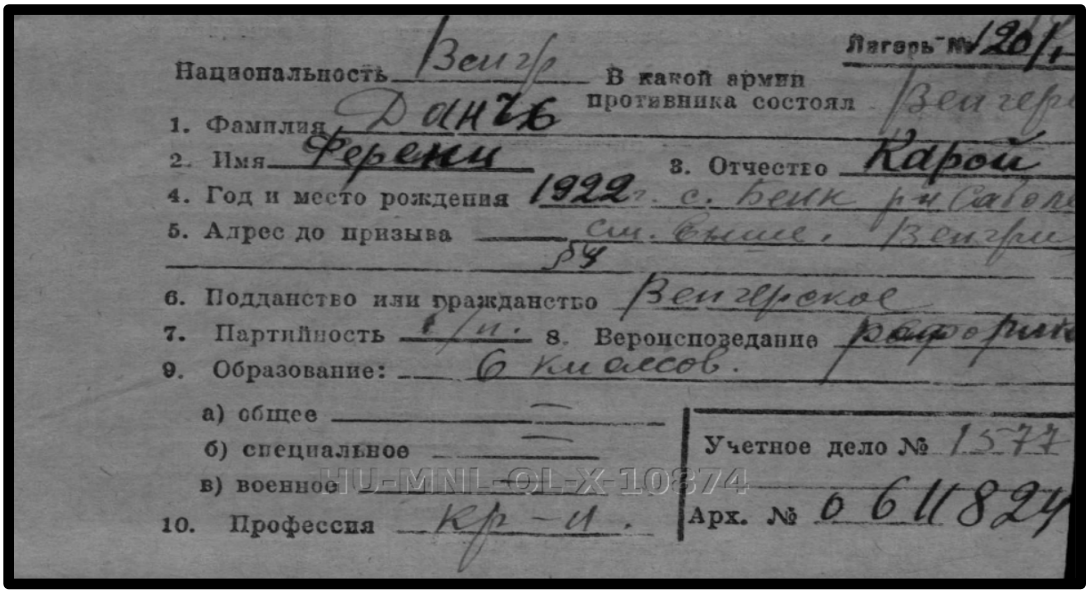


Figura 1: Ficha de um prisioneiro de guerra húngaro num campo de detenção da União Soviética.

O resultado deste esforço foi a criação, em poucos meses, de uma base de dados pública[1], disponível *online* e de acesso gratuito, onde se pode pesquisar pelos nomes das pessoas feitas prisioneiras pelo exército russo, saber das circunstâncias da sua detenção e qual o seu destino dentro do sistema prisional da União Soviética. A base informa quais os dados obtidos por processo automático, dando aos utilizadores a oportunidade de sugerir correções.

No mesmo seminário, a Universidade do Luxemburgo apresentou um projeto que tem como base a transcrição de registos, no caso os da imigração. Partindo dos formulários que todos os cidadãos que pretendiam entrar no Luxemburgo tinham de preencher à chegada (com o âmbito temporal de final do séc. XIX ao séc. XX), foi possível construir uma base de dados que permite pesquisar por nomes, nacionalidades, país de origem, motivos invocados para a imigração, etc. Tal como no caso húngaro, a equipa luxemburguesa usou o software Transkribus com dois objetivos: criar um template do formulário, para dele retirar os dados segmentados por categoria e reconhecer o texto manuscrito, transpondo-o para a forma digital (Relicovschi, 2026).

Em Portugal, tivemos conhecimento do Projeto TraPrInq[2], uma iniciativa que juntou académicos de várias universidades e centros de investigação do nosso país, para criar um modelo de transcrição baseado nos registos dos processos da Inquisição Portuguesa, produzidos entre 1536 e 1821. Este projeto resultou na criação do modelo *Portuguese Handwriting 16th-19th c.*, aquele que é, neste momento, o modelo de transcrição para a língua portuguesa mais abrangente entre os que estão disponíveis na plataforma da Transkribus, e o que serviu de base aos primeiros ensaios do projeto de transcrição do AHPEGD.

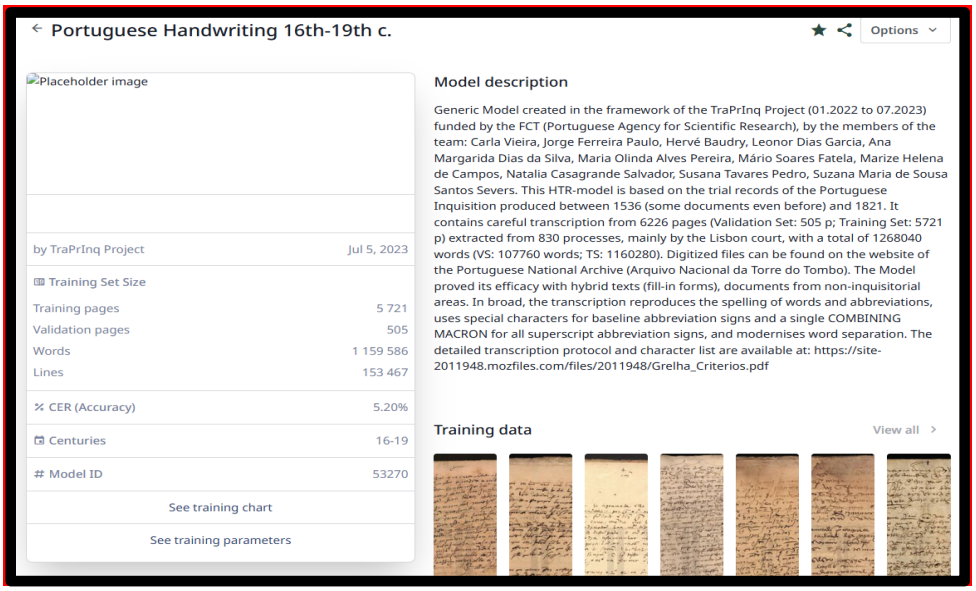


Figura 2: Ficha técnica do modelo de transcrição Portuguese Handwriting 16th-19th c, desenvolvido no âmbito do projeto TraPrInq.

[1] Site do projeto do Arquivo Nacional da Hungria sobre prisioneiros de guerra na União Soviética (https://adatbazisokonline.mnl.gov.hu/adatbazis/szovjet-taborok-magyar-foglyai).
[2] Site do Projeto TraprInq ().https://traprinq.mozellosite.com/

Quando começámos a dar os primeiros passos do nosso projeto, contactámos a equipa que conduziu o projeto TraPrInq, nomeadamente o professor Hervé Baudry, que nos deu uma importante ajuda na compreensão dos procedimentos a adotar para a transcrição de documentos e para o treino de um novo modelo de transcrição.

Método

No início de 2024, o Arquivo Histórico Parlamentar – AHPEGD iniciou o seu projeto para a transcrição automática de documentos manuscritos, utilizando tecnologia de IA. O objeto do projeto é um lote de cerca de 30 mil documentos históricos, digitalizados em formato PDF. Estes documentos estão identificados e descritos ao nível da série documental, com cota na base de dados do arquivo, mas nunca foram descritos individualmente, o que dificulta a sua localização e recuperação, quando necessárias. Neste conjunto documental encontramos cartas, requerimentos, representações e petições enviados à Câmara dos Deputados durante o período da Monarquia Constitucional (1821-1910) por cidadãos particulares e por instituições. Os documentos abrangem as mais variadas temáticas, sendo, na sua maioria das vezes, pedidos sobre assuntos do interesse direto dos remetentes, tais como impostos, atribuição de pensões e compensações financeiras a pessoas, progressão nas carreiras do aparelho do Estado, organização administrativa do território, pedido de realização de obras públicas, etc. Este projeto tem como objetivo a produção de transcrições destes documentos, com vista a facilitar a localização dos originais em depósito e apoiar o respetivo processo de descrição documental

O projeto de transcrição automática de documentação tem como base tecnológica a plataforma Transkribus, desenvolvida pela cooperativa Read Coop[3] (sedeada em Innsbruck, na Áustria, e que congrega instituições universitárias e culturais de toda a Europa). Consiste num software de transcrição automática de manuscritos, com base numa aplicação disponível *online*. O texto obtido pode ser corrigido na própria plataforma e depois exportado em diferentes formatos (PDF, TXT, DOC, etc.). Em complemento a estas funcionalidades de base, a versão paga de subscrição da plataforma disponibiliza aos utilizadores uma API (*Application Programming Interface*) que faz ponte entre o servidor do Transkribus e a base de dados do AHPEGD, permitindo o envio em massa de documentos para transcrição automática.

Havendo outras soluções no mercado com características próximas das que o Tranksribus oferece, a escolha desta plataforma deve-se a vários fatores. A facilidade de utilização, aliada à possibilidade de utilização inicial sem ser necessário subscrever a versão paga, levou-nos a explorar a plataforma. Outra componente importante é o espírito colaborativo, que leva a que os modelos criados pelos utilizadores fiquem, se estes assim o autorizarem, disponíveis para serem utilizados por todos. A relativa facilidade com que se podem criar modelos próprios de transcrição, a partir das coleções de cada utilizador, constituiu outro dos fatores que nos levou a optar por esta plataforma. Uma vez que esta tem algumas especificidades técnicas que não dominávamos, foi necessário realizar reuniões com os responsáveis técnicos e editoriais da Read Coop. Assim, realizaram-se várias sessões por videoconferência, durante as quais foram expostas as principais funcionalidades da plataforma, bem como as possibilidades de ajuste do modelo de transcrição às características da documentação e aos nossos objetivos.

[3] Site da plataforma Transkribus e da cooperativa Read Coop (https://www.transkribus.org/about).

Resultados

Os primeiros testes de aplicação aos documentos do AHPEGD dos modelos de transcrição disponíveis na plataforma Transkribus revelaram-se pouco eficazes. No início de 2024, tanto os "supermodelos" oferecidos a todos os utilizadores, supostamente preparados para transcrever a maioria das línguas europeias, como os dois modelos específico de língua portuguesa, não funcionavam com os documentos do nosso fundo documental. As peculiaridades da escrita manual do século XIX, não só na forma de desenho das letras, mas também da pontuação não eram reconhecidas pelos modelos disponíveis (situação que, entretanto, sofreu uma evolução muito significativa no ano de 2025).

Em 2024 existiam dois modelos na plataforma para a língua portuguesa (*Transkribus Portuguese handwriting M2* e *General Portuguese M1*), mas a sua utilidade para a documentação em causa revelou-se praticamente nula. Os textos obtidos nesta primeira experiência eram praticamente ilegíveis (Figura 3).

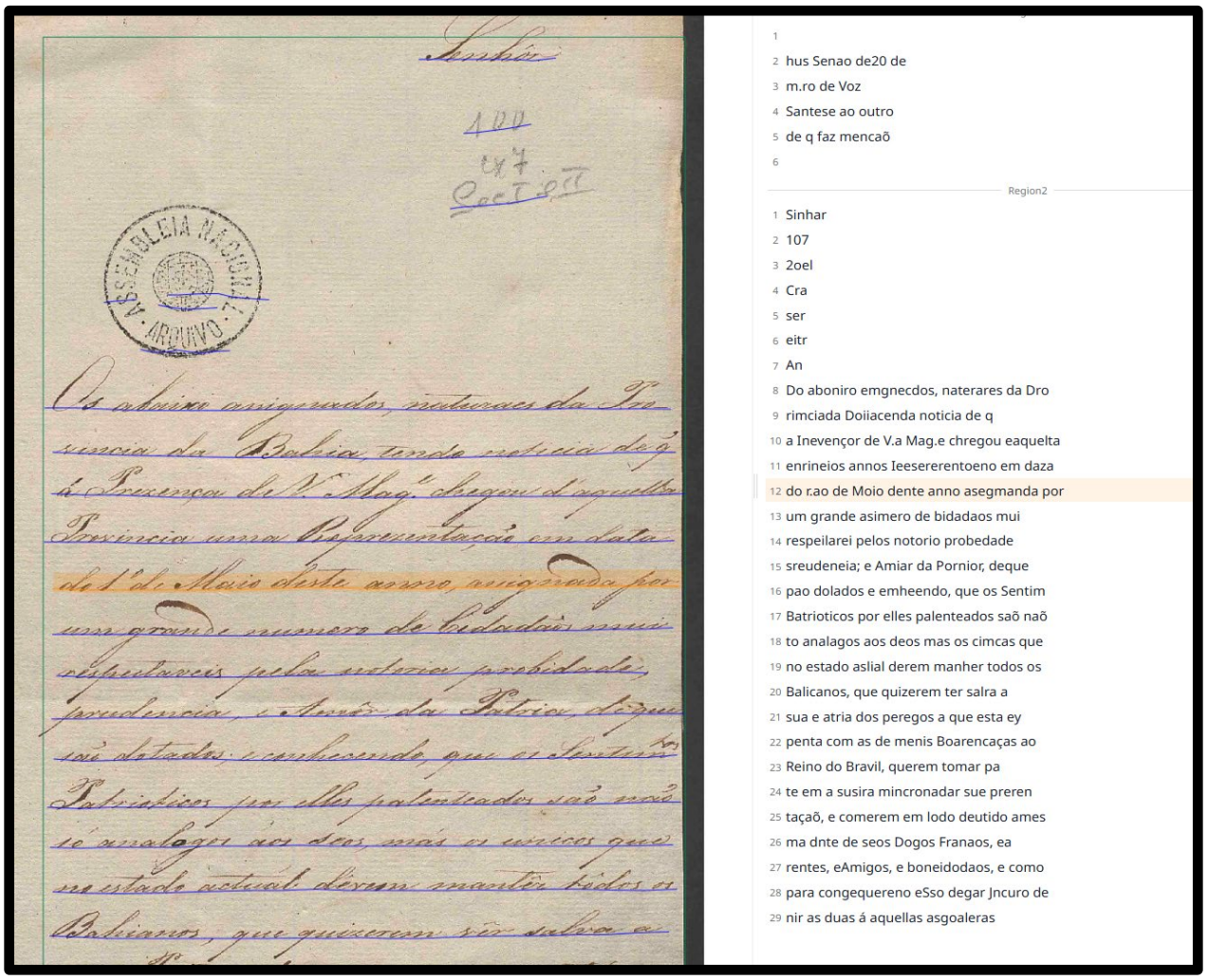
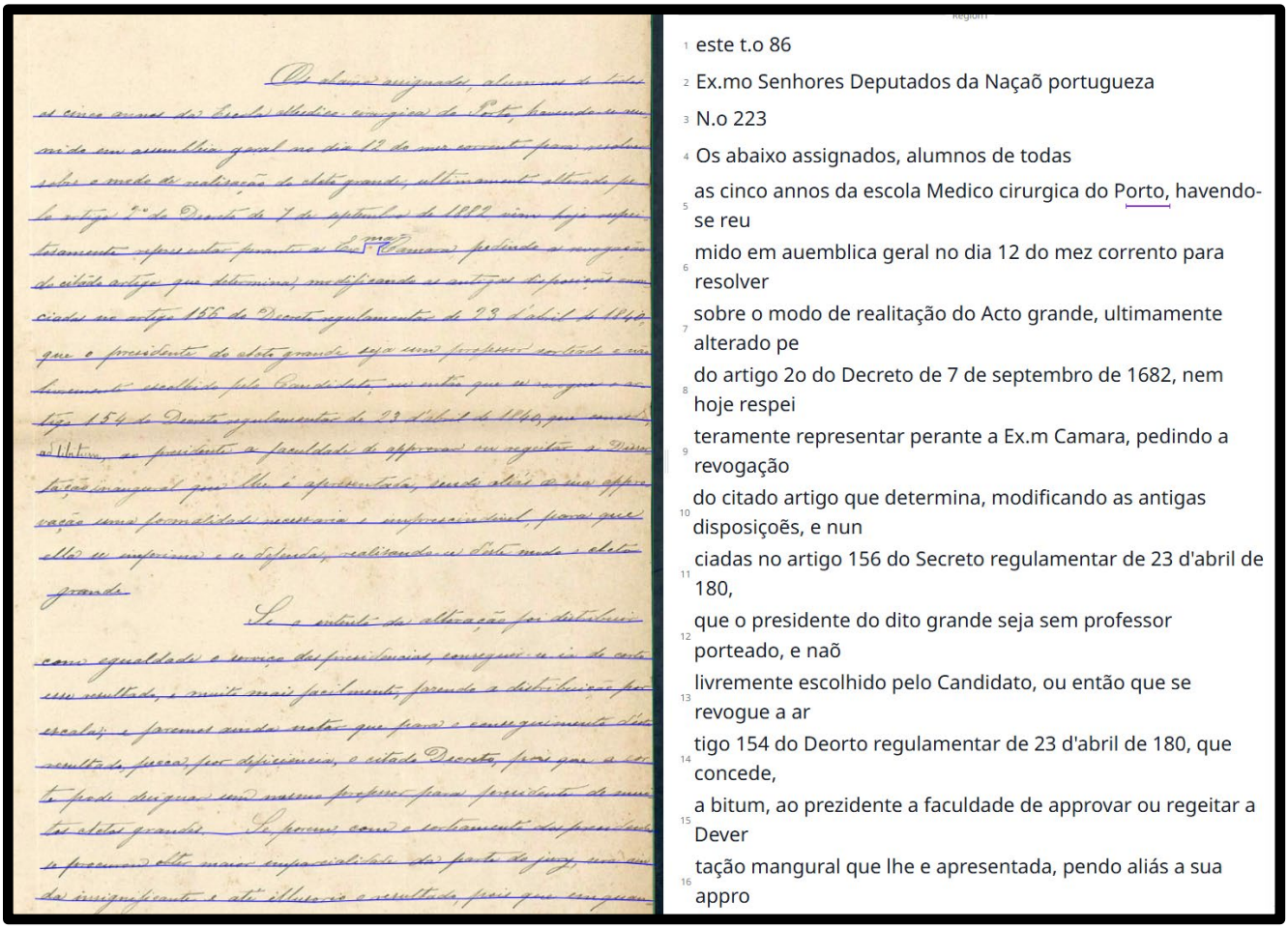


Figura 3: Exemplo de transcrição de um documento com o modelo General Portuguese M1, da Transkribus.

Decidimos então adotar o modelo do projeto TraPrInq, de acesso público na mesma plataforma, que nos permitiu transcrever os documentos do AHPEGD com um grau de fiabilidade muito razoável. Este modelo tornou-se a base inicial do projeto. As transcrições geradas permitiram-nos obter textos que seriam depois submetidos a correção manual na plataforma. Nesta fase, os textos processados pelo algoritmo continham ainda muitos erros, mas, na maioria dos casos, eram percetíveis elementos cruciais como os nomes dos autores das missivas enviadas à Câmara dos Deputados ou o tema principal da comunicação. A maior dificuldade que registámos foi a identificação de algarismos, o que comprometia a identificação de datas ou de valores monetários constantes na documentação.



este t.o 86
Ex.mo Senhores Deputados da Naçaõ portugueza
N.o 223
Os abaixo assignados, alumnos de todas
as cinco annos da escola Medico cirurgica do Porto, havendo-se reu
mido em auemblica geral no dia 12 do mez corrento para resolver
sobre o modo de realitação do Acto grande, ultimamente alterado pe
do artigo 2o do Decreto de 7 de septembro de 1682, nem hoje respei
teramente representar perante a Ex.m Camara, pedindo a revogação
do citado artigo que determina, modificando as antigas disposiçoẽs, e nun
ciadas no artigo 156 do Secreto regulamentar de 23 d'abril de 180,
que o presidente do dito grande seja sem professor porteado, e naõ
livremente escolhido pelo Candidato, ou então que se revogue a ar
tigo 154 do Deorto regulamentar de 23 d'abril de 180, que concede,
a bitum, ao prezidente a faculdade de approvar ou regeitar a Dever
tação mangural que lhe e apresentada, pendo aliás a sua appro

Figura 4: Transcrição de um documento com o modelo Portuguese Handwriting 16th-19th (do projeto TraprInq). De realçar os erros nos valores numéricos transcritos.

Tem do em conta as lacunas destas primeiras transcrições, apercebemo-nos de que haveria margem de progressão para se criar um modelo de transcrição próprio. Isto porque, por coincidência histórica, a Inquisição terminou em Portugal precisamente durante os trabalhos das Cortes Gerais, Extraordinárias e Constituintes de 1821-22, que deram origem à primeira Constituição portuguesa e à abertura da primeira assembleia parlamentar em Portugal, então designada como Câmara dos Deputados. Assim, a documentação histórica à guarda do AHPEGD tem origem cronológica precisamente quando terminam os

documentos da Inquisição. Esta coincidência de datas revelou-se problemática, uma vez que notámos grandes mudanças na grafia dos manuscritos desde o fim da Inquisição para as décadas seguintes, o que explica o elevado número de erros ao aplicar este modelo aos documentos do nosso arquivo. Chegámos à conclusão de que havia espaço para melhorias significativas.

Assim, durante um ano, a nossa equipa iniciou um processo de transcrição integral de cerca de 200 documentos, com o objetivo de treinar o algoritmo do Transkribus e criar um modelo de transcrição. No decorrer deste processo, realizámos duas sessões de treino com um número relativamente reduzido de documentos e de páginas transcritas, com resultados modestos. As taxas de erro destes dois modelos iniciais (ver Figura 5) revelaram-se bastante elevadas, pelo que continuámos a acumular transcrições integrais de documentos, com o objetivo de realizar novos treinos de modelos.

A terceira sessão de treino foi realizada em abril de 2025, com base em 178 documentos. Foram transcritas um total de 1 063 páginas manuscritas, o que corresponde a 186 342 palavras e 27 008 linhas de texto. O resultado desta nova sessão revelou-se muito positivo e levou à criação do modelo *AHP Handwritten Portuguese 19th-20th Centuries*. Este modelo tem uma Taxa de Erro de Caracteres (CER – *Character Error Rate*) de 2,53%, o que é muito significativo, dado que a própria plataforma considera fiáveis os modelos que conseguem atingir uma CER de até 8%. Uma das grandes melhorias que notámos com a criação de um modelo próprio foi a identificação de algarismos e de valores numéricos, sobretudo os de natureza monetária, que, anteriormente, surgiam invariavelmente errados, muitas vezes até sem serem identificados como valores numéricos. Este modelo é agora público, o que significa que está disponível para todos os utilizadores da plataforma Transkribus.

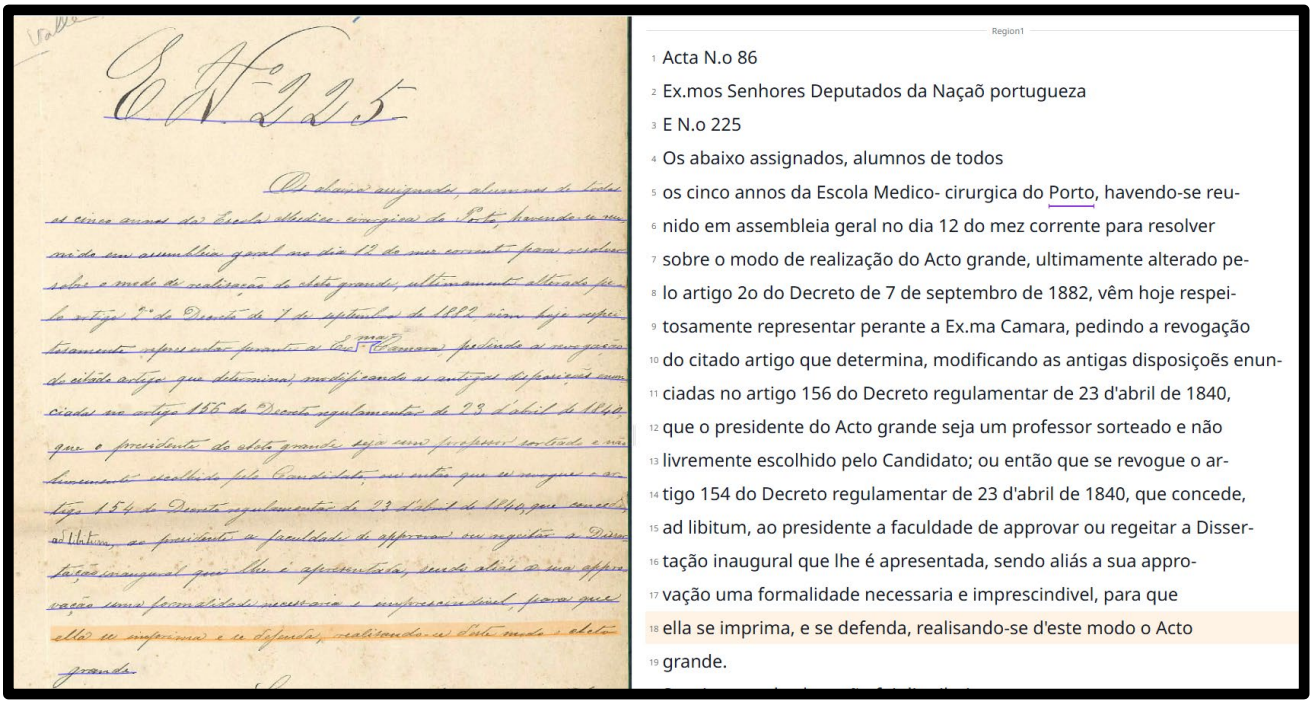


Figura 5: Transcrição de um documento com a versão final do modelo AHP Handwritten Portuguese 19th-20th Centuries.

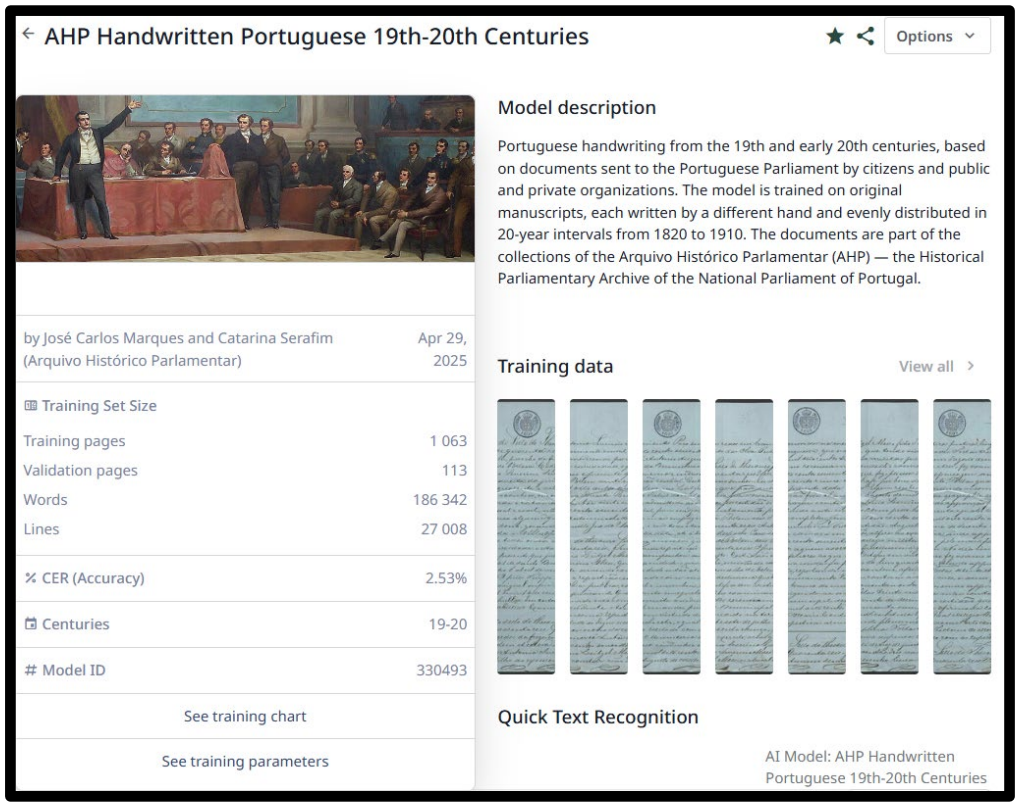


Figura 6: Ficha técnica do modelo de transcrição criado pelo AHPEGD na plataforma Transkribus.

Discussão

Dadas as limitações de tempo e de recursos, optámos por construir um modelo que compreendesse a totalidade do período da Monarquia Constitucional (1821-1910). Constatámos, no entanto, que as variações na escrita num período tão dilatado justificariam a criação de modelos mais circunscritos no tempo (consideramos razoável o estabelecimento de intervalos de 30 a 40 anos). Será um trabalho a desenvolver num futuro próximo.

O AHPEGD encontra-se neste momento a desenvolver uma segunda fase do projeto, que consiste em incorporar o Transkribus na base de dados do arquivo histórico, através da API disponibilizada pelo programa. O arquivo fez uma subscrição organizacional do programa, que dá acesso a esta ferramenta. A implementação das transcrições automáticas na base de dados prevê, para já, diferentes níveis de uso entre o *backoffice* e o *front office*, disponível para o público, não estando prevista, para já, a disponibilização pública das transcrições automáticas. Nesta fase do projeto, o objetivo não é tanto obter um texto totalmente legível – estamos cientes de que ainda é muito difícil evitar erros significativos –, mas antes tornar o conteúdo dos ficheiros digitalizados legível e pesquisável.

Foi criado um fluxo que permite enviar um documento em formato PDF para uma pasta localizada no servidor do serviço. Este ficheiro é então automaticamente remetido ao Transkribus para transcrição, sendo depois publicados na base de dados tanto o ficheiro PDF original como o texto dele extraído. Este texto serve, numa primeira fase, para apoiar o cumprimento de dois objetivos: permitir a recuperação dos documentos através de pesquisas em linguagem natural e auxiliar a criação manual de descrições documentais. Este processo semiautomático de transcrição começou em janeiro de 2026. No início de maio

de 2026, tinham sido transcritos 7.009 documentos, o que corresponde a 27.708 páginas de texto manuscrito.

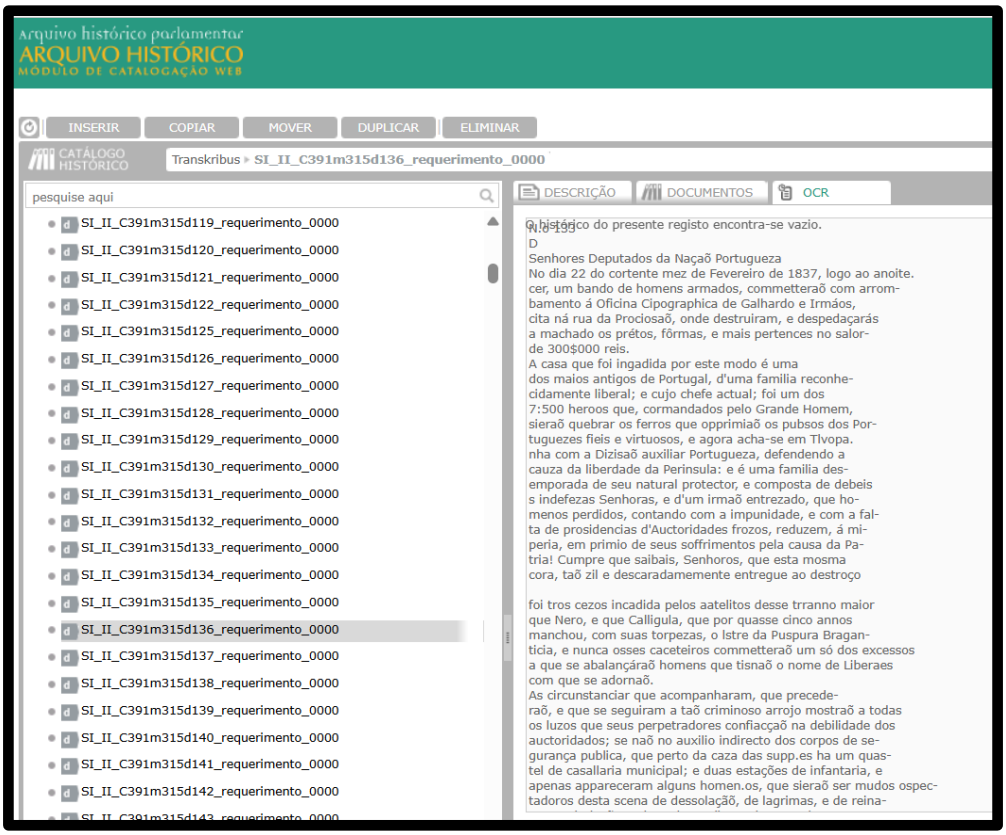


Figura 7: Documento transcrito automaticamente no backoffice da base dados do Histórico do AHPEGD.

A médio prazo, com um eventual aperfeiçoamento do modelo do AHPEGD ou recorrendo aos modelos cada vez mais precisos que são disponibilizados na plataforma, pretendemos chegar então a um ponto em que seja possível aos utilizadores externos usar esta tecnologia para obter transcrições de documentos *on demand*, a partir dos ficheiros digitais da nossa base de dados.

Outras experiências realizadas, com resultados significativos, consistem na aplicação de outros motores de IA, como o Gemini ou o Chat GPT, para corrigir as transcrições automáticas do Transkribus.

O potencial que os diferentes sistemas de IA oferecem aos utilizadores no tratamento da informação está a operar mudanças profundas em todos os campos do saber, fenómeno ao qual os arquivos não escapam. Mas é frequente acontecer que, com o entusiasmo dos resultados obtidos, nos esqueçamos dos problemas que os usos destas tecnologias trazem. Desde logo, é fácil tomarmos como certas as informações que nos são dadas, no caso, as transcrições que o programa entrega. Será sempre necessário confrontar o texto obtido automaticamente com o original manuscrito, procurando aferir da sua fiabilidade.

Num outro aspeto, também é frequente ignorar-se o impacto ambiental que o uso da IA tem. Quando processamos um documento ou, sobretudo, quando treinamos um modelo de transcrição, estamos a ativar computadores (provavelmente centenas) alojados em *data centers* que consomem quantidades muito

significativas de energia elétrica e água. Uma estimativa do custo ambiental do treino de um modelo de rede neuronal durante 24 horas aponta que se produz um nível de emissões de carbono semelhante ao de um voo transatlântico (Nockels et al., 2024). Recomenda-se, assim, um uso racional da tecnologia, procurando-se evitar exageros na transcrição de documentos que poderão ter escassa ou nula importância para o fim pretendido.

Conclusões

O projeto do AHPEGD revelou a utilidade das tecnologias de Inteligência Artificial na transcrição automática de textos manuscritos. Embora os textos obtidos a partir do modelo criado pelo AHPEGD ainda contenham muitos erros, estes representam uma significativa melhoria nos processos de localização e descrição dos documentos. Estando esta tecnologia ainda em fase de aperfeiçoamento, é de esperar que futuras versões de programas como o Transkribus possam trazer resultados ainda mais apurados. O exemplo da nossa experiência mostra que é possível, com uma equipa relativamente reduzida, criar modelos de transcrição adaptados às especificidades de cada acervo. As modernas ferramentas de HTR são especialmente eficazes no caso de conjuntos de documentos em que a caligrafia é uniforme (por exemplo, o espólio de um escritor, em que a letra da maioria dos documentos será a mesma).

Consideramos, no entanto, que deve haver prudência no uso desta tecnologia. A interpretação da caligrafia presente nos manuscritos é sempre propícia à dúvida, que a máquina resolve com aplicação de regras que se baseiam na aprendizagem de padrões. O que significa que é certo que continuarão a ocorrer erros de transcrição, que podem ter impacto na identificação de datas, nomes de pessoas ou nomes de locais – elementos cruciais para a correta interpretação do documento. Assim, será sempre recomendável manter a supervisão humana no processo, para garantir a fiabilidade dos resultados. Daí que o AHPEGD não esteja, para já, a equacionar a disponibilização destas transcrições para o público. Nesta primeira fase do projeto, os principais objetivos quanto ao uso do Transkribus relacionam-se com a melhoria da capacidade de recuperação da informação e com a criação de descrições dos documentos. A experiência tem sido muito positiva.

Apesar das dificuldades apresentadas, a grande mais-valia desta tecnologia consiste na possibilidade de identificação e localização de documentos que, de outra forma, provavelmente não seriam lidos nem descritos, permanecendo no anonimato de uma caixa de arquivo. A aplicação desta tecnologia a grandes conjuntos documentais representa uma expansão muito significativa da informação disponível para os serviços de informação e, sobretudo, para os utilizadores.

Destacamos também o facto de o nosso projeto ter resultado na criação de um novo modelo de transcrição para manuscritos em língua portuguesa. O modelo *AHP Handwritten Portuguese 19th-20th Centuries* está agora disponível em livre acesso para todos os utilizadores da plataforma Transkribus e pode servir de base para novos treinos do algoritmo. O caminho pela frente ainda é longo, mas os passos já trilhados revelam que a IA abre todo um mundo de novas possibilidades no tratamento de documentos históricos.

Referências bibliográficas

Agrawal, V. J., Jayant, K., & Prasad, M. V. (2024). Exploration of advancements in handwritten document recognition techniques. *Intelligent Systems with Applications, 22*. https://doi.org/10.1016/j.iswa.2024.200358

Asadi, N. S. (2026). *What lies beyond the recognition wall open-source tools for HTR post-processing and publication in digital scholarly editing*. [Tese de Doutoramento], Universidade de Antuérpia. https://hdl.handle.net/10067/2199420151162165141

Bazzaco, S. (2024). Revolucionar el acceso al patrimonio librario: Los sistemas de HTR entre humanidades digitales y ciencia de la información, *Philologia Hispalensis, 38*(2), 59-77. https://dx.doi.org/10.12795/PH.2024.v38.i02.03

Cummings, J. (2025). The future is already here: Navigating the new frontiers of digital scholarly editing in an age of HTR and AI. *Philologia Hispalensis, 39*(2), 179-199. https://dx.doi.org/10.12795/PH.2025.v39.i02.07

Garrido-Munoz, C., Rios-Vila, A., & Calvo-Zaragoza, J. (2025). *Handwritten text recognition: A survey*. Cornell University. arXiv:2502.08417. https://doi.org/10.48550/arXiv.2502.08417

Gomes, S. A. (2018). Paleografia, passado e presente. In Lose, A. D., & Souza, A. S. (Ed.), *Paleografia e suas interfaces* (pp. 286-293). Memória & Arte. https://repositorio.ufba.br/handle/ri/26224

Lacerda, M. F., & Araújo, S. S. (2026). *Língua, história e tecnologia: Em comemoração aos 25 anos do Núcleo de Estudos de Língua Portuguesa da UEFS*. Pontes Editores. https://doi.org/10.29327/5824500

Matos, A., Almeida, P., Correia, P. L., & Pacheco, O. (2025). iForal: Automated handwritten text transcription for historical medieval manuscripts. *Journal of Imaging, 11*, 36. https://doi.org/10.3390/jimaging11020036

Nockels, J., Godding, P., & Terras, M. (2024). The implications of handwritten text recognition for accessing the past at scale. *Journal of Documentation, 80*(7), 148–167. https://doi.org/10.1108/JD-09-2023-0183

Plamondon, R., & Srihari, S. N. (2000). Online and off-line handwriting recognition: A comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(1), 63-84. https://doi.org/10.1109/34.824821

Relicovschi, A. (2026). *Enregistrer, quantifier et contrôler les migrations dans une ville industrielle du Luxembourg: Dudelange, fin XIXe-XXe siècles*. In Doctoriales du Centre d'Histoire du XIXe siècle 2025, Paris (France). https://hdl.handle.net/10993/64969

Silva, A. D. (2025). A Inteligência Artificial no acesso à informação em documentos manuscritos. *Cadernos BAD*, 1-2. https://doi.org/10.48798/cadernosbad.3095

Silva, A. Dias, & Borges, L. C. (2018). A transcrição e a leitura de manuscritos entre o crowdsourcing e a participação cidadã. *Congresso BAD N.º 13: Sustentabilidade e Transformação / Comunicações: II – Redes, Comunidades e Literacias.* https://publicacoes.bad.pt/revistas/index.php/congressosbad/article/view/1792

Vanshika, G., Shruti, J., Shreya, S., & Aaryan, R. (2025). A review on handwritten text recognition. *International Journal for Research Trends and Innovation, 10*(5), 601-608. https://www.ijrti.org/viewpaperforall?paper=IJRTI2505071

Websites

Site do projeto do Arquivo Nacional da Hungria sobre prisioneiros de guerra na União Soviética: https://adatbazisokonline.mnl.gov.hu/adatbazis/szovjet-taborok-magyar-foglyai

Informação sobre a plataforma Transkribus e a cooperativa Read Coop: https://www.transkribus.org/about

The Brief History of OCR Technology: https://www.docsumo.com/blog/optical-character-recognition-history