LHTR.br: em Busca de um Conjunto Anotado de Textos Manuscritos em Português*

Gabriel Henrique Coelho da Silva[1], **Daniel de Oliveira**[1], **Isabel Rosseti**[1], **Aline Paes**[1]

[1]Instituto de Computação – Universidade Federal Fluminense (UFF)
Caixa Postal 24210-310 – Niterói, RJ, Brasil

gabrielhcs@id.uff.br, {danielcmo,rosseti,alinepaes}@ic.uff.br

Abstract. *Various activities rely on handwritten records, such as medical prescriptions and patient records, and security services. Although technological resources, such as tablets and cell phones, enable handwriting using digital means, people still use paper to document their writing. In all cases, automating the transcription of such records into a digital format involves recognizing their textual content. While methods based on deep neural networks aid in this process, they lack annotated datasets for specific languages. However, the data available is predominantly in the English language, which does not utilize accentuation symbols. Additionally, the writing may incorporate cultural styles that may not be familiar to speakers of other languages. To address this problem, this article contributes to* `LHTR.br` *(Labeled Handwritten Text Recognition in Brazilian Portuguese), a dataset with text markings in images and transcriptions in Portuguese. It is expected that this dataset can be used for training neural network-based models.*

Resumo. *Atividades diversas utilizam registros escritos à mão, tais como receituários e prontuários médicos e serviços de segurança. Embora recursos tecnológicos, como tablets e celulares, permitam a escrita à mão usando meios digitais, muitos ainda utilizam papel para registrar sua escrita. Em todos os casos, automatizar a transcrição de tais registros para um formato digital implica no reconhecimento de seus conteúdos textuais. Embora métodos baseados em redes neurais profundas auxiliem este processo, eles carecem de conjuntos de dados anotados de idiomas específicos. Porém, majoritariamente, os dados disponibilizados estão na língua inglesa, que não faz uso de símbolos de acentuação. Também, a escrita pode conter estilos culturais que podem não ser parte de falantes de outros idiomas. Para abordar este problema, este artigo contribui com o* `LHTR.br` *(Labeled Handwritten Text Recognition in Brazilian Portuguese), um conjunto de dados com demarcações de textos em imagens e transcrição do texto em Português. Espera-se que esse conjunto de dados possa ser utilizado para o treinamento de modelos baseados em redes neurais.*

1. Introdução

Mesmo com a modernização das possibilidades comunicativas por meio de recursos tecnológicos e novos formatos de reproduzir a escrita, registros feitos à mão propagam-

*O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Os autores gostariam ainda de agradecer ao CNPq e FAPERJ pelo apoio financeiro.

se ao redor do mundo em diversos idiomas, contextos variados e diferentes formatos [Guimarães 2005]. Na área jurídica [Chakraborty et al. 2023], encontramos documentos em cartório e ocorrências policiais provenientes da escrita manual; na área da saúde [Bouh et al. 2023], os atendimentos realizados resultam em prescrições médicas manuscritas em muitos casos; na educação [Aqab and Tariq 2020, Sharma et al. 2023], o modelo instituído para grande parte dos exames de avaliação é aplicado majoritariamente de forma manual; entre outros campos que realizam suas atividades de tal forma.

Considerando a informação como um dos artefatos mais importantes na era digital, ter acesso à ela por meio de dispositivos eletrônicos não só diminui o seu processo de propagação e recepção, como também propicia manipulá-la de forma mais automatizada. Tal acesso foi ampliado pelo avanço da tecnologia ao longo dos anos. Porém, como consequência da pandemia de 2019, que acarretou em um período de atividades efetuadas remotamente de forma virtual, a transmissão de informação em formato digital atingiu um novo patamar, obrigando instituições de diversos setores a transferir seus procedimentos e registros de materiais físicos para um modelo virtual adequado.

Entretanto, há casos em que o acesso digital ainda não se encontra disponível, dado que a força de trabalho humana é insuficiente para atender a todo o volume de informações que precisam ser compreendidas e transcritas para este formato e, ainda, analisadas e validadas posteriormente. Além disso, a manipulação de dados textuais implicam em tratamentos sobre especificidades do idioma. Para a língua portuguesa, é preciso se atentar, por exemplo, a atributos que modificam a estrutura da palavra, como uso de acentuação (*e.g.*: "*dê*", "*disposição*", "*órgão*") e símbolos de pontuação (*e.g.*: "*V.Sas.*", "*coloco-me*", "*e/ou*", "*veja:*"). Em se tratando de conteúdos provenientes da escrita manual, os desafios são ainda maiores. Além de considerar a complexidade da língua, é preciso atentar-se às características como estilo, material e velocidade de escrita, que influenciam na informação sendo produzida. Outro fator a ser considerado é o processo de transformação destes materiais físicos para meios digitais, por meio, por exemplo, de digitalização, registros fotográficos, ou transcrição manual. Tais aspectos evidenciam as barreiras encontradas no processamento destes conteúdos.

A tarefa de Reconhecimento de Texto Escrito à Mão (do inglês *Handwritten Text Recognition*/HTR) [Kim et al. 1999] une conceitos da área de Visão Computacional (por meio do Reconhecimento Óptico de Caracteres/OCR) e Inteligência Artificial (com Aprendizado de Máquina e Processamento de Linguagem Natural/PLN [Chowdhary 2020]) para lidar com os desafios supracitados. Entretanto, para o correto funcionamento dos métodos de HTR, são necessários dados que abranjam os símbolos e as particularidades existentes no vocabulário adotado e que expressem a heterogeneidade das formas de escrita possíveis (*e.g.*: "*pure cursive script writing*" e "*boxed discrete character*" [Tappert et al. 1990], Figura1.b). Tais dados também precisam ser anotados com a transcrição digital, para que métodos baseados em Aprendizado de Máquina consigam encontrar padrões no conteúdo das imagens e transcrevê-lo corretamente para um formato textual. A anotação deve conter os dois aspectos principais do processo, ou seja, onde estão os símbolos textuais na imagem e a transcrição léxica de tais símbolos. Por conta disso, o processo de HTR demanda uma grande quantidade de dados e processamentos custosos em valor e tempo [Sanches et al. 2022] para alcançar o desempenho esperado nos mecanismos utilizados.

Este artigo tem como objetivo disponibilizar um *dataset*, chamado `LHTR.br`, em português para tarefa de HTR e apresentar o processo conduzido para a obtenção de tal *dataset* – coleta, seleção, pré-processamento e anotação de amostras provenientes de fontes de dados abarcadas na literatura [Freitas et al. 2008]. Os dados compreendem amostras de imagens de símbolos do alfabeto e de linhas de textos oriundos da língua portuguesa, com respectivos arquivos texto contendo anotações relevantes sobre as informações textuais. Também é proposta a distribuição dos dados em subconjuntos de treino, validação e teste, para o treinamento de modelos de HTR, que se baseiam em informações textuais contidas em imagens, para abranger aspectos relativos às especificidades da língua portuguesa. O *dataset* encontra-se disponível por meio da plataforma OSF na URL `https://l1nq.com/osf-lhtr-br`.

Este artigo organiza-se da seguinte forma. Na Seção 2, são elencados trabalhos relacionados. Na Seção 3, são dados detalhes sobre as etapas para composição do *dataset* e sua organização, e propriedades quantitativas compreendidas a partir do conjunto resultante. Na Seção 4, são listadas as possibilidades de utilização, desafios e limitações encontrados e melhorias previstas para trabalhos futuros. Por fim, na Seção 5, é apresentada uma contextualização final do escopo de trabalho explorado no presente artigo.

2. Trabalhos Relacionados

A tarefa de HTR se ramifica em diferentes frentes dependendo dos atributos considerados. Dentre os aspectos que distinguem tais frentes, destacam-se: (i) o idioma alvo (de domínio específico ou multi-língua); (ii) o nível de compreensão da informação (nível de palavras, linhas de texto, folhas de texto, *etc*.); e (iii) o tipo de aquisição da informação (*offline*, quando as imagens provém de textos empregados em artefatos físicos, como papel, placas e *outdoors*, ou *online*, quando as imagens provém da escrita empregada em telas de dispositivos eletrônicos, como celulares e *tablets*). A escolha dentre tais características implica nas estratégias de transcrição a serem adotadas, pois cada atributo compreende tratamentos específicos. Considerando o cenário apresentado na Seção 1, foram relacionados trabalhos no contexto de aquisição *offline* e preferencialmente para língua portuguesa, podendo variar no nível de informação sendo compreendida.

Para a língua portuguesa, destacam-se dois trabalhos. [Freitas et al. 2008] apresentam um conjunto de 945 amostras de folha de texto digitalizada, escritas por 315 autores (três amostras por autor). O texto base compreende um vocabulário de 131 palavras e apresenta especificidades como letras acentuadas (*e.g.*: “ã”, “ó”, “ç”) e símbolos especiais (*e.g.*: “.”, “-”, “ª”). [Pereira et al. 2021] fornecem um conjunto de 570 amostras de folha de texto (285 em português e 285 em japonês) escritas por 57 autores, utilizando dez textos base (cinco para cada idioma) e cada um reproduzido uma vez pelos autores.

Existem trabalhos que disponibilizam *datasets* em outros idiomas. O *dataset* em língua inglesa proposto por [Marti and Bunke 2002] é amplamente utilizado como *baseline* de diversos trabalhos de HTR. Sua última versão[1] inclui 1.539 imagens de folhas de texto digitalizadas, produzidas por 657 autores e transformadas em 5.686 amostras de sentenças, 13.353 de linhas de texto e 115.320 de palavras, todas anotadas. Além disso, foram relacionados trabalhos mais recentes e que apresentam variações internas ao

[1]**IAM Handwriting Database:** `https://fki.tic.heia-fr.ch/databases/iam-handwriting-database`

cenário descrito anteriormente. [Joshi et al. 2023] apresentam um conjunto extenso de 1.865.134 amostras de imagem a nível de palavras em inglês, de um vocabulário com 10.711 termos, extraídas de documentos do Departamento de Censo dos Estados Unidos dos anos de 1930 e 1940. Já [Rahman et al. 2022] apresentam um conjunto de dados no idioma bengali contendo 786 amostras de folhas de texto escritas por 150 autores distintos, com anotações em diferentes níveis de informação além de folhas de texto, sendo 14.383 de linhas de texto e 1.008.018 de palavras, com um vocabulário de 23.115 termos.

Dois trabalhos dão suporte à metodologia apresentada no presente artigo, com detalhes sobre a aquisição e composição dos dados descritos na seção a seguir. [Bertolini et al. 2013] utilizam as amostras de [Freitas et al. 2008] e de [Marti and Bunke 2002] para a tarefa de identificação de autores a partir da escrita. [Souibgui et al. 2021] propõem um modelo com aprendizado a partir de poucas imagens, pré-treinado com manuscritos sintéticos e avaliado usando dois *datasets*[2] [3]. Isto posto, o presente trabalho contribui com recursos relevantes para enriquecimento do cenário de HTR para língua portuguesa e colabora para o estado da arte no refinamento e treinamento de modelos envolvidos neste cenário, considerando informações compreendidas por meio de linhas de texto (diferenciando-se de [Freitas et al. 2008] e [Pereira et al. 2021]) e o uso de acentuação e símbolos especiais (diferenciando-se de conjuntos em outros idiomas).

3. Metodologia para Criação do *Dataset*

Neste trabalho, é proposto um *dataset* para tarefas que exploram o reconhecimento de textos escritos à mão a partir de imagens, compreendendo o uso de modelos de aprendizado de máquina supervisionado e o foco na língua portuguesa. Para a construção de tal *dataset*, são requeridos dois componentes básicos, arquivos de imagem com conteúdo dimensionado para o nível de informação a ser processada, e arquivos de texto com anotações relevantes sobre os conteúdos da imagem, como coordenadas das regiões onde eles se encontram e os respectivos símbolos destas regiões.

Como *baseline*, foi utilizado o trabalho de [Souibgui et al. 2021][4], no intuito de, *a posteriori*, aplicar o presente conjunto de dados no *Fine-Tuning* do modelo fornecido no *baseline*. Tal aplicação foi escolhida dada a precariedade de dados anotados em português, como uma alternativa ao treinamento do zero de modelos que demandam alto custo de processamento, fazendo uso, portanto, de uma quantidade reduzida de dados para o aprendizado e considerando que a maior parte dos componentes já foram aprendidos no desenvolvimento do modelo [Sanches et al. 2022].

Os procedimentos foram efetuados a nível de linhas de texto e considerando duas rotinas principais: coleta e seleção das amostras (Seção 3.1) e pré-processamento e anotação dos dados (Seção 3.2). Todos os procedimentos foram efetuados em uma máquina com sistema operacional Windows 11. Ferramentas e métodos específicos de cada etapa são mencionados ao longo da Seção.

3.1. Concepção das Amostras

Dadas as instruções fornecidas no *baseline*, o *dataset* deve ser composto por dois grupos de imagens: linhas de texto e símbolos do alfabeto relativo ao texto, obtidas como segue.

[2]**Borg Dataset**:`https://cl.lingfil.uu.se/~bea/borg/`
[3]**Copiale dataset**:`https://cl.lingfil.uu.se/~bea/copiale/`
[4]**HTRbyMatching**: `https://github.com/dali92002/HTRbyMatching`

Coleta de Dados. Como fonte de dados, foi utilizada a *Brazilian Forensic Letter Database*/BFL[5] [Freitas et al. 2008]. A base contém 945 imagens de papel A4 digitalizado como folhas de texto (Figura 1.a). Todas as imagens foram consideradas para concepção das amostras.

Geração das Amostras. Para as linhas de texto, foi aplicado um procedimento de segmentação[6] nas 945 imagens, resultando em 16.744 imagens de linhas (Figura 1.b). Para os símbolos do alfabeto, foi aplicada a função *image_to_boxes* do repositório *Python Tesseract*[7], que retorna um vetor com coordenadas da posição de cada símbolo detectado, as quais foram utilizadas para extração de 205.098 imagens de símbolos (Figura 1.d).

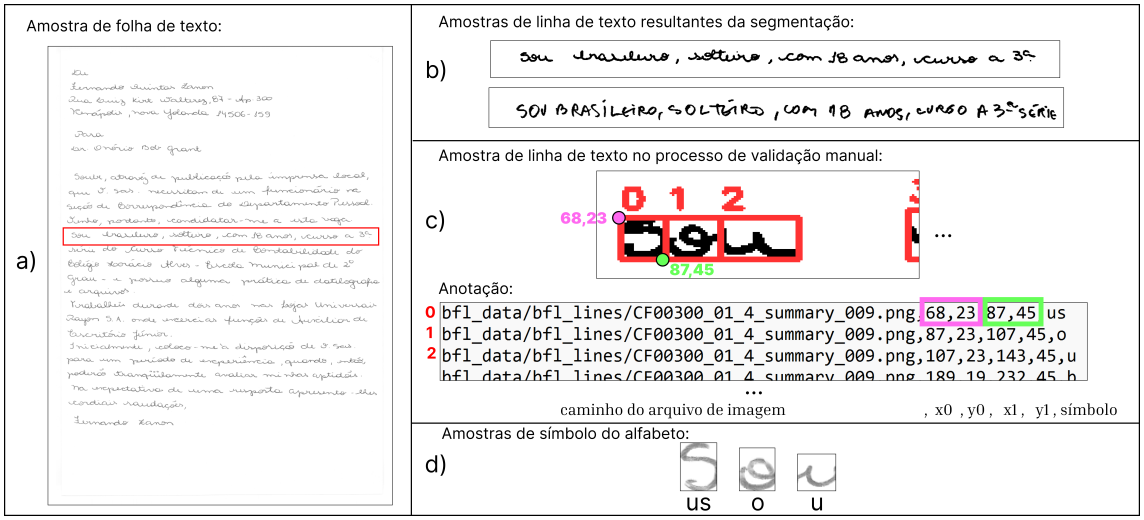

Figura 1. Representação dos dados que correspondem ao *dataset*

Filtragem dos Documentos. As amostras de linhas de texto foram separadas pelo autor manualmente entre 11.483 imagens *processáveis* (68,58%) e 5.261 *não processáveis* (31,42%), quando identificados visualmente ruídos da digitalização ou problemas na segmentação do texto em linhas. Para os símbolos do alfabeto, foram escolhidas pelo menos 10 imagens de cada símbolo, totalizando 750 amostras e considerando estilos distintos de escrita [Tappert et al. 1990] para ampliar a variedade dos dados.

3.2. Pré-processamento das Amostras

Dado o caráter de aprendizado supervisionado e características definidas no trabalho de [Bertolini et al. 2013], que também utiliza o conjunto de dados BFL, 36,5% das amostras de linha de texto foram separadas para o subconjunto de teste, restando 63,5% das amostras para os subconjuntos de treino e de validação. Métodos distintos de anotação dos dados de treino e validação foram adotados, conforme as instruções do *baseline*.

Distribuição das Amostras. A distribuição das amostras nos subconjuntos foi efetuada aleatoriamente, apenas considerando como premissa a seleção tanto de amostras com maior potencial de extração de informações (8.539 amostras) quanto de amostras

[5]**BFL Database:** `https://web.inf.ufpr.br/vri/databases/brazilian-forensic-letter-database/`

[6]**Text-Segmentation:** `https://github.com/arthurflor23/text-segmentation`

[7]**Python Tesseract**: `https://github.com/madmaze/pytesseract`

que denotam maior dificuldade de extração (2.944 amostras) devido à falta de legibilidade das informações. Essa categorização levou em conta a quantidade de linhas extraídas na segmentação das 945 imagens de folhas de texto. Assumiu-se que folhas com número de segmentações menor do que a média (13 linhas) denotam uma maior dificuldade de extração de informações. Esta estratégia evita que algum subconjunto eventualmente abarque apenas amostras desafiadoras a partir da escolha aleatória, garantindo, portanto, que cada subconjunto abranja ambos os cenários.

Anotação das Amostras de Treino. Para o subconjunto de treino, foi adotado um processo automático de inferência dos símbolos nas imagens, visando agilizar e aliviar a anotação manual de todos os símbolos. Inicialmente, foram selecionadas 100 amostras de linhas de texto. A anotação de cada símbolo foi obtida também utilizando a função *image_to_boxes* do repositório *Python Tesseract*, que retorna a inferência do símbolo como primeira posição do vetor: *[inferência, x0, y, x1, h]*.

No *baseline*, um arquivo texto único, a nível de símbolos, deve ser constituído para as anotações do subconjunto de treino. As informações devem ser registradas no formato CSV (*Comma-separated values*), com cada linha do arquivo indicando: i) o caminho do arquivo da imagem corrente, ii) as coordenadas $x0$, $y0$, $x1$, $y1$ da região contendo um símbolo da imagem e iii) o valor que aquele símbolo representa. Por limitações de nomeação de arquivos – não diferenciação entre maiúscula e minúscula e restrição no uso de símbolos, o valor em (iii) compreende o seguinte formalismo: números e letras minúsculas seguem a nomenclatura convencional, [*0-9*] e [*a-z*]; letras maiúsculas utilizam a nomenclatura de letras minúsculas precedida da letra "*u*": [*ua-uz*]; letras acentuadas e símbolos especiais adotam nomenclaturas particulares, descritas nas Tabelas 1 e 2, respectivamente. Assim, são previstas 75 nomenclaturas distintas para o registro dos valores dos símbolos.

Tabela 1. Nomenclatura para letras acentuadas

Símbolo	á	é	í	ó	ú	ã	õ	ê	à	ç
Nomenclatura	aagd	eagd	iagd	oagd	uagd	atll	otll	ecrc	acrs	ccdl

Tabela 2. Nomenclatura para símbolos especiais

Símbolo	.	,	-	ª	º
Nomenclatura	dtsmb	vglsmb	hfnsmb	grfsmb	grmsmb

Validação Manual da Anotação das Amostras de Treino. Após gerado o arquivo texto com as anotações do subconjunto de treino, foi efetuada pelo autor uma validação manual dos dados registrados no arquivo. Foram utilizadas ferramentas simples no processo, de modo que não fossem necessários especialistas para efetuar esta etapa. Para a revisão das coordenadas, foi executada uma rotina para redesenhar as regiões de cada símbolo e, por meio do programa MSPaint, ao posicionar o cursor sobre os vértices da região desenhada, comparou-se os valores exibidos na posição do cursor com os valores registrados no arquivo texto (Figura 1.c). Em paralelo, o símbolo contido em cada região foi comparado com o valor do último campo de cada linha do arquivo texto. Ao longo deste processo, optou-se por remover registros que representavam símbolos corrompidos na etapa de segmentação ou sobrepostos por outros símbolos em decorrência

da escrita. Símbolos com acentuação inexistente na imagem foram substituídos para a letra correspondente sem acentuação. Ao final, 3.805 símbolos foram anotados no arquivo texto.

Anotação das Amostras de Validação. Para o subconjunto de validação, também foram selecionadas 100 amostras dos 63,5%. Para cada imagem, foi criado um arquivo texto e registrado manualmente todos os símbolos contidos na imagem separados por espaço, seguindo o *baseline*. Os valores registrados assumiram o mesmo formalismo da anotação das amostras de treino. Caso a legibilidade de algum símbolo no meio do texto estivesse comprometida, era registrado o valor "*ivld*". Símbolos de linhas de texto adjacentes, exibidos na amostra corrente devido ao processo de segmentação, também foram anotados. Símbolos de amostras escritas no estilo "*boxed discrete character*" (segundo a Figura 2 do artigo [Tappert et al. 1990]) foram anotados como letra maiúscula apenas no início de termos previamente sabidos, a partir do texto de 131 palavras do *baseline*. Foram anotados 4.237 símbolos ao longo dos 100 arquivos texto das amostras de validação.

Anotação das Amostras do Alfabeto. A anotação das amostras de símbolos do alfabeto consiste apenas em nomear cada arquivo de imagem como números em sequência e nomear o respectivo diretório dos símbolos utilizando o mesmo formalismo adotado para anotação das amostras de treino, conforme indicado no *baseline*. Como resultado, foram selecionadas 10 amostras de 75 símbolos do alfabeto, totalizando 750 imagens. Mais detalhes sobre o armazenamento desses dados são apresentados na Seção 3.4.

3.3. Descrição Quantitativa

A escolha aleatória de 100 amostras de linhas de texto para compor cada subconjunto resultou em 3.805 símbolos anotados para o treino e 4.237 símbolos anotados para validação. Além disso, 750 imagens de símbolos foram agrupadas em correspondência aos 75 símbolos do alfabeto. Conforme demonstrado na Figura 2, apesar da escolha aleatória, a maior parte dos símbolos apresenta ocorrências em ambos os subconjuntos com proporções próximas ao equilíbrio (50%). Dada a aleatoriedade da seleção de amostras, não foram incluídas ocorrências do símbolo "*Q*" (em maiúsculo) em ambos os subconjuntos, e ocorrências dos símbolos "*B*" e "*O*" (em maiúsculo) só aparecem no subconjunto de treino. Pelo vocabulário adotado, contendo 131 termos apresentados por [Freitas et al. 2008], não são definidas ocorrências dos símbolos "*k*" e "*w*" (em minúsculo) e nem de símbolos especiais divergentes aos que são elencados na Tabela 2. Além disso, a ocorrência de letras acentuadas (Tabela 1) é diretamente afetada pela utilização do acento corretamente no momento da escrita e pela segmentação do texto.

3.4. Organização e Armazenamento dos Dados

Cumprindo os requisitos de armazenamento do *baseline*, os dados foram divididos em quatro diretórios: `/bfl_alphabet/` (para as amostras de símbolos do alfabeto), `/bfl_data/` (para as amostras de treino), `/data_validation/` (para as amostras de validação) e `/bfl_lines/` (para as amostras de teste). Para fins de organização, todos os diretórios estão contidos no diretório raiz `/bfl_dataset/` e se encontram disponível por meio da plataforma OSF.

As amostras do alfabeto estão distribuídas entre 75 subpastas nomeadas a partir do formalismo explicitado na Seção 3.2 (*e.g.*: `/bfl_alphabet/aagd/1.png`).

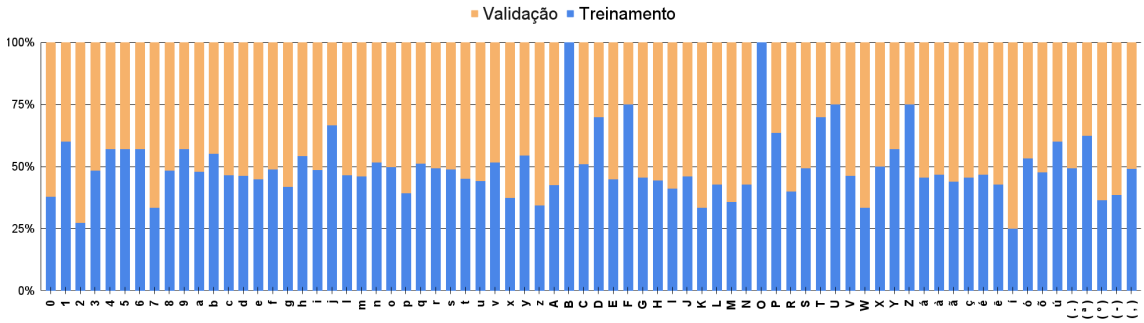

Figura 2. Proporção de símbolos anotados por subconjunto

As imagens de linhas de texto do subconjunto de treino são armazenadas na subpasta `/bfl_data/bfl_lines/` e o arquivo texto (*train.txt*) com as anotações é armazenado na subpasta `/bfl_data/annotation/`. As imagens de linhas de texto do subconjunto de validação são armazenadas na subpasta `/data_validation/lines/` e os arquivos texto com a respectiva anotação são armazenados na subpasta `/data_validation/gt/`, registrados com o mesmo nome do arquivo de imagem (*e.g.*: "*amostra1_linha_1.png*" → "*amostra1_linha_1.txt*"). Foram armazenadas 200 imagens de linha de texto do subconjunto de teste no diretório `/bfl_dataset/bfl_lines/`, sem necessitar de nenhum critério extra para o armazenamento.

4. Aplicação

O principal objetivo para construção deste dataset consiste em futuramente aplicá-lo para o *Fine-Tuning* de um modelo que experiencia o aprendizado supervisionado para a tarefa de HTR utilizando de poucos dados de entrada [Souibgui et al. 2021]. Tais considerações contribuem, de modo geral, para aplicação do *dataset* em tarefas que envolvam análises, aplicações, refinamento e treinamento sobre modelos que utilizam informações textuais provenientes de imagens. Em especial, para o arcabouço de dados oriundos da língua portuguesa [Freitas et al. 2008, Pereira et al. 2021], que apesar de consolidado na literatura, carece de recursos anotados disponíveis.

Além disso, a especificação de procedimentos simples de geração de dados com tais características possibilita a replicação para conjuntos de dados em outros idiomas, como contribuição para trabalhos voltados para multi-línguas ou idiomas com poucos recursos disponíveis [Ignat et al. 2022]. Por fim, as amostras de símbolos e linhas de texto disponibilizadas permitem seu aproveitamento para o aumento do volume de dados usado em tarefas como reconhecimento de autor [Freitas et al. 2008, Pereira et al. 2021, AL-Qawasmeh et al. 2023] e de entidades nomeadas [Adak et al. 2016].

4.1. Desafios

Por meio da metodologia apresentada na Seção 3 para a construção do dataset, foram identificados três desafios principais:

Textos Escritos à Mão. A espessura do traço empregado na escrita reflete no desempenho dos procedimentos de aquisição de informações textuais em imagens. Espessuras muito finas podem dificultar a inferência do conteúdo, necessitando de métodos complementares para prepará-las para a extração de informações corretamente.

Ainda sobre a escrita, dado que se trata de uma particularidade de cada indivíduo [Xing and Qiao 2016], podem ser empregados diferentes estilos que, ao apresentarem características como inclinação e sobreposição de símbolos, acarretam problemas de legibilidade e interpretação do conteúdo. Além disso, a diversificação do estilo resulta em possibilidades distintas de escrita do mesmo símbolo, dificultando o processo de diferenciação entre símbolos distintos e de generalização pela inferência (*e.g.*: "5" e "s", "0" e "o").

Particularidades do Idioma de Escrita. A composição do alfabeto está intimamente relacionada com a configuração estabelecida para escrita dos seus símbolos. Logo, alfabetos com configurações de símbolo ambíguas ou complexas – que incluem, por exemplo, aspectos de acentuação para denotar características fônicas de pronúncia – também contribuem para problemas de interpretação dos dados textuais.

Tais desafios também impactam no processo de anotação dos dados, pois apesar da familiaridade com a língua, casos de ambiguidade e falta de legibilidade tornam a anotação dos dados, e posterior validação, mais custosa.

Textos Embutidos em Imagens. Refere-se ao formato em que os dados são registrados. A utilização de imagens como amostras reflete na necessidade de métodos que expressem o conteúdo textual com qualidade, desde a etapa de aquisição, por exemplo, através de digitalização ou registros fotográficos, até o processo de inferência, utilizando de filtros apropriados para destacar informações relevantes nos pixels e minimizar ruídos que comprometam a extração de informações corretamente.

4.2. Limitações

Dado o contexto do conjunto de amostras resultante, foram identificados os seguintes aspectos que caracterizam limitações no dataset disponibilizado:

Alfabeto Limitado. O alfabeto empregado na composição do dataset é relativo ao conjunto de dados utilizado como fonte (BFL [Freitas et al. 2008]), logo, não considera todos os símbolos existentes que compõem um texto, por exemplo ";", "$", "&" e demais itens inexistentes na Tabela 2, e letras não usuais na língua portuguesa, como "*k*" e "*w*".

Anotações Divergentes ao Vocabulário. Símbolos comprometidos pela segmentação do texto em linhas ou com baixa legibilidade foram removidos do arquivo de anotação do treino. Dessa forma, representam ruídos na imagem, pois a falta de registro sobre determinado símbolo na anotação denota sua falta de relevância para a informação textual. Letras com acentuação mal posicionada ou inexistente são anotadas sem o acento. Com isso, a concatenação de símbolos contendo tais inconsistências, para composição de termos, afeta a concepção semântica dos conteúdos das linhas de texto. Símbolos inclinados podem não corresponder em sua totalidade à região definida pelas coordenadas anotadas, dado o espaço necessário para denotar os símbolos adjacentes. É necessário um pós-processamento para mapear os dados com nomenclaturas não convencionais (Tabelas 1 e 2) para os símbolos que de fato correspondem ao alfabeto (*e.g.*: " *ccdl* " $\rightarrow$ "*ç*").

Amostras Anotadas para Linhas de Texto. Os conteúdos das amostras são apresentados a nível de linhas de texto, gerando implicações no uso em aplicações que consideram apenas outros níveis de granularidade, como palavras. Por fim, dada a quantidade limitada de amostras anotadas para treino e validação, podem haver implicações no uso em aplicações que exigem grandes volumes de dados.

4.3. Melhorias

Para aperfeiçoar o *dataset* criado e contribuir com outras formas de utilização, destacam-se as seguintes melhorias a serem exploradas em trabalhos futuros:

Inclusão de mais Amostras de Linhas de Texto resultantes da segmentação do texto, para o incremento de símbolos não capturados, dado que a quantidade selecionada aleatoriamente a princípio considerou o critério mínimo para as experimentações das fases posteriores de aplicação dos dados, e algumas amostras, intituladas *não processáveis*, foram desconsideradas por necessitarem de procedimentos complementares não explorados até o momento. Mais aplicações podem ser consideradas ampliando o volume de dados e possibilita um comparativo do desempenho utilizando diferentes quantidades.

Inclusão de Amostras de Outras Fontes de Dados em Português, como [Pereira et al. 2021] ou repositórios[8] disponibilizados com documentos manuscritos digitalizados, ainda não exploradas dada a complexidade de anotação e tratamento de novos dados com tais características para incorporação ao *dataset*. A replicação da metodologia em outras fontes contribui não só para o aumento do volume de dados como também para a diversificação do conteúdo (ao englobar outros estilos de escrita) e para um aumento no vocabulário, com ocorrências de símbolos desconsiderados atualmente. Possibilita um comparativo do desempenho utilizando diferentes bases de dados nas aplicações.

Aperfeiçoamento da Metodologia Proposta, principalmente na etapa de anotação, de modo a otimizar o tempo gasto na validação manual através de um tratamento aprimorado dos dados na etapa de inferência. Além da inclusão de métodos para ajustes nas propriedades das imagens, como tamanho e aplicação de filtros, proporcionando adequações específicas de requisitos para dados de entrada de diferentes modelos.

5. Considerações Finais

Este artigo fornece um *dataset*[9] anotado de textos manuscritos em português, provenientes de imagens. E apresenta detalhes sobre as estratégias adotadas para alcançar tal feito, com os respectivos desafios e limitações considerando a coleta, seleção, pré-processamento e anotação de amostras oriundas de fonte de dados abarcada na literatura [Freitas et al. 2008]. Foi utilizado como *baseline* o trabalho de [Souibgui et al. 2021], adotando os requisitos especificados para posterior aplicação do dataset no *Fine-Tuning* do modelo desenvolvido partindo de uma quantidade reduzida de dados para o aprendizado da tarefa de HTR em português. Foram selecionadas aleatoriamente 100 amostras de linhas de texto para subconjuntos de treino e validação, totalizando 3.805 símbolos anotados para o treino e 4.237 símbolos anotados para validaçao. Além disso, 750 imagens de símbolos foram agrupadas em correspondência a 75 símbolos do alfabeto fonte.

Deste modo, o presente trabalho contribui para que a área de PLN tome proveito das estratégias apresentadas e do conjunto de dados disponibilizado (constituído a nível de linhas de texto e voltado para língua portuguesa), abrindo espaço para execução de pesquisas futuras que exploram aplicações e análises em modelos a partir destes cenários. Além disso, é fornecida uma metodologia simples de aquisição dos dados, permitida de ser replicada para outros idiomas com poucos recursos disponibilizados [Ignat et al. 2022].

[8] https://digitarq.arquivos.pt/

[9] https://l1nq.com/osf-lhtr-br

Referências

Adak, C., Chaudhuri, B. B., and Blumenstein, M. (2016). Named entity recognition from unstructured handwritten document images. In *12th DAS*, pages 375–380. IEEE Computer Society.

AL-Qawasmeh, N., Khayyat, M., and Suen, C. Y. (2023). Novel features to detect gender from handwritten documents. *Pattern Recognition Letters*, 171:201–208.

Aqab, S. and Tariq, M. U. (2020). Handwriting recognition using artificial intelligence neural network and image processing. *IJACSA*, 11(7).

Bertolini, D., Oliveira, L. S., Justino, E. J. R., and Sabourin, R. (2013). Texture-based descriptors for writer identification and verification. *Expert Syst. Appl.*, 40(6):2069–2080.

Bouh, M. M., Hossain, F., and Ahmed, A. (2023). A machine learning approach to digitize medical history and archive in a standard format. In *9th ICT4AWE*, pages 230–236.

Chakraborty, S., Harit, G., and Ghosh, S. (2023). TransDocAnalyser: A framework for offline semi-structured handwritten document analysis in the legal domain. *CoRR*, abs/2306.02142.

Chowdhary, K. (2020). Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649.

Freitas, C., Oliveira, L. S., Sabourin, R., and Bortolozzi, F. (2008). Brazilian forensic letter database. In *11th International workshop on frontiers on handwriting recognition, Montreal, Canada*.

Guimarães, E. (2005). A língua portuguesa no brasil. *Ciência e Cultura*, 57(2):24–28.

Ignat, O., Maillard, J., Chaudhary, V., and Guzmán, F. (2022). OCR improves machine translation for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1164–1174.

Joshi, C., Sorenson, L., Wolfert, A., Clement, M. J., Price, J., and Buckles, K. (2023). CENSUS-HWR: a large training dataset for offline handwriting recognition. *CoRR*, abs/2305.16275.

Kim, G., Govindaraju, V., and Srihari, S. N. (1999). An architecture for handwritten text recognition systems. *Int. J. Document Anal. Recognit.*, 2(1):37–44.

Marti, U. and Bunke, H. (2002). The iam-database: an english sentence database for offline handwriting recognition. *Int. J. Document Anal. Recognit.*, 5(1):39–46.

Pereira, L. F. M., Pinhelli, F., Cizeski, E. M. A., Uber, F. R., Bertolini, D., and Costa, Y. M. G. (2021). Japanese kana and brazilian portuguese manuscript database. In *25th CIARP*, volume 12702, pages 173–183. Springer.

Rahman, M. A., Tabassum, N., Paul, M., Pal, R., and Islam, M. K. (2022). Bn-htrd: A benchmark dataset for document level offline bangla handwritten text recognition (HTR) and line segmentation. *CoRR*, abs/2206.08977.

Sanches, M., de Sá, J., Foerste, H., Souza, R., Reis, J. D., and Villas, L. (2022). Textual datasets for portuguese-brazilian language models. In *IV DSW*, pages 1–12.

Sharma, A., Katlaa, R., Kaur, G., and Jayagopi, D. B. (2023). Full-page handwriting recognition and automated essay scoring for in-the-wild essays. *MTAP*, pages 1–24.

Souibgui, M. A., Fornés, A., Kessentini, Y., and Megyesi, B. (2021). Few shots is all you need: A progressive few shot learning approach for low resource handwriting recognition. *CoRR*, abs/2107.10064.

Tappert, C. C., Suen, C. Y., and Wakahara, T. (1990). The state of the art in online handwriting recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(8):787–808.

Xing, L. and Qiao, Y. (2016). Deepwriter: A multi-stream deep CNN for text-independent writer identification. In *15th International Conference on Frontiers in Handwriting Recognition, ICFHR 2016, Shenzhen, China, October 23-26, 2016*, pages 584–589. IEEE Computer Society.