Eduardo Leite Ribeiro

Análise sobre o uso de ferramentas de OCR para reconhecimento de textos em periódicos com alfabeto ucraniano

(*versão pré-defesa, compilada em 26 de agosto de 2024*)

Trabalho apresentado como requisito parcial à conclusão do Curso de Bacharelado em Ciência da Computação, setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação, OCR*.

Orientador: Dr. Eduardo Todt.

Curitiba PR
2024

Resumo

O trabalho consiste na avaliação das ferramentas de OCR OCR4all e Tesseract para o reconhecimento e transcrição de páginas escaneadas de periódicos ucraniano-brasileiros impressos, em circulação durante as décadas de 1910-20. Foi possível encontrar os principais problemas no uso de ambas as ferramentas, aplicar uma boa metodologia para o reconhecimento do texto e determinar a configuração que resulta na melhor acurácia.

Palavras-chave: OCR, redes neurais, datasets, documentação histórica.

Abstract

The project consists in the evaluation of the OCR tools OCR4all and Tesseract for the recognition and transcription of scanned pages of printed Ukrainian-Brazilian periodicals, in circulation during the decades of 1910-20. It was possible to find the main problems in the use of both tools, to apply a good methodology for recognition of the text and to determine the configuration setting that gives the best accuracy.

Keywords: OCR, neural networks, datasets, historical documentation.

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

DINF	Departamento de Informática
PPGINF	Programa de Pós-Graduação em Informática
UFPR	Universidade Federal do Paraná
OCR	Optical Character Recognition (reconhecimento ótico de caracteres)
LAREX	Layout Analysis and Region Extraction
API	Application Programming Interface (interface de programação de aplicações)
LSTM	Long Short-Term Memory (memória de curto longo prazo)
CC	Componente Conexo
TIFF	Tagged Image File Format
dpi	Pontos por polegada (Dots Per Inch)
PNG	Portable Network Graphics
EM	Engine Model

Sumário

Capítulo 1

Introdução

”Eternity is in love with the productions of time” - Willian Blake

A documentação digital atualmente possui um enorme valor como ferramenta de pesquisa e para a preservação histórica e cultural. Através dela, somos capazes de acessar documentos de manuseio delicado e localidade restrita. Ela também proporciona a busca e seleção de textos separados por tópicos com palavras-chave de forma rápida e simples, que se fossem feitas com cópias físicas levariam muito mais tempo para pesquisar e manusear. Além disso, a digitalização dos caracteres permite a acessibilidade para pessoas com deficiências visuais, já que o texto gerado é conversível em voz.

O presente trabalho se propõe a investigar maneiras efetivas já disponíveis de processar e transcrever periódicos em formato digital do jornal Pracia (trabalho, em ucraniano). Esse jornal é uma importante fonte para conhecer a história e cultura de Prudentopólis, um município no centro-sul do Paraná em que a grande maioria da população possui ascendência ucraniana (cerca de 75%). A primeira edição do jornal foi lançada em dezembro de 1912 na Tipografia dos padres basilianos de Prudentopólis, e ele foi circulado continuamente até hoje, excetuando uma interrupção de sete anos causada por censuras durante o governo do presidente Vargas (1930-45). Pode-se afirmar que durante esse centenário de existência o jornal foi capaz de "oferecer suporte e informações de caráter político, cultural e religioso para o povo ucraniano no Brasil e elevar o seu nível cultural, fornecendo notícias internacionais, nacionais e da terra de origem desse povo – Ucrânia", como explica o Pe. Zaluski em entrevista de 2010 [Burnat (2010)].

Pracia é um jornal bilíngue em ucraniano e português, e usa por vezes uma linguagem coloquial típica dos colonos, que mescla o ucraniano nativo com o português. O jornal é uma importante fonte antropológica, contendo os aspectos de religiosidade e costumes próprios da comunidade dos colonos. Uma maneira de preservar essa documentação centenária e torná-la facilmente acessível ao público geral é uma forma de manter vivo o conhecimento e interesse pela história e cultura da região.

Nesse contexto, a possibilidade de digitalizar o texto desses documentos é bastante oportuna, dada as vantagens citadas sobre o manuseio de mídias digitais. O uso de OCR

possibilita a automatização do processo de transcrição que levaria um tempo considerável de trabalho de datilografia humana, sujeita a erros e imprecisões. Porém, espera-se que esse tipo de documento tenha problemas inerentes para o processamento de OCR, dentre os quais podemos citar a qualidade de resolução, o contraste entre fonte e fundo, o uso de diferentes fontes, o enquadramento adotado pelo periódico, a presença de figuras, que só podem ser solucionados através de um tratamento prévio da página digital.

Sendo assim, nesse trabalho é feita uma investigação do uso de softwares de código livre já disponíveis para o reconhecimento de caracteres, o OCR4all e o Tesseract, tentando encontrar uma boa metodologia para o tratamento da imagem, a definição de layouts e o reconhecimento dos textos do jornal Pracia. Busca-se também catalogar os principais erros associados ao reconhecimento e sugerir alternativas que possam solucioná-los.

Na seção de Fundamentação, aborda-se a origem e particularidades do alfabeto ucraniano, o atual estado-da-arte dos softwares empregados e a função usada para analisar a acurácia dos resultados. Na seção seguinte 'Experimentação e Resultados' explica-se a metodologia usada nos experimentos e a validação dos resultados, e procura-se também catalogar os principais erros gerados no reconhecimento. A seção posterior de Conclusão sintetiza a interpretação dos resultados do trabalho.

Capítulo 2

Fundamentação

2.1 Sobre a escrita cirílica

O alfabeto cirílico é o 6. mais popular do mundo e é adotado na escrita de 50 línguas da Eurásia. Em 2019, haviam cerca de 250 milhões de pessoas usando-o. Ele foi o 3. alfabeto adotado pela União Europeia após o ingresso da Bulgária, em 2007 [Dimitrov (2023)].

Foi desenvolvido pela escola literária Preslav de monges bizantinos no Primeiro Império Búlgaro durante o século X, formada por seguidores dos teólogos Cirilo e Metódio, para missões evangelizadoras dos povos eslavos. Foi adaptado do alfabeto grego e da escrita glagolítica, outra língua evangelizadora empregada anteriormente por Cirilo, procurando adequá-lo às necessidades fonéticas desses povos.

Teve aceitação e difusão entre os povos eslavos orientais (russos, bielo-russos, ucranianos, búlgaros), onde a cristianização bizantina foi bem sucedida. Eslavos ocidentais adotaram o alfabeto latino em uso pela igreja romana [Bidwell (1967)].

Derivado da escrita cirílica, o alfabeto ucraniano consiste em 33 letras que representam 38 fonemas. O apóstrofo ’ também é adotado na escrita, mas não é incluso como letra.

2.2 Sobre o OCR4all

O OCR4all é um software que reune várias soluções open-source para a automatização do workflow de todas as etapas do processo de OCR de documentos antigos, dentro de um ambiente acessível e intuitivo. Pode-se selecionar as etapas desejadas de pré-processamento, segmentação de Layout com LAREX, segmentação de linha e reconhecimento com Calamari, bem como ao final a produção de gabarito (ground truth) para a avaliação dos resultados e uso em treinamento [Reul e Christ (2019)].

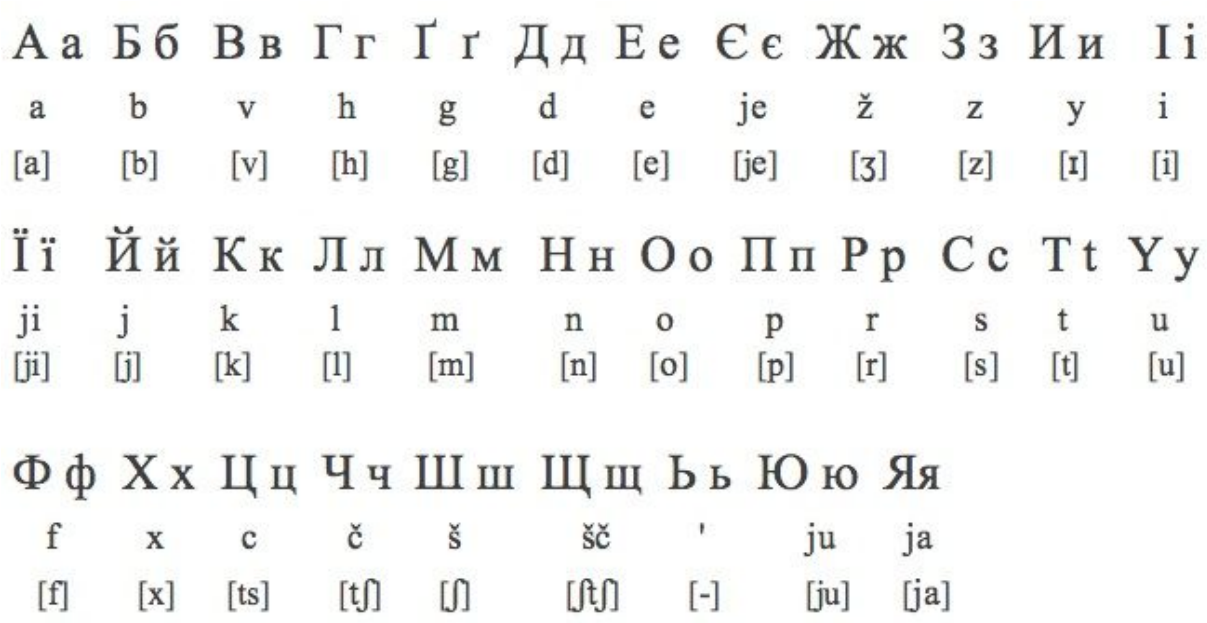


Figura 2.1: Alfabeto ucraniano

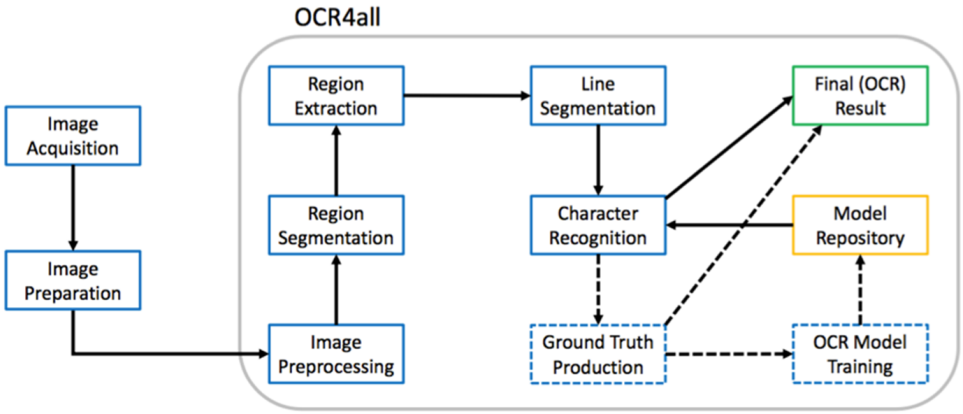


Figura 2.2: Componentes do workflow do OCR4all, extraída de [Reul (2024)]

2.3 Sobre o Tesseract

O Tesseract é um OCR completo, pois realiza a binarização, segmentação e reconhecimento óptico de uma imagem. Foi inicialmente desenvolvido como software proprietário pela HP durante os anos 80, e obteve notabilidade no Teste Anual de Acurácia de OCR em 1995. Em 2005, teve sua licença trocada para *open source* e é mantida pela Google desde 2006, sobre a licença Apache 2.0. Está disponível como ferramenta de linha de comando para as plataformas Linux, macOS e Windows. As versões mais recentes possuem APIs para as linguagens C/C++ e Python. A partir da versão 4.0 foi adicionado um EM baseado em redes neurais LSTM, que funciona bem no x86/Linux com os dados de Modelo de Linguagens oficiais.

O processamento segue um pipeline passo-a-passo, em que as versões mais recentes fazem a binarização de imagens coloridas usando a método de Otsu pela biblioteca de processamento de imagens Leptonica, que encontra um limiar global para a separação entre os caracteres e

o fundo. O tratamento para a correção da inclinação da imagem e a redução de ruído, importantes para o desempenho da leitura, não são realizados pelo Tesseract.

Para a segmentação, a análise de layouts se baseia na tabulação do documento, as tab-stops. Uma tab-stop é uma posição na horizontal que delimita o início ou o fim de uma linha, ou outro elemento não retangular, como uma imagem. A análise é feita de maneira híbrida por uma abordagem top-down, que encontra as colunas do layout e realiza o alinhamento, e outra bottom-up, que classifica os pixels como componentes conexos (CC). Os CCs são classificados em pequenos, médios ou grandes e tratados como ruído, texto ou título, respectivamente.

O reconhecimento de caracteres é realizado com unidades classificáveis, os 'blobs', compostos por um ou mais CCs sobrepostos na horizontal. A palavra é levada ao reconhecedor, e se o resultado for considerado insatisfatório, os caracteres com baixa acurácia serão divididos pelo *Character Chopper*. Se a taxa de reconhecimento for baixa para a palavra, os fragmentos de carácteres são combinados e inseridos num espaço de busca A*(*best-first*). Os resultados da classificação de caracteres são combinados com um dicionário de palavras do modelo de idioma, escolhendo a string com melhor classificação [Araujo (2019)].

Atualmente o Tesseract produz saídas em .txt, .pdf, .tsv e .hocr. A versão usada nos experimentos é a 5.4.1, a mais atual no momento.

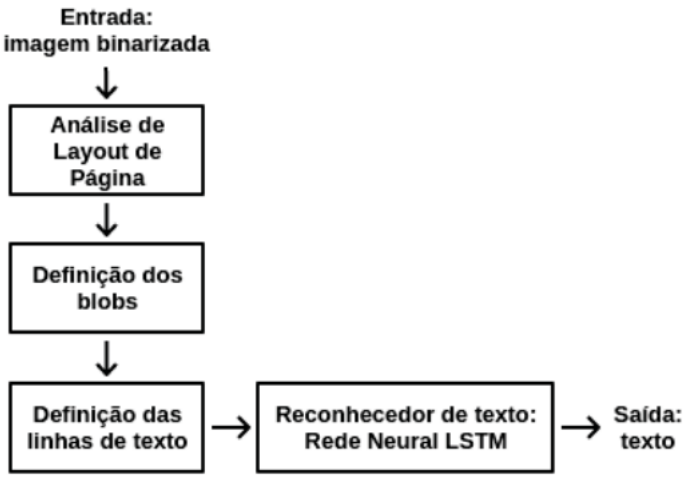


Figura 2.3: Componentes do workflow sem segmentação do Tesseract, extraída de [Araujo (2019)]

2.4 Sobre a metodologia de acurácia do OCR

Para medir a acurácia do OCR usamos a distância de Levenshtein como métrica de String, segmentando o texto por linhas no jornal. Isso é, compararemos a saída do OCR linha por linha com o gabarito do texto presente na imagem.

Dessa forma obtemos palavras de tamanho comparável entre si e evitamos saltos de indexação na comparação caso haja um lapso no reconhecimento de uma palavra, o que causaria uma incorrespondência nas comparações seguintes.

A distância de Levenshtein é uma métrica em que duas strings são comparadas identificando o número total mínimo de substituições (troca de caractere), remoções (ausência de caractere) e inserções para transformar uma string na outra.

Formalmente em $d(a, b)$ temos:

$$d(a,b) = \begin{cases} |a| & \text{se } |b| = 0, \\ |b| & \text{se } |a| = 0, \\ d(a_{1\ldots|a|-1}, b_{1\ldots|b|-1}) & \text{se } a_{|a|} = b_{|b|} \\ 1 + min \begin{cases} d(a_{1\ldots|a|-1}, b) \\ d(a, b_{1\ldots|b|-1}) \\ d(a_{1\ldots|a|-1}, b_{1\ldots|b|-1}) \end{cases} & \text{caso contrário,} \end{cases}$$

Para exemplificar, a palavra ’gaita’ está a distância 3 da palavra ’gatos’, pois precisamos substituir o caractere ’a’ por ’o’, remover ’i’ e inserir ’s’ para transformar a primeira string na segunda, o que totaliza 3 operações com caracteres.

Capítulo 3

Experimentação e Resultados

3.1 Tratamento das imagens

Selecionou-se uma amostra de páginas dos periódicos em formato TIFF com alta resolução de 600 dpi para tratamento da imagem com o OCR4all. As imagens são convertidas em PNG pelo software, e passam por etapas de conversão em níveis de cinza, binarização e remoção de ruídos. Foram recortadas manualmente amostras de texto binarizado e binarizado com remoção de ruídos usando ferramentas de edição de imagem. Como o alfabeto latino só é usado em seções de título e anúncio, foram escolhidas seções de texto de fonte homogênea onde só o alfabeto ucraniano era utilizado. Terminologia própria do português é transliterado nas seções de texto para o alfabeto ucraniano pelos redatores, como por exemplo 'Diário oficial do Estado to Paraná' (Діярійо офісіял до Естадо до Парана).

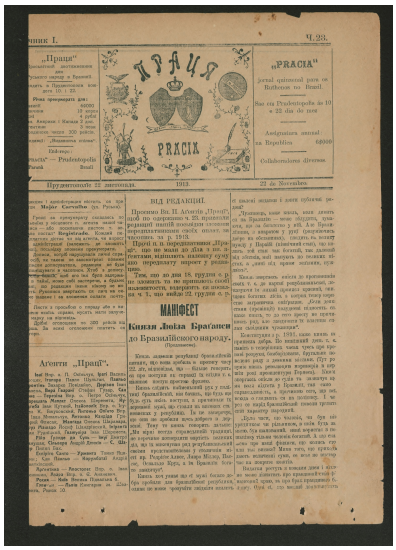


Figura 3.1: Imagem escaneada

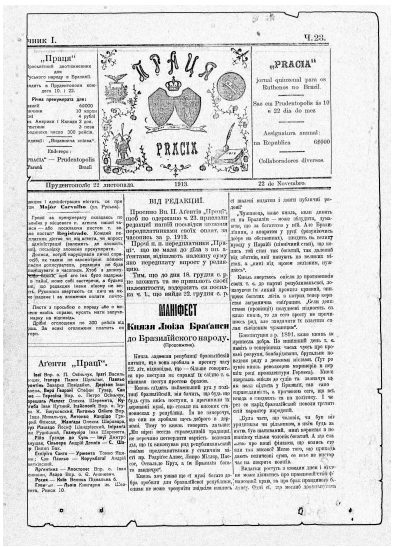


Figura 3.2: Imagem em cinza

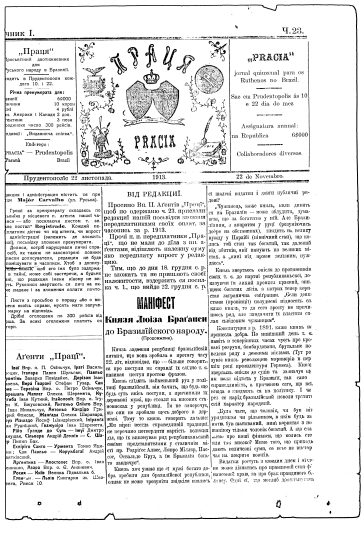


Figura 3.3: Imagem binarizada

Присяга.

Красна Ти, прекрасна руська Мамо!
Золотистими липами в літі,
Як пишна рожонька у цьвіті
Ти красуєш ся! . . . У твого храму
Нині урочисто присягаю.
Що тебе одну, одну кохаю!
Любий ти, народе руський, добрий,
Мова в Тебе пісонька прекрасна,
Серце в Тебе, зірка в небі ясна,
Дух Твій, витяз гарний і хоробрий,
Нині урочисто присягаю,
Що по вік, по вік Тебе кохаю!
В злиднях, серед болю й громів тучі,
В щастю, в ясному душі спокою
Ти одна, Вкраїно, перед мною
І Твої пісні сьвяті, могучі,
Тож Тобі на вік присягаю
Що Тебе о рідная, кохаю!

Сильвестер Яричевский.

(a) Recorte binarizado

(b) Recorte binarizado com remoção de ruído

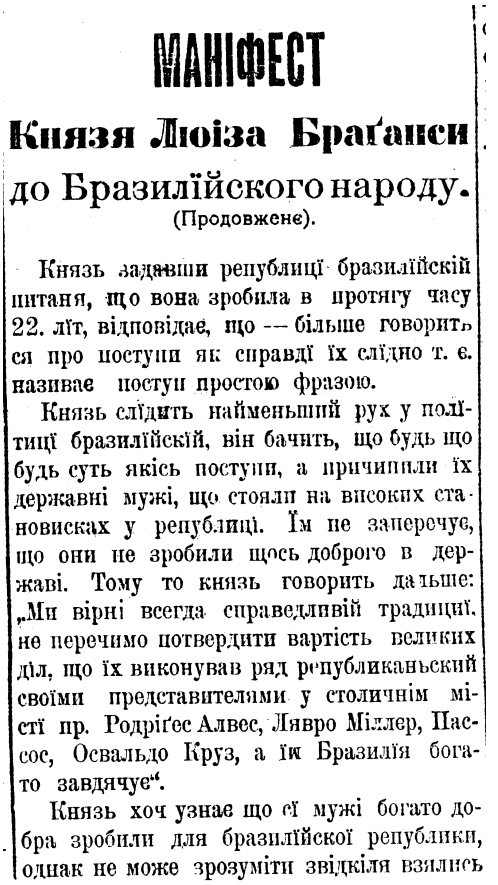
Figura 3.4: Comparação entre imagem binarizada e com tratamento de remoção de ruído

3.2 Reconhecimento e validação

Os recortes selecionados foram processados pelo Tesseract 5.4.1. Os arquivos de saída foram comparados com arquivos gabarito (ground truth) em que cada linha do jornal foi comparada usando a distância de Levenshtein, gerando logs com informações sobre o número de erros de caractere, o número de linhas com erro, as linhas em que houve discrepância entre a saída e o gabarito e a distância entre elas. Descobriram-se assim os erros mais frequentes de reconhecimento.

A seguir temos o arquivo de log gerado pela análise do recorte da figura 3.5:

Razão de erros de caracter: 27 / 888 = 0.03
Razão de linhas erradas: 13 / 26 = 0.5
[Князя Люіза Браганси,]-[Киязя Люїза Браганеи,] -> 3
[до Бразилїйского народу.]-[до Бразилійского народу.] -> 1
[Князь задавши републиці бразилїйскій]-[Князь задавши републиці: бразилійскій] -> 3
[22. лїт, відповідає, що -- більше говорить]-[29, діт, відповідає, що -- більше говорить] -> 4
[ся про поступи як справдї їх слїдно т. є.]-[ся про поступи як справді їх слідно т. є.|] -> 3
[Князь слїдить найменьший рух у полі-]-[Князь слідить найменьший рух у полі-] -> 1
[тиці бразилїйскій, він бачить, що будь що]-[тиці бразилійскій, він бачить, що будь що] -> 2
[новисках у републиці. Їм не заперечуе,]-[новисках у републиці. Їм пе заперочув,] -> 3
[діл, що їх виконував ряд републиканьский]-[діля, що їх виконував ряд републиканьский] -> 1
[стї пр. Родріґес Алвес, Ляврo Міллер, Пас-]-[сті пр. Родрігес Алвес, Ляврo Міллер, Пас-] -> 2
[сос, Освальдо Круз, а їх Бразилїя бога-]-[сос, Освальдо Круз, а їх Бразилія бога-] -> 1
[то завдячує".]-[Їто завдячує".] -> 1
[бра зробили для бразилїйскої републики,]-[бра зробили для бразплійскої републики,] -> 2



МАНІФЕСТ

Князя Люіза Браґанси

до Бразилїйского народу.

(Продовженє).

Князь задавши републицї бразилїйскій питаня, що вона зробила в протягу часу 22. лїт, відповідає, що — більше говорить ся про поступи як справдї їх слїдно т. є. називає поступи простою фразою.

Князь слїдить найменьший рух у политицї бразилїйскій, він бачить, що будь що будь суть якісь поступи, а причинили їх державні мужі, що стояли на високих становисках у републицї. Їм не заперечує, що они не зробили щось доброго в державі. Тому то князь говорить дальше: „Ми вірні всегда справедливій традициї, не перечимо потвердити вартість великих дїл, що їх виконував ряд републиканьский своїми представителями у столичнім містї пр. Родріґес Алвес, Лявро Міллер, Пассос, Освальдо Круз, а їм Бразилїя богато завдячує".

Князь хоч узнає що сї мужі богато добра зробили для бразилїйскої републики, однак не може зрозуміти звідкіля взялись

Figura 3.5: Exemplo de amostra recortada

3.3 Desempenho em relação aos diferentes estágios de pré-processamento

Buscou-se analisar a taxa de erros de caracteres de recortes de imagens em todos os estágios de tratamento oferecidos pelo OCR4all e obteve-se os seguintes resultados por estágio, com a configuração padrão LSTM e dataset ukr-fast: recortes das imagens escaneadas tiveram 3,9%; imagens em tons de cinza tiveram 3,6%; imagens binarizadas tiveram 3,4%; imagem com remoção de ruído tiveram 6,1%. Foram introduzimos mais erros de reconhecimento com o tratamento de ruído, e esse mau desempenho pode ser explicado por remoções que alteram a forma de alguns caracteres e que até os eliminam completamente. Testes subsequentes usaram a imagem binarizada como referencial, já que obteve melhores resultados de reconhecimento.

3.4 Desempenho das opções de Engine Model de redes neurais

A partir da versão 4.0 o Tesseract faz uso de EM de redes neurais no processo de reconhecimento. É possível usar o EM LSTM, Legacy ou a combinação dos dois. Foi possível encontrar uma taxa de erros do reconhecimento dos textos recortados para cada versão disponível de redes neurais e datasets preparados para essas redes. Os modelos ukr-fast e ukr-best foram feitos

Figura 3.6: Exemplo de página com layouts complexos, textos com fontes variadas e língua mista

para o EM baseado em LSTM e só podem ser usados por ele. Os resultados do reconhecimento estão apresentados na tabela 3.1.

Tabela 3.1: Taxa de erros (%) na variação de datasets x redes neurais

	Legacy	LSTM	Legacy+LSTM
ukr-fast	–	3,4	–
ukr-best	–	3,9	–
ukr-best+ukr-fast	–	3,3	–
ukr	8,4	3,9	4,8
ukr+ukr-best	8,4	3,8	4,8
ukr+ukr-fast	8,4	3,2	4,8
ukr+ukr-best+ukr-fast	8,4	3,3	4,8

Conclui-se pelos resultados que a rede padrão LSTM é a mais adequada a tarefa, que o dataset ukr-fast obteve o melhor desempenho adotado isoladamente, e que se adotado em combinação com os outros produz taxa de acertos comparáveis entre si.

3.5 Erros

Obteve-se uma taxa de erros de caracteres de 3,2% em relação a todos os caracteres reconhecidos usando a configuração EM LSTM e os datasets ukr e ukr-fast combinados, a menor encontrada. Encontramos uma média de erros por linha de 1,07 com desvio padrão de 1,36, ou seja, os erros estão relativamente dispersos entre as linhas e são poucos, mas frequentes. Muitos deles são causados por especifidades tipográficas e facilmente reconhecíveis mesmo para alguém sem fluência em ucraniano. Algumas substituições e remoções foram causadas pela perda de tinta e deterioração da página escaneada. Os problemas mais recorrentes na leitura dos caracteres foram listados a seguir:

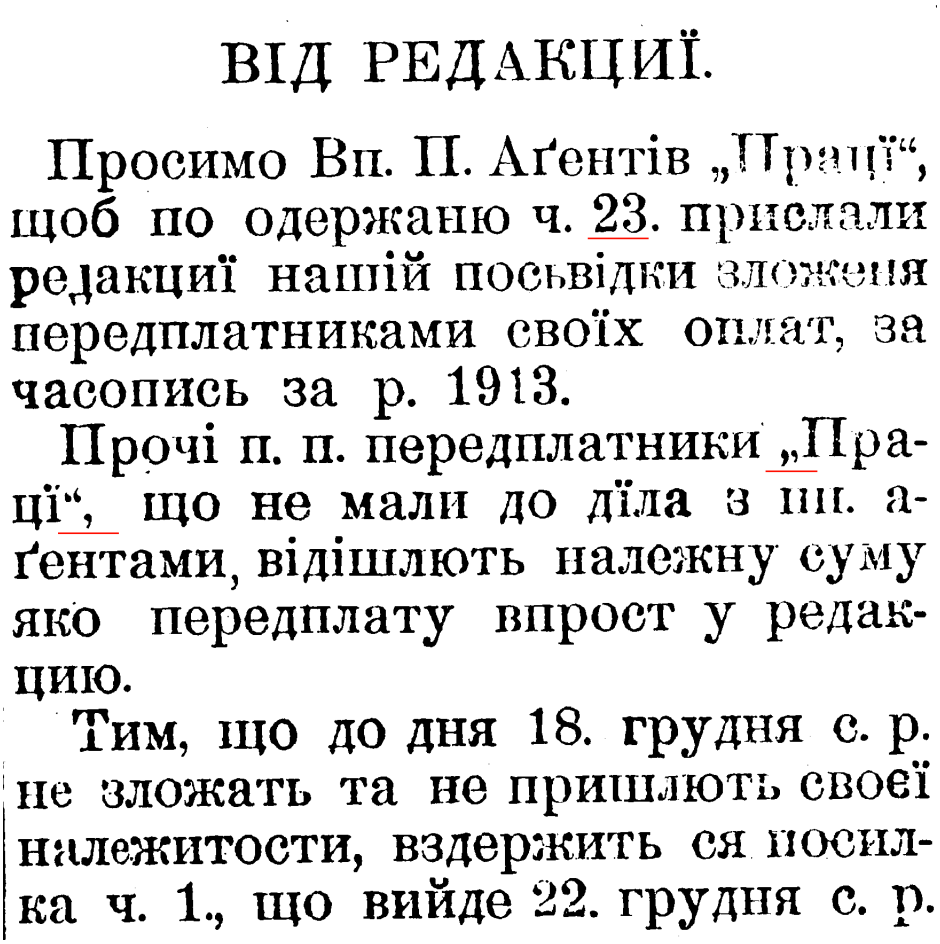
3.5.1 Substituições comuns do Tesseract

O erro mais recorrente foi a troca da letra ’ ї ’ por ’ і ’. Já que o Tesseract realiza a classificação por palavras completas pode-se especular que essa recorrência seja causada pelo fato de que a escrita do jornal seja datada em relação a novas reformas ortográficas que justifiquem essa substituição, consideradas pelo dicionário do Tesseract, ou até mesmo como um estilo ortográfico adotado volitivamente pelos redatores do jornal para reforçar a identidade cultural ucraniana, já que a letra ’ ї ’ é própria do alfabeto ucraniano e atualmente serve inclusive como símbolo de resistência e independência política [Kuznetsov (2022)].

A letra ґ foi reconhecido sempre como г , mas como ela é pouco usado na grafia e empregada geralmente em palavras estrangeiras, ela aparece muito raramente nos textos. Portanto ela pode não ser reconhecida no dicionário do Tesseract.

O algarismo 2 e 3 também foram substituídos com frequência respectivamente por 9 e 8, talvez pela formato encurvado da tipografia usada pela prensa do jornal. As escolhas tipográficas de pontuação das aspas e ponto introduziu substituições com caracteres de formato digital similar.

Para muitas letras do alfabeto ucraniano o que diferencia as maiúsculas das minusculas é apenas o tamanho, o que causou leituras errôneas. A substituição de ’н ’ por ’и ’ e vice-versa, letras de formato similar, também ocorreu com frequência. O símbolo de parágrafo ’ § ’ não foi reconhecido.



ВІД РЕДАКЦИЇ.

Просимо Вп. П. Аґентів „Праці“, щоб по одержаню ч. 23. прислали редакції нашій посьвідки зложеня передплатниками своїх оплат, за часопись за р. 1913.

Прочі п. п. передплатники „Праці“, що не мали до дїла з ш. аґентами, відішлють належну суму яко передплату впрост у редакцию.

Тим, що до дня 18. грудня с. р. не зложать та не пришлють своєї належитости, вздержить ся посилка ч. 1., що вийде 22. грудня с. р.

Figura 3.7: Recorte destacando caracteres com formas de difícil reconhecimento

3.5.2 Adições em bordas do layout

Outro problema recorrente ocorre quando o Tesseract considera a linha vertical que separa as bordas de layout como parte da string da linha, em especial em imagens inclinadas, e a identifica como ’ | ’, ’ I ’ ou ’ Ï ’.

Capítulo 4

Conclusão

Conseguimos obter transcrição dos textos de fonte homogênea dos periódicos com taxas de erro de caracteres na ordem de unidades de porcentagem. Podemos concluir que a revisão humana de alguém fluente em ucraniano seria necessário para eventuais correções do texto e que um OCR que não use reconhecedor por palavra é mais adequado no caso de documentos históricos, já que o dicionário de palavras geralmente considerara a grafia mais atual como correta. A necessidade de encontrar um processo automatizado de definição de layout dos textos e recorte solucionaria o maior desafio encontrado, visto que os métodos de pré-processamento, segmentação das linhas e reconhecimento oferecem bons resultados.

Referências Bibliográficas

Araujo, A. B. (2019). Análise de layout de página em jornais histórico germano-brasileiros. Dissertação de Mestrado, Universidade Federal do Paraná.

Bidwell, C. E. (1967). *Alphabets of the Modern Slavic Languages*. University of Pittsburgh. https://ia600508.us.archive.org/14/items/ERIC_ED016193/ERIC_ED016193.pdf . Acessado em 17/07/2024.

Burnat, F. A. (2010). O JORNAL PRUDENTOPOLITANO [PRACIA] uma experiência jornalística católica ucraniana de 98 anos em folkcomunicação. *Revista Internacional de Folkcomunicação*, 8(15). Disponível em https://www.redalyc.org/articulo.oa?id=631768778008.

Dimitrov, B. (2023). Book exhibition dedicated to the day of the cyrillic alphabet. https://blogs.eui.eu/library/cyrillic-alphabet/. Acessado em 17/07/2024.

Kuznetsov, S. (2022). In occupied ukraine, a letter in chalk symbolizes resistance. https://www.politico.eu/article/ukraine-resists-russia-letter-chalk/. Acessado em 06/08/2024.

Reul, C. (2024). User guide – introduction. https://www.ocr4all.org/guide/user-guide/introduction. Acessado em 06/08/2024.

Reul, C. e Christ, D. (2019). Ocr4all — an open-source tool providing a (semi-) automatic ocr workflow for historical printings. *Applied Sciences*.