Character Identification in Images Extracted from Portuguese Manuscript Historical Documents

Gustavo Cunha Lacerda and Raimundo C. S. Vasconcelos
Instituto Federal de Brasília, Taguatinga - DF, Brazil

Keywords: OCR, Text Recognition, Historical Manuscripts, Neural Networks.

Abstract: The creation of writing has facilitated the humanity's accumulation and sharing of knowledge, being a vital part of what differentiates humans from other animals and has a high importance for the culture of all peoples. Thus, the first human records (manuscripts), historical documents of organizations and families, began to have new perspectives with the digital age accumulation. These handwritten records remained the primary source for the history of countries, including Brazil before the period of independence, until the Gutenberg movable type printing press dominated the archival world. Thus, over the decades, these handwritten documents, due to their fragility, became difficult to access and manipulate. This has changed, with the possibility of digitization and, consequently, its distribution over the internet. Therefore, this work shows a solution for transcribing historical texts written in Portuguese, bringing accessibility, searchability, sharing and preservation to these records, which achieved a result of 97% of letters recognized in the database used.

1 INTRODUCTION

Writing is the symbolic process that allowed human beings to expand the information generated by the species far beyond what was possible in time and space for a single individual. In this way, given the importance of the emergence of this process, which marked the beginning of Ancient History around 4000 B.C., there was the beginning of the documentation and recording of human actions, beginning the era of the accumulation of knowledge, which is properly the creation of the culture of today's civilizations. Thus, writing and documentation play an extremely important role in human history, as they are fundamental tools for the preservation and propagation of knowledge throughout the generations. This production of information has a unique importance in the construction of the human being and is much deeper than we think, being the foundation for culture, technological and scientific knowledge.

The process of civilization has relied heavily on human records, which are mainly generated from social, family, and personal organizations. These records were traditionally written, copied, and edited by hand until the advent of primitive printing methods. The oldest known printed texts date back to Japan in the years 764-770, specifically Buddhist prayers. However, it was Johannes Gutenberg's invention of the movable type printing press in Europe during the 15^{th} century that popularized printing worldwide. Prior to Gutenberg's press, records were created by scribes or literate individuals who used their own handwriting, resulting in significant variations in the appearance of characters. This is evident in historical Portuguese documents, where cursive writing leads to different letter connectors and shapes depending on neighboring letters and the writer's consistency.

Before the invention of the Gutenberg press, records were written by hand by scribes and literate individual, resulting in non-standardization of characters and variations in cursive writing. This is evident in historical Portuguese documents. Additionally, there is a problem with storing these valuable documents, as many are incomplete and damaged. Incorrect storage practices include using unsuitable materials, lack of backup copies, and storing in hard-to-access locations. To preserve the integrity of these documents and make them easily accessible to researchers, proper storage is crucial.

The widespread use of computers has led to a shift towards digital documents, with software replacing handwriting. This trend has sparked debates about the abolished of handwriting in some European countries. In Finland, handwriting was discontinued in 2016, as typing was deemed more relevant in today's

Lacerda, G. and Vasconcelos, R.
Character Identification in Images Extracted from Portuguese Manuscript Historical Documents.
DOI: 10.5220/0012353900003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 3: VISAPP*, pages 401-408
ISBN: 978-989-758-679-8; ISSN: 2184-4321

context. Minna Harmanen, president of Finland's National Board of Education, acknowledged that while handwriting aids in motor coordination and memory development, the personal nature of cursive handwriting can harm literacy skills.

Through the use and spread of digitization, there is the process of collecting and storing non-original texts digitally through manual transcription or more modern techniques such as Optical Character Recognition (OCR), which is a machine learning application for recognizing characters in images. This technique is common in standardized typography and is currently widely used for text recognition in digital fonts such as signs and printed books (Edwards III, 2007).

This paper suggests an approach using image processing and neural networks to process and recognize the content of handwritten texts in Portuguese. These texts come mainly from images collected by the digital databases of the national libraries of Portugal, Brazil and France, each of which has an initiative to digitize documents physically available at their headquarters with the mission of cataloguing and reproducing handwritten historical documentation relating to the human history of various countries in various languages, including the official language of Portugal and several other nations such as Brazil.

2 ANALYSIS OF THE PROBLEM

Trying to recognize characters from handwritten or printed historical documents Insertionsare full of various types of problems, as they differ greatly from the more standardized structure of magazines, newspapers or scientific articles (Edwards III, 2007). According to (Yanque, 2018) these problems arise mainly because of characteristics related to:

- Individual's personal handwriting (Figure 1): This problem represents a significant challenge challenge for OCR systems because different people's writing styles vary a lot. Unlike standardized fonts, individual handwriting has unique idiosyncrasies with letter formats, sizes, spacing, and slant angles, making it difficult for OCR systems to accurately transcribe handwritten text.
- Marks of the tools used for writing (Figure 2): This problem mainly occurs when is used a of low-quality ink and paper, on which stains and paper marks appear on manuscripts. These problems are separate classified as faded ink (Ink fading over time), bleed-through (Stains that go through the paper) and ghosting (Letter stains from other text on the page).
- Deterioration (Figure 3): This problem is caused by time and various factors such as improper storage, age the archive and various material degradation processes that lead to holes, stains, tears and the total loss of documents.

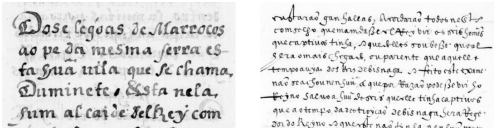

Figure 1: Individual's personal handwriting from two authors.

Figure 2: Examples of marks due to the material used at the time.

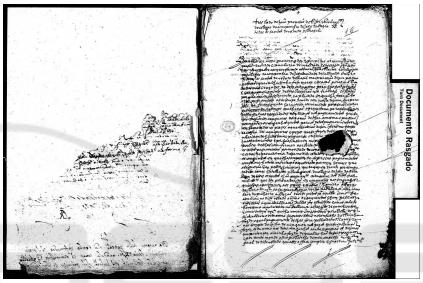

Figure 3: Example of a document that has been damaged.

3 RELATED WORKS

In (Ströbel et al., 2022), the authors develop a search for TrOCR on historical documents, proposed by (Li et al., 2021), using a neural network pre-trained with the dataset *Image-Net-1K*, containing 1.2 million images. Synthetic handwritten text creation techniques were used to augment the data and improve the training of the TrOCR architecture in the English language. Using the fine-tuning process with real data from other datasets, it was possible to obtain a CER of 2.55% in validation over the 15 fine-tuning epochs. The study concludes that the data shows that better accuracies come from methods using fine-tuning to improve accuracy.

In (Silvério Costa et al., 2022), the computational Lusophone language processing software *Lapelinc Framework* ((Costa et al., 2022)) was used. This software is made available through JSP and JPA. In addition, backend uses the *TensorFlow* library from *Python* to produce handwritten text recognition. The program accepts training based on the letters and words found and has proven to be of great value for recognizing and segmenting the words in documents.

In (Martínek et al., 2019), ideas are presented about how to train a neural network to recognize texts in historical documents in the best possible way using the state of the art. A way of generating new data is proposed, as well as three training approaches: using some real handwritten texts, using only manuscripts generated by software and the third approach combines the use of real and synthetic manuscripts. In this way, using old newspapers from Germany, the results showed that using only synthetic texts is sufficient to obtain results in OCR programs, but combining the two types of data is the most suitable way to guarantee better accuracy with little real data.

Another work by the same author, (Martínek et al., 2020), described methods for segmenting historical document structures and an OCR method for digitizing the information. In this project, *Kraken*, CNN and LSTM were used to produce a model that could be trained with data generated by software and adjusted using the fine-tuning technique with real data. The results proved to be superior to the state of the art in *Porta Fortium* dataset of old German newspapers, reaching an WER and CER of 0.118 and 0.024, respectively.

The *eScriptorium* browser program is described in (Kiessling et al., 2019). This project arose from the absence, in the state of the art, of a web program with a user-friendly interface and open-source to recognize texts from historical manuscripts. The software has an interface that allows both automatic and manual transcriptions and uses the same input sector. Thus, using the trainable OCR library for historical documents from *Python*, *Kraken* ((UNIVERSITÉ PSL, 2015)), it was possible to obtain an CER of between 2% and 8.9% on historical Arabic manuscripts.

In (Salimzadeh, 2019), a method is proposed for treating errors in the output of OCR algorithms, improving the quality of text digitization. Using natural language processing to correct the errors generated in the OCR output, it was possible to correct 92% of the errors generated in the output of the character recognition algorithms. Despite having used the English language for testing and training, the author proposed transfer-learning so that the system would work satisfactorily in other European languages.

The work (Neudecker et al., 2021) presents an evaluation of the metrics used to assess OCR models for historical documents. In this way, an experiment was carried out to see if common methods and metrics for evaluating OCR agree or disagree when evaluating the results for different historical documents, using two different data sets. Several common metrics such as CER and WER and alternatives such as the *GT-free heuristic* were used, comparing the efficiency, use and comparability of each.

In (Aguilar and Jolivet, 2022) a system for recognizing handwritten texts from medieval documents is described. The author used a pre-trained R-CNN with around 5.9 million adjustable parameters with a batch size of 1 and a learning rate of 0.0001 to train the network. It was thus possible to obtain an accuracy of 85% and, through fine-tuning, this accuracy increased to values above 90%.

In (Milioni, 2020), an analysis of the transcription program for historical documents called *Transkribus* is presented. A qualitative study was carried out based in two research methods: a distribution questionnaire and an experimental evaluation of the tool. The author concluded that the user interface of *Transkribus* is quite simple and user-friendly for humanities scholars, including all the processes and steps necessary to achieve an automatic transcription. However, the quality of the transcriptions produced is not yet the highest so that libraries and archives can use this tool systematically and for different types of historical material.

4 TECHNOLOGIES

For this work, the *Python* language was chosen, which is a high-level, general-purpose programming language created in 1989 by Dutch mathematician Guido van Rossum. It has a simple and intuitive syntax, which makes it an easy language to learn and use.

The ecosystem chosen to train an OCR model is Kraken, a system developed at the French Université PSL, optimized for recognizing texts from historical documents. This system has or is easily coupled with various tools ranging from data collection to publication of the generated model, and is widely used in the work found in systematic mapping. This choice is also due to the fact that *Kraken* is *open-source* and has a native option for *fine-tuning* through the use of a customized database, as in the case of this work.

The tools provided and supported natively by the ecosystem are:

- *Ketos*: this is the main *Kraken* tool, which has commands for managing data, compiling and generating model training and tests. Using *Ketos*, it is possible to create new data with the augmentation option, in which an alphabet is formed from the symbols found, which are then recombined into new word sequences.
- *eScriptorium*: is a software that facilitates certain OCR processes, mainly the creation of databases. It includes tools for processing, seg-

menting and creating the masks needed for transcription, which generates an ALTO file used by *Kraken* for training and testing. This file contains the coordinate information of the delimiters of the words and letters found with the appropriate transcription made by the user through a very simple website.

- *Zenodo*: this is a repository of neural network research and models in which *Kraken* stores its base models and is ideal for finding solutions made by other researchers and developers.

5 METHOD

The research method is shown in Figure 4 and defines the general tasks that make up the work. The first step is to define the problem, which consists of clearly and objectively defining the question that will be addressed throughout the project. In addition to presenting the context in which the problem is inserted and raising important definitions for the scope of the topic.

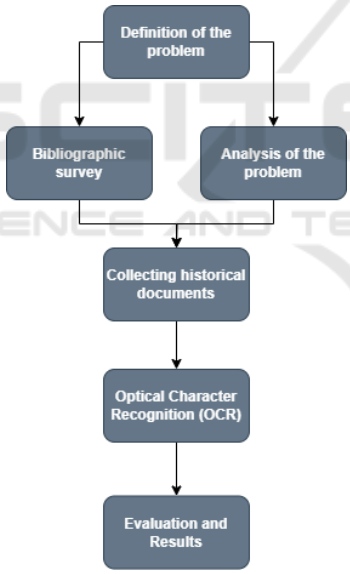

Figure 4: Steps of the research used in this work.

After, the process of analyzing the problem and surveying the literature begins in parallel. The problem analysis stage involves gathering information, assessing the context and expectations in order to find a possible solution. In the bibliographic survey stage, in the case of this research, a systematic mapping technique was used to search for, agglomerate and synthesize ideas, procedures and solutions from similar works.

Subsequently, data was collected to be used in the OCR stage, in which images are used as input for the system that generates the prediction of the texts in the images. In the case of historical documents, this process is carried out by collecting the original document and digitizing it using scanners and cameras. Once the digitized documents have been collected, they are agglomerated into datasets such as those used in this research. This was done through national digitization projects in Portugal and France. It was then transcribed by the author of this research to obtain properly cataloged data for use in supervised training.

In the stage of using OCR, the key steps for analyzing historical documents presented by (Dixit and Shirdhonkar, 2015) were used and adapted to the needs of this research. The adapted sequence of steps for the OCR stage is described in Figure 5 and has the following specifications:

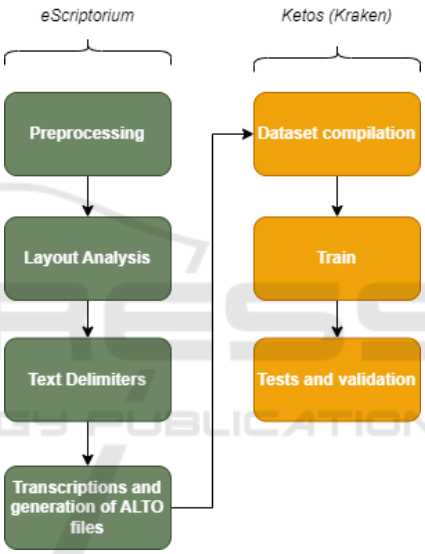

Figure 5: Specific OCR steps used in this work separated by tool.

- Image and Document Capture: The first stage of the process is understood as the collection and selection of data. In the case of historical documents, this process involves collecting the original document and digitizing it using scanners and cameras. Once the scanned documents have been collected, they are agglomerated and catalogued in datasets like those of the digital libraries used in this research.
- Binarization and pre-processing: at this stage the images are filtered to reduce wear and noise, with a focus on highlighting the textual components of the documents collected. An example of a crucial filter for text analysis is binarization, which consists of a process of selecting areas of interest in a given image using a cut-off threshold defined by the technique used. The result of binarization

is an image highlighted in binary pixels of intensity 0 or 1, usually white and black respectively. This technique is used in character recognition to highlight text in relation to the environment and material on which it is printed or written.

- Layout analysis: at this stage the document is separated by areas of interest for analysis. In the case of this research, texts are extracted from the documents and separated from images, drawings, graphics and other non-textual elements.
- Text analysis and recognition: once the texts have been properly filtered and separated, with the best possible highlighting, some recognition algorithm is applied to each character of the texts found in the document. The technique used in this research uses machine learning to recognize handwritten texts.
- Text description prediction: after recognizing each character, a statistical value between 0 and 1 is assigned, which corresponds to the chance of the character being a particular letter of the portuguese alphabet, making changes if necessary to form words correctly and maintain the original content of the document. This stage results in the completely digitized text of the initial manuscript.

In these more specific stages, related to optical character recognition, two main technologies from the *Kraken* ecosystem were used: *eScriptorium* and *Ketos*. At first, after choosing the documents, the steps that make up the construction of the database are carried out, in which the software web *eScriptorium*. It pre-processes, analyzes and recognizes the text and also creates transcription files for each word and letter found. Figure 5 shows how the stages are divided according to each technology.

Next, the ALTO files are compiled by *Ketos* into a *Arrow* file type used by the program to carry out training faster than directly using the documents in XML pairs, with coordinates and transcription, and the images. This compilation stage divides the data into personalized parts for training, validation and testing, which in the case of this work follow the proportions of 0.75, 0.15 and 0.15 respectively. Finally, after compilation and division, training is carried out using the *Ketos* training command with a pre-trained base neural network. We also used the SGD optimizer, learning rate of 0.001 and the tool's own data augmentation.

For the number of training epochs, a earling stopping solution was chosen, which stops training when there are a number of drops in accuracy in the validation at the end of each epoch, in the case of this research. This technique avoids overtraining (Overfitting is a common problem in machine learning and neural networks, in which a model overfits the training data. As a result, the model becomes very specific to the training data and is often unable to generalize well to new, unseen data, leading to poorer performance in real-world situations.) with the data used and works well in cases of fine-tuning (Fine-tuning is the process of adjusting a pre-trained machine learning model to a specific task or data set, taking advantage of the model's prior knowledge.)

5.1 Database and Transcripts

An extremely important part of defining the effectiveness of a neural network is the data on which it is trained. For this research, we couldn't find a dataset that had an available catalog and transcriptions of historical documents handwritten in Portuguese. Therefore, it was necessary to create a database of transcriptions to test how the model would adapt to Old Manuscript Portuguese.

The first step was to choose the documents needed to carry out the procedures. The criteria were files that had good reading quality and few faults to facilitate and speed up the data cataloging process, since the main objective of the research is not to create a large complex database that can be used by other OCR training models, although it is possible, as this would require increasing the overall scope of the research.

Using the search engines and filters of the website libraries, the few documents with a good reading quality were chosen so that there would be no doubts at the time of transcription, thus avoiding divergences caused by wrong classification.

After the selection, the documents were processed to remove artifacts, deterioration, stains and other problems that make it difficult to extract only the text. This part is left to non-linear binarization, which uses methods that take into account the relationship between neighbouring pixels and the distribution of intensities in the image to carry out the binarization instead of a fixed threshold value. This stage of the process highlights the letters written in focus on the page and eliminates most of the problems of historical manuscripts such as ghosting and bleed-through.

After the binarization stage, the images go through a layout analysis and segmentation process in order to eliminate the other non-textual components and ensure that only the text is catalogued and used in the transcriptions. This is done by another internal neural network and pre-trained by the *eScriptorium*, which has a very interactive abstraction in the website, which requires no configuration other than file selection and segmentation initialization. The infor-

mation found by the algorithm is saved in an ALTO file and has the coordinates of the limiters that form the mask of the text found by the program, the number of lines, text direction and transcription, as can be seen in Figure 6.

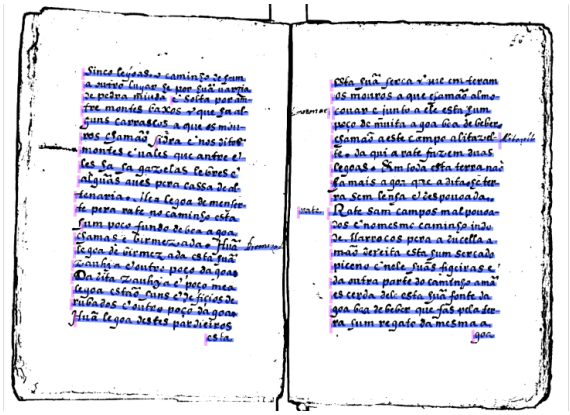

Figure 6: Result of segmentation and layout analysis on a part of a document history.

The documents are then transcribed via input by the user, who is able to make the relationship between the line mask and the written text. An example of how the process of transcribing a page of a document takes place can be seen in Figure 7. Finally, the transcriptions are saved in the ALTO files in pairs with the respective images.

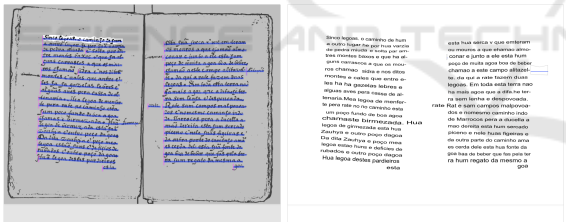

Figure 7: Example of transcription input based on lines in eScriptorium.

The process of creating the transcriptions resulted in 20,961 characters being transcribed into four documents totaling 307 pages. Not all the characters were transcribed, only those that there was no doubt about in order to maintain cohesion in the database. In addition, stains and other problems made it impossible for the segmentation algorithm to recognize some lines, which were also discarded from the generated files. The document with the most transcribed data is "Copia do emperio e reinos dos xarifes na Berberia em Africa e de algumas terras de negros, comessando da emperial cidade de Marrocos, cabessa do dito emperio e sua comarca." from the National Digital Library of France, with 16,349 letters corresponding to approximately 78% of the dataset.

6 RESULTS

Firstly, a curation process was carried out to see which model would have the best base performance to carry out the end-tuning (Results in Table 1). To do this, they were selected from the newest models available on the *Zenodo* platform, which *Kraken* recommends and uses as a solution repository. Thus, four models were chosen and tested on the base dataset.

Considering the accuracy results, the *Generic CREMMA* and *UB Mannheim German* models were chosen as the basis for fine-tuning.

The model using the dataset *CREMMA* reached a peak in character accuracy on the 24th epoch, with a value of 97.38% over 34 total epochs before early stopping stopped running (Figure 8). This fine-tuning result had a gain of 18.93% compared to the base model tests, which is a leap that justifies this type of training. The final generated model is 22MB and the process took around 6 hours to complete. In addition, the specific OCR metrics were used to measure the quality of the generated model in which the CER and WER values were 2.62% and 24.7%, respectively.

The results for insertion, deletion and addition errors can be seen in Table 2. Finally, Table 3 shows the relationship between the ten substitutions made by the model in the character tests, which explains which characters were the most difficult to transcribe correctly.

The *UB Mannheim German* model, on the other hand, reached its accuracy peak in the dataset created by this article in epoch 19, with a character accuracy value of 97.38% over the 29 total epochs before stopping its execution (Figure 8). This fine-tuning result represented an increase of 18.72% over the base model tests. The model was finalized with the same 16 MB size as at the start, which is much smaller than that of *Generic CREMMA* with a similar result.

In addition, the CER and WER values were 2.62% and 24.7%, respectively. The results for insertion, deletion and addition errors can be seen in Table 4. Table 5 shows the ten most frequent substitutions made by the model in the character tests, highlighting which characters were the most challenging to transcribe correctly.

Thus, the results were satisfactory compared to the state of the art of OCR in terms of letter matching. Words had a higher error rate because no treatment was carried out on the output of the predictions, as in related work. This absence is explained by the fact that applying a spelling correction to the text would require a specialized corrector for colonial Portuguese, something that was not easily found and that the construction was not in the initial proposal of

Table 1: Results of the curation of the OCR methods.

Modelo	Base(s) de Dados	Tamano	Acurácia de Letras
Generic CREMMA	CREMMA Medieval; CREMMA Medieval Latin; Eutyches.	22.8 MB	78.45%
Medieval Latin and French 12th-15th c. expanded	CREMMA Medieval; Oriflamms; Saint-Victor.	21.62 MB	71.39%
UB Mannheim German	Urfehdenbuch X.	16.4 MB	78.60%
Joseph Hooker HTR Model	Teanscrições de um pequeno grupo de voluntários.	16.2 MB	57.18%

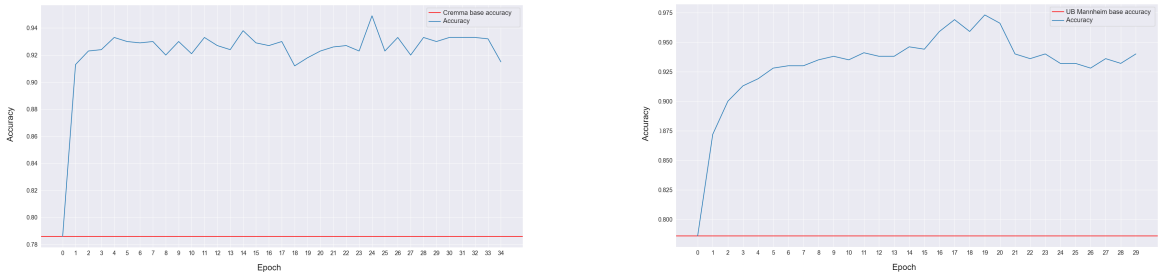

Figure 8: Accuracy in fine-tuning the Generic CREMMA and he UB Mannheim German models.

Table 2: Results of the specific OCR errors in the fine-tuning of the *Generic CREMMA* model.

Metric	Absolute Value	Percentage
Insertions	126	22,95%
Deletions	234	42,62%
Insertions	189	34,42%
Total	**549**	**100%**

Table 3: Results of the ten common errors in the predictions of the fine-tuning of the *Generic CREMMA* model.

No. of errors	Correct character	Prediction
294	”Nothing”	”Space”
99	u	v
36	”Space”	”Nada”
24	”Nothing”	o
24	f	s
6	-	”Nothing”
6	”Nothing”	s
3	g	b
3	a	”Nothing”
3	a	.

Table 4: Results of the specific OCR errors in the fine-tuning of the *UB Mannheim German* model.

Metric	Absolute value	Percentage
Insertions	198	35,3%
Deletions	135	24,06%
Insertions	228	40,64%
Total	**561**	**100%**

Table 5: Results of the ten common errors in the predictions of the fine-tuning of the *UB Mannheim German* model.

No. of errors	Correct character	Prediction
72	”Space”	”Nothing”
66	”Nothing”	”Space”
36	-	”Nothing”
18	”Nothing”	a
18	.	”Nothing”
18	a	”Nothing”
12	s	”Nothing”
12	S	s
12	t	s
12	ã	9

the research due to the non-mandatory nature of the transcriptions and the difficulty.

Another test was carried out with another document that had not been trained in the model, in which the accuracy results dropped considerably for a completely new type of writing and author. This data reveals that fine-tuning served to add the handwritings of the authors used in the constructed dataset rather than generalizing to Portuguese as a whole. The reason for this divergence is that de-standardization is colossal when it comes to cursive writing and the solutions based on *Kraken* were not good at universalizing with the limited amount of data provided.

Similarly, the *Generic CREMMA* and *UB Mannheim German* solutions declined in letter accuracy when using the dataset developed by this research using the same Latin characters. There is also great similarity in the initial accuracy and errors of the two base models, which reinforces this idea, but a larger amount of data and a better constructed database are needed to corroborate this information.

7 FUTURE WORKS

To continue the work presented, it would be ideal to segment the lines of research present, bringing more specific improvements to each aspect of the solution:

- Use or create some method of spell-checking the result of the predictions to improve the quality of the transcription of words from colonial Portuguese and other eras.
- Create another optical character recognition architecture that is better at generalizing the transcription of manuscripts with a smaller amount of data.

In this way, the solution could be more complete in all respects and would gain independence, especially in the creation of an improved database, to help and proliferate the identification of characters from images extracted from historical Lusophone manuscript documents.

8 CONCLUSION

Using a small training database to perform OCR on certain documents, using models trained in other languages with the same character, is not only possible but also very advantageous. Although the blind use of the model does not guarantee high accuracies, with a small number of transcriptions, it is possible to add the author's handwriting and the algorithm takes care of the rest of the translation. However, in this specific case of falling accuracy and the lack of treatment of the output with the use of spell-checkers, fully automatic transcription is impossible. Therefore, using *Kraken* with *eScriptorium* proved to be a better alternative to help digitize a handwritten historical document.

ACKNOWLEDGMENT

We gratefully to the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the support.

REFERENCES

Aguilar, S. T. and Jolivet, V. (2022). Handwritten text recognition for documentary medieval manuscripts. *HAL Open Science*.

Costa, B. S., Santos, J. V., Namiuti, C., and Costa, A. S. (2022). The systematic construction of multiple types of corpora through the lapelinc framework. In *International Conference on Computational Processing of the Portuguese Language*, pages 401–406. Springer.

Dixit, U. and Shirdhonkar, M. (2015). A survey on document image analysis and retrieval system. *International Journal on Cybernetics & Informatics*, 4:259–270.

Edwards III, J. A. (2007). *Easily adaptable handwriting recognition in historical manuscripts*. University of California, Berkeley.

Kiessling, B., Tissot, R., Stokes, P., and Ezra, D. S. B. (2019). escriptorium: an open source platform for historical document analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 19–19. IEEE.

Li, M., Lv, T., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., and Wei, F. (2021). Trocr: Transformer-based optical character recognition with pre-trained models. *arXiv preprint arXiv:2109.10282*.

Martínek, J., Lenc, L., and Král, P. (2019). Training strategies for ocr systems for historical documents. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 362–373. Springer.

Martínek, J., Lenc, L., and Král, P. (2020). Building an efficient ocr system for historical documents with little training data. *Neural Computing and Applications*, 32(23):17209–17227.

Milioni, N. (2020). Automatic transcription of historical documents: Transkribus as a tool for libraries, archives and scholars.

Neudecker, C., Baierer, K., Gerber, M., Clausner, C., Antonacopoulos, A., and Pletschacher, S. (2021). A survey of ocr evaluation tools and metrics. In *The 6th International Workshop on Historical Document Imaging and Processing*, pages 13–18.

Salimzadeh, S. (2019). *Improving OCR Quality by Post-Correction*. PhD thesis, PhD thesis, Universiteit van Amsterdam.

Silvério Costa, B., Viana Santos, J., and Namiuti, C. (2022). Transcrição manual e automática de textos históricos manuscritos através do software lapelinc transcriptor. *Colóquio do Museu Pedagógico-ISSN 2175-5493*, 14(1):2885–2890.

Ströbel, P. B., Clematide, S., Volk, M., and Hodel, T. (2022). Transformer-based htr for historical documents. *arXiv preprint arXiv:2203.11008*.

UNIVERSITÉ PSL (2015). Kraken. https://kraken.re/master/index.html.Online; acessed 26/12/2022.

Yanque, N. Y. A. (2018). *Um estudo comparativo de metodos de segmentaçao de documentos antigos*. PhD thesis.