Towards the Creation of a Collao Quechua–Spanish Parallel Corpus Using Optical Character Recognition

Gian Carlo Orcotoma **Lida Leon** **Hugo Espetia**
Universidad Andina del Cusco
{016101111f, lleon, hespetia}@uandina.edu.pe

Abstract

The Quechua language stands as a fundamental element of Peru's social and cultural identity, carries linguistic and cultural significance. However, it faces substantial challenges in terms of digital representation. One major limitation is the scarcity of resources such as a parallel corpus, which limits the development of technological resources for its analysis and practical application. This study addresses this gap through a methodology for building a parallel corpus using Optical Character Recognition (OCR). We digitized a collection of texts from a common origin to create a corpus that enables reliable access. The resulting corpus serves as a valuable asset for linguistic and Natural Language Processing (NLP) research, as well as for Quechua speakers. The source material derives from works produced by graduate students from the *Academia Mayor de la Lengua Quechua*, validated by academic staff, ensuring grammatical, syntactic and semantic integrity.

1 Introduction

Data from the 2017 Peruvian national census conducted by the National Institute of Statistics and Informatics (INEI) indicate that 13.6% of the Peruvian population identify themselves as Quechua speakers. The Cusco region has the fourth highest density of Quechua speakers, with 54.32% of its population who use Quechua as their first or secondary language in Peru.

Despite this, many government services are still not offered in Quechua, which makes it a crucial task to close this gap. In this study, we propose a technological approach using OCR to aid in the development of technology in this language.

Ortega et al. (2020) among other researchers in the field of natural language processing (NLP), categorize the Quechua language as Low Resource Language (LRL) due to the lack of available digital information. Technologies such as automatic translation, speech recognition, or natural language processing in general need a large amount of information and examples to be successfully trained.

The necessity for a parallel corpus made us think of the official website of Jehovah's Witnesses (https://www.jw.org/es/), who publish their magazines in almost 300 languages, including Quechua. This initially led to the idea of collecting this information using web scraping techniques. But most of these magazines, repositories, and websites had already been utilized in previous projects. In particular, the Jehovah's Witnesses website was comprehensively addressed in the JW300 project by Agić and Vulić (2019). Consequently we decided to get information from a different, yet equally reliable source.

We established contact with a Quechua language school based in Cusco city in Perú called Academia Mayor de la Lengua Quechua (AMLQ) which kindly granted us access to explore their library with an extensive collection of books. After reaching an agreement, we proposed a methodology for collecting the texts from this physical source. Our approach comprises several stages: i) Identification and collection of texts, ii) Photo environment setup, iii) Book digitizing, iv) Text labeling, v) Image preprocessing, vi) Text recognition and extraction, vii) Correction and evaluation of the OCR.

The importance of the result of this study lies in providing a methodology for the development of new technologies that require Collao Quechua - Spanish parallel corpora, such as automatic translators and sentence auto-completion systems. Furthermore, it contributes to the Quechua language preservation, by compiling and capturing examples of its usage in a digital medium that can be easily consulted.

Proceedings of the Workshop on Beyond English: NLP for all Languages in an Era of LLM from Texts associated with RANLP 2025, pages 1–6, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-105-9-001

2 Method

2.1 Identification and collection of texts

At the start of the project, the source of the texts was unclear. However, we had a clear starting point: Texts should be in a two-column format, aligned side by side in Quechua and Spanish, to facilitate the alignment and cleaning tasks in later stages.

At the end of their 8-month course, the students write a set of literary products, including stories, legends, poems, articles, academic essays, songs, etc. All of them are written in Quechua, with their respective side-by-side translation in Spanish and then printed in a book. After revision and correction by the school's own professors, the books are stored to be available in the school's library. The diversity in the book content arises from the many backgrounds of the students who learn this language. In a meeting for material identification, we counted approximately 30 pages per book and 10 books per year.

2.2 Photo environment setup

Following the identification of the texts, the procedure for digitizing the books was coordinated with the school. Ten books from the year 2021 were selected to be digitized; however, only five were finally processed with OCR.

The academy's library provided an ideal environment for capturing the photographs; The books are easily accessible and organized by year on shelves, and a central reading table is located in the available in the room. The photographs were taken on this table, using a Xiaomi Poco X3 Pro smartphone with a top-down angle for each page of the books.

To ensure optimal photo quality, and minimize the possibility of blurred or out of focus shots. We used a tripod, which stabilized the phone 20 cm above the book's pages, this distance was found to be the most suitable to capture the entire content of most pages. Some books were larger, so the tripod column needed to be raised a few centimeters to elevate the phone and increase the distance, in order to widen the field of view.

2.3 Book digitizing

Once the equipment was in place, the capture of the books began; maintaining consistent lighting and focus was essential to ensure image quality. To achieve this, the camera's "Pro" mode, which most phones have, was tested. While it is true that "Pro" mode has enough options to adjust camera exposure time, it does not have the option to numerically measure the focus distance. The Open Camera application was used as an alternative because it allowed these parameters to be adjusted and locked. The values used during capture were: exposure time: 1/50 s, aperture: f/1.79, ISO: 300 and focus distance: 20 cm (the same distance from the book to the phone's camera).

For the organization of the photographs, the prefix "AMLQ" was adopted. This prefix is configurable within the Open Camera application. The full file name of each photograph conforms to the format: AMLQ + date of the photograph + time of the photograph, for example: "AMLQ_20230103_123037.png". The photographs were stored in a folder created for each book, which was labeled with the year and the order in which the book was photographed. For example, "2021_L1", where "L" means "'Libro" (Book 1).

2.4 Photo labeling

A manual labeling approach was adopted, using the LabelImg tool, developed by Tzutalin (2024). Essentially, this involved enclosing the text areas on the pages within bounding boxes or rectangles. To distinguish between the two languages present, specific labels were applied: "que" label for areas containing Quechua text and "esp" label for those with Spanish text. We followed the YOLO standard described by Redmon et al. (2016) which utilizes four values (x, y, w, h) to define the center and size of each box in pixel units.

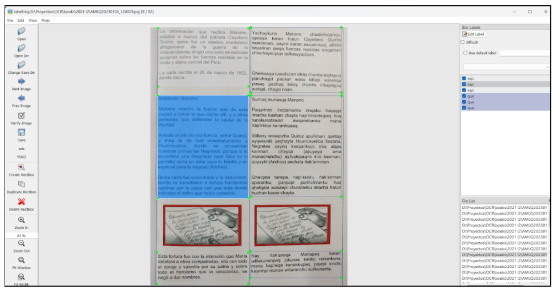

Figure 1: Labeling a page with Spanish and Collao Quecuha texts side by side with LabelImg software

2.5 Photo preprocessing

This step was developed using the Python programming language in conjunction with the OpenCV 2 library. The primary objective of this stage is twofold: first, to mitigate various errors that may arise from digitizing the physical books through photography, and second, to enhance the efficiency

of the Tesseract OCR (Optical Character Recognition) system. To achieve these goals, two specific types of filters were applied to the captured images:

Median Filter (Blur): Its main purpose is to smooth the edges of the characters and reduce noise present in the images. It also helps to lessen the prominence of serifs often found in fonts like Times New Roman, which could potentially cause confusion during the OCR application process.

Binarization Filter: The aim of this filter is to transform the image into a black and white representation. This binary conversion is fundamental because it significantly facilitates text recognition and serves to eliminate variations in lighting that might exist within the image. For this specific case, the Otsu binarization algorithm, described by Otsu (1979), was chosen. Despite efforts to maintain consistent lighting during photography, slight variations did occur, But overall the Otsu algorithm is well-suited for handling images with such light variations.

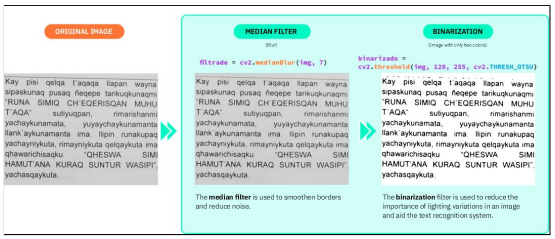

Figure 2: Photo preprocessing chain (median filter and binarization) software

2.6 Text recognition and extraction

Once the images are clean and ready, the text extraction process is carried out using Python and the Tesseract OCR engine. Tesseract was chosen due to its ease of use and its support for language recognition in both Quechua and Spanish which helped to minimize text recognition errors. Although the specific variety of Quechua supported by Tesseract OCR is unknown, it has proven to be better than the default spanish configuration in our photo database, accordingly, the corresponding configurations for both languages were used.

The output generated by the Tesseract OCR engine is in digital text format. For the purpose of this study, the Spanish and Quechua versions of the scanned text were saved in two separate .txt files. The naming convention for these files follows a defined pattern, it begins with the publication year of the book followed by the letter ”L” to denote ”libro” (book). Then, the order in which that book was published in that year is indicated. Finally, the language abbreviation of the content is specified (”esp” for Spanish or ”que” for Quechua).

An example of this naming convention is ”2021 l2_esp.txt” for the Spanish content of the second book from the 2021 set, and ”2021 l2_que.txt” for its corresponding Quechua content. Each of these files contains the complete text of the book in the respective language. Additionally, within each file, the original pages are also separated and indicated using brackets to denote the page number, e.g.: [1] for page 1, [2] for page 2, etcetera.

2.7 Correction and evaluation of the OCR

Following the text recognition step, an analysis of the scanned texts revealed few scanning errors for Spanish. However, this was not the case with the Quechua language, where common character identification errors were detected, such as letter confusions (e.g., mistaking ”q” for ”g”), character duplications, and unrecognized characters.

To address these errors and effectively evaluate the proposed method, we took a sample from the corpus. Specifically, we chose the Quechua version scan of the second book from 2021 (2021_l2_que.txt) to generate a second corrected file which would serve as ground truth. This corrected version represents the most faithful transcription of the original text and was used as a reference to evaluate the accuracy of the OCR output.

To create this ground truth file, we collaborated with a Quechua professor from the Universidad Andina del Cusco. Utilizing the photos of the original book as a reference, the professor manually corrected the text extracted from the Quechua section of Book 2 from 2021 using Microsoft Word.

Once completed, the evaluation constituted the final step; we adopted the approach described by Rice (1996) who defines OCR quality evaluation as: ”manipulation of character strings, which are transformed by an edit distance algorithm”. Following a review of OCR evaluation methods and metrics presented in the work of Neudecker et al. (2021), we selected CER (Character Error Rate) and WER (Word Error Rate) metrics for evaluating the OCR accuracy. CER is calculated as follows:

$$\text{CER} = \frac{S + D + I}{N}$$

Where N is the total number of characters, I

is the number of insertions, S is the number of substitutions, and D is the number of deletions needed in the OCR file in order to match the ground truth. WER is calculated similarly but at a word level.

For the evaluation, both CER and WER metrics were employed, as both represent the inverse precision of text recognition. After reviewing available tools, the open-source software ocrevaluation by Carrasco (2014), also described in the work of Neudecker et al. (2021) emerged as the most suitable option due to its comprehensive feature set, including the calculation of both CER and WER metrics, along with a comparative table of differences between the scanned text and the ground truth. ocrevaluation is available through a desktop Java application.

3 Results

Five graduation books from the Academia Mayor de la Lengua Quechua served as the data source for this study. These books contain texts in Spanish and Collao Quechua across diverse themes, including stories, poetry, history, science, lyrics, and personal narratives from the authors. To digitize the books, we set up an environment with uniform and constant lighting, a tripod, and a smartphone camera. Manual labeling and image preprocessing techniques were also employed to enhance the results of text recognition with the Tesseract OCR library. Subsequently, the text was stored in digital formats, which can be classified and located by year, book order, and language.

The corpus consists of a total of 44,263 words distributed across two languages. As shown in Table 1, the majority of the words are in Spanish, with 26,084 tokens (58.9%), while Quechua accounts for 18,179 tokens (41.1%). This distribution highlights the predominance of Spanish words in the dataset. However, it should be noted that the relatively lower word count in Quechua does not necessarily indicate less linguistic content, since Quechua is an agglutinative language in which a single word often carries the information that would require several words in Spanish.

The scanned texts exhibit certain errors, such as the insertion of unwanted characters (e.g., punctuation marks, hyphens, and alphanumeric characters in incorrect positions throughout the corpus), character confusions (where one character is mistaken for a similar-looking one), and deletions or omissions of some characters. The recognized Quechua texts present these issues more frequently than the Spanish ones, preventing the corpus from being a 100% accurate reproduction of the original books.

Language	Word count	Percentage
Spanish	26,084	58.9%
Quechua	18,179	41.1%

Table 1: Word distribution by language

For this reason, the Quechua text of book 2 from 2021 served as a sample to test the quality of the applied OCR. The evaluation, comparing the scanned text to the corrected text (or ground truth), revealed that 1.82% of characters were incorrectly detected according to the CER analysis, and 6.59% of words were incorrectly detected according to the WER analysis.

CER	1.82%
WER	6.59%
WER (order independent)	5.61%

Table 2: CER and WER results

In addition to the Tesseract-based pipeline, we evaluated transformer-based OCR architectures, specifically TrOCR and DONUT, using the following pretrained models:

- microsoft/trocr-small-printed
- naver-clova-ix/donut-base-finetuned-cord-v2

TrOCR achieved satisfactory results for English text but consistently failed to recognize the Spanish and Quechua texts in the photos, producing incoherent outputs. This behavior is expected given that its base model lacks multilingual training for these languages. DONUT, on the other hand, recognized both Quechua and Spanish words, but failed to correctly identify the character “ñ” and produced substitutions, likely because this character was absent from its original vocabulary. However, it returned the output as a structured JSON object rather than plain text. This demonstrates its document understanding capability but also indicates the need for fine-tuning to align its output with the parallel corpus structure required in this work.

4 Discussion

The present study introduces a Collao Quechua - Spanish corpus along with the method employed for its construction. This corpus includes books from the Academia Mayor de la Lengua Quechua's library, featuring a broad spectrum of themes. This diversity contributes to the variability and richness of the corpus, making it suitable for future research.

To evaluate the quality of the method, the CER (Character Error Rate) and WER (Word Error Rate) metrics were calculated on a Quechua sample from the corpus, producing errors of 1.82% and 6.59%, respectively. These results, while revealing text recognition errors, are encouraging, especially considering that Tesseract's default configuration for Quechua was used. Such errors were anticipated, and many were mitigated thanks to the preprocessing step.

In the research made by Cordova and Nouvel (2021), the scope extended to digitizing and correcting a dictionary for the Ancash Quechua variant, in addition to training an OCR model adapted to the specificities of that material. Their work compared three OCR software programs, with Tesseract emerging as the most accurate; however, similar errors were observed. This suggests that default configuration precision is often insufficient for low-resource languages with numerous variants, such as the Quechua family.

The work of Agarwal and Anastasopoulos (2024) highlights that incorporating OCR adaptation stages for each particular case significantly improves text quality in languages with limited digital resources. For instance, Cordova and Nouvel (2021) trained their own OCR model, while the present work included a photo preprocessing phase. The quality and resolution of images, font type (handwritten or computerized), lighting, color, and other factors can drastically affect OCR results, therefore adapting each OCR method to the specific problem presented is important.

In this work, the labeling and post-OCR correction phases were performed manually. However, Agarwal and Anastasopoulos (2024) highlight the existence of automatic processes based on machine learning algorithms, which reduce manual labor and cost.

Transformer-based models like DONUT and TrOCR offer greater robustness and contextual understanding, compared to traditional OCR methods, yet they require adaptation and fine-tuning for Quechua. This represents a possible future development path for this project, given that the manual post-correction stage only covered the Quechua sample from Book 2, leaving the possibility of its application to the remaining books.

The corpus provides a comprehensive and thematically rich collection that will serve as a valuable resource for future research in NLP and linguistics for the Collao Quechua variant. It is worth reiterating that no post-OCR processing (cleaning) of the texts has been performed; addressing this problem surely presents an opportunity for future research.

References

Milind Agarwal and Antonios Anastasopoulos. 2024. A concise survey of ocr for low-resource languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, page 88–102. Association for Computational Linguistics.

Željko Agić and Ivan Vulić. 2019. Jw300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Rafael C. Carrasco. 2014. An open-source ocr evaluation tool. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, DATeCH 2014, page 179–184. ACM.

Johanna Cordova and Damien Nouvel. 2021. Toward creation of ancash lexical resources from ocr. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, page 163–167. Association for Computational Linguistics.

Clemens Neudecker, Konstantin Baierer, Mike Gerber, Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. 2021. A survey of ocr evaluation tools and metrics. In *The 6th International Workshop on Historical Document Imaging and Processing*, HIP '21, page 13–18. ACM.

John E. Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

Nobuyuki Otsu. 1979. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection.

Stephen Vincent Rice. 1996. *Measuring the accuracy of page-reading systems*. Ph.D. thesis.

Tzutalin. 2024. Labelimg. https://github.com/HumanSignal/labelImg. Accessed: 2025-06-03.