AILLA-OCR: A First Textual and Structural Post-OCR Dataset for 8 Indigenous Languages of Latin America

Milind Agarwal, Antonios Anastasopoulos
George Mason University
Correspondence: magarwa@gmu.edu

Abstract

It is by now common knowledge in the NLP community that low-resource languages need large-scale data creation efforts and novel contributions in the form of robust algorithms that work in data-scarce settings. Amongst these languages, however, many have a large amount of data, ripe for NLP applications, except that this data exists in image-based formats. This includes scanned copies of extremely valuable dictionaries, linguistic field notes, children's stories, plays, and other textual material. To extract the text data from these non machine-readable images, Optical Character Recognition (OCR) is the most popular technique, but it has proven to be challenging for low-resource languages because of their unique properties (uncommon diacritics, rare words etc.) and due to a general lack of preserved page-structure in the OCR output. So, to contribute to the reduction of these two big bottlenecks (lack of text data and layout quality), we release the first textual and structural OCR dataset for 8 indigenous languages of Latin America. We hope that our dataset will encourage researchers within the NLP and Computational Linguistics communities to work with these languages.[1]

1 Introduction

Latin America is home to a linguistically diverse set of hundreds of indigenous languages. Many of these are low-resource in terms of text and audio resources, and generally lack basic natural language applications such as spell checkers, part of speech (POS) taggers, etc. However, these languages have a large number of digital resources (not machine-readable) in the form of recordings, plays, stories, and dictionaries. One major repository of such materials is the Archive of the Indigenous Languages of Latin America (AILLA), whose raw materials and digitizations form the core of the dataset in our paper (Agarwal and Anastasopoulos, 2024).

[1]Relevant code and data are available here

Figure 1: The AILLA-OCR corpus covers 8 indigenous languages spoken across 6 countries in Latin America. Languages differ in terms of vitality, with only South Bolivian Quechua with over a million speakers and some official status, but most others exist as minority languages in the respective countries (Table 1).

Of particular interest to us are linguistic materials such as grammars, dictionaries, ethnographies, and field notes, that can serve as training data for NLP applications and Optical Character Recognition (OCR). The goal of releasing this digitized and corrected dataset is to preserve invaluable linguistic materials, promote research on downstream tasks such as language identification and machine translation, and encourage better OCR techniques that allow for more accurate extraction of data from such corpora at scale (Nguyen et al., 2021; Agarwal et al., 2023). Modern OCR systems specialize in extracting text from such documents, but this requires high-quality layout detection to make the extracted text usable for downstream NLP tasks (Bustamante et al., 2020; Neudecker et al., 2021). While progress has been made on correcting the OCR *text* outputs after extraction, no work has focused on automatically correcting the *layouts* themselves either before/after text post-correction due to lack of annotated data. We aim to address

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 120–127
March 4-5, 2025 ©2025 Association for Computational Linguistics

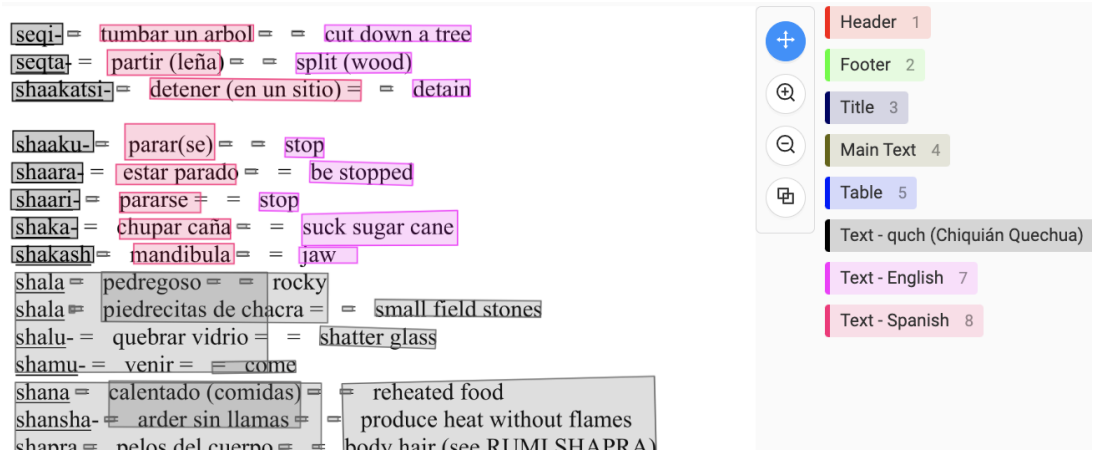

Figure 2: An in-progress annotation of a Chiquián Quechua language document (multilingual with Spanish and English) in our Annotation Workflow Portal. Here, the annotator is not only readjusting the detected bounding boxes (light grey), but is also correcting the textual errors in the new boxes, and labeling them (if language is known). Note that not all corrected bounding boxes need to be phrase or line-level. However, such organized post-corrected structure and text allows us to extract text more consistently.

this research gap by creating the first textual and structural OCR dataset for indigenous languages of Latin America. To summarize, our main contributions are:

1. OCR extractions from 8 Latin American indigenous languages from the AILLA collection.
2. Human-annotated text corrections for a sample of the digitized data, which can be used to model supervised post-correction of first-pass OCR output.
3. Structural post-corrections and associated metadata, including standard transformations like scaling, horizontal or vertical shifts, and creation of new gold-standard bounding boxes.

2 Language Profiles

South Bolivian Quechua (QUH) is a Quechuan language variety spoken primarily in Bolivia, but is also indigenous to some northern parts of Chile and Argentina. It is an agglutinative, polysynthetic language with a rich derivational morphology, is one of the most spoken indigenous languages in Bolivia with over 1.5 million speakers, and is constitutionally recognized. Ethnologue classifies South Bolivian Quechua's development as vigorous with standardized literature beginning to take shape. It is written in an extended Latin-based alphabet.

Mísqito/Mískito (MIQ) is a Misumalpan language spoken by more than 150K people (primarily Miskito) in Nicaragua and eastern Honduras. While orthographic conventions are not fully standardized, Miskito uses a subset of the Latin script for writing. Ethnologue's language vitalization hierarchy pegs Mísqito as threatened, since it is used for face-to-face communication within all generations, but it is losing young speakers to more dominant languages like Spanish and English.

Mam (MAM) belongs to the Eastern branch of the Mayan language family and is spoken by over 600K people mainly in Guatemala, where it is a recognized minority language. It is also called Qyo:l or Qyol Mam by its own speakers. Ethnologue classifies Mam's development as vigorous with standardized literature being steadily circulated. It is written in an extended Latin-based alphabet. Efforts to revitalize and preserve Mam have been ongoing, with initiatives such as bilingual education programs and the creation of written materials to strengthen literacy in both Mam and Spanish.

Chuj (CAC) is a Western Mayan (Q'anjob'alan branch) language spoken by about 60K people primarily in Guatemala. It uses the Latin alphabet and has two main dialects: San Mateo Ixtatán dialect and San Sebastián Coatán. It is heavily influenced by Spanish, the dominant and official language in Guatemala, and Chuj features heavy code-mixing and Spanish loan words. Ethnologue classifies Chuj's vitality as developing with standardized literature being developed due to language conservation and revitalization efforts taking place in San Mateo Ixtatán, through groups like the Academia de Lenguas Mayas de Guatemala.

Chimalapa/Oaxaca Zoque (ZOH) is an indigenous language primarily spoken in Oaxaca, Mexico

Language	693-3	Main Country	Speakers	Resource/Collection
South Bolivian Quechua	QUH	Bolivia	1.6M	Kalt (2016)
Mísqito	MIQ	Nicaragua, Honduras	150K	Bermúdez Mejía (2015)
Mam	MAM	Guatemala	600K	England (1972-1985)
Chuj	CAC	Guatemala	60K	Hopkins (1964)
Chimalapa Zoque	ZOH	Mexico	75K	Johnson (2000-2005)
Chiquián Quechua	QXA	Peru	<5K	Proulx (1968)
Sharanahua	MCD	Peru	<1K	Déléage (2002)
Tzeltal	TZH	Mexico	600K	Kaufman (1960-1993)

Table 1: A brief description of the 8 languages in our dataset, including their ISO 693-3 codes and other information about the primary country where it is spoken, and number of speakers. Along with this, we have also included references to the resources that are being released as part of the AILLA-OCR corpora's first release.

by about 75K speakers as per the 2020 report from the Mexican National Institute of Statistics and Geography. It is called Tzunitzame by its speakers and it belongs to the Zoquean language family. While it is written in the Latin script, there is no digital support for Chimalapa Zoque. As per Ethnologue, it's vitality is considered threatened as its face-to-face use among speakers is growing slowly.

Chiquián Quechua (QXA) belongs to the Central Quechuan language family and is spoken by less than 5K people primarily in Central Peru in the Bolognesi province. It does not have a standardized orthography and remains primarily oral. In AILLA records, it is transcribed in the Latin script like other American indigenous languages. Ethnologue classifies the language's vitality as *shifting* which means the language is no longer being consistently passed on to new generations, and speakers are instead shifting to Spanish.

Sharanahua (MCD) is an indigenous Panoan language spoken by less than 1000 people in Madre de Dios and Ucayali regions and the upper Purús river area in Peru. It is written in the Latin script and is spoken by all members of the small indigenous language community, who are also often bilingual in Spanish. Ethnologue classifies Sharanahua's vitality as developing with standardized literature being slowly developed due to low literacy rates and the small community size.

Tzeltal (TZH) is a Cholan–Tzeltalan Mayan language (also called Bats'il K'op Tseltal) spoken by about 600K people in Mexico. According to Ethnologue, it is a developing language, with increasing digital support, and a small amount of literature in its Latin-based orthography. Its usage is currently almost exclusively oral, and there is almost universal bilinguality in Spanish for younger speakers.

3 AILLA-OCR Corpora Creation

Language and Document Selection We selected 8 languages that have permissive licenses, use the Latin alphabet, whose special diacritics were available on the English keyboard, and which had typed documents (as opposed to handwritten ones) for this phase of the AILLA-OCR corpus. A uniform sample of pages, covering different layouts, is chosen for annotation per language.

Annotation Setup Annotators are trained to use the annotation platform using standardized guidelines (§A), and are allowed to label each corrected bounding box from several semantic categories (header, footer, title, main text, table, text - *lang_label* etc.), as shown in Figure 2.

Annotators When working with data in small indigenous languages for language documentation purposes, it can be extremely challenging to find native speaker data annotators. Previous work has shown that annotators without knowledge of the indigenous language can be reasonably adept at performing OCR corrections, provided they can read the script or are trained to read it (Rijhwani et al., 2023). So, for our 8 languages, we recruited 14 computer science graduate students as our annotators. The authors timed themselves annotating a small sample of pages and calculated an estimated commitment of 30 mins per 5 pages. Based on this, the payout rate was set at $20/10 pages (1 hour of work). Cumulatively, the annotation process itself costed ~$750 (~40hrs), not including time for recruitment, outreach, training, quality control etc.

In the current stage of the corpus, due to limited budget, we have *one* annotation per page, therefore inter-annotator agreement was not computed.

Manual Audit The lead author manually audited all annotators' annotations for all 8 languages. The author can easily identify Spanish, French, and English text in the documents. Moreover, since each multilingual document has document-level language identifiers, indigenous language text on a page was inferred and labeled by process of elimination and additionally confirmed by matching with the language's Universal Declaration of Human Rights text.

Annotated Corpus Table 2 shows the distribution of the annotated pages and other metadata. Overall, the annotators completed 340 pages. Previous work has used 10-30 pages (we share 50 for most languages) to train post-correction models and the first-pass OCR for unannotated pages can be used for pre-training (Rijhwani et al., 2020).

4 OCR Post-Correction

First-Pass OCR We use a high-quality commercial OCR system, Google Vision, that is known to work well on endangered-language documents (Fujii et al., 2017; Rijhwani et al., 2020). We define a document $\mathcal{C}$ as follows:

$$\mathcal{C} = \{p_i\}_{i=1}^{K} \quad (1)$$

where p_i denotes the i-th page of a K page document. Performing OCR on page p_i gives us a first-pass output, f_i in the form of n_i bounding boxes x and the texts within them a. Each x contains the set of coordinates for the bounding box, and the corresponding string a represents the text within the box.

$$f_i = [(x_1, a_1), (x_2, a_2), .., (x_{n_i}, a_{n_i})]$$

Structural Corrections Annotators are required to first structurally correct the first-pass OCR outputs. This would involve scaling, translating, merging, or splitting bounding boxes, while keeping the text within faithful to the each box's new coordinates. We frame the structure post-correction task as follows. For every OCR'd input page f_i, we output a corrected page

$$q_i = [(y_1, b_1), (y_2, b_2), .., (y_{m_i}, b_{m_i})]$$

where m_i denotes the number of new bounding boxes after post-correction (may be different from n_i). We consider human-corrected q_i as the ground-truth text and layout. Note that while this step mainly transforms the structure, it also involves transferring the first-pass text (x_i, x_{i+1}, etc) from the first-pass boxes that now make up the corrected box b_i, and therefore, the texts are labeled as y_i.

Text Corrections We frame the text post-correction task to follow the structural corrections made in the previous step. For every structure-corrected page q_i, we output a corrected page:

$$r_i = [(z_1, b_1), (z_2, b_2), .., (z_{m_i}, b_{m_i})],$$

where m_i indicates the gold bounding boxes, and z_i indicates the transformed and corrected text in box b_i as compared to the first-pass text in structure-corrected gold boxes, y_i. We use character and word-level error rates (CER and WER) to report the quality of the first-pass OCR and the post-corrected outputs from the annotators.

5 Correction Results

Text Corrections Based on the gold dataset created by our annotators, Table 2 shows an evaluation of the text quality of the first-pass OCR by Google Vision. We see that for almost all languages, the CER (character-level error rate) and WER (word-level error rate) are both reasonable (<10%, with the exception of MAM and MIQ). This range is to be expected for low-resource languages written in extensions of the Latin-script (even with diacritics or new characters) and those that don't have available language models for decoding in Google Vision (all selected languages). Since desired error rates for readability are usually less than 2%, the first-pass results are a great starting point and with efficient post-OCR correction modeling or alignment improvements, this error could be reduced further.

Structure Corrections We have included detailed statistics on structural annotations (Table 2) and the raw data contains detailed metadata. To the best of our knowledge, no previous work has explored modeling techniques for structure post-correction, and so we did not include a benchmark for this task. Classically, structure is learned and predicted as a first-step and more emphasis is laid on post-correcting the extracted text. We anticipate that with better alignment and structure, the CER/WER scores in Table 2 will decrease further and consistently across languages with post-correction.

				Structure				Text				
693-3	Multiling	P_{total}	P_{ann}	μ_1	μ_2	μ_3	$\mu_{\Delta b}$	$\mu_{\Delta l}$	μ_d	μ_i	CER	WER
ZOH	SPA,ENG	3744	50	1.02	4.85	4.93	-0.24	5.61	0.73	6.34	3.56	6.15
CAC	SPA,ENG	564	50	1.76	5.59	4.71	-1.20	-4.34	11.95	7.61	4.12	5.33
MAM	SPA,ENG	144	50	0.94	3.98	7.36	-7.74	7.55	17.34	24.89	10.56	19.66
MIQ	SPA,ENG	61	50	0.40	2.26	3.78	-7.16	8.04	16.20	24.24	10.47	12.34
MCD	FRA	209	50	1.45	4.08	4.72	-7.17	10.65	2.73	13.38	7.13	9.15
QUH	SPA,ENG	216	50	1.24	3.76	3.98	0.36	1.46	0.46	1.92	2.72	3.76
QXA	SPA,ENG	29	20	2.88	17.06	20.53	-41.00	7.06	60.82	67.88	6.64	9.60
TZH	SPA	38	20	1.69	6.77	4.62	-8.85	14.92	8.08	23.00	1.43	2.73
AVG				1.42	6.04	6.83	-9.13	6.37	14.79	21.16	5.82	8.59

Table 2: For each of the 8 indigenous languages, we report the number of pages that we have selected to be part of the first release of the AILLA-OCR corpora (P_{total}) and number of human-annotated pages (P_{ann}). Along with this, we report some metrics to gauge the quality of the first-pass OCR outputs and the corrections. For structural annotations, we report some metadata including transforming involving one, two, three coordinates of a first-pass bounding box (μ_1, μ_2, μ_3). Annotators reduced the aggregate number of boxes detected across languages, to simplify the detected layout to different extents ($\mu_{\delta b}$). For text-corrections, we report average change in length of page text ($\mu_{\delta l}$), character-level deletions (μ_d), and character-level insertions (μ_i), in addition to the achieved character and word-level error rates.

6 Related Work

OCR Resource Creation Text or image-based datasets and corpora are most commonly created by scraping or crawling the web; however, we would like to highlight a few OCR-created datasets, especially those that work with indigenous languages. Cordova and Nouvel (2021) addresses the lack of resources for Central Quechua, since resources exist mostly in the dominant Southern variety, using OCR technologies. Hunt et al. (2023) digitizes an Akuzipik (indigenous language spoken in Alaska and parts of Russia) dictionary parallel with Russian text, which is very valuable for downstream NLP tasks. Other relevant but non-OCR dataset creation efforts include Guarani-Spanish news articles' (Góngora et al., 2021), Nahuatl speech translation (Shi et al., 2021), and Mazatec and Mixtec translations (Tonja et al., 2023).

Post-Correction An ideal post-OCR text correction algorithm would model the error distribution of the OCR algorithm's output text and systematically correct it (Berg-Kirkpatrick et al., 2013; Schulz and Kuhn, 2017). This can be an extremely valuable tool when digitizing indigenous language documents because the OCR pipeline's decoder language model is often of low-quality due to the low-resource nature of indigenous and endangered languages. Across the digitization efforts that we've highlighted and amongst others, it is quite common to perform text-based automatic/human post-correction (Maxwell and Bills, 2017; Cordova and Nouvel, 2021; Rijhwani et al., 2021). However, as mentioned in § 5, for structure and layout detection, previous work has focused on layout detection as a first-step (Bustamante et al., 2020) and it has not been explored as a post-processing step. This is primarily because there is a lack of ground-truth structural data (which our dataset provides). Previously, two major studies (Blecher et al., 2023; Zhong et al., 2019) have used existing large-scale corpora like arXiv to extract large-scale ground truth (source-code); but, this approach is not scalable to resource-creation efforts involving low-resource languages.

7 Conclusion

We present the AILLA-OCR corpus covering 8 indigenous languages of Latin America spoken across 6 countries. Our dataset is the first textual and structural corrections dataset. All data has been audited carefully by the authors to maintain high-quality annotations and rich metadata for future researchers to build modeling approaches on top off our dataset. We train a popular post-correction model to benchmark the text-corrections that highlights the utility of our dataset and associated gaps in structure modeling approaches. We hope this dataset will serve as a starting point to researchers to build and test new modeling approaches for the unexplored task of structure post-correction. Future work can also explore what methods would work best for reducing the error rates (both text and structure). This could involve classic post-OCR neural correction methods or utilize current advances in multimodal large language models.

Limitations

The main contribution of this paper is a new resource for textual and structure OCR post-correction in 8 low-resource indigenous languages of Latin America. Since such a contribution is best suited to a short paper, we did not include more extensive benchmarking.

Ethics Statement

The raw data digitized and corrected as part of the AILLA-OCR corpus initiative is entirely hosted by AILLA. The data is freely available to the general public, with some files shareable through request. The data can be used without asking for permission, and without paying any fees, as long as the resource and collection is cited appropriately. We acknowledge the linguists, native and heritage speakers, and the AILLA team for creating such a valuable repository of raw data in indigenous languages of Latin America. Our dataset, by design, digitizes and augments the raw data, to allow researchers and language community members to utilize it for modeling, and for educational purposes. An ethical implication of this work is that it will allow for more sustainable and equitable work in language resource creation and natural language processing.

Acknowledgments

This work was generously supported by the National Endowment for the Humanities under award PR-276810-21 and the George Mason University's Doctoral Research Scholars Award 2024-25 The authors are also grateful to the anonymous reviewers for their valuable suggestions, feedback, and comments.

References

Milind Agarwal, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2023. LIMIT: Language identification, misidentification, and translation using hierarchical models in 350+ languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14496–14519, Singapore. Association for Computational Linguistics.

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)*, pages 88–102, Mexico City, Mexico. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, Greg Durrett, and Dan Klein. 2013. Unsupervised transcription of historical documents. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 207–217, Sofia, Bulgaria. Association for Computational Linguistics.

Tulio Bermúdez Mejía. 2015. Miskitu dance, food, and traditions: traditional miskitu food, dance, songs, festivities. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID ailla:119700. Accessed February 15, 2024. Other Contributors include Waldan Peter, Wanda Luz (Speaker), Bermúdez Mejía, Tulio (Transcriber), Waldan Peter, Wanda Luz (Translator).

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents.

Gina Bustamante, Arturo Oncevay, and Roberto Zariquiey. 2020. No data to crawl? monolingual corpus creation from PDF files of truly low-resource languages in Peru. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2914–2923, Marseille, France. European Language Resources Association.

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*, pages 163–167, Online. Association for Computational Linguistics.

Pierre Déléage. 2002. Sharanahua language collection of pierre déléage. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public.Accessed February 15, 2024.

Nora England. 1972-1985. Mam language stories and grammars. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID ailla:119520, ailla:119520,ailla:119520,ailla:119520,ailla:119520. Accessed February 15, 2024.

Yasuhisa Fujii, Karel Driesen, Jonathan Baccash, Ash Hurst, and Ashok C. Popat. 2017. Sequence-to-label script identification for multilingual OCR. In *14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, November 9-15, 2017*, pages 161–168. IEEE.

Santiago Góngora, Nicolás Giossa, and Luis Chiruzzo. 2021. Experiments on a Guarani corpus of news and social media. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 153–158, Online. Association for Computational Linguistics.

Nicholas Hopkins. 1964. A dictionary of the chuj (mayan) language language community. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID ailla:119647. Accessed February 15, 2024.

Benjamin Hunt, Lane Schwartz, Sylvia Schreiner, and Emily Chen. 2023. Community consultation and the development of an online akuzipik-English dictionary. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–143, Toronto, Canada. Association for Computational Linguistics.

Heidi Anna Johnson. 2000-2005. A grammar of san miguel chimalapa zoque. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID: ailla:119500 Accessed February 15, 2024.

Susan Kalt. 2016. Entrevista con tomas castro v y santusa quispe de flores. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID ailla:119707, ailla:119707 . Accessed February 15, 2024. Other Contributors include Waldan Peter, Wanda Luz (Speaker), Bermúdez Mejía, Tulio (Transcriber), Waldan Peter, Wanda Luz (Translator).

Terrence Kaufman. 1960-1993. Colección de idiomas mayenses de terrence kaufman. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID ailla:119707, ailla:119707 . Accessed February 15, 2024.

Michael Maxwell and Aric Bills. 2017. Endangered data for endangered languages: Digitizing print dictionaries. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 85–91, Honolulu. 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 *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing*, HIP '21, page 13–18, New York, NY, USA. Association for Computing Machinery.

Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. 2021. Survey of post-ocr processing approaches. *ACM Comput. Surv.*, 54(6).

Paul Proulx. 1968. Chiquian quechua vocabulary. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public.Accessed February 15, 2024.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR Post Correction for Endangered Language Texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5931–5942, Online. Association for Computational Linguistics.

Shruti Rijhwani, Daisy Rosenblum, Antonios Anastasopoulos, and Graham Neubig. 2021. Lexically aware semi-supervised learning for OCR post-correction. *Transactions of the Association for Computational Linguistics*, 9:1285–1302.

Shruti Rijhwani, Daisy Rosenblum, Michayla King, Antonios Anastasopoulos, and Graham Neubig. 2023. User-centric evaluation of OCR systems for kwak'wala. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 19–29, Remote. Association for Computational Linguistics.

Sarah Schulz and Jonas Kuhn. 2017. Multi-modular domain-tailored OCR post-correction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2716–2726, Copenhagen, Denmark. Association for Computational Linguistics.

Jiatong Shi, Jonathan D. Amith, Xuankai Chang, Siddharth Dalmia, Brian Yan, and Shinji Watanabe. 2021. Highland Puebla Nahuatl speech translation corpus for endangered language documentation. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 53–63, Online. Association for Computational Linguistics.

Atnafu Lambebo Tonja, Christian Maldonado-sifuentes, David Alejandro Mendoza Castillo, Olga Kolesnikova, Noé Castro-Sánchez, Grigori Sidorov, and Alexander Gelbukh. 2023. Parallel corpus for indigenous language translation: Spanish-mazatec and Spanish-Mixtec. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 94–102, Toronto, Canada. Association for Computational Linguistics.

Xu Zhong, Jianbin Tang, and Antonio Jimeno-Yepes. 2019. Publaynet: Largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 1015–1022. IEEE.

A Annotation Setup

Annotation Guidelines We shared annotatator assignments over email and the guidelines were shared using a YouTube video due to the visual nature of the task. We share the email template below with anonymity-compromising links redacted temporarily.

Subject: Annotation Assignments - [[NAME]] (Annotator #[[ID]])

Hi [[NAME]],

Thank you for being a part of this annotation effort for AILLA (Archive of the Indigenous Languages of Latin America). We appreciate you taking out the time to help us digitize and document these valuable resources. From the information you shared with us on the Google Form, you have been assigned **[[N]] labeling tasks**. Once you've completed your annotation assignment, please let us know (by replying to this email) and I will send you a **$[[AMOUNT]]** Amazon gift card. If you like doing the annotations, you can also always request more assignments.

Assignments:
Your unique ID is still **Annotator [[ID]]**
(Example) Language assignments:

- **mam [Mam]**. 7 pages. Task IDs: 40743-40749
- **cac [Chuj]**. 8 pages. Task IDs: 40280-40287
- **zoh [Chimalapa Zoque].** 15 pages. Task IDs. 39457-39471

While you only need your ID and language codes (mam, cac, zoh) to find your assignments, I will encourage you to check your tab before annotating to make sure you're actually seeing the tasks I've assigned you. If you notice anything off, just let me know.

Setup Instructions:
To enable swift annotation, we will be utilizing a open-source data labeling platform, [[redacted]]. If you haven't already, we invite you to create a Community Edition account through the signup link given below. We request that you not share the link publicly. [[redacted]]

Get Started:
Once you have created your account, you can can use [[redacted] to login and begin your annotations! We've made a short 5-minute video to guide you through the interface, how the annotation process works, and our expectations. Please watch it here **[[redacted]]** before you start annotation. The video is English closed-captioned (CC).

If you have any followup questions (about a specific assignment, the process, account setup etc.), please feel free to contact us on this thread.