A Concise Survey of OCR for Low-Resource Languages

Milind Agarwal
George Mason University
magarwa@gmu.edu

Antonios Anastasopoulos
George Mason University
antonis@gmu.edu

Abstract

Modern natural language processing (NLP) techniques increasingly require substantial amounts of data to train robust algorithms. Building such technologies for low-resource languages requires focusing on data creation efforts and data-efficient algorithms. For a large number of low-resource languages, especially Indigenous languages of the Americas, this data exists in image-based non-machine-readable documents. This includes scanned copies of comprehensive dictionaries, linguistic field notes, children's stories, and other textual material. To digitize these resources, Optical Character Recognition (OCR) has played a major role but it comes with certain challenges in low-resource settings. In this paper, we share the first survey of OCR techniques specific to low-resource data creation settings and outline several open challenges, with a special focus on Indigenous Languages of the Americas. Based on experiences and results from previous research, we conclude with recommendations on utilizing and improving OCR for the benefit of computational researchers, linguists, and language communities.

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).[1] Of the documents in AILLA's collection, particularly interesting to NLP researchers 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). Releasing digitized versions of such a repository of hundreds of datasets can preserve invaluable linguistic materials and accelerate research in NLP. Modern OCR can extract text from such documents, but this requires accurate layout detection and post-processing to make the extracted text usable for downstream NLP tasks (Bustamante et al., 2020). OCR is a well-established field, with its advances mostly drawing from innovations in Computer Vision. More recently, OCR has been increasingly used for resource-creation for low-resource languages in NLP contexts (Ignat et al., 2022a). There are also several excellent surveys and tutorials (Nguyen et al., 2021; Neudecker et al., 2021; Memon et al., 2020) on building and using OCR for broad applications, however, there is a dearth of specialized surveys for low-resource language OCR. Therefore, the aim of this paper is to fill this gap and acquaint researchers and language

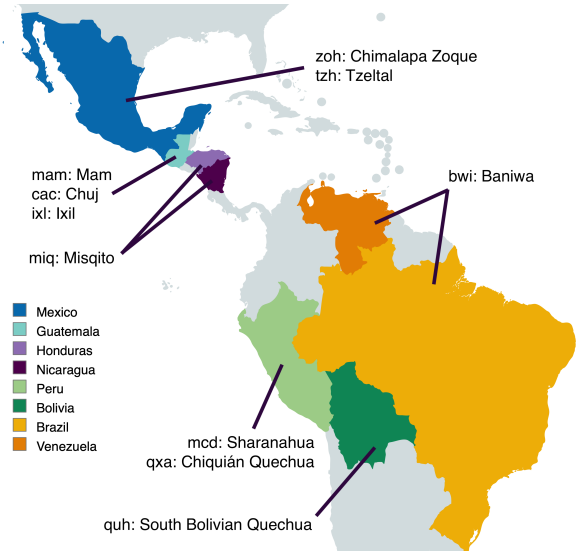

Figure 1: We highlight 10 Indigenous Languages from Central and South America with large amounts of undigitized resources to anchor our survey and workflow recommendations for researchers and linguists.

[1] A joint effort of the LLILAS Benson Latin American Studies and Collections and UT Austin.

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 88–102

Language	693-3	Family	Main Country	Speakers	Pages	Undigitized Resource
S. Bolivian Quechua	QUH	Quechuan	Bolivia	1.6M	216	Kalt (2016)
Mísqito	MIQ	Misumalpan	Nicaragua, Honduras	150K	61	Bermúdez Mejía (2015)
Mam	MAM	Mayan	Guatemala	600K	144	England (1972-1985)
Chuj	CAC	Mayan	Guatemala	60K	564	Hopkins (1964)
Chimalapa Zoque	ZOH	Mixe-Zoquean	Mexico	<10K	3744	Johnson (2000-2005)
Chiquián Quechua	QXA	Quechuan	Peru	100K	29	Proulx (1968)
Sharanahua	MCD	Panoan	Peru	<10K	209	Déléage (2002)
Tzeltal	TZH	Mayan	Mexico	600K	38	Kaufman (1960-1993)
Baniwa	BWI	Maipurean	Brazil, Venezuela	12K	310	Wright et al. (2000)
Ixil	IXL	Mayan	Guatemala	120K	2	Adell et al. (2016)

Table 1: A brief description of the 10 languages that we focus on to highlight the amount of data in Indigenous Languages of the Americas that requires high-quality OCR. We include their ISO 693-3 codes, primary country, number of speakers, and references to the resource that requires digitization. Overall, this data includes 5317 pages to be transcribed which *if digitized* can be sufficient to train many downstream NLP tasks.

communities with techniques and adaptions necessary for high-quality digitization in low-resource settings. To summarize, this paper makes the following contributions:

1. Highlights undigitized resources in 10 American Indigenous languages (§2).
2. First concise survey of OCR for low-resource settings and languages (§3).
3. Discussion on major open problems in scaling digitization for low-resource languages (§4).
4. Recommendations for researchers, linguists, and language communities on the entire resource curation and digitization pipeline (§5).

2 Undigitized Data

Over the past decade, many researchers, linguists, and consortiums have worked closely with native speakers and language communities to create datasets including digitized text, audio, transcriptions, translations, stories, etc. Some of these resources may not be machine-readable but include extremely valuable resources from an NLP perspective, such as multilingual lexicons, pronunication guides, plain text from a wide-variety of domains such as stories, essays, plays, news, linguistics etc. A comprehensive guide with resources in all low-resource languages (over 6000+) would be valuable but is out of the scope of this paper, so we focus on highlighting relevant resources in 10 American Indigenous languages. The AILLA collection contains several textual corpora in non-machine-readable image format for the selected languages in Table 1, as well as in hundreds of other indigenous languages of the Americas. The selected languages together cover over 5000 pages of undigitized data in these 10 languages. Each page contains multilingual textual data that needs high-quality extraction.

A large majority of OCR work for low-resource settings includes preservation and digitization of historical data, early printed books (Reul et al., 2017), palm-leaf manuscripts (Prusty et al., 2019; Sharan et al., 2021; Alaasam et al., 2019) etc. Pre-existing repositories (such as PubMed or arXiv) are also widely used for training OCR systems (Zhong et al., 2019; Blecher et al., 2023), however, note that this approach is not scalable to low-resource settings which often lack such ready-to-use datasets.

For the Americas, due to widespread adoption of extended Latin alphabets in writing, texts from the last couple of centuries are often typed, but several collections include partially or completely written handwritten documents and annotations. Historically used typing fonts may be challenging to decipher or out of use due to orthographic reforms (Naoum et al., 2019; Klaiman and Lehne, 2021; Jiang et al., 2019), and handwriting varies widely across individuals, making extraction challenging (Déjean and Meunier, 2019; Alaasam et al., 2019; Sharan et al., 2021). Over time, language communities may even adopt new orthographies, which might require researchers to build new keyboards and transcription systems to make the digitized corpora readable by community members (Rijhwani et al., 2023). Digitizing these resources can allow for more accessible linguistic research, training language models, translation systems, POS taggers, etc. The AILLA collection constitutes of a healthy

mix of both typed and handwritten text. As evident from Table 1, the highlighted languages will require sustained OCR efforts to digitize their respective resources. With access to machine-readable text, downstream NLP tools can then begin to be built.

Note that in our concise survey paper, our aim is not to digitize these specific books - that would warrant separate carefully designed studies as each resource is bound to have unique challenges and is connected to language communities with possibly different language technology needs. This paper highlights, for researchers unfamiliar with these languages and domain, different resources available for experimenting with OCR modeling approaches and recommended workflows for achieving such digitization.

3 A Concise Survey of OCR

Now that we've seen the data resources available for our 10 selected Indigenous languages (§2), we will highlight useful and practical OCR adaptations and innovations necessary for digitization of such low-resource language data. We cover techniques in four major parts of the digitization pipeline: preparation of the data and model, active training, decoding or generation, and post-processing. To ground the following discussion, we will define an example dataset $\mathcal{C}$, with K pages, where p_i represents separate pages. L represents the paired labels for each $p_i \in C$ (with each l_i representing the the ground-truth words and characters for the page p_i).

$$\mathcal{C} = \{p_i\}_{i=1}^{K}; \mathcal{L} = \{l_i\}_{i=1}^{K}$$

For an OCR experimental setup, we would usually have four different datasets: C_{pretrain} (unlabeled pages), C_{train} (with labels L_{train}), C_{val} (validation/development set used for evaluation during training along with labels L_{val}), C_{test} (for reporting model performance along with labels L_{test}).

3.1 Preparation: Setting the Stage

Data Augmentation Due to lack of data in low-resource indigenous languages, data augmentation should be the first step for any digitization pipeline, to increase the utility of the small labeled gold dataset (Shorten and Khoshgoftaar, 2019). For an OCR system, this means that the images themselves must go through several transformations such as skewing, binarization, scaling, cropping, blurring, etc. to ensure that the final model can handle such variations in-the-wild and still be able to extract text from the image. Data augmentation is well-studied in literature (Liu et al., 2018; Khan et al., 2021) and incorporating it into OCR pipelines has shown to increase robustness and performance by making the best use of a small training set (Storchan and Beauschene, 2019; Namysl and Konya, 2019).

More precisely, a set of augmentation operations, $O = \{o_1, o_2, .., o_j\}$ where j denotes the number of operations can be applied to each image p_i. o can denote functions like binarization, greyscale, gaussian blur, cropping etc. C_{train} can be augmented using any combination of operations from set O, to generate a new set $C_{\text{train-aug}}$, which would serve as the newly expanded training corpus. For each new augmented page, $p_{i,j} = o_j(p_i)$ and it's label would be $l_i \in L_{\text{train}}$.

Pretraining with General Unlabeled Data For data that is not labeled, self-supervised pretraining techniques are often used to better initialize the network (Li et al., 2023; Bugliarello et al., 2021). In case of encoder-decoder models, pretraining has been applied to both components separately and has been shown to be successful (Lyu et al., 2022; D'hondt et al., 2017), when large amounts of unlabeled images or text are available. Similarly, in case of the in-house pretraining set C_{pretrain}, ground-truth *text* labels are not available. So, the images from this pretraining set can be used to pretrain the OCR model, and the first-pass text can incorporated into pretraining the post-correction model (with learned denoising rules) (Rijhwani et al., 2020).

Transfer Learning from Related Languages For certain minority and low-resource languages, it is shown that a base system that is trained to identify a similar language or similar character-set generally leads to performance increases downstream(Lin et al., 2019; Zhuang et al., 2021; Rijhwani et al., 2019). As an illustration, in our selection of 10 American Indigenous languages, choosing a corpus in a high-resource language of Central and South America i.e. Spanish or Portuguese, might be appropriate for transfer learning. In the OCR domain, transfer learning has been applied to better enhance the quality of low-resource OCR output at the decoding step (Todorov and Colavizza, 2020; Jaramillo et al., 2018). However, Tjuatja et al. (2021) investigates transfer learning for OCR post-correction for indigenous and endangered lan-

guages and points at mixed results. They say that for downstream performance improvements, transfer learning is not straightforward and may require getting data from a larger set of domains. Gunna et al. (2021) investigated transfer at the text detection level for Indian languages, and observed positive outcomes when transferring from other Indian languages that look *visually* similar, even if they are from different language families.

3.2 Training: Learning Quickly and Better

For training OCR systems, supervised techniques are usually preferred in low-resource settings. Unsupervised methods have shown some promise recently (Gupta et al., 2021; Dong and Smith, 2018; Garrette and Alpert-Abrams, 2016), but they often require larger datasets for training. Since our focus is on low-resource indigenous languages, we restrict our discussion to supervised techniques. In this setup, there are usually two options - using off-the-shelf systems like Google Vision, Tesseract etc. or training from scratch. For Indigenous languages of the Americas, using off-the-shelf OCR systems can give an excellent starting point (Rijhwani et al., 2023), and since they are the focus languages of our paper, we will discussing training strategies on top of the first-pass OCR output obtained from such systems. Post-OCR processing aims to rectify mistakes made by OCR systems in text extraction, and can be extremely valuable for low-resource languages. Post-processing is valuable because it makes little to no assumptions about the first-pass OCR system itself (helpful when the system is commercial or closed-source) and instead focuses on improving the quality of the output (Kolak and Resnik, 2005).

First-Pass OCR For the first-pass, a high-quality OCR system, such as Google Vision or Tesseract, that is known to work well on endangered-language documents (Fujii et al., 2017; Rijhwani et al., 2020) is commonly used. 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.

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

Text Corrections 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 semi-automatic or human post-correction (Maxwell and Bills, 2017; Cordova and Nouvel, 2021; Rijhwani et al., 2021). For every first-pass page f_i, we output a corrected page:

$$q_i = [(x_1, b_1), (x_2, b_2), .., (x_{n_i}, b_{n_i})]$$

where x indicates the boxes from the first-pass, and b indicates corresponding corrected text. In human post-correction, an annotator (preferably a speaker of the language), would edit the first-pass OCR output to match with the ground-truth text as evident from the image. In semi-automatic setups, several consistent OCR errors may be identified from a small number of corrections and automatically applied to the remaining first-pass prediction to reduce the burden on the annotator.

Coverage Mechanism Since OCR is seen a generation task, it can be important for the model's attention distribution to pay attention to different parts of the input string. To ensure that this happens, a coverage mechanism is often introduced (Tu et al., 2016; Mi et al., 2016). This mechanism has empirically been shown to greatly improve OCR accuracy and seq2seq performance (See et al., 2017; Rijhwani et al., 2021; Klaiman and Lehne, 2021). A coverage vector at at time step t will be

$$c_t = \sum_{t'=0}^{t'=t-1} \alpha^a_{t'}$$

where α^a_t represents the attention distribution for the input a at time step t. This coverage vector c_t can be weighted and included in the attention computation for the next α_{t+1}, and be added to the base cross entropy loss as follows:

$$\sum_t \sum_{i=0}^{\text{len}(a)} \min(\alpha^a_{t,i}, c_{t,i})$$

Diagonal Attention Since post-correction from first-pass OCR output is mostly a copying step and reordering rarely occurs (Schnober et al., 2016), the model can mostly focus on generating the elements close to the diagonal. Therefore, under this paradigm, off-diagonal entries outside a certain radius are penalized more heavily by including them

in the training loss (Cohn et al., 2016). This simplifies the decoding step and encourages the model to maximize attention on items within the diagonal attention range. The modified loss function at time step t for a diagonal range d and attention distribution α would be:

$$\sum_{t'=1}^{t-d} \alpha^{a}_{t,t'} + \sum_{t'=t+d}^{\text{len}(a)} \alpha^{a}_{t,t'}$$

Diagonal attention shown to empirically improve OCR performance for low-resource languages (Rijhwani et al., 2021, 2020) and can be easily incorporated in OCR post-correction modeling.

Active Learning Data labeling is an expensive task for low-resource languages and especially so for a non-trivial annotation task such as OCR correction or image labeling. To select only those pages to annotate that would help the OCR model the most, a systematic paradigm called Active Learning can been utilized (Settles, 2012). For the low-resource OCR domain and for layout analysis, active learning has shown to be empirically quite valuable (Reul et al., 2018; Shen et al., 2022; Monteleoni and Kaariainen, 2007; Abdulkader and Casey, 2009; Gupta et al., 2016). It can help select which part of the C_{pretrain} to annotate and add into C_{train} using *query by committee* which trains several learner models on the current C_{train} and each model casts its vote/prediction on a set of V unlabeled examples from C_{pretrain}. In the equations below, $\text{uq}(\cdot)$ counts the number of unique characters in a list of predictions, M represent the independently trained models (m in total), s_v represents the v^{th} sentence in C_{pretrain}, and $V = \text{len}(C_{\text{pretrain}})$.

$$\text{ag}_{s_v} = \text{uq}([M_1(s_v), M_2(s_v), ..., M_m(s_v)])$$

$$v* = \text{argmax}^{V}_{v=0}(\text{ag}_{s_v})$$

Sample $v* \in C_{\text{pretrain}}$ is the sample that models disagree on most so it is actively added into the training set C_{train} (principle of maximal disagreement) since it would benefit from human annotation and improve the OCR model the most (Settles, 2012).

3.3 Decoding: To Generate or Not?

In this subsection, we'll discuss some recently proposed and empirically useful strategies to improve OCR decoding under low-resource settings.

Copy Mechanism Since at the decoding step, it is highly likely that most of the corrected text would be identical to the input, it is shown to be useful (Gu et al., 2016) to have two different probability distributions for decoding - *copy* and *generation*. At decoding, the model can choose, whether to sample from the attention distribution (P_{copy}) or generate the output through generation (See et al., 2017; Sutskever et al., 2014).

$$P_{\text{copy}}(y_t) = \sum_{t'=0}^{t} \alpha_{t,t'}$$

This can reduce the OCR character and word error rates by 2-5 times under low-resource settings (Rijhwani et al., 2020; Gu et al., 2016). Krishna et al. (2018) also use a copying mechanism for Sanskrit OCR and gain about 10% points over the base model with the copy mechanism, demonstrating that incorporating copying into an OCR pipeline for low-resource indigenous languages can be extremely beneficial. The copy probability can be weighted for each time step based on a $p_{\text{copy}} \in (0, 1)$ which can be generated as a weighted sum of the context vector, decoder state, and the previous time step's decoder probability. Therefore, we get the following copy-generation probability for a particular time step t and output string y:

$$p(y_t) = (1 - p_{\text{copy}}) * P(y_t) + p_{\text{copy}} * P_{\text{copy}}(y_t)$$

Lexical Decoding In order to counter the noise that self-training from the previous training step is bound to introduce i.e. reinforcing the errors from the first-pass, *lexical* adaptations have been successfully introduced in the OCR decoding step to improve the quality of the prediction (Schulz and Kuhn, 2017; Rijhwani et al., 2021). This proposed approach has shown to empirically benefit the decoding because it assumes that the correct forms of a word appear more frequently (assuming OCR errors to be inconsistent) and biases the output towards such observed forms.

3.4 Evaluation: How to Measure Progress?

Prediction Scoring and Evaluation Metrics When building an OCR system from scratch, mean-average-precision (mAP) and intersection-over-union (IoU) are the most commonly used metric to evaluate the quality of the bounding boxes. For the predicted bounding boxes $P = \{x_1, x_2, ..., x_e\}$, researchers commonly use IoU over all pairs of boxes to generate a ranked list of the best possible bounding box prediction and reference pairs (Girshick, 2015; Prasad et al., 2019; Prieto and Vidal, 2021). Then, a range of IoU thresholds can be used generate a confusion matrix from which we can

get a pair of precision and recall values for that threshold. Plotting these two values for all thresholds, we can get a precision-recall curve, the area under which is called AP i.e. average precision. We can get an AP for each reference box x_e, and averaging them all will generate a mAP for that page. This can indicate the quality of alignment of the predictions P with the true reference labels.

However, for many Indigenous Languages of the Americas, off-the-shelf systems and commerical systems will produce a reasonable first-pass prediction since they use extensions of the Latin alphabet (Rijhwani et al., 2020). In this case, evaluation needs to match two text strings: the prediction and the gold reference. For this, character error rate (CER) and word error rate (WER) are the most popular evaluation metrics. Depending on the language, both CER and WER may not be indicative - for polysynthetic languages where a large amount of vocabulary would be unseen at test-time, character-level error rate has been shown to be more indicative of OCR performance (Rijhwani et al., 2023).

$$\text{CER} = \frac{s_c + d_c + i_c}{n_c}; \text{WER} = \frac{s_w + d_w + i_w}{n_w}$$

where s, d, and i represent substitutions, deletions, and insertions at the $c$haracter or $w$ord level over the reference text which has n characters/words.

Loss Functions If using an off-the-shelf system for first-pass output, researchers only need to train post-correction models. In this case, a cross entropy loss is essential in addition to several other adaptive losses discussed in §3.2 such as diagonal loss and coverage loss (Cohn et al., 2016; Tu et al., 2016). To optimize a combination of these losses, common optimizers like Stochastic Gradient Descent (SGD) or Adam are often used (Rijhwani et al., 2020). In situations where the OCR system needs to be trained from scratch, per-pixel sigmoid or softmax losses are employed due to the pixel-level nature of the predictions from common models like Mask R-CNN and Fast R-CNN (Girshick, 2015; He et al., 2017). Multiple losses are generally computed if different branches of a network analyze and predict different aspects of the recognition task, and total loss in such cases can be computed by using a convex combination of these individual losses (Prusty et al., 2019).

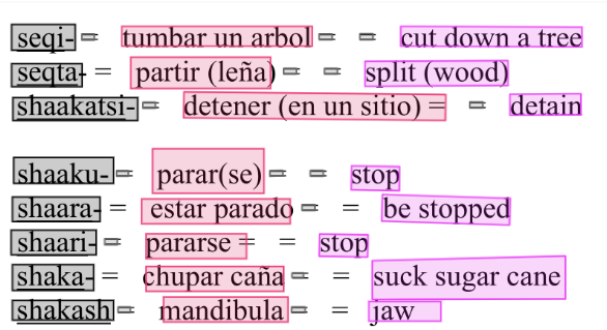

Figure 2: A post-corrected OCR document in Chiquián Quechua (multilingual with Spanish and English) from the AILLA collection (§2). Here, the annotator readjusted he detected bounding boxes, corrected the textual errors in the new boxes, and colored boxes belonging to the 3 languages differently.

4 Open Problems

Layout Preservation One of the most pressing issues remaining largely unsolved in OCR literature is that of structure preservation. OCR tools, especially those off-the-shelf may not good preserve the layout of the page in the output OCR text accurately and might require manual post-OCR alignment (Tafti et al., 2016; Rijhwani et al., 2020). The detected bounding boxes may not follow a logical layout as would be expected by human inspection. This means that researchers need to perform some level of alignment after getting the OCR outputs (Xie and Anastasopoulos, 2023), before applying OCR models (Ignat et al., 2022a), or cropping each image into separate line-level images (may be financially impractical if using commercial systems). From a resource-creation perspective for indigenous languages, preserving structure in the final output is extremely important, so we recommend that researchers think about how to design their experiments early on to address this issue.

To the best of our knowledge, while previous work has focused on layout detection as a first-step (Bustamante et al., 2020), it has not been explored as a post-processing step, primarily due to a lack of ground-truth structural data. 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. To build such a structure post-correction model, annotators would be required to not only correct the text in the OCR but also structurally correct the first-pass OCR outputs in some kind of graphical user inter-

face (as shown in Figure 2). This would involve scaling, translating, merging, or splitting bounding boxes, while keeping the text within faithful to the each box's new coordinates. Such a task could be framed as follows: for every text-corrected page q_i, we output a corrected page

$$r_i = [(y_1, c_1), (y_2, c_2), .., (y_{m_i}, c_{m_i})]$$

where m_i denotes the number of new bounding boxes after post-correction (may be different from n_i). We consider human-corrected r_i as the ground-truth text and layout. Note that while this step mainly transforms the structure, it would also involve transferring the initially corrected text (b_i, b_{i+1}, etc) from the first-pass boxes that now make up the corrected box y_i, and therefore, the texts are labeled as c_i. However, since such structural post-correction ground-truth data may be expensive to obtain, researchers may also consider getting this ground-truth from a dedicated layout detection model automatically, and post-correcting output from the best first-pass OCR system to adapt to this automatically-extracted desired layout.

Atypical Characters, Fonts, and Words Modeling historical orthographic variations with modern-day LMs, trained on current spelling conventions, can prove challenging during the decoding step (Poncelas et al., 2020). Work on better text extraction from historical documents from the printing press era resulted in the development of the popularly used unsupervised Ocular model (Berg-Kirkpatrick et al., 2013). Synthetic data has been successfully used before to offset the effect of atypical characters and typefaces (Borenstein et al., 2023; Drobac et al., 2017), and unsupervised techniques have been used to automatically learn the font style of a document in the context of historical document recognition and OCR (Berg-Kirkpatrick and Klein, 2014). However, research is still limited in the low-resource domain and researchers would need to ensure that their fonts and character sets are supported by their chosen OCR model (if training one from scratch) or are reconstructed using recent work in visual representation learning (Srivatsan et al., 2021; Vogler et al., 2022). For off-the-shelf systems (interfaced through APIs), it is not possible to directly include support for unique characters and researchers will need to add a post-correction step that includes a mix of post-processing scripts for easily resolvable errors and dedicated trained supervised models to correct first-pass output.

Linguistic Diversity For researchers working with several low-resource or indigenous languages at the same time, it can be desirable to train one model that is capable of handling different writing systems, diacritics, differing image qualities, and unique document formatting (Joshi et al., 2020). While one approach may be to develop 'language-agnostic' methods, previous work has shown that in practice such models are far from language-agnostic (Joshi et al., 2020; Bender, 2011) and tend to have high-performance only for a handful of languages. High OCR accuracy is usually desirable for all the low-resource languages under consideration, and in such a scenario, it may be best to train separate OCR or post-correction models.

5 Workflow Recommendations

In this section, we share recommendations based on the most successful strategies followed in the surveyed papers. This can serve as a starting point for computational researchers, linguists, and students new to the low-resource domain. We acknowledge that these recommendations, while grounded in our survey, are still subjective and researchers may need to modify some elements to suit their specific cases.

Language and Document Selection To anchor our survey, we selected 10 languages that have permissive licenses, use the Latin alphabet, whose special diacritics were available on the English keyboard, and which had typed documents in common fonts. Similarly, when selecting documents for their languages of interest, researchers should consider licensing, need for special keyboards, quality of the document, and layout diversity.

Evaluation Techniques For evaluation of the OCR quality, we recommend simply looking at the final output i.e. text. Provided there is a reference or gold text, these predictions can be compared to them and a CER/WER can be obtained. There is no gold standard target CER/WER so researchers will have to inspect the quality of the output and decide, with feedback from language community members, the CER/WER they would like to target in any potential modeling. Note that a line-aligned version of the extracted text will be required and this may either be obtained by only OCRing at the line-level (after cropping) or by aligning the predicted text with the reference text using metrics such as Levenshtein distance.

Preliminary Experiments For preliminary experiments, we recommend that 1-2 lead researchers manually annotate and audit a modest sample of the dataset themselves. This can help ensure that the researchers and language community members are familiar with the annotation workflow and can better guide any future annotators. Conducting some annotation before running OCR experiments is crucial because there needs to be some standard set for evaluation of all the models we will now experiment with. Once a few pages have been annotated, researchers can begin the OCR process using a common off-the-shelf OCR method like Google Vision or Ocular.

Data Annotation Now that the researchers have an idea of the quality of off-the-shelf OCR systems, we recommend that they recruit annotators to annotate a larger sample of the data if the OCR quality was found to be *low*. Annotators don't need to be speakers of the indigenous languages selected; however, they should have basic pattern recognition, data annotation, and typing skills. Previous work has shown that annotators without knowledge of the indigenous language are fairly adept at performing OCR corrections, if they can read the language's script and distinguish between any new diacritics (Rijhwani et al., 2023). Annotators should be trained to use the annotation platform using standardized guidelines, and a manual audit should be conducted by lead researchers to ensure compliance.

Post-Correction If the performance from preliminary experiments is satisfactory, we recommend post-correcting to further improve the results. A post-correction model should ideally help reduce character-level errors down to less than 5% (Maxwell and Bills, 2017; Cordova and Nouvel, 2021; Rijhwani et al., 2021). As discussed in §3, we recommend that researchers use a combination of copy mechanism, coverage, diagonal/positional attention, and active learning to improve performance. Rijhwani et al. (2021) implements most of the necessary post-correction features and their code can be used directly to train post-correction models for low-resource settings.

Training from Scratch On the other hand, if the preliminary experiments reveal that the error rates are quite high, researchers can consider simply training a custom OCR system from scratch. This will require a sizeable amount of annotated pages for training in addition to computational expertise in settling on the best hyperparameters and setting up the training pipeline. Human-annotations collected in the data annotation phase can be used to train OCR models from scratch and first-pass outputs can be used to train further with post-correction models. We recommend using open-source tools like Tesseract (Smith, 2007) or Ocular (Berg-Kirkpatrick et al., 2013) to train custom models due to their efficiency, optimizations, and active user community. More advanced researchers may also consider writing their own architecture and training pipeline from scratch. However, note that training systems from scratch is not straightforward, and researchers are bound to run into challenges. For instance, Tesseract has a high setup time and learning curve, doesn't have any graphical user interface, and requires high-quality images which may not be available for certain low- languages or image collections.

Deployment and Improvements For in-house use, the final trained model can be used directly to digitize the entire corpus and any other collections in that language. We recommend that computational researchers and language community members stay in touch throughout the training, annotation, and deployment process, and flag any issues with OCR quality and modeling. In some cases, if sufficient OCR quality is not being achieved despite trying the aforementioned techniques, some concessions and further data selection and annotation may be required. For instance, the quality of the data itself might need improvement (redoing scanning of the original source text), another phase of annotation may need to be conducted for substantially more data, or some unique algorithmic techniques may need to be developed to achieve quality OCR for the particular documents. We recommend using LabelStudio (Tkachenko et al., 2020-2022), which is an open-source labeling and annotation platform. The user interface is high-quality, user-friendly, and quite simple to setup and share with collaborators and annotators. There is also an active LabelStudio Slack where issues get resolved relatively quickly.

6 Related Work

Optical Character Recognition OCR has been studied as a research problem for decades, and today, commercial and open-source OCR systems can extract text quite accurately from most images

and can even be used in real-time due to test-time efficiency (Smith, 2007; Blecher et al., 2023; Berg-Kirkpatrick et al., 2013). OCR can involve extracting characters, words, paragraphs, and even preserving the layout of text on a page or in an image. OCR is widely used in the digital humanities (Reul et al., 2017; Rijhwani et al., 2021, 2020) and in businesses since it is a necessary step for digitization of rare manuscripts, books, linguistic field notes, invoices, business documents etc. It is also an invaluable technique in creating new data for low-resource languages for downstream NLP tasks and applications (Ignat et al., 2022b). In the last two decades, several excellent surveys from the computer vision community have been published covering OCR developments (Nguyen et al., 2021; Neudecker et al., 2021; Memon et al., 2020). In the low-resource domain, Hedderich et al. (2021)'s survey covers broad NLP advances but it does not cover optical character recognition. To the best of our knowledge, other than ours, no previous work has surveyed OCR for low-resource languages.

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 additional OCR-created datasets, especially those that work with American indigenous languages other than those reported in Table 1. Cordova and Nouvel (2021) addresse the lack of resources for Central Quechua, since resources exist mostly in the dominant Southern variety, using OCR technologies and share a successfully digitized corpus. Hunt et al. (2023) digitizes an Akuzipik (indigenous language spoken in Alaska and parts of Russia) dictionary parallel with Russian text, which was shown to be 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), which can serve as valuable pretraining corpora for OCR.

7 Conclusion

In this paper, we have presented a concise survey of optical character recognition (OCR) techniques shown to be most applicable to low-resource languages in the OCR literature. The survey is focused and similar work has not been published before due to the small community of OCR researchers working with low-resource and Indigenous language communities. We also highlight undigitized datasets in 10 Central and South American Indigenous languages, mostly from the AILLA collection, that can be extremely valuable to digitize for downstream NLP applications. Based on our own experiences and on findings from our literature review, we conclude with recommendations on utilizing and improving OCR for the benefit of computational researchers, linguists, and language communities. We hope that our paper can be used as a starting point for researchers or language community members wishing to digitize their resources but unaware of what OCR adaptations have become absolutely necessary to move towards a high-quality OCR output as well as what the open challenges in the field are.

Limitations

We acknowledge that even with the page limit provided by a long paper, fitting all details even for a focused topic like ours is not possible. Where possible, we have included the most relevant details, including mathematical equations, figures, and tables, in order to keep the survey concise and relevant to the AmericasNLP community. In addition, experimental results for the 10 Indigenous languages we selected to anchor our survey are out of the scope of this paper and would easily warrant a separate study.

Ethics Statement

The raw data resources shared for the 10 selected Indigenous languages are 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. 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. However, we don't foresee any negative ethical concerns with our work, which hopes to encourage open-source development of OCR models to allow researchers to move away from relying on commercial systems to process low-resource and Indigenous language data.

Acknowledgments

This work was generously supported by the National Endowment for the Humanities under award PR-276810-21. The authors are also grateful to the anonymous reviewers for their valuable suggestions, feedback, and comments.

References

Ahmad Abdulkader and Mathew R. Casey. 2009. Low cost correction of ocr errors using learning in a multi-engine environment. In *2009 10th International Conference on Document Analysis and Recognition*, pages 576–580.

Eric Adell, Antonio Moisés Toma Cruz, and Edelmira Catarina Sánchez Toma. 2016. Kam nib'anax tu ma'l xhemaana (a brief description of a typical week). In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID: ailla:119533. Accessed March 21, 2024. IXIL-CTZ-DES-EST-2016-06-23-0507.

Reem Alaasam, Berat Kurar, and Jihad El-Sana. 2019. Layout analysis on challenging historical arabic manuscripts using siamese network. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 738–742. IEEE.

Emily M. Bender. 2011. On achieving and evaluating language-independence in nlp. In *Linguistic Issues in Language Technology*.

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.

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, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 207–217. The Association for Computer Linguistics.

Taylor Berg-Kirkpatrick and Dan Klein. 2014. Improved typesetting models for historical OCR. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 118–123, Baltimore, Maryland. 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.

Nadav Borenstein, Phillip Rust, Desmond Elliott, and Isabelle Augenstein. 2023. PHD: Pixel-based language modeling of historical documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 87–107, Singapore. Association for Computational Linguistics.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994.

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 12th Language Resources and Evaluation Conference*, pages 2914–2923, Marseille, France. European Language Resources Association.

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. 2016. Incorporating structural alignment biases into an attentional neural translation model. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 876–885, San Diego, California. Association for Computational Linguistics.

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.

Hervé Déjean and Jean-Luc Meunier. 2019. Table rows segmentation. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 461–466. IEEE.

Eva D'hondt, Cyril Grouin, and Brigitte Grau. 2017. Generating a training corpus for OCR post-correction using encoder-decoder model. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1006–1014, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Rui Dong and David Smith. 2018. Multi-input attention for unsupervised OCR correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2363–2372, Melbourne, Australia. Association for Computational Linguistics.

Senka Drobac, Pekka Kauppinen, and Krister Lindén. 2017. OCR and post-correction of historical Finnish texts. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 70–76, Gothenburg, Sweden. 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.

Dan Garrette and Hannah Alpert-Abrams. 2016. An unsupervised model of orthographic variation for historical document transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 467–472, San Diego, California. Association for Computational Linguistics.

Ross B. Girshick. 2015. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1440–1448. IEEE Computer Society.

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.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Sanjana Gunna, Rohit Saluja, and C. V. Jawahar. 2021. Transfer learning for scene text recognition in indian languages. In *Document Analysis and Recognition, ICDAR 2021 Workshops, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part I*, volume 12916 of *Lecture Notes in Computer Science*, pages 182–197. Springer.

Anshul Gupta, Ricardo Gutierrez-Osuna, Matthew Christy, Richard Furuta, and Laura Mandell. 2016. Font identification in historical documents using active learning. *CoRR*, abs/1601.07252.

Harsh Gupta, Luciano Del Corro, Samuel Broscheit, Johannes Hoffart, and Eliot Brenner. 2021. Unsupervised multi-view post-OCR error correction with language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8647–8652, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2980–2988. IEEE Computer Society.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, 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.

Oana Ignat, Jean Maillard, Vishrav Chaudhary, and Francisco Guzmán. 2022a. OCR improves machine translation for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1164–1174, Dublin, Ireland. Association for Computational Linguistics.

Oana Ignat, Jean Maillard, Vishrav Chaudhary, and Francisco Guzmán. 2022b. OCR improves machine translation for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1164–1174, Dublin, Ireland. Association for Computational Linguistics.

José Carlos Aradillas Jaramillo, Juan José Murillo-Fuentes, and Pablo M. Olmos. 2018. Boosting handwriting text recognition in small databases with transfer learning. In *16th International Conference on Frontiers in Handwriting Recognition, ICFHR 2018, Niagara Falls, NY, USA, August 5-8, 2018*, pages 429–434. IEEE Computer Society.

Zhaohui Jiang, Zheng Huang, Yunrui Lian, Jie Guo, and Weidong Qiu. 2019. Integrating coordinates with context for information extraction in document images. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 363–368. IEEE.

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.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6282–6293. Association for Computational Linguistics.

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.

Umar Khan, Sohaib Zahid, Muhammad Asad Ali, Adnan Ul-Hasan, and Faisal Shafait. 2021. Tabaug: Data driven augmentation for enhanced table structure recognition. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part II*, volume 12822 of *Lecture Notes in Computer Science*, pages 585–601. Springer.

Shachar Klaiman and Marius Lehne. 2021. Docreader: Bounding-box free training of a document information extraction model. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part I*, volume 12821 of *Lecture Notes in Computer Science*, pages 451–465. Springer.

Okan Kolak and Philip Resnik. 2005. OCR post-processing for low density languages. In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*, pages 867–874. The Association for Computational Linguistics.

Amrith Krishna, Bodhisattwa P. Majumder, Rajesh Bhat, and Pawan Goyal. 2018. Upcycle your OCR: Reusing OCRs for post-OCR text correction in Romanised Sanskrit. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 345–355, Brussels, Belgium. Association for Computational Linguistics.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13094–13102. AAAI Press.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Manfei Liu, Zecheng Xie, Yaoxiong Huang, Lianwen Jin, and Weiyin Zhou. 2018. Distilling gru with data augmentation for unconstrained handwritten text recognition. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 56–61.

Pengyuan Lyu, Chengquan Zhang, Shanshan Liu, Meina Qiao, Yangliu Xu, Liang Wu, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. 2022. Maskocr: Text recognition with masked encoder-decoder pretraining. *CoRR*, abs/2206.00311.

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.

Jamshed Memon, Maira Sami, Rizwan Ahmed Khan, and Mueen Uddin. 2020. Handwritten optical character recognition (ocr): A comprehensive systematic literature review (slr). *IEEE Access*, 8:142642–142668.

Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. 2016. Coverage embedding models for neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 955–960, Austin, Texas. Association for Computational Linguistics.

Claire Monteleoni and Matti Kaariainen. 2007. Practical online active learning for classification. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8.

Marcin Namysl and Iuliu Konya. 2019. Efficient, lexicon-free ocr using deep learning. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 295–301.

Andrew Naoum, Joel Nothman, and James R. Curran. 2019. Article segmentation in digitised newspapers with a 2d markov model. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 1007–1014. IEEE.

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).

Alberto Poncelas, Mohammad Aboomar, Jan Buts, James Hadley, and Andy Way. 2020. A tool for facilitating OCR postediting in historical documents. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 47–51, Marseille, France. European Language Resources Association (ELRA).

Animesh Prasad, Hervé Déjean, and Jean-Luc Meunier. 2019. Versatile layout understanding via conjugate graph. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 287–294. IEEE.

José Ramón Prieto and Enrique Vidal. 2021. Improved graph methods for table layout understanding. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part II*, volume 12822 of *Lecture Notes in Computer Science*, pages 507–522. Springer.

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.

Abhishek Prusty, Sowmya Aitha, Abhishek Trivedi, and Ravi Kiran Sarvadevabhatla. 2019. Indiscapes: Instance segmentation networks for layout parsing of historical indic manuscripts. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 999–1006. IEEE.

Christian Reul, Uwe Springmann, and Frank Puppe. 2017. LAREX - A semi-automatic open-source tool for layout analysis and region extraction on early printed books. *CoRR*, abs/1701.07396.

Christian Reul, Uwe Springmann, Christoph Wick, and Frank Puppe. 2018. Improving OCR accuracy on early printed books by combining pretraining, voting, and active learning. *J. Lang. Technol. Comput. Linguistics*, 33(1):3–24.

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.

Shruti Rijhwani, Jiateng Xie, Graham Neubig, and Jaime G. Carbonell. 2019. Zero-shot neural transfer for cross-lingual entity linking. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6924–6931. AAAI Press.

Carsten Schnober, Steffen Eger, Erik-Lân Do Dinh, and Iryna Gurevych. 2016. Still not there? comparing traditional sequence-to-sequence models to encoder-decoder neural networks on monotone string translation tasks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1703–1714, Osaka, Japan. The COLING 2016 Organizing Committee.

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.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Burr Settles. 2012. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.

Prema Satish Sharan, Sowmya Aitha, Amandeep Kumar, Abhishek Trivedi, Aaron Augustine, and Ravi Kiran Sarvadevabhatla. 2021. Palmira: A deep deformable network for instance segmentation of dense and uneven layouts in handwritten manuscripts. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part II*, volume 12822 of *Lecture Notes in Computer Science*, pages 477–491. Springer.

Zejiang Shen, Weining Li, Jian Zhao, Yaoliang Yu, and Melissa Dell. 2022. OLALA: Object-level active learning for efficient document layout annotation. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 170–182, Abu Dhabi, UAE. 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.

Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *J. Big Data*, 6:60.

R. Smith. 2007. An overview of the tesseract OCR engine. In *9th International Conference on Document Analysis and Recognition (ICDAR 2007), 23-26 September, Curitiba, Paraná, Brazil*, pages 629–633. IEEE Computer Society.

Nikita Srivatsan, Si Wu, Jonathan Barron, and Taylor Berg-Kirkpatrick. 2021. Scalable font reconstruction with dual latent manifolds. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3060–3072, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Victor Storchan and Jocelyn Beauschene. 2019. Data augmentation via adversarial networks for optical character recognition/conference submissions. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 184–189.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Ahmad Pahlavan Tafti, Ahmadreza Baghaie, Mehdi Assefi, Hamid R. Arabnia, Zeyun Yu, and Peggy L. Peissig. 2016. OCR as a service: An experimental evaluation of google docs ocr, tesseract, ABBYY finereader, and transym. In *Advances in Visual Computing - 12th International Symposium, ISVC 2016, Las Vegas, NV, USA, December 12-14, 2016, Proceedings, Part I*, volume 10072 of *Lecture Notes in Computer Science*, pages 735–746. Springer.

Lindia Tjuatja, Shruti Rijhwani, and Graham Neubig. 2021. Explorations in transfer learning for ocr post-correction. In *Fifth Widening Natural Language Processing Workshop (WiNLP)*, volume 6.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Konstantin Todorov and Giovanni Colavizza. 2020. Transfer learning for historical corpora: An assessment on post-ocr correction and named entity recognition. In *Proceedings of the Workshop on Computational Humanities Research (CHR 2020), Amsterdam, The Netherlands, November 18-20, 2020*, volume 2723 of *CEUR Workshop Proceedings*, pages 310–339. CEUR-WS.org.

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.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85, Berlin, Germany. Association for Computational Linguistics.

Nikolai Vogler, Jonathan Allen, Matthew Miller, and Taylor Berg-Kirkpatrick. 2022. Lacuna reconstruction: Self-supervised pre-training for low-resource historical document transcription. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 206–216, Seattle, United States. Association for Computational Linguistics.

Robin M. Wright, Manuel da Silva, and José Felipe Aguiar. 2000. Baniwa history: Uapui cachoeira, aiary river (1970s - 2000). In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID: ailla:119657 Accessed March 21, 2024.

Ruoyu Xie and Antonios Anastasopoulos. 2023. Noisy parallel data alignment. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1501–1513, Dubrovnik, Croatia. 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.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2021. A comprehensive survey on transfer learning. *Proc. IEEE*, 109(1):43–76.