User-Centric Evaluation of OCR Systems for Kwak'wala

Shruti Rijhwani,[1] **Daisy Rosenblum,**[2] **Michayla King,**[3]
Antonios Anastasopoulos,[4] **Graham Neubig**[1]
[1]Language Technologies Institute, Carnegie Mellon University
[2]University of British Columbia
[3]Ḵ'waḵ'wala Language Program
[4]Department of Computer Science, George Mason University
srijhwan@cs.cmu.edu, daisy.rosenblum@ubc.ca, michayla.g3@gmail.com
antonis@gmu.edu, gneubig@cs.cmu.edu

Abstract

There has been recent interest in improving optical character recognition (OCR) for endangered languages, particularly because a large number of documents and books in these languages are not in machine-readable formats. The performance of OCR systems is typically evaluated using automatic metrics such as character and word error rates. While error rates are useful for the comparison of different models and systems, they do not measure whether and how the transcriptions produced from OCR tools are useful to downstream users. In this paper, we present a human-centric evaluation of OCR systems, focusing on the Kwak'wala language as a case study. With a user study, we show that utilizing OCR reduces the time spent in the manual transcription of culturally valuable documents – a task that is often undertaken by endangered language community members and researchers – by over 50%. Our results demonstrate the potential benefits that OCR tools can have on downstream language documentation and revitalization efforts.[1]

1 Introduction

Documentation and revitalization efforts for endangered languages frequently lead to the creation of textual documents in these languages. These include cultural materials such as folk tales and poetry; linguistic documentation like speech transcriptions and vocabulary lists; and other archival material (Himmelmann, 1998; Grenoble and Whaley, 2005). However, even though a substantial number of such documents have been created for endangered languages around the globe, the vast majority are not widely accessible because they exist only as printed books and handwritten notes.

Although some of these documents are digitally available as scanned images, the text contained in the images is not machine-readable, inhibiting several use cases that are important to communities that speak endangered languages. For example, (1) the text is not searchable for speakers and researchers of these languages; (2) it cannot be reformatted, indexed, or adapted to various needs; and (3) it cannot be used to build datasets for training NLP models. Machine-readable transcriptions of documents are typically produced by a human transcriber, who looks at the document and retypes the text present in it. Like other manual transcription tasks (e.g., speech transcription), this process is time-consuming and requires significant effort.

That said, there are computational approaches to producing machine-readable text from scanned documents, specifically through optical character recognition (OCR). Training a high-performance OCR system is challenging given the small amount of data that is typically available in endangered languages. However, there has been recent interest (Rijhwani et al., 2020, 2021; Tjuatja et al., 2021; Disbray et al., 2022) in improving OCR even in very low-resourced settings using the technique of *automatic post-correction*. Post-correction models correct errors in existing OCR transcriptions (Kolak and Resnik, 2005; Dong and Smith, 2018; Krishna et al., 2018). The post-correction methods presented in Rijhwani et al. (2021) demonstrated substantial performance gains for multiple low-resourced endangered languages – reducing character error rates (CER) by 32–58% and word error rates (WER) by 29–59% relative to off-the-shelf OCR systems.[2]

---

[1]Code, models, and datasets are available at https://shrutirij.github.io/ocr-el/.

[2]Character error rate (CER) and word error rate (WER) are based on edit distance and are standard metrics for evaluating OCR systems (Berg-Kirkpatrick et al., 2013; Schulz and Kuhn, 2017). CER is the edit distance between the predicted and

While error rates are useful to quantify the performance of various OCR technologies, they do not measure whether the produced transcriptions are useful to the primary audience for these transcriptions: community language learners, teachers, and researchers. In this paper, we look beyond error rates and take a human-centered approach to evaluating OCR and understanding whether the automatically produced transcriptions are beneficial to downstream users. More specifically, we analyze whether OCR is effective in lowering the time and effort spent in manually creating accurate transcriptions of scanned documents which, as discussed above, is a task that is frequently undertaken in language documentation and preservation programs.

As a case study, we focus on Kwak'wala, an endangered language spoken in North America, because of its long tradition of written documentation and active community engagement in accessing the knowledge contained in these texts (detailed in Section 2). We conduct a user study where we compare the time spent by human transcribers on producing an accurate transcription of typewritten Kwak'wala documents with and without the use of an OCR system.[3] We demonstrate that there is a statistically significant reduction in the time needed for manual transcription when an OCR system is used beforehand. Our results indicate that further research and development of improved OCR tools for endangered languages can add valuable efficiency to language preservation and revitalization efforts.

2 Documents in the Kwak'wala Language

To conduct our proposed human-centric evaluation of OCR, we focus on documents in the Kwak'wala language, while noting that our user study does not involve any language-specific components and can be extended to other languages.

Kwak'wala is a member of the Wakashan language family spoken on the Northwest North American Coast. Heritage learners and teachers are ac-

Figure 1: An excerpt from the Hunt-Boas publications documenting the community's method for picking viburnum berries. As seen, the Hunt-Boas orthography is complex – it uses several digraphs and diacritics that are challenging for an OCR system to recognize.

tively engaged in the revitalization of Kwak'wala. Written documentation of the language extends back over 120 years, including a collection of documents produced by anthropologist Franz Boas in collaboration with George Hunt, a native speaker of Kwak'wala (Boas, 1897; Boas and Hunt, 1902; Boas, 1911; Boas and Hunt, 1921; Boas, 1934, *inter alia*). The Hunt-Boas documents include 14 published volumes and several more unpublished manuscripts. The documents encompass a grammar of the language; word lists; stories; recipes; procedural texts; descriptions of practices, beliefs, and customs; descriptions of dialectal differences; maps and lists of placenames; and more. For Kwak'wala communities and language researchers today, these texts are rich troves containing knowledge that has special value to community-led projects focused on teaching, learning, strengthening, and reclaiming their language, cultural practices, and territorial sovereignty (Lawson, 2004).

However, to the extent the Hunt-Boas documents have been digitized, they are still 'trapped' in scanned images. The texts are not searchable and researchers potentially need to look at tens or hundreds of images to locate relevant information. Moreover, the Hunt-Boas orthography is technical and somewhat idiosyncratic and is primarily used in archival research contexts – because the texts are not machine-readable, they cannot be automatically transliterated to modern, community-preferred orthographies. Researchers who draw on these materials often resort to retyping excerpts (sometimes into a different writing system), a time-consuming process that introduces a tight bottleneck to sharing

---

the gold transcriptions of the document, divided by the total number of characters in the gold transcription. WER is similar but is calculated at the word level.

[3]Similar user studies have been carried out to determine the effectiveness of machine translation in reducing human post-editing effort (Specia and Farzindar, 2010; Gaspari et al., 2014; Koponen, 2016), but none for OCR or endangered languages, to the best of our knowledge. Kettunen et al. (2022) measure user-perceived (qualitative) utility of OCR transcripts based on information gain, as opposed to our quantitative study on reducing transcription time.

and accessing this knowledge.

Therefore, extracting the Hunt-Boas texts into a machine-readable format can serve the community in many ways. Our user study, thus, focuses on evaluating the utility of existing OCR techniques as applied to these culturally important documents. We select OCR systems based on the experiments in Rijhwani et al. (2021) which describe two models that worked particularly well on the challenging Hunt-Boas orthography (Figure 1 has an example):

- **Ocular** is an unsupervised OCR system that uses a generative model to transcribe scanned documents (Berg-Kirkpatrick et al., 2013; Garrette et al., 2015). Ocular's transcription model relies on a character n-gram language model trained on the target language. Rijhwani et al. (2021) use a small amount of Kwak'wala text data to train the language model and show that Ocular's OCR system resulted in a CER of 7.90% and a WER of 38.22% on the Hunt-Boas texts.

- **Post-correction** involves correcting the errors made by an existing OCR system to improve overall accuracy. Rijhwani et al. (2021) present a neural encoder-decoder model (Bahdanau et al., 2015) trained with semi-supervised learning to improve post-correction performance in low-resource scenarios. Relative to Ocular, the post-correction method reduces the CER by 52% and the WER by 41% on the Kwak'wala data.

In the following sections, we describe a user study focused on evaluating the two OCR pipelines (Ocular and post-correction) to understand whether the automatically produced transcriptions are beneficial to downstream users that access the information in the Hunt-Boas publications.

3 Evaluation with a User Study

Traditionally, accurate transcriptions of the Hunt-Boas documents are produced by a human transcriber (often a Kwak'wala community member, linguistic researcher, or archivist). The transcriber looks at the scanned image of each document and types out the text present in it – a time-consuming process. To evaluate the utility of the outputs from OCR models, we conduct a user study where we compare the time spent by transcribers on producing an accurate transcription in various settings

with and without the use of an OCR system. We attempt to answer two primary questions:

1. Is it faster for a human transcriber to correct the errors in an OCR output as compared to typing out the text from scratch?

2. Does adding a post-correction model affect transcription speed beyond existing off-the-shelf OCR tools such as Ocular?

We design controlled experiments to measure human transcription speed on a subset of images from the Hunt-Boas texts and evaluate how the speed is affected in various settings to understand whether there is utility in introducing OCR into the process. Additionally, we obtain subjective feedback on how having OCR outputs affected the transcription task through a survey sent to participating transcribers after tasks were completed.

3.1 Participants

We employed nine participants for the user study, all of whom had some transcription experience. Of the nine, two participants had familiarity with the Kwak'wala language as well as the Hunt-Boas texts and the orthography – one is a heritage Kwak'wala language learner and the other is an academic linguist working with Kwak'wala language materials.

We also employed seven participants that had no experience or familiarity with Kwak'wala. Three of these participants are computer science graduate students at a university and four participants were employed through Upwork,[4] a marketplace for freelance professionals. We selected them based on prior transcription experience, knowledge about data annotation for machine learning, and linguistic training as well as a high job success rate on the Upwork platform.[5] Including participants with varying degrees of prior knowledge of the Kwak'wala language also allowed us to evaluate whether this is a factor that affects transcription speed and the overall experience with the user study tasks.

3.2 Transcription Interface and Keyboard

We use Label Studio,[6] an open-source data annotation interface for setting up transcription tasks for the user study. We customized the interface for the

---

[4] https://www.upwork.com/

[5] Full IRB approval was obtained for the user study; all participants signed a consent form before working on the transcription tasks; and all data collected was anonymized.
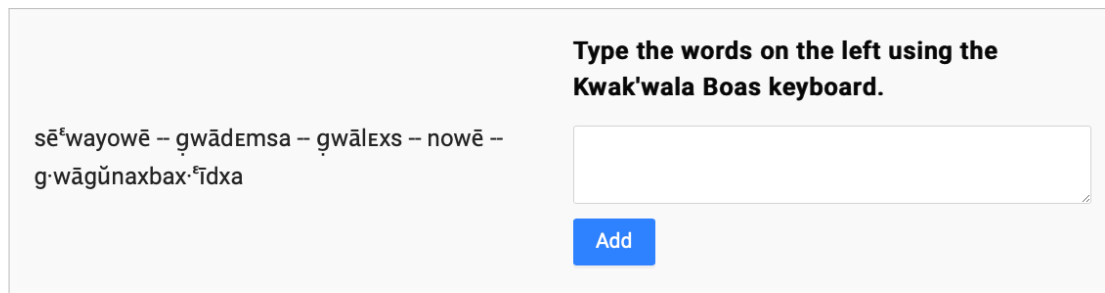
[6] https://labelstud.io

Figure 2: Practice task for transcribers to become familiar with the Boas keyboard. We included eight practice tasks in the Label Studio interface to cover all special character combinations in the Boas orthography multiple times. Users could repeat tasks as many times as they wanted to before moving on to the main transcription task.
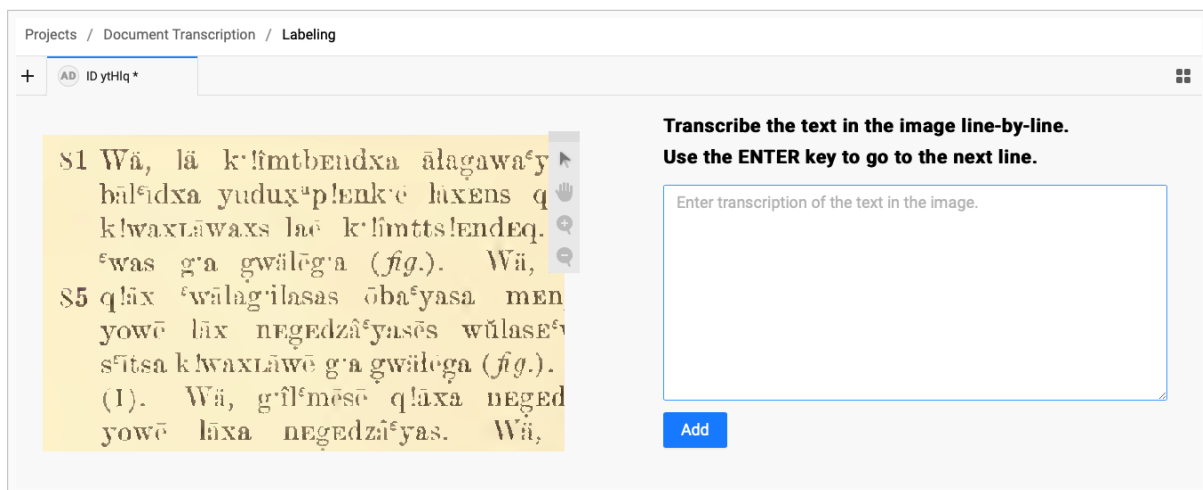

Figure 3: Transcription task interface, designed in Label Studio. The interface displays the image of a page and a text box to enter the transcription. It also has zoom and pan tools for the image, allowing users to zoom in on characters that might be hard to identify. The figure depicts a cropped image for clarity. When an OCR system is used before the manual transcription task, the text box on the right is pre-filled with the output transcription from the model and the user's task is to correct any remaining errors.

transcription task and additionally modified it to record information necessary for our analysis of transcription speed, including timestamps for when transcribers operate on each task.

Many characters and diacritics in the Hunt-Boas orthography are not present on a standard computer keyboard. To increase transcription efficiency, we used Keyman Developer[7] (an open-source toolkit) to create a keyboard for representing the characters in the orthography. The keyboard maps standard US English keyboard keystrokes to characters in the Hunt-Boas orthography. A detailed description of the keyboard layout and usage is in Section A.1. All participants were required to use this virtual keyboard to ensure consistency in terms of typing efficiency across all transcribers.

To train participants before the user study experiments, we designed a keyboard practice task, which

presents a few sentences of text in the Hunt-Boas orthography that the transcriber has to type using the keyboard. The practice texts were selected such that all the different diacritic and digraph keystroke combinations were covered multiple times. The practice tasks were also added to the Label Studio web interface – a screenshot of the interface for the practice task is shown in Figure 2. Participants were able to repeat the practice tasks as many times as needed to gain familiarity with the keyboard. Additionally, we added keystroke mapping information to the interface for all tasks (transcription and practice tasks) for users to quickly reference.

3.3 Transcription Task Settings

The primary objective for the participants was to produce an accurate transcription of the image presented to them in each task. In the Label Studio interface, as seen in Figure 3, the image is dis-

---

[7]https://keyman.com/developer/

played alongside a text box for the user to enter the transcription. To evaluate whether using OCR is useful in reducing transcription speed, we have three different setups for the tasks:

- **Baseline**: This setup does not include the use of any OCR system. The transcriber must type out the text seen in the image from scratch – they are presented with the image and an empty text box in the interface (see Figure 3). This setup represents our baseline for measuring transcription speed, as this is the method currently used by Kwak'wala researchers and community members.

- **Ocular**: In this setup, we use the off-the-shelf OCR tool Ocular on the image for each task before manual annotation. The transcriber is presented with the image and a text box containing the OCR output – that is, the text box on the right in Figure 3 will be pre-filled with the OCR output. The task here involves looking at the text present in the image (which is the target text) and editing the OCR output in the text box to correct all the errors and produce an accurate transcription.

- **Post-correction**: This is similar to the previous setup, but we use a pipeline that includes applying the OCR post-correction method from Rijhwani et al. (2021), and as described in Section 2), it improves OCR performance (CER and WER) on Kwak'wala text as compared to Ocular. The transcriber is presented with the image alongside a text box containing the post-corrected transcription. The task is to correct any remaining errors.

3.4 Experiment Design

While measuring transcription speed for a single page is relatively straightforward, determining whether there is a statistically significant difference in speed between the three different setups described above requires consideration of several factors. For example, a single transcriber cannot be assigned the same page multiple times with different setups as they would become familiar with the page's content, potentially leading to incorrect estimation of speed differences. Additionally, some participants may be faster at transcription in general and some pages in the document may be more challenging than others – these factors need to be

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

A	B	C	D
B	A	D	C
C	D	B	A
D	C	A	B

Figure 4: Two 4x4 Latin Squares. Each symbol appears only once in each row and each column. The number of symbols is the same as the number of rows and columns. Figure adapted from Dean and Voss (1999).

accounted for when measuring transcription time across the task setups.

In statistics, such factors are known as sources of variability (or nuisance factors). We design the transcription tasks to control the variability introduced by these factors using the Latin Square Design (Dean and Voss, 1999) to assign tasks to each transcriber. The Latin Square has the same number of rows and columns (square-shaped), with a specific symbol appearing exactly once in each row and exactly once in each column; Figure 4 shows two examples of a Latin Square design that has 4 rows and 4 columns. This design allows control of two sources of variability – one along the rows and one along the columns.

Since we have three task setups, we choose a 3x3 Latin Square – each setup appears only once in each row and column. The two sources of variability we control are (1) the user doing the transcription and (2) the page being transcribed. We randomly divide the nine participants into three groups of three users each (to fit the 3x3 square) and choose a fixed set of nine pages from the documents that all participants will transcribe in their tasks. For each group of three users, we form three squares (since we have nine pages). The task setups – i.e., baseline, Ocular, post-correction – are randomly assigned within the Latin Square constraints. Adding randomization for all factors (user, page, task setup) is aimed at spreading out the effect of undetectable or unsuspected characteristics. An example of task setup assignments for one group of three users for the nine pages is in Figure 5.[8]

Therefore, each user has nine transcription tasks with the task setups evenly distributed so all users are sufficiently timed on each setup. The user does not transcribe the same page more than once, but all users transcribe the same set of nine pages (with

---

[8]We follow `https://online.stat.psu.edu/stat503/lesson/4/4.4` and randomize Latin Squares separately for each group of users and each set of pages, so task setup assignments may not look identical across groups.

	page1	page2	page3		page4	page5	page6		page7	page8	page9
user1	ocu	base	post		ocu	post	base		post	ocu	base
user2	post	ocu	base		base	ocu	post		base	post	ocu
user3	base	post	ocu		post	base	ocu		ocu	base	post

Figure 5: Task setup assignments for a group of three users using the Latin Square design. We use 3x3 Latin Squares because we have three task setups: Baseline (*base*), Ocular (*ocu*), and post-correction (*post*). We need three squares for each group of users because we have nine pages for transcription. All users transcribe the same set of pages, but with the Latin Square framework, they have different task setups for each page which helps control sources of variability. All user identifiers and page identifiers are randomized before applying the Latin Square design.

varied task setups). The Latin Square Design, thus, introduces randomness across the factors to reduce variance and improve the generalization of the statistical analysis.

Dataset selection We selected nine pages from the Hunt-Boas volumes for the user study experiments, which were randomly chosen from a larger subset of 50 pages that community-based researchers deemed representative of the volumes and important to transcribe.

3.5 Evaluation Procedure

The nine transcription tasks were designed to take approximately 7 hours to complete. The participants accessed the Label Studio interface remotely through any web browser and first completed the keyboard practice tasks described above. Then, the participants began the transcription tasks and the interface recorded all timestamps for when transcriptions were edited and submitted. After the participants completed all tasks, we collected the timestamp information and computed how long it took to complete each task – with nine users transcribing nine pages each, we have 81 measurements of transcription speed to be used for quantitatively evaluating the utility of the OCR systems. We also calculated the character error rate (CER) of each transcription with respect to the transcription for the same page by our most experienced participant (a Kwak'wala heritage language learner who is very familiar with the orthography and had transcribed parts of the Hunt/Boas volumes before the user study), and discarded time measurements for transcriptions with CER $\geq 1\%$. Across all 81 transcriptions, only one had an error rate higher than this threshold, and thus, the quantitative analysis below is conducted with 80 time measurements.[9]

We also obtained qualitative feedback through a short survey that the participants filled out after completing the transcriptions. The survey asked several questions about the experience with the user study, including if the transcribers found specific tasks more difficult than others, whether they preferred typing from scratch or correcting OCR outputs (and which they thought was faster) as well as general feedback on the task and interface.

3.6 Quantitative Analysis

To quantify the effect of introducing OCR into the transcription process, we analyze the measurements of transcription speed that were collected from the user study tasks. As stated previously, we cannot use the time values directly to make a generalized conclusion because transcription time is not independent of the sources of variability. Instead, we use the statistical technique of Linear Mixed Effects (LME) modeling (Bates, 2007) to describe the relationship between the response variable (the transcription time) and the factors that contribute to variance. The term "mixed effects" refers to a combination of random effects and fixed effects. We have two random effects:

1. *transcriber identity*, which can take values from user1 to user9;

2. *page number*, which can take values from page1 to page9.

We also have two fixed effects:

1. *transcriber group*, which can either be yes or no indicating prior familiarity with the Kwak'wala language or not;

2. *task setup*, which can be one of the three setups described above – baseline, Ocular, or post-correction.

---

[9]There was no statistical difference between character error rates of from-scratch and corrected transcriptions as well as between participants with and without prior knowledge of Kwak'wala. The participants chosen for the user study had ex-

perience in transcription tasks, and all except one transcription were highly accurate (CER $< 1\%$).

Task Setup	Time Est. (*min.*)	*p*-value
Baseline	61.65	3.04e-07 *
With OCR	28.21	4.80e-08 *

Table 1: Per-page transcription time estimates in minutes from the LME model comparing the baseline, which does not use any OCR, with the task setups that use some form of OCR (either Ocular or post-correction). The time estimate for producing an accurate transcription of a page is reduced by 33.44 minutes when OCR technologies are used beforehand. The *p*-value is < 0.05, indicating statistical significance (*).

Task Setup	Time Est. (*min.*)	*p*-value
Ocular	31.67	2.55e-05 *
Post-correction	24.98	0.0121 *

Table 2: Per-page transcription time estimates in minutes from the LME model comparing task setups using an off-the-shelf OCR system (Ocular) with an OCR post-correction method. The time estimate is reduced by 6.69 minutes for a page, indicating the utility of post-correction to downstream users over using Ocular. The *p*-value is < 0.05, indicating statistical significance (*).

The LME estimation models the transcription time as a function of the above random and fixed effects. Using the estimations, our primary analysis attempts to identify whether the task setup affects transcription time in a statistically significant manner. We additionally look at whether the transcriber group (i.e., whether the participant has prior knowledge of Kwak'wala) plays a role in how fast the user completes tasks.

Does having some form of OCR help reduce transcription time? In Table 1, we present transcription time estimates from the LME model comparing two settings: (1) the baseline setup which does not use any OCR and the user types the transcription from scratch, and (2) having some form of OCR before the transcription process which the user can correct to produce error-free text (either Ocular or post-correction). As is evident from the results, having some form of OCR greatly improves transcription speed, reducing the time estimate by over 50% (from 61.65 to 28.21 minutes) and consequently, reducing the manual effort needed to produce an accurate machine-readable version of the documents.

Does post-correction help reduce transcription time beyond using an off-the-shelf OCR tool?

Group	Time Est. (*min.*)	*p*-value
Not familiar	43.60	8.12e-05 *
Familiar	25.74	0.228

Table 3: Per-page transcription time estimates in minutes from the LME model comparing transcribers that had prior familiarity with Kwak'wala with those that did not. The time estimate is reduced by 17.86 minutes for a page when the user is familiar with Kwak'wala, indicating that target knowledge language might be useful to have in image transcription tasks. The *p*-value is > 0.05 for the estimate, which indicates that it is not statistically significant, likely because we only had two users that were familiar with the language.

From the previous results, it is evident that using OCR is beneficial in reducing manual transcription time. We also evaluate whether using the post-correction model is useful or just using an off-the-shelf tool like Ocular is sufficiently useful for transcribers. The LME model estimates for this comparison are in Table 2. We see that using post-correction, as proposed in Rijhwani et al. (2021), in the transcription pipeline reduces manual correction time by 21%, indicating its utility to the downstream task of manually correcting the text.

Does prior familiarity with Kwak'wala and the Boas script affect transcription time? Beyond our primary analysis of the effect of using OCR, we also try to evaluate the extent to which the user's knowledge of the Kwak'wala language affects the speed of transcription. Table 3 demonstrates this comparison with results across all three task setups. The estimates show that this factor does play a role with the LME model estimate with a 40% reduction in transcription time for the group familiar with Kwak'wala. However, the p-value of this estimate is > 0.05, indicating that the result is not statistically significant – this is likely because only two transcribers in the user study had prior knowledge of the language and more data is needed to draw a statistically significant conclusion.

3.7 Subjective Feedback

After participants completed the transcription tasks, we asked them to fill out a short survey to describe their experience with the task. Note that, to avoid any bias, the participants were not told which OCR setup (Ocular or post-correction) was used for each task. Therefore, the survey focused on understanding whether users observed any dif-

ferences between typing from scratch or correcting transcriptions, but the questions did not distinguish between the two OCR-based setups. The full list of questions contained in the survey is in Section A.2.

We asked which of the setups led to faster completion of the tasks, and 100% of the participants perceived that correcting an OCR output was faster than typing the transcription from scratch. Some participants also provided feedback:

> *"Correcting is faster, as there is much less typing involved which requires most of the time"*
> (user7, from Upwork, not familiar with Kwak'wala)

> *"Correcting felt far more efficient!"*
> (user2, linguistic researcher, familiar with Kwak'wala)

However, even though it was slower, two out of the nine participants preferred typing out the text without the aid of an OCR output:

> *"I preferred typing the text from scratch, as searching for any editable text is difficult. You need more effort for editing."*
> (user8, from Upwork, not familiar with Kwak'wala)

However, the remaining seven transcribers provided strong feedback that correcting OCR outputs was the preferable task setup, for various reasons:

> *"I vastly preferred correcting OCR outputs. It was so much faster, and also required less investment of attention."*
> (user2, linguistic researcher, familiar with Kwak'wala)

> *"I preferred correcting text - it's much faster. I can spend more mental energy making sure the characters are correct rather than wasting time on transcribing trivially-easy letters."*
> (user5, computer science student, not familiar with Kwak'wala)

> *"I prefer correcting text because typing from scratch is somehow tricky to follow line by line."*
> (user9, from Upwork, not familiar with Kwak'wala)

Overall, transcribers participating in the user study identified a reduction in time spent when the OCR outputs were utilized and the majority preferred the task setup not only because of the speed improvement but also because the OCR outputs allowed them to zoom in and fix specific errors rather than spending time on the entire image.

Additionally, we asked participants if any tasks seemed to be easier or more difficult than others. While several described correction as easier than typing from scratch, some transcribers focused on interesting language-specific and document-specific challenges:

> *"A few alphabets were difficult to annotate from the images. For example, it was difficult to differentiate between l and ł."*
> (user6, from Upwork, not familiar with Kwak'wala)

> *"image text was with small fonts."*
> (user4, computer science student, not familiar with Kwak'wala)

> *"the hardest thing for me was identifying a particular character (ł) that is very faint in the original PDF. It is often difficult to tell if a character is ł or l. Because I have some knowledge of the language, I relied on that background knowledge at times, but this slowed down the correction process."*
> (user2, linguistic researcher, familiar with Kwak'wala)

In giving feedback about the keyboard practice tasks, all participants indicated that the practice task helped them learn the Hunt-Boas orthography and the keystroke mappings. Moreover, 100% of the participants stated that as they completed more tasks, they became faster at transcription. One participant (user7, from Upwork, not familiar with Kwak'wala) stated *"After transcribing a few pages, I became faster at typing with the keyboard and noticing the different accents and letters."* While the ordering of the tasks was not taken into account in our LME model because of the small amount of data in the current user study, we hope to understand the effect of task order on transcription time in future, larger-scale research.

4 Conclusion

In this paper, we evaluate the utility of OCR and post-correction models in a user-centric manner. We conduct a case study on Kwak'wala, an endangered language with a long history of written documentation that is currently not widely accessible to community-based speakers and researchers. With a user study, we highlight the utility of incorporating OCR to make these texts easier to manually transcribe into machine-readable formats. Our statistical analysis shows that OCR can reduce the time taken by a human transcriber in producing an accurate retyping of the texts by over 50%. While we focus on a single language in this case study, our results demonstrate the immense potential impact that OCR technologies can have on global language documentation and revitalization efforts. Our work, however, is limited in scale and scope – we do not make statistically significant conclusions on the effect of prior knowledge of the language; whether the order of pages transcribed has an impact on measured time; and the effect of general familiarity with computers and technology. In the future, we hope to conduct a larger-scale evaluation that accounts for these factors; includes transcriptions from a variety of state-of-the-art OCR systems; and expands to more languages, documents, and orthographies.

Acknowledgements

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Douglas Bates. 2007. Linear mixed model implementation in lme4. *Manuscript, University of Wisconsin*, 15.

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.

Franz Boas. 1897. *The Social Organization and the Secret Societies of the Kwakiutl Indians: Smithsonian Institution. United States National Museum. By Franz Boas. With 51 Plates*. Washington: G.P.O.

Franz Boas. 1911. *"Kwakiutl." Pp. 423–557 in Handbook of American Indian Languages, vol. 40.1, Bureau of American Ethnology Bulletin, edited by Franz Boas*. Washington: G.P.O.

Franz Boas. 1934. *Geographical Names of the Kwakiutl Indians*. New York: Columbia University Press.

Franz Boas and George Hunt. 1902. *Kwakiutl Texts*. Leiden, New York: E.J. Brill; G.E. Stechert & Co.

Franz Boas and George Hunt. 1921. *Ethnology of the Kwakiutl: Based on Data Collected by George Hunt*. Washington: G.P.O.

Angela Dean and Daniel Voss. 1999. *Design and analysis of experiments*. Springer.

Samantha Disbray, Ben Foley, Shruti Rijhwani, and Meladel Mistica. 2022. Reading it right: A case study in pintupi-luritja. In *Digital Approaches to Multilingual Text Analysis*.

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.

Dan Garrette, Hannah Alpert-Abrams, Taylor Berg-Kirkpatrick, and Dan Klein. 2015. Unsupervised code-switching for multilingual historical document transcription. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1036–1041, Denver, Colorado. Association for Computational Linguistics.

Federico Gaspari, Antonio Toral, Sudip Kumar Naskar, Declan Groves, and Andy Way. 2014. Perception vs. reality: measuring machine translation post-editing productivity. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, pages 60–72, Vancouver, Canada. Association for Machine Translation in the Americas.

Lenore A Grenoble and Lindsay J Whaley. 2005. *Saving languages: An introduction to language revitalization*. Cambridge University Press.

Nikolaus P Himmelmann. 1998. Documentary and descriptive linguistics.

Kimmo Kettunen, Heikki Keskustalo, Sanna Kumpulainen, Tuula Pääkkönen, and Juha Rautiainen. 2022. Ocr quality affects perceived usefulness of historical newspaper clippings–a user study. *arXiv preprint arXiv:2203.03557*.

Okan Kolak and Philip Resnik. 2005. OCR postprocessing for low density languages. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 867–874, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Maarit Koponen. 2016. Is machine translation postediting worth the effort? a survey of research into post-editing and effort. *The Journal of Specialised Translation*, 25:131–148.

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.

Kimberley L. Lawson. 2004. *Precious fragments: First Nations materials in archives, libraries and museums*. Ph.D. thesis, University of British Columbia.

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 postcorrection. *Transactions of the Association for Computational Linguistics*, 9:1285–1302.

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.

Lucia Specia and Atefeh Farzindar. 2010. Estimating machine translation post-editing effort with hter. In *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, pages 33–43.

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

A   Appendix

A.1   Keyboard for the Boas/Hunt Orthography

For the user study described in Section 3, we designed a keyboard for the Boas/Hunt orthography to make transcription more efficient.

The keyboard is developed using open-source software Keyman[10] and it maps characters in the Boas orthography to the user's computer keyboard. Keyman also provides an on-screen keyboard to see the mapped layout. We briefly describe the layout and usage of the keyboard below:

- Standard English keyboard alphabet and numbers remain in the same position (A-Z, a-z, 0-9) because the Boas orthography uses several Latin script characters.

- The special characters, diacritics, and digraphs of the Boas orthography have been assigned to various punctuation keys according to their frequency of use, estimated with a small sample of manually transcribed text (10 pages from Boas and Hunt (1921)).

- All accents are typed after the base character. Examples are shown below:

  ⋆ ä is typed *a then square bracket]*
  ⋆ k· is typed *k then slash /*
  ⋆ ō is typed *o then single quote '*
  ⋆ â is typed *a then shift + comma ,*
  ⋆ ă is typed *a then shift + period .*
  ⋆ g̣ is typed *g then shift + square bracket]*
  ⋆ q´ is typed *q then option (alt key) + 1*

- Other special characters are:

  ⋆ ᴇ is assigned to *semicolon ;*
  ⋆ ł is assigned to *square bracket [*
  ⋆ Ł is assigned to *shift + square bracket [*
  ⋆ ᵋ is assigned to *option (alt key) + e*
  ⋆ ᵘ is assigned to *option (alt key) + u*
  ⋆ ʟ is assigned to *option (alt key) + l*

- All changed punctuation keys can type their original value by holding down the Alt or Option key. For example, to get the original value of the square bracket [, type *Alt + [* (Windows) or *Option + [* (Mac).

---

[10] https://keyman.com/developer/

A.2 Kwak'wala Transcription: Post-Completion Survey

In Section 3, we describe a user study to evaluate the utility of OCR and post-correction models in reducing the time and effort needed for manual transcription. After participants completed transcriptions tasks, we also asked them to fill out a survey to get subjective feedback on their experience with the tasks. Discussion and analysis of the answers from the survey are in Section 3.7. We provide a complete list of the questions asked in the survey here:

1. Were there specific tasks you found easier or more difficult to annotate?

2. Did you prefer typing the text from scratch or correcting predictions from a model? Why?

3. If you are a Kwak'wala language learner, did the annotation help your language learning? How?

4. Did the practice task help you become familiar with the keyboard?

5. After annotating a few pages, do you feel like you become faster at annotation?

6. Which do you feel is faster: typing from scratch or correcting predictions?

7. Any other feedback or thoughts on the task?