Developing a Mixed-Methods Pipeline for Community-Oriented Digitization of Kwak'wala Legacy Texts

Milind Agarwal[1], Daisy Rosenblum[2], Antonios Anastasopoulos[1],
[1]George Mason University, [2]University of British Columbia,
Correspondence: magarwa@gmu.edu

Abstract

Kwak'wala is an Indigenous language spoken in British Columbia, with a rich legacy of published documentation spanning more than a century, and an active community of speakers, teachers, and learners engaged in language revitalization. Over 11 volumes of the earliest texts created during the collaboration between Franz Boas and George Hunt have been scanned but remain unreadable by machines. Complete digitization through optical character recognition has the potential to facilitate transliteration into modern orthographies and the creation of other language technologies. In this paper, we apply the latest OCR techniques to a series of Kwak'wala texts only accessible as images, and discuss the challenges and unique adaptations necessary to make such technologies work for these real-world texts. Building on previous methods, we propose using a mix of off-the-shelf OCR methods, language identification, and masking to effectively isolate Kwak'wala text, along with post-correction models, to produce a final high-quality transcription.[1]

1 Introduction

In this work, we focus on the Kwak'wala language (Wakashan, ISO 639.3 kwk), spoken on Northern Vancouver Island, nearby small islands, and the opposing mainland. Kwak'wala and several other Indigenous languages in this region have over a century of of legacy documentation created by early anthropologists, primarily in orthographies developed by Franz Boas to capture complex and typologically unusual phonetic and phonological inventories (Himmelmann, 1998; Grenoble and Whaley, 2005). Kwak'wala, for example, has 42 consonant phonemes represented with a selection of characters from the North American Phonetic Alphabet (cousin to the IPA), and over 13 possible vowel pronunciations represented with a heavy dose of diacritics and digraphs in all its scripts. During the first half of the 20th-century, scripts such as these were created and used by ethnologists, researchers, and collectors to transcribe the languages spoken in communities across North America. Between 1897 and 1965, an extensive series of texts in Indigenous languages was published by the United States Bureau of American Ethnology (BAE, now the Smithsonian). The collaboration between Franz Boas and George Hunt generated 11 volumes of published texts over 50 years, as well as extensive unpublished documentation. This script is difficult for anyone to read, amplifies phonetic complexity, and is primarily considered a legacy script, limiting access only to a few. However, many precious documents with detailed information of cultural value, were created in this script (see Figure 1), necessitating their accurate digitization and transliteration into modern Kwak'wala writing systems. Note that while Kwak'wala is classified as an endangered language with most first-language speakers over the age of 70, it has thriving language revitalization programs focused on creating new speakers among children and adults. Research progress for Kwak'wala and its three scripts (U'mista in the Northern communities, SD-72 in the Southern communities, and the legacy Boas-Hunt script) is urgent to better support revitalization and educational efforts led by community members. Currently, Kwak'wala, like many other 'low-resource' endangered and Indigenous languages, lags behind in the number of available computational tools (Agarwal and Anastasopoulos, 2024).

To remove this disadvantage and enable greater online community participation, in our project, we focus on digitization of valuable Kwak'wala texts, prioritized according to community needs, to enable building tools such as word processing, speech to text, predictive typing, etc. We create these resources by applying existing optical character recognition and language identification techniques,

[1]Relevant code and data resources are available here.

and making necessary modifications to suit them to Kwak'wala. We use grapheme-to-phoneme technology (Pine et al., 2022) to transliterate texts into the U'mista orthography, one of two community-preferred modern Kwak'wala writing systems. A draft of the 1921 Boas-Hunt text produced through a previous collaboration was distributed to 50 community and academic experts for review, and the feedback we gathered through surveys and conversations informed our production of a second draft PDF for publication and distribution. This feedback assisted us in prioritizing highly-valued elements of the texts which had originally been overlooked or erased through the process, such as the text-referenced line numbers cited by Boas in his dictionary and grammar, creating an analog concordance and networking these Kwak'wala texts into the prototypical 'Boasian trilogy'. This research, conducted in consultation with community-based language programs and guided by community priorities, will greatly increase access to culturally significant documents, thus empowering the community to draw on these resources to propagate the language and culture to future generations (Lawson, 2004).

2 Data

We focus our effort on digitizing five books that include Kwak'wala text and, often, parallel translations in English. We chose these books due to their similar fontfaces, clean layout, typed content (as opposed to handwritten), and high-quality scans.

1. **Jesup Volume 5, Part 1 (Franz Boas and George Hunt, January 1902)**: This 280 page book is part of the Jesup North Pacific Expedition publication series and contains an anthology of Kwak'wala texts in Hunt-Boas orthography. The book primarily contains dictated Kwak'wala texts (with parallel running English translation), and an appendix with grammatical information, stems, vocabulary, and traditional songs sung by Kwak'wala communities (Boas and Hunt, 1902a).

2. **Jesup Volume 5, Part 2 (Franz Boas, December 1902)**: This 144 page book is mostly formatted similarly to Volume I, but this particular volume doesn't contain interlinear text, and instead has an abundance of monolingual single-column Kwak'wala prose (Boas and Hunt, 1902b).

3. **Jesup Volume 5, Part 3 (Franz Boas, 1902)**: This is the third and final part of Volume 5, and is formatted similarly to Part II. It also contains a substantial appendix with vocabulary and stems (Boas and Hunt, 1902c).

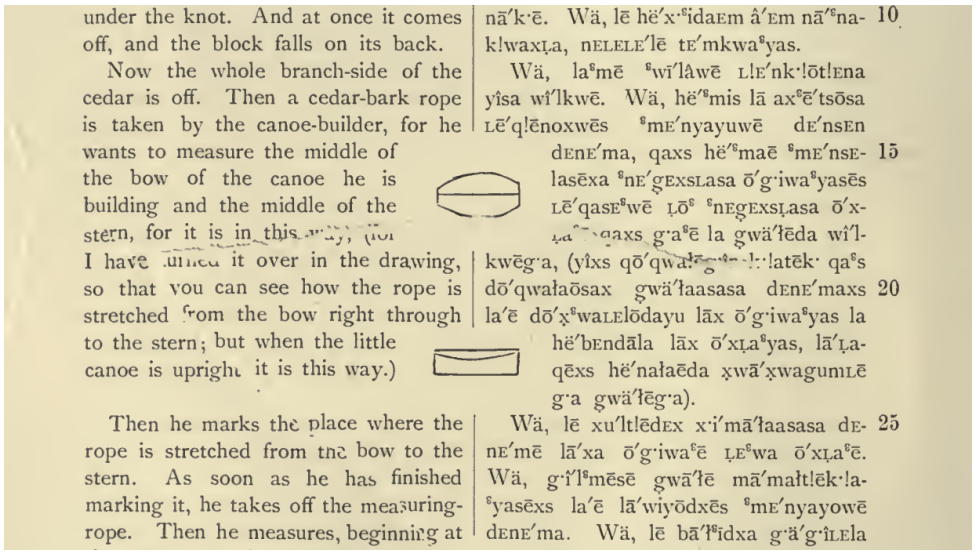
4. **Jesup Volume 10 (Franz Boas and George Hunt, 1906)**: This book contains valuable texts from the North Pacific Expedition in Kwak'wala and Haida (Masset dialect) languages. For the purposes of this project, we use only the first part of the first 282 pages of this book that contains the Kwak'wala texts (Boas and Hunt, 1906).

5. **The Kwakiutl Of Vancouver Island Volume II (Franz Boas, 1909)**: This book contains valuable texts in Kwak'wala on woodworking, weaving, hunting, fishing, clothing, measurements etc. Most of the descriptions are in English, with plenty of inline figures (that disrupt the layout extraction of the OCR), but there are also tens of pages of Kwak'wala dictated text (with parallel running English). The book alternates between a single and double column layout (Boas, 1909).

3 Related Work

Optical character recognition (OCR) is a multi-label classification problem, where a patch of pixels is shown to an OCR model, and its task it to classify it into one of n classes (usually the alphabet + punctuation). When extended to entire pages or documents containing textual material, this can allow us to digitize previously inaccessible materials. Since it is crucial for digitization of manuscripts, linguistic field notes etc., it is widely used in the humanities to render such texts accessible to researchers and to language community members (Reul et al., 2017; Rijhwani et al., 2021, 2020).

This technique has, over time, developed into a discipline, with many excellent surveys written covering the technical and applied aspects of OCR (Agarwal and Anastasopoulos, 2024; Nguyen et al., 2021; Neudecker et al., 2021; Memon et al., 2020; Hedderich et al., 2021). Today, many open-source (Tesseract and Ocular) and commercial systems (Google Vision and Microsoft OCR) exist for OCR and they can extract text from most images quite effectively, as long as they are in a language it has seen during training (Smith, 2007; Blecher et al., 2023; Berg-Kirkpatrick et al., 2013). Several research efforts before have tried to address the lack



under the knot. And at once it comes off, and the block falls on its back.

Now the whole branch-side of the cedar is off. Then a cedar-bark rope is taken by the canoe-builder, for he wants to measure the middle of the bow of the canoe he is building and the middle of the stern, for it is in this way, (for I have turned it over in the drawing, so that you can see how the rope is stretched from the bow right through to the stern; but when the little canoe is upright it is this way.)

Then he marks the place where the rope is stretched from the bow to the stern. As soon as he has finished marking it, he takes off the measuring-rope. Then he measures, beginning at

nā′k·ē. Wä, lē hë′x·ˢidaem â′em nā′ˢna- 10
k!waxɬa, nelele′lē te′mkwaˢyas.

Wä, laˢmē ˢwï′lâwē l!e′nk·!ōt!ena yîsa wï′lkwē. Wä, hë′ˢmis lā axˢē′tsōsa le′q!ēnoxwēs ˢme′nyayuwē de′nsen dene′ma, qaxs hë′ˢmaē ˢme′nse- 15
lasēxa ˢne′gexsɬasa ō′g·iwaˢyasēs le′qaseˢwē ɬōˢ ˢnegexsɬasa ō′x-ɬaˢ·qaxs g·aˢē la gwä′łēda wï′l-kwēg·a, (yîxs qō′qwa·ˢ·!·latēk· qaˢs dō′qwałaōsax gwä′łaasasa dene′maxs 20
la′ē dō′xˢwalelōdayu lāx ō′g·iwaˢyas la hë′bendāla lāx ō′xɬaˢyas, lā′ɬa-qēxs hë′nałaēda x̣wā′x̣wagumle g·a gwä′łēg·a).

Wä, lē xu′ltlēdex x·ï′mā′łaasasa de- 25
ne′mē lā′xa ō′g·iwaˢē ɬeˢwa ō′xɬaˢē. Wä, g·ï′lˢmēsē gwā′łē mā′małtlēk·laˢyasēxs la′ē lā′wiyōdxēs ˢme′nyayowē dene′ma. Wä, lē bā′lˢidxa g·ä′g·îlela

Figure 1: Example two-column text from the Kwakiutl of Vancouver Island (1909) collection. Notice the abundance of inline figures in this text that interfere with Google Vision's OCR pipeline.

of resources in certain indigenous languages using OCR to create machine-readable texts such as Central Quechua (Cordova and Nouvel, 2021) and Akuzipik (Hunt et al., 2023).

4 Methodology

4.1 First-Pass OCR

Google Vision is a well-maintained modern OCR tool that tends to work well on Latin/Roman orthographies and their extensions (Fujii et al., 2017; Rijhwani et al., 2020). Additionally, since our collections are composed of multilingual texts, it is vital to use a tool that can handle multilinguality within documents. It is a paid (per page) service at the rate of $1.25/1000 pages, but the first 1000 pages every month are free. Since our project and its digitization was conducted over several months, we did not incur any first-pass OCR charges. Open-source alternatives like Ocular or Tesseract may also be used for OCR, especially when data restrictions require local processing, instead of sending data through APIs to Google servers. However, note that they require manual training, computational expertise, preparation of training and evaluation data, and have a higher learning curve (Smith, 2007; Berg-Kirkpatrick et al., 2013) .

4.2 Language Identification

We use language identification (langID) to distinguish English from Kwak'wala as proxy for structure identification in our collections. LangID is also extremely important to enable masking of non-Kwak'wala text. To the best of our knowledge, no off-the-shelf language identification model supports Kwak'wala in the Hunt-Boas orthography. So, we use fastText as it allows easily training on custom data from scratch on CPU (Joulin et al., 2017; Agarwal et al., 2023). Our final model, trained on first-pass Kwak'wala and English texts (binary model, default fastText parameters, 1000 sentences per language), achieves a sentence-level accuracy of 99.84%. This model is applied on each page's bounding boxes, which allows us to reorganize text with improved layout.

4.3 Masking

The texts are diversely formatted, and contain additional information in illustrations, figures, line numbers and the like. Since the post-correction model (see Section 4.4) is trained to correct Kwak'wala text alone, we quickly realized that real-world digitization projects like ours require the development of a masking pipeline. Additionally, masking is preferable as post-OCR correction models are best trained for a single language, and English first-pass OCR quality is often extremely good without requiring post-OCR correction. Following the first-pass OCR, we apply a masking layer that temporarily hides/masks all English text (as labeled by the langID model), numbers, and certain punctuation like parentheses, that were impacting subsequent steps in the pipeline. This allows us to isolate, to the best ability of the language identification model and based off our overall structural cropping, the first-pass text in Kwak'wala that needs

post-correction. For each line, token-level indices of the masked tokens are stored in a separate file at this stage. This allows us to track exactly what tokens were masked so we can reintroduce them in the same spots after post-correction.

4.4 Post-Correction

Post-correction can allow us to automatically correct errors in very low-resource OCR settings, by training a correction model on a small sample of first-pass and reference text pairs (Kolak and Resnik, 2005; Dong and Smith, 2018). The post-correction model has a multi-source neural architecture, based on Rijhwani et al. (2021), which has been shown to reduce character error rates by 32–58%. We use the model from this paper directly for post-correction, with the weighted finite-state transducer setting for lexical induction turned off, as it was shown in Rijhwani et al. (2021) not to improve Kwak'wala post-correction in contrast to other low-resource languages. This is likely due to the polysynthetic structure of Kwak'wala words, leading to low lexical frequency of any one token. We train the model from scratch on the labeled Boas-Hunt dataset shared in the paper, with pre-training conducted on the unlabeled first-pass OCR outputs for the collection. We first replicated the character error rate results from the original paper to ensure reliability of the model. Then we applied our trained model to our test text. The unmasked Kwak'wala text from the previous stage is fed line-by-line to the post-correction model to obtain post-corrected Kwak'wala text.

4.5 Reconstruction

Next, we reinsert the masked tokens (English text, punctuation, line numbers in the margins, etc.) into the post-corrected sentence at the appropriate indices. This gives us the final reconstructed multilingual output, along with crucial indexical cross-referencing information such as page and line numbers. At this stage, the Kwak'wala text is also transliterated into the desired modern orthography (ex. U'mista or SD-72) using grapheme-to-phoneme conversion to allow for better readability and accessibility of the text.

4.6 Evaluation

We compare the reconstructed output to gold reference texts to evaluate the digitized texts' quality. We do this for two books at two levels:

	Jesup 5.1, 1902		Kwakiutl, 1909	
	CER	SER	CER	SER
First Pass	0.43	25	0.33	18
Corrected	0.18	2	0.15	3

Table 1: For both books, we find that using our pipeline greatly reduces not only textual errors (CER) but also greatly improves the layout and structure (SER)

- **Textual Errors**: To capture textual errors, such as misspellings, missing diacritics, tokenization etc., we use Character Error Rate (CER). This is a popular metric to understand character-level variations and error distributions in the output text, as compared to the gold-reference. For morphologically complex and polysynthetic languages like Kwak'wala, CER is a much better metric than word-level scores because a large amount of vocabulary would be unseen at test-time (Rijhwani et al., 2023).

- **Structural Errors**: We use the metric from Kanai et al. (1995) that measures insertion, deletion, and maximal move operations required across the output page to make it identical with the reference text. A weighted sum of these operations gives us the overall error, allowing us to quantify the structural quality of out outputs, and we normalize it to be between 0-100, with less being better.

Gold reference pages are created by inspecting the post-corrected output, comparing it with the source image, and manually correcting any errors. This is the most expensive part of the overall process and requires expertise in the language. So, for the moment, we evaluate on a few representative sample pages for two books. We showcase these results in Table 1, where we can observe a 50% decrease in character error rate and 87.5% reduction in structural error with our pipeline of language identification, masking, and automatic post-correction.

5 Conclusion

We apply the latest OCR techniques to a series of previously undigitized Kwak'wala texts, and demonstrate the challenges and unique adaptations necessary to make OCR work for real-world texts and collections. We propose using a mix of off-the-shelf OCR methods, language identification and

masking to effectively isolate Kwak'wala text, and post-correction models to produce a high-quality transcription. We plan to disseminate the digitized documents directly to the community members. Additionally, with consent of the community partners and data annotators, we plan to share the digitized and transliterated text (in three orthographies) with the data hosting institutions, such as the American Philosophy Society and Columbia University Rare Books and Special Collections, where a large collection of Boas-Hunt manuscripts have recently been digitized (Schlottmann, 2023). We hope to explore ways that this work in improving OCR for Kwak'wala and developing reliable digitization workflows for legacy texts can be transferable to other legacy orthographies, directly benefitting other language communities.

Limitations

Since our contribution type is best suited to a short paper, at the moment, we did not include more extensive benchmarking for language identification. As we continue to work with our language community collaborators, we will continue to add more gold reference texts for comparison and better evaluation of the transcriptions.

Ethics Statement

Though they derive from material in the public domain, the first-pass, gold reference texts, and corrected transcriptions of the selected Kwak'wala texts will only be released publicly with the consent of the language community members. 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, under the guidance of the language community members and their immediate and long-term needs for effective Kwak'wala revitalization.

Acknowledgments

This work was generously supported by the National Endowment for the Humanities under award PR-276810-21, George Mason University's Doctoral Research Scholars Award 2024-25 and the Stanford Initiative on Language Inclusion and Conservation in Old and New Media (SILICON) Practitioners 2024-25 Award. The authors are also grateful to the anonymous reviewers for their valuable suggestions, feedback, and comments.

References

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

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

Taylor Berg-Kirkpatrick, Greg Durrett, and Dan Klein. 2013. Unsupervised transcription of historical documents. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 207–217. The Association for Computer Linguistics.

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

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *Preprint*, arXiv:2308.13418.

Franz Boas. 1909. *The Kwakiutl of Vancouver Island*. Leiden, New York: E.J. Brill; G.E. Stechert & Co.

Franz Boas and George Hunt. 1902a. *Volume 5, Part 1. Kwakiutl Texts - Memoirs of The American Museum of Natural History*. Leiden, New York: E.J. Brill; G.E. Stechert & Co.

Franz Boas and George Hunt. 1902b. *Volume 5, Part 2. Kwakiutl Texts - Memoirs of The American Museum of Natural History*. Leiden, New York: E.J. Brill; G.E. Stechert & Co.

Franz Boas and George Hunt. 1902c. *Volume 5, Part 3. Kwakiutl Texts - Memoirs of The American Museum of Natural History*. Leiden, New York: E.J. Brill; G.E. Stechert & Co.

Franz Boas and George Hunt. 1906. *Jesup North Pacific Expedition - Kwakiutl Texts, Second Series, Volume 10*. Leiden, New York: E.J. Brill; G.E. Stechert & Co.

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.

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.

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.

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

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.

Nikolaus P Himmelmann. 1998. Documentary and descriptive linguistics.

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.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

J. Kanai, S.V. Rice, T.A. Nartker, and G. Nagy. 1995. Automated evaluation of ocr zoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(1):86–90.

Okan Kolak and Philip Resnik. 2005. OCR post-processing 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.

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

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.

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).

Aidan Pine, Patrick William Littell, Eric Joanis, David Huggins-Daines, Christopher Cox, Fineen Davis, Eddie Antonio Santos, Shankhalika Srikanth, Delasie Torkornoo, and Sabrina Yu. 2022. G_i2P_i rule-based, index-preserving grapheme-to-phoneme transformations. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 52–60, Dublin, Ireland. Association for Computational Linguistics.

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.

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.

Kevin Schlottmann. 2023. Description and digitization of the george hunt kwak'wala ethnographic manuscripts. Accessed: 2025-01-10.

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.