Towards a Community-accessible Cahuilla corpus: Developing HTR for J.P. Harrington's handwritten fieldnotes on Mountain Cahuilla

Ray Huaute *
University of California, Los Angeles
ray.huaute@ucla.edu

Jacqueline Brixey *
University of Wisconsin-Madison
brixey@wisc.edu

Abstract

This paper describes ongoing work to develop a corpus of Cahuilla language from the John Peabody Harrington collection, which contains linguistic and ethnographic fieldnotes documenting Indigenous languages of California and other regions across the Americas. Handwritten notes present numerous processing challenges, including scratch-outs, multilingual entries in Spanish and other Indigenous languages, unique abbreviations, and varying script orientations. We compare the efficacy of deep learning text recognition models to convert images of the notes into a machine-readable format, with a focus on respecting tribal data sovereignty in our methods. We find that Pylaia is the most accurate model for our data. Finally, we present the preliminary findings and indicate future directions for developing a Cahuilla corpus.

1 Introduction

The John Peabody Harrington (J.P., for short) collection represents a monumental record of over 100 Indigenous languages (Harrington, 1907). While the collection has been digitized by the Smithsonian Institution[1], most of the handwritten notes have not been converted into a machine-readable format. Of interest to this paper is the estimated 6,000 pages of notes on the Cahuilla language, an endangered language of Southern California (Simons and Fennig, 2018).

The goal of this work is to create and share a searchable corpus of the Harrington notes on Cahuilla with the Cahuilla community. A searchable corpus will support linguistic and downstream Natural Language Processing (NLP) research, as well as support community goals in the urgent work of language revitalization. A secondary objective is to create a broadly applicable and replicable model for transcribing the remainder of the J.P. Harrington fieldnotes. This initiative has the potential to substantially assist the more than 100 Indigenous communities with whom Harrington collaborated in gaining access to a wealth of knowledge about their languages. This paper presents preliminary work to identify the most accurate and efficient Handwritten Text Recognition (HTR) approach for converting the Cahuilla notes in the collection into a machine-readable format. We also describe a data sovereignty framework for working with the Cahuilla community that can be applied more generally to similar projects with other Indigenous language communities.

2 Overview of Cahuilla Language and Tribal Communities

Cahuilla (chl - ISO 639-3, henceforth referred to by the endonym: ˈiviʌuʔat) is a Native American language of Southern California with few first-language speakers remaining. Today, ˈiviʌuʔat is being spoken and reclaimed across many Cahuilla reservations and communities[2]. ˈiviʌuʔat, along with Luiseño, Acjachemem (Juaneño), and Cupeño, comprise the Cupan sub-group of languages that are part of the larger Takic branch of Uto-Aztecan languages (Hill and Hill, 2019). There are three dialects of ˈiviʌuʔat: Mountain, Desert, and Pass or Wanakik. ˈiviʌuʔat is an agglutinative, head-final language with SOV word order (Seiler, 1977).

The orthography in the J.P. Harrington collection is the Americanist Phonetic Notation (APN)[3].

*Equal contribution

[1] https://sova.siedu/record/naa.1976-95/contents

[2] There are currently nine federally recognized tribes that identify themselves as Cahuilla: Agua Caliente Band of Cahuilla Indians, Augustine Band of Cahuilla Indians, Cabazon Band of Mission Indians, Cahuilla Band of Indians, Los Coyotes Band of Cahuilla and Cupeño Indians, Morongo Band of Mission Indians, Ramona Band of Cahuilla Indians, Santa Rosa Band of Cahuilla Indians, and Torres Martinez Desert Cahuilla Indians.

[3] Also known as the North American Phonetic Alphabet

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 107–114

For this project, these characters will be converted into a code suitable for integration into text files within our database, and later transliterated into the community-preferred orthography (see (Huaute, 2023) for orthographies).

3 Overview of J.P. Harrington Collection

Recognized as one of the most prolific documentarians of California Indian languages, J.P. Harrington compiled over one million pages of cultural and linguistic fieldnotes covering more than 135 languages in California and the Far West from 1915 to 1954 (Mills and Ann, 19876). However, the communities whose ancestors collaborated with him have yet to achieve comprehensive access to, or benefit from, this valuable knowledge, a gap this project seeks to address.

Numerous scholars and workshops over the years aimed to make the data easily searchable[4] (Golla, 1991). A large-scale effort to manually transcribe, annotate, and format the collection into a database was undertaken at the University of California, Davis (Macri, 2010), resulting in the transcription of 67 reels of data, or roughly 235,000 sentences, representing 16 languages[5]. The main work concluded in 2013, and while the resulting database is not publicly available, the transcribed files are available upon request[6]. To the best of our knowledge, no prior work has attempted to convert any portion of the JP Harrington images into a machine-readable format using HTR techniques.

For this project, we aim to convert approximately 6,000 pages on the Mountain Cahuilla dialect. The bulk of the entries from this series were provided by Adan (Adam) Castillo, a Mountain Cahuilla speaker from the Soboba Indian Reservation.

3.1 Challenges of the collection

A significant challenge is that the notes are handwritten, with characters of varying shapes and sizes. An example page is shown in Figure 2. Harrington also wrote sporadically in cursive and in varying orientations, both of which automatic recognition approaches often struggle with (Khan et al., 2023; Pavlenko and Blackledge, 2004).

An additional challenge is that the collection is multilingual, with English, Spanish (Anderton, 1991), and multiple Indigenous languages present. Finally, Harrington's prolific use of abbreviations (Woodward and Macri, 2005) and inconsistencies in orthographic representation raise interpretability issues.

4 Review of Relevant Literature

4.1 Data sovereignty

In response to Indigenous communities' concerns about the development of large data centers and the expansion of AI technologies (Cox, 2025), we adopt data sovereignty as a foundational principle guiding this project. Central to the Indigenous Data Sovereignty movement is the idea of self-determination for Indigenous peoples and their sovereign right to own and control their data, which, in this work, includes linguistic data (Holton et al., 2022). The development of international standards for data sovereignty governance, such as the CARE principles (Collective benefit, Authority to control, Responsibility, and Ethics) (Carroll et al., 2023), are also informative for our data sovereignty policy. Finally, a recent publication (Holton et al., 2022) notes potential licensing issues associated with "Terms of use" clauses when using third-party apps, such as Google Drive and iCloud. This was a consideration as we reviewed large platforms for our project, such as Transkribus (Kahle et al., 2017), eScriptorium (Kiessling et al., 2019), and OCR4ALL (Reul et al., 2019). However, many platforms we reviewed were not explicit and transparent in how shared data is stored and protected, leading us to not pursue these platforms as a viable approach.

4.2 Text recognition

Deep learning models, such as convolutional neural networks (CNNs) (Alam et al., 2025), recurrent neural networks (RNNs) (Keshri et al., 2018), and, more recently, the two combined as convolutional recurrent neural networks (CRNNs) (Shi et al., 2016), are effective at HTR tasks (Idris and Taha, 2022; AlKendi et al., 2024; Balci et al., 2017; Dash et al., 2024). Important aspects of the Harrington collection are multilingualism, abbreviations, and cursive, which deep learning models have demonstrated the ability to recognize (Alam et al., 2025; Romein et al., 2025; Al-Saffar et al., 2021).

Some popular online platforms that utilize HTR have been developed specifically for converting

(NAPA)

[4] http://www.rock-art.com/jph/nl04.htm

[5] Obtained via personal communication

[6] https://nas.ucdavis.edu/jp-harrington-database-project

archival documents to machine-readable text (e.g., Transkribus, OCR4all) but may not be universal, as we find that they have mostly been trained on large Indo-European languages. Automatic text recognition has meaningfully aided in the documentation of other Indigenous languages with digitized text resources (Carrera et al., 2024; Agarwal and Anastasopoulos, 2025, 2024).

5 Methods

Due to the collection's volume and the time and effort required for manual transcription, we propose developing an HTR approach to efficiently convert the scanned images into a machine-readable format. In this current work, we implemented three deep learning models that align with our data sovereignty policy. We compare the models to determine the most accurate approach, potentially reducing the need for additional post-processing steps.

5.1 Data Sovereignty Policy

To ensure maximal collective benefit of our project, our primary goal is to generate a machine-readable and searchable database of the J.P. Harrington ˈiviʌuʔat fieldnotes that can be provided to Cahuilla community members in a format they can easily access and understand.

Given the potential issues indicated in Section 4.1 and (Holton et al., 2022), we utilized only locally hosted models for this stage of the project. At later stages, we will work closely with the Cahuilla communities to develop a plan for data storage, curation, and access protocols. This approach aligns with the second CARE principle, authority to control (Carroll et al., 2023).

5.2 Preprocessing Images

We used a sample of 66 pages for our experiments. All downloaded PDF files[7] were converted into JPEG files of each page. We implemented a Python interface to create sliced images by selecting just a word, line, or single letter from a given JPEG. This approach allowed us to resolve issues in orientation changes and to omit scratch-outs. However, this resulted in slices of different dimensions.

The Pylaia and Jax libraries automatically correct image size differences. For our baseline CRNN model, we added a padding strategy to standardize the sizes. Figure 1 illustrates padding for 3 different slices.

[7] https://sova.si.edu/record/naa.1976-95/search?q=cahuilla&t=W&o=doc_position

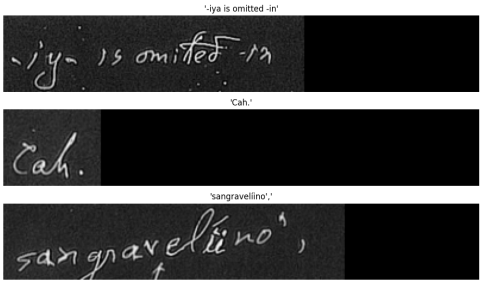

Figure 1: Example of padding to standardize image "slices" sizes. Slices can be lines, words, or singular characters.

5.3 Training and testing data

Next, we created ground-truth labels for each slice. Decomposed characters, such as s̥, were substituted for a composed character for ease and accuracy.

We then reviewed the dataset for word and character frequencies. There are 390 unique and 845 total words in the training set. "Cah", shorthand for "Cahuilla", appears the most (46 times). The majority of words are in English, 46 in Cahuilla (5% of total), and 21 in Spanish (roughly 2%).

The prediction character set contains at least 86 glyphs, including upper and lowercase English, Spanish, Cahuilla, and some IPA characters. The most commonly occurring characters are: space (369), e (310), a (284), o (262), and t (230). Thirteen characters appear only once. Not present in the data are: U, V, X, Z, z, 6, and Cahuilla ʌ.

The 66-page sample resulted in 494 slices. We allocated roughly 5% of the slices as a limited evaluation set. Given the small set of training examples at this preliminary stage of the research, we proposed comparing the models' performance on an unbalanced (raw) data set to that on a balanced data set. The merits of using the real-world representation (i.e., unbalanced) of characters are that the models will focus on learning the most highly occurring characters. In contrast, oversampling, or creating a balanced data set, is thought to encourage models to learn all characters (Kaur et al., 2019). As it is crucial to recognize Cahuilla, which occurs less frequently, we proposed testing the models on both balanced and unbalanced datasets. We created a balanced dataset using a Python script, ensuring that each character appears at least 60 times.

5.4 Models

All experiments were completed on a MacBook Pro M2. Training and testing sets were manually configured so that model results would be compara-

ble. At this early research stage, we did not explore additional fine-tuning of the models.

1. Baseline CRNN We implemented a CRNN model in Python using the TensorFlow (Abadi et al., 2015) and Keras (Chollet et al., 2015) libraries. The model is a three-block CNN with ReLU activations. The recurrent component is a single bidirectional LSTM.

2. Pylaia: Pylaia (Tarride et al., 2024) is a popular Python text recognition library that uses the deep learning framework PyTorch (Paszke, 2019). Pylaia powers historical text documentation platforms like Transkribus (Park, 2025).

Our Pylaia model comprises a four-block CNN, each block incorporating LeakyReLU activations and batch normalization. The recurrent component consists of three bidirectional LSTM layers.

3. Jax: Jax (Bradbury et al., 2018) is a Python library recently created by Google that is optimized for memory usage for large machine learning model development (Sapunov, 2024). Limited previous research using Jax has shown promising performance on computer vision tasks, such as medical imaging classification (Bećirović et al., 2025) and recognizing handwritten structured medical notes (Kale et al., 2025). Given that it can be run locally, we proposed to include it in our comparison experiments.

Our model consists of a 3 CNN layers followed by a bidirectional GRU and a linear classification layer trained with CTC loss. The model is optimized with Adam using a warmup schedule and gradient clipping.

6 Results

We found that the baseline CRNN approach was the most time-efficient of the models; the Pylaia and Jax models took 6-10 hours to train and test on the balanced dataset. Pylaia also frequently ran into issues with exceeding memory allocation. It may be a consideration for language communities with fewer technological resources to know the time and memory requirements for each model. We also reviewed the models' performance for overall accuracy and by language.

1. Overall performance: The results are given in Table 1. Pylaia was the best model in terms of both overall word error rate (WER) and character error rate (CER), and then the baseline CRNN model and Jax on the balanced data. Performance across the models mostly declined on the unbalanced data. This indicates that the balanced data had a positive effect on the models.

2. Performance by language: Next, we reviewed each model's performance by language (also in Table 1). Again, there was an improvement by all the models when using the balanced data. No model had zero errors on the unbalanced data. It is assumed that a language model is produced as a result of the RNN layer(s) in a CRNN (Dash et al., 2024). Both the Pylaia and baseline CRNN use LSTMs as the RNN layers; it is notable how much better the Pylaia model recognizes letters overall and performs on the two languages that are underrepresented in the data than the baseline CRNN.

7 Conclusions and Future Work

This initial work towards creating a corpus of the Cahuilla language compared the performance of deep learning models in converting handwritten text into a machine-readable format. We found that the Pylaia model achieved lower CER than a baseline CRNN and Jax models when trained and tested on the same data.

In future work, we will prepare additional training data covering the missing letters indicated in Section 5.3. Our results indicate that we should use the balanced data for the rest of the HTR work. We will explore computing power resources (such as servers or a more powerful computer) that align with our data sovereignty policy in the next step using Pylaia, as we anticipate that the computer used for the experiments in this paper will be insufficient for a larger training set. We will also consider post-processing correction approaches.

Our data sovereignty policy will continue to be refined and determined in future steps. *Collaborative consultation* with Cahuilla communities for guidance, reflection, and sharing throughout the project will ensure that researchers behave ethically and follow cultural protocols throughout the project and post-project, ensuring responsibility to the community (Leonard and Haynes, 2010). Our data management plan will include provisions for using cloud and server storage solutions that respect data sovereignty. Finally, we acknowledge the importance of maintaining flexibility and will revise the project's data sovereignty policy in close consultation with the Cahuilla community.

Limitations

Our data sovereignty policy helped to determine the models selected. LLMs were necessarily excluded because they did not align with the policy. We note that the computer used for the experiments was a limitation, as it was not the most powerful or the most recent MacBook. The performance of the models, especially with regard to running times, would differ on a different (more powerful) computer. However, we also recognize that technological limitations may be a factor for other Indigenous communities considering this type of work, who may have limited financial capacity and access to powerful computing resources.

As there are few HTR publications using Jax to compare our results to, we are limited in our ability to hypothesize about the model's performance; it may be attributable to some aspect of our limited sample data.

Acknowledgments

The authors wish to acknowledge the significant contributions of renowned linguist and ethnologist JP Harrington and his diligent and knowledgeable Cahuilla speakers and collaborators whose efforts provided the data utilized in this paper. We also thank the anonymous reviewers for their helpful feedback.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, and 21 others. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

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.

Milind Agarwal and Antonios Anastasopoulos. 2025. Ailla-ocr: A first textual and structural post-ocr dataset for 8 Indigenous Languages of Latin America. In *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 120–127.

Ahmed Al-Saffar, Suryanti Awang, Wafaa Al-Saiagh, Ahmed Salih Al-Khaleefa, and Saad Adnan Abed. 2021. A sequential handwriting recognition model based on a dynamically configurable crnn. *Sensors*, 21(21):7306.

Mahanur Alam, Md Johirul Islam Tutul, Md Anwar Hussen Wadud, Md Jakir Hossen, and MF Mridha. 2025. Bilingual Bangla ocr for rural empowerment: Detecting handwritten queries and agricultural assistance. *IEEE Open Journal of the Computer Society*.

Wissam AlKendi, Franck Gechter, Laurent Heyberger, and Christophe Guyeux. 2024. Advancements and challenges in handwritten text recognition: A comprehensive survey. *Journal of Imaging*, 10(1):18.

Alice J Anderton. 1991. Kitanemuk: Reconstruction of a dead phonology using John P. Harrington's Transcriptions. *Anthropological Linguistics*, pages 437–447.

Batuhan Balci, Dan Saadati, and Dan Shiferaw. 2017. Handwritten text recognition using deep learning. *CS231n: convolutional neural networks for visual recognition, Stanford University, Course Project Report, Spring*, pages 752–759.

Merjem Bećirović, Amina Kurtović, Nordin Smajlović, Medina Kapo, and Amila Akagić. 2025. Performance comparison of medical image classification systems using tensorflow keras, pytorch, and jax. *arXiv preprint arXiv:2507.14587*.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Yash Katariya, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. 2018. JAX: composable transformations of Python+NumPy programs.

Shadya Sanchez Carrera, Roberto Zariquiey, and Arturo Oncevay. 2024. Unlocking knowledge with ocr-driven document digitization for Peruvian indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 103–111.

Stephanie Russo Carroll, Ibrahim Garba, Oscar L Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, and 1 others. 2023. The care principles for indigenous data governance. *Open Scholarship Press Curated Volumes: Policy*.

François Chollet and 1 others. 2015. Keras. `https://keras.io`.

Evelyn Cox. 2025. AI in a tribal context: A brief review of the literature.

Saswata Kumar Dash, Sompalli Pranay, Pervela Hemanth, and Aravindkumar Sekar. 2024. Multi-lingual

handwritten recognition using convolutional recurrent neural networks. In *2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, pages 1–7. IEEE.

Victor Golla. 1991. Introduction: John P. Harrington and his legacy. *Anthropological Linguistics*, pages 337–349.

John Peabody Harrington. 1907. Volume three: A guide to the field notes: Native american history, language, and culture of Southern California/basin. *The Papers of John Peabody Harrington in the Smithsonian Institution*, 1957.

Jane H Hill and Kenneth C Hill. 2019. Comparative takic grammar.

Gary Holton, Wesley Y Leonard, and Peter L Pulsifer. 2022. Indigenous peoples, ethics, and linguistic data. *The open handbook of linguistic data management*, pages 49–60.

Incamu Ray Huaute. 2023. *Topics in the phonology and morphology of Torres Martinez Desert Cahuilla*. Ph.d. dissertation, University of California, San Diego.

Ahmed A. Idris and Dujan B. Taha. 2022. Handwritten text recognition using crnn. In *2022 8th International Conference on Contemporary Information Technology and Mathematics (ICCITM)*, pages 329–334.

Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. Transkribus-a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th iapr international conference on document analysis and recognition (icdar)*, volume 4, pages 19–24. IEEE.

Apoorwa Kale, Yash Khandelwal, Vibhor Pandhare, Atreyee Ghosh, Nidhi Pathak, Bhure Singh Saitya, and Bhupesh Kumar Lad. 2025. A scalable, low-cost framework for multilingual intelligent document processing for continuity of care. In *IET Conference Proceedings CP942*, volume 2025, pages 161–166. IET.

Harsurinder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. 2019. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM computing surveys (CSUR)*, 52(4):1–36.

Pooja Keshri, Prabhat Kumar, and Rajib Ghosh. 2018. Rnn based online handwritten word recognition in Devanagari script. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 517–522. IEEE.

Sulaiman Khan, Shah Nazir, and Habib Ullah Khan. 2023. Analysis of cursive text recognition systems: A systematic literature review. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(7):1–30.

Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. 2019. eScriptorium: an open source platform for historical document analysis. In *2019 international conference on document analysis and recognition workshops (icdarw)*, volume 2, pages 19–19. IEEE.

Wesley Y Leonard and Erin Haynes. 2010. Making "collaboration" collaborative: An examination of perspectives that frame linguistic field research.

Martha J Macri. 2010. Working with language communities in unarchiving: Making the JP harrington notes accessible. In *Language Documentation: Practice and values*, pages 213–220. John Benjamins Publishing Company.

Elaine Mills and J Brickfield Ann. 19876. The papers of john peabody harrington in the smithsonian institution, 1907-57. *Millwood NY: Kraus International*.

Fiona Park. 2025. What are super models and how do they work?

A. et al. Paszke. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Aneta Pavlenko and Adrian Blackledge. 2004. Introduction: New theoretical approaches to the study of negotiation of identities in multilingual contexts. In Aneta Pavlenko and Adrian Blackledge, editors, *Negotiation of Identities in Multilingual Contexts*, pages 1–33. Multilingual Matters LTD.

Christian Reul, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner, and Frank Puppe. 2019. Ocr4all—an open-source tool providing a (semi-) automatic ocr workflow for historical printings. *Applied Sciences*, 9(22):4853.

Christel A Romein, Achim Rabus, Gundram Leifert, and Phillip Benjamin Ströbel. 2025. Assessing advanced handwritten text recognition engines for digitizing historical documents: Romein et al. *International journal of digital humanities*, 7(1):115–134.

Grigory Sapunov. 2024. *Deep learning with JAX*. Simon and Schuster.

Hansjakob Seiler. 1977. *Cahuilla grammar*. Banning: Malki Museum Press.

Baoguang Shi, Xiang Bai, and Cong Yao. 2016. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304.

Gary F. Simons and Charles D. Fennig, editors. 2018. *Ethnologue: Languages of the World*, twenty-first edition. SIL International, Dallas, Texas.

Solène Tarride, Yoann Schneider, Marie Generali, Melodie Boillet, Bastien Abadie, and Christopher Kermorvant. 2024. Improving automatic text recognition with language models in the pylaia open-source library. In *Submitted at ICDAR*.

Lisa L Woodward and Martha J Macri. 2005. JP harrington database project: an archival resource for anthropologists, archaeologists, and Native communities. *Journal of California and Great Basin Anthropology*, 25(2):235–240.

A Appendix

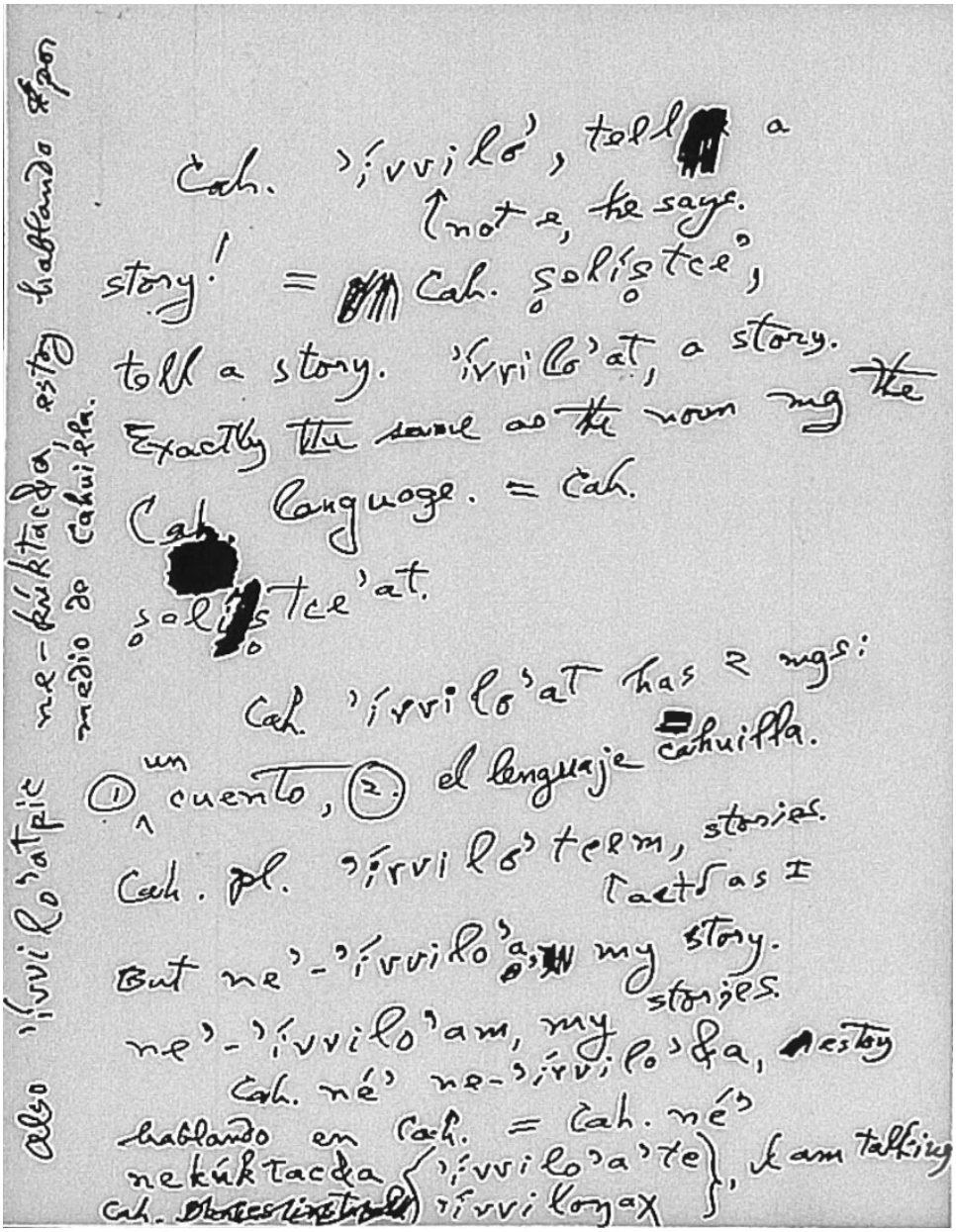

Figure 2: A page from the J.P. Harrington collection, demonstrating some of the variations in orientation, scratch-out, multilingualism, and cursive challenges unique to the data.

	CER				WER			
	Overall	English	Spanish	Cahuilla	Overall	English	Spanish	Cahuilla
Balanced								
CRNN	75.64	72.29	69.23	90	128.57	95	300	200
Pylaia	9.62	25.61	0	0	21.43	30	0	0
Jax	83.97	83.33	84.62	83.13	107.14	242.86	100	100
Unbalanced								
CRNN	80.13	78.31	84.62	86.67	103.57	95	200	114.29
Pylaia	61.69	51.22	69.23	81.67	75	48.78	100	100
Jax	96.15	95.18	92.31	98.33	100	100	100	100

Table 1: Results for each model of the character error rate (percent) and word error rate (percent) in each language.