Unlocking Knowledge with OCR-Driven Document Digitization for Peruvian Indigenous Languages

Shadya Sánchez **Roberto Zariquiey** **Arturo Oncevay**
Chana Research Group, Pontificia Universidad Católica del Perú, Perú
{shadya.sanchez,rzariquiey,arturo.oncevay}@pucp.edu.pe

Abstract

The current focus on resource-rich languages poses a challenge to linguistic diversity, affecting minority languages with limited digital presence and relatively old published and unpublished resources. In addressing this issue, this study targets the digitalization of old scanned textbooks written in four Peruvian indigenous languages (Asháninka, Shipibo-Konibo, Yanesha, and Yine) using Optical Character Recognition (OCR) technology. This is complemented with text correction methods to minimize extraction errors. Contributions include the creation of an annotated dataset with 454 scanned page images, for a rigorous evaluation, and the development of a module to correct OCR-generated transcription alignments.

1 Introduction

Natural Language Processing (NLP) has prompted the development of diverse language technologies, including machine translation, spell checkers, and information extraction tools. Given this impact, there is an urgent need to democratize these technologies, making them available for speakers of the more than 7,000 languages spoken worldwide.

Currently, such technologies are restricted to languages with ample linguistic resources that are easily exploitable (Ataa Allah et al., 2023). This presents a challenge for minority languages due to their limited digital presence and the prevalence of their resources in less accessible formats, hindering their incorporation into the development of these technologies (Bustamante et al., 2020). Consequently, speakers of minority languages are forced to adopt languages with greater technological access, leading to a loss of cultural, historical, and linguistic knowledge.

To address this situation, multiple efforts are underway to diversify these technologies to minority languages and their speakers, who face the challenge of overcoming data availability limitations. In some cases, synthetic data has been generated (Oncevay et al., 2022), translations to languages with more resources have been utilized (Ko et al., 2021; Rijhwani et al., 2020), and technologies have been adapted for dataset extraction and processing (Bustamante et al., 2020).

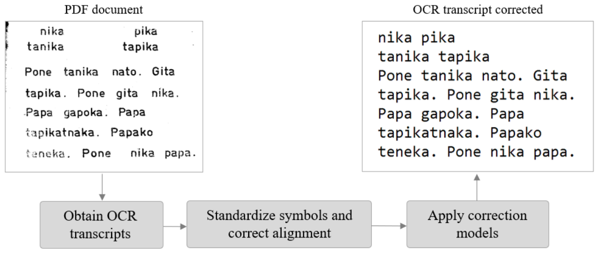


Figure 1: OCR process

However, identifying digital sources of knowledge for endangered languages is a very challenging task, as they are not usually available on the web (Bustamante et al., 2020), which is the case for several indigenous languages of Peru. In this context, Optical Character Recognition (OCR) models have been useful to extract information and new resources for endangered language texts (Rijhwani et al., 2020, 2021).

For this reason, we extend the application of OCR for digitizing old documents, with typewritten texts, in four Peruvian languages (Asháninka, Shipibo-Konibo, Yanesha, and Yine), using Optical Character Recognition (OCR), and followed by correction methods to minimize extraction errors (see Figure 1).

2 Language context

According to official statistics, 48 languages are spoken in Peru. 44 out of these 48 languages are Amazonian languages (de Educación del Peru, 2013). The four languages this paper is focused on (Shipibo-Konibo, Asháninka, Yanesha, and Yine) are Amazonian languages.

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 103–111
June 21, 2024 ©2024 Association for Computational Linguistics

Asháninka, belonging to the Nijagantsi branch of the Arawak language family, is primarily spoken in the central Peruvian Amazonia, along the Low Perené, Tambo, Ene, Urubamba, and Apurímac rivers (Pedrós, 2018). Although the Asháninka population is estimated to be around 70,000 speakers (Pedrós, 2018), it remains unclear if this count includes speakers of Ashéninka, a closely related language.

Yine and Yanesha are also languages of the Arawak family. Yine is spoken by approximately 3,000 people living near the Ucayali and Madre de Dios rivers. Yanesha, in turn, is spoken by 1,142 people in the Peruvian department of Pasco. Yanesha people generally express concern for their language, since very few children speak it and speakers are mostly over 30 years old. Both Yine and Yanesha are classified as "definitely endangered" according to the UNESCO Atlas of the World's Languages in Danger (Moseley, 2010).

With an estimated 40,000 speakers, Shipibo-Konibo is by large the most vital language in the Pano language family. It is predominantly spoken in the Peruvian regions of Ucayali and Loreto, along the Ucayali river and its tributaries (Valenzuela, 2003). It is important to mention that there is a relatively large Shipibo-Konibo community in Lima.

3 Related work

The correction of OCR transcripts has seen the application of various methodologies, ranging from manual and resource-intensive approaches to more recent and prevalent machine learning models, particularly those based on neural networks (see Nguyen et al. (2021) for further details). The effectiveness of applications such as language models, translation models, and spell checkers in rectifying OCR errors is well-established. For instance, Afli et al. (2016) employed a statistical machine translation (SMT) model, while Schulz and Kuhn (2017) combined such models with spell checkers.

Furthermore, sequence-to-sequence neural networks have emerged as successful models in correcting OCR transcripts, especially in scenarios with limited data availability. Rijhwani et al. (2020) developed a model that effectively learned from limited data for languages like Ainu, Griko, and Yakkha by leveraging existing translations. This approach was further enhanced in Rijhwani et al. (2021) through the incorporation of lexical decoding and self-training strategies, achieving significant improvements (up to 29%). For Sanskrit texts, Maheshwari et al. (2022) obtained favorable results by considering both phonetic encoding and the language's official alphabet.

4 Methodology for dataset creation

4.1 Data selection

We sourced documents from the SIL International[1] repository, targeting materials written in four languages: Asháninka, Shipibo-Konibo, Yanesha and Yine. These languages were chosen for their availability of resources within the repository compared to other Peruvian languages [2].

The documents, primarily in PDF format, present a wide range of contents, attributes, and layouts, including typewritten and handwritten text, tables, and images. Additionally, the content may be organized in multiple columns and vary in font sizes, sometimes presented in multiple languages.

To ensure dataset consistency, we focused on a subset of monolingual, typewritten documents with uniform font sizes. From each document, we selected a sample of pages (10%) for annotation and evaluation, based on criteria such as readability, resolution, tilt, and content alignment.

4.2 Data annotation

We manually annotated the documents following the workflow depicted in Figure 2. This process involved two key roles: an annotator and a reviewer. The annotator, possessing prior annotation experience, utilized a free online OCR tool[3] to generate a preliminary transcription. This initial step facilitated the annotation process, minimizing the time, effort, and potential errors associated with manual transcription.

Subsequently, the preliminary transcription was rectified by the annotator and reviewed by another annotator, who double-checked the annotations and addressed any discrepancies or errors encountered.

4.3 Data preprocessing

Although most document pages generally exhibit good quality, certain defects, such as ink stains overlapping with characters and low scanning resolution, can significantly affect fine text details,

[1] SIL International: https://www.sil.org/

[2] See Figure 4 in Appendix B for details about the data availability.

[3] Free online OCR tool: https://www.onlineocr.net/es/

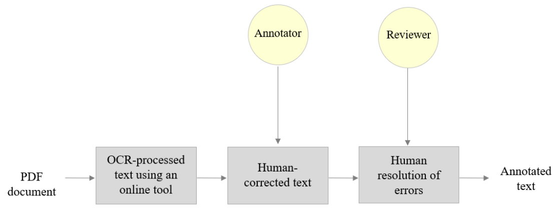

Annotator
Reviewer
PDF document
OCR-processed text using an online tool
Human-corrected text
Human resolution of errors
Annotated text

Figure 2: Annotation Process.

degrading page quality, and potentially impacting OCR performance. To address these issues, we conducted preprocessing steps including noise and attribute removal, as well as image enhancement, aimed at improving OCR accuracy.

Noise removal We manually cleaned documents using the Nitro PDF[4] tool to eliminate elements adding noise, such as images and page numbers.

Attribute removal Text delimited by boxes can impact text recognition (see Figure 5 in the Appendix). Consequently, we categorized pages into two groups: those containing only text (Group 1) and those with text delimited by boxes (Group 2).

For Group 2 pages, we applied image correction to automatically detect and remove the boxes when needed (for one of the tools we experimented with, it did not provide any benefit).

Image enhancement After converting the pages to Portable Network Graphics (PNG) format, we applied corrections focused on removing irregularities and improving the contour of the characters to achieve more effective segmentation.

Language	NoP	NoS	NoT	NoUT	NoTP
Asháninka	134	2239	8103	2309	61
Shipibo-Konibo	89	1495	7251	1685	82
Yanesha	91	1468	6315	1574	70
Yine	140	2246	9754	2449	70

Table 1: Corpora description: NoP = Number of pages, NoS = Number of sentences, NoT = Number of tokens, NoUT = Number of unique tokens, NoTP = Number of tokens per page.

4.4 Dataset description

The resulting dataset comprises 454 scanned pages from 89 books written in the four indigenous languages: Asháninka, Shipibo-Konibo, Yanesha, and Yine. This dataset[5] comprises 31,423 tokens, distributed almost equally across the languages (see Table 1). Significantly, compared to the dataset previously generated by Bustamante et al. (2020), our work expands the vocabulary by incorporating an average of 3,110 unique tokens per language.

Beyond standard alphanumeric characters, the dataset includes digits (0-9), diacritics, punctuation marks, and various compound characters like **m̃**, **c̈**, and **t̃**. Approximately 36% of the characters appear fewer than 10 times. Moreover, nearly 45% of the employed characters deviate from the contemporary official alphabets of these languages.[6] This issue arises from the fact that the analyzed documents were written before the establishment of the official alphabets for these languages.

5 OCR Process

We employed two OCR systems, Google Vision[7] (version 3.4.4) and Tesseract[8] (version 5.3.3), to generate initial text transcriptions. Although neither system directly supports the languages studied, they recognize the common Latin script shared by these languages. Previous research has demonstrated their effectiveness in low-resource language settings, including Sanskrit (Maheshwari et al., 2022), Ainu, Griko, Yakka (Rijhwani et al., 2020), Tamil, and Sinhala (Vasantharajan et al., 2022).

After the initial OCR transcriptions, a two-step preprocessing stage was implemented to enhance output quality. We evaluated the results using two standard metrics: Character Error Rate (CER) and Word Error Rate (WER). These metrics quantify OCR accuracy based on the Levenshtein distance, which measures the minimum number of edit operations (substitutions, deletions, insertions) required to transform the original text into the OCR-generated text (Neudecker et al., 2021).

5.1 Preprocessing of initial OCR transcripts

Both OCR systems faced challenges in differentiating between various forms of similar punctuation marks, such as hyphens (-, —, _) and single quotation marks (', ’). To address this ambiguity, we standardized analogous punctuation marks and converted all text to lowercase while preserving the original content.

[4] Nitro PDF: https://www.gonitro.com/

[5] The dataset is available at https://github.com/iapucp/OCR-Peru

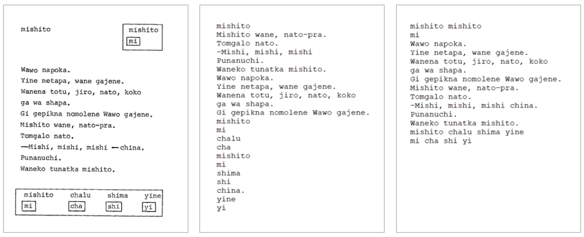
[6] The complete list of characters are included in Table 3 in the Appendix.

[7] Google Vision OCR: https://cloud.google.com/vision/docs/ocr

[8] Tesseract OCR: https://tesseract-ocr.github.io/

5.2 Alignment correction

Additionally, we observed that Google Vision OCR recognizes texts but fails to maintain the correct order, particularly affecting the text recognition of Group 2 pages, as depicted in Figure 3. To address this challenge, we developed a module to automatically align the initial transcriptions based on their vertical and horizontal positions, resulting in a reduction of approximately 9% in CER and 12% in WER. This enhancement was unnecessary for Tesseract OCR transcripts, as it effectively detects text order using text block segmentation.



```
mishito                         mishito
                                mi

Wawo napoka.
Yine netapa, wane gajene.
Wanena totu, jiro, nato, koko
ga wa shapa.
Gi gepikna nomolene Wawo gajene.
Mishito wane, nato-pra.
Tomgalo nato.
—Mishi, mishi, mishi —china.
Punanuchi.
Waneko tunatka mishito.

mishito     chalu     shima     yine
mi          cha       shi       yi
```

(a) Scanned document

```
mishito
Mishito wane, nato-pra.
Tomgalo nato.
-Mishi, mishi, mishi
Punanuchi.
Waneko tunatka mishito.
Wawo napoka.
Yine netapa, wane gajene.
Wanena totu, jiro, nato, koko
ga wa shapa.
Gi gepikna nomolene Wawo gajene.
mishito
mi
chalu
cha
mishito
mi
shima
shi
china.
yine
yi
```

(b) Google Vision OCR transcription

```
mishito mishito
mi
Wawo napoka.
Yine netapa, wane gajene.
Wanena totu, jiro, nato, koko
ga wa shapa.
Gi gepikna nomolene Wawo gajene.
Mishito wane, nato-pra.
Tomgalo nato.
-Mishi, mishi, mishi china.
Punanuchi.
Waneko tunatka mishito.
mishito chalu shima yine
mi cha shi yi
```

(c) Google Vision OCR transcription aligned using the developed automatic alignment module

Figure 3: Alignment of Google Vision OCR transcriptions

5.3 Types of errors

We identified three main types of errors in the OCR predictions:

Misprediction of characters with diacritics Characters such as b̃, c̈, m̃, p̃, t̃, and s̈ primarily found in ancient texts, were frequently misrecognized. This likely stems from their relative scarcity in modern Latin-script training data used in the OCR model training. This limited exposure led to inefficient recognition, contributing to approximately 60% of OCR errors, particularly in Shipibo-Konibo and Yanesha languages.

Insertion of non-existent characters Both OCR engines introduced orthographically similar characters not present in the original texts. Tesseract was more prone to this error (2.7 times more frequently), adding an average of 65 additional characters compared to Google Vision's 24. Tesseract showed repetitive patterns in adding characters, combining similar ones like **cç** and **ií**, often at sentence boundaries. Additionally, it misrecognizes small stains as characters. In contrast, Google Vision demonstrated better stain filtering but tended to replace similar characters like **š** with **ŝ**. This error type represented approximately 12% of the errors made by Google Vision OCR and 20% by Tesseract.

Incorrect word boundary detection Predominantly observed in Google Vision OCR, this involved adding extra spaces between words. It accounted for 47% of text identification errors in Asháninka and Yine languages but only 8% in Shipibo-Konibo and Yanesha languages.

6 Post-OCR process

6.1 Correction models

We applied five post-OCR methods to correct the errors made by the OCR systems:

SingleSource (Rijhwani et al., 2020) A sequence-to-sequence model tailored to effectively learn from limited data. We employed the single-source model.

Denorm (Oncevay et al., 2022) A spell checker trained to correct misspelling errors in Asháninka, Shipibo-Konibo, Yanesha, and Yine languages, normalizing sentences according to each language's grammar and norms.

Ensemble (Oncevay et al., 2022) An ensemble spell checker addressing five types of errors: character replacements, insertions, or deletions; errors from using a QWERTY keyboard; errors due to syllable similarity or ambiguity between phonemes and graphemes; and characters not included in the standardized alphabets of the languages.

SingleSource+Denorm A cascaded approach applying the SingleSource model followed by the Denorm model.

SingleSource+Ensemble A cascaded approach applying the SingleSource model followed by the Ensemble model.

6.2 Model training

Since only the model proposed by Rijhwani et al. (2020) required training, we trained it for each language using the basic hyperparameters configuration suggested. We employed five different random initializations on a system with 45 GB of RAM and 8 CPUs, and it required a total of 98 hours to complete. Subsequently, we evaluated all models and tools using the annotated test set.

OCR	Model	CER				WER			
		Asháninka	Shipibo-Konibo	Yanesha	Yine	Asháninka	Shipibo-Konibo	Yanesha	Yine
Tesseract	Baseline OCR	1.65	4.30	8.11	1.61	13.42	19.74	41.89	9.53
	SingleSource	1.34	1.55	3.75	**1.12**	9.85	8.21	20.66	**6.83**
	Denorm	7.32	15.83	11.32	3.68	35.84	50.33	52.28	15.25
	Ensemble	4.92	11.85	9.91	4.49	29.35	44.5	48.63	18.77
	SingleSource + Denorm	6.95	14.52	8.31	2.87	32.74	42.91	43.17	12.46
	SingleSource + Ensemble	4.2	9.48	6.85	3.74	25.81	34.17	38.8	16.28
Google Vision	Baseline OCR	**0.76**	2.61	5.53	1.49	9.00	12.98	39.16	10.26
	SingleSource	0.92	**0.88**	**2.32**	1.45	**8.61**	**4.85**	**17.41**	7.74
	Denorm	6.86	14.61	10.01	3.45	33.19	44.77	54.1	15.69
	Ensemble	4.04	10.55	8.09	4.47	25.52	39.07	50.46	20.09
	SingleSource + Denorm	6.56	13.63	7.09	3.02	31.42	40.13	41.89	12.76
	SingleSource + Ensemble	3.83	8.75	5.66	4.01	24.48	31.52	37.89	17.16

Table 2: Results of applying the correction methods to the transcripts of the Tesseract and Google Vision OCRs

7 Results

Table 2 presents the results of applying correction methods to OCR transcripts. The SingleSource model proved most effective in rectifying OCR errors due to several factors. Firstly, pre-training the model with the languages' characters facilitated the removal of non-existent characters from the transcripts. Secondly, it reduced errors from incorrect word boundary identification by 33% in Asháninka and Yine languages. Lastly, it significantly enhanced the recognition of characters with diacritics by 35% for Shipibo-Konibo and 65% for Yanesha, achieving a 99% accuracy in identifying these characters.

Regarding errors introduced by this model, they primarily involved character deletion but were significantly fewer compared to successfully corrected words. The ratio of successfully corrected words to unsuccessfully corrected words was 5:1 for Shipibo-Konibo, 4:1 for Yanesha, and 2:1 for Yine. However, in the case of Asháninka, this ratio shifted to 2:1 only for the Tesseract OCR-generated transcripts, but reversed to 3:5 for Google Vision OCR-generated transcripts. This reversal led to a higher number of degraded words than enhanced ones, evidenced by a 0.16 increase in CER attributed to minimal errors in OCR transcription that the correction model cannot rectify. An important consideration of this model arises when the text contains low-frequency characters. Uneven distribution during dataset partitioning may result in some characters being absent from training but present in evaluation sets, impacting performance.

On the other hand, the correction models based on spell-checkers approached OCR transcription errors by standardizing the texts. Given that most of these texts were old and did not adhere to the official alphabet and language rules, this method was ineffective, resulting in more errors introduced by the model than words successfully corrected. These errors primarily consisted of omitted diacritics and character replacements aimed at conforming the text to standardization norms. More analysis about the standardization is discussed in Appendix A.

8 Conclusions and future work

This work digitized textbooks in four Peruvian languages using OCR systems. We contributed an annotated dataset to assess the performance of Google Vision and Tesseract OCRs. Google Vision demonstrated higher accuracy in character recognition, while Tesseract excelled in maintaining text order across multiple columns. To address Google Vision's limitation in maintaining text order, we developed an alignment module. Additionally, we evaluated five error correction methods and found that the SingleSource model, designed for learning from limited data, was the most effective, particularly in correcting characters with diacritics.

Future efforts aim to optimize the hyperparameters of the SingleSource model and implement the multi-source model by Rijhwani et al. (2020) to leverage Spanish translations available for 89% of the books.

References

Haithem Afli, Loïc Barrault, and Holger Schwenk. 2016. Ocr error correction using statistical machine translation. *Int. J. Comput. Linguistics Appl.*, 7:175–191.

Michael Arrigo, Stephanie Strassel, Nolan King, Thao Tran, and Lisa Mason. 2022. CAMIO: A corpus for OCR in multiple languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1209–1216, Marseille, France. European Language Resources Association.

Fadoua Ataa Allah, Siham Boulaknadel, and Seth Darren. 2023. New trends in less-resourced language processing: Case of amazigh language. 12.

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

Ministerio de Educación del Peru. 2013. *Documento nacional de lenguas originarias del Perú*. Lima.

Wei-Jen Ko, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Naman Goyal, Francisco Guzmán, Pascale Fung, Philipp Koehn, and Mona Diab. 2021. Adapting high-resource nmt models to translate low-resource related languages without parallel data.

Ayush Maheshwari, Nikhil Singh, Amrith Krishna, and Ganesh Ramakrishnan. 2022. A benchmark and dataset for post-ocr text correction in sanskrit.

Christopher Moseley, editor. 2010. *Atlas of the world's languages in danger*, 3rd edition. UNESCO, Paris, France.

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).

Arturo Oncevay, Gerardo Cardoso, Carlo Alva, César Lara Ávila, Jovita Vásquez Balarezo, Saúl Escobar Rodríguez, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Juan López Bautista, Nimia Acho Rios, Remigio Zapata Cesareo, Héctor Erasmo Gómez Montoya, and Roberto Zariquiey. 2022. SchAman: Spell-checking resources and benchmark for endangered languages from amazonia. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 411–417, Online only. Association for Computational Linguistics.

Toni Pedrós. 2018. Ashéninka y asháninka: ¿de cuántas lenguas hablamos? *Cadernos de Etnolingüística*, 6(1):1–30.

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.

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.

Pilar M. Valenzuela. 2003. *Transitivity in Shipibo-Konibo Grammar*. Ph.D. thesis, Eugene: University of Oregon, Eugene, OR.

Charangan Vasantharajan, Laksika Tharmalingam, and Uthayasanker Thayasivam. 2022. Adapting the tesseract open-source ocr engine for tamil and sinhala legacy fonts and creating a parallel corpus for tamil-sinhala-english. In *2022 International Conference on Asian Language Processing (IALP)*. IEEE.

A Spell checker assessment with standardized texts

Due to the spell checker's limitations in correcting OCR transcripts using the standardization approach, we assessed a small set of 50 sentences. This evaluation compared the spell checker's corrections with the original texts in their standardized versions. We manually standardized 25 sentences in both Yanesha and Shipibo-Konibo languages with the support of native speakers to ensure accuracy. Despite this fair comparison to the standardized texts, we noted no improvement in the CER and WER values. Moreover, opportunities for enhancement remain in both the Denorm and Ensemble models proposed in Oncevay et al. (2022).

B Dataset additional information

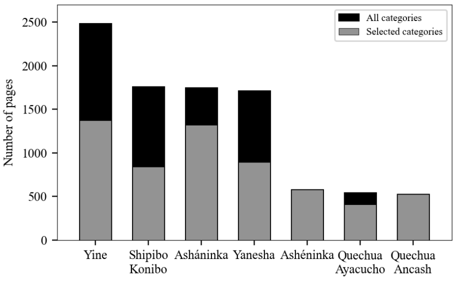
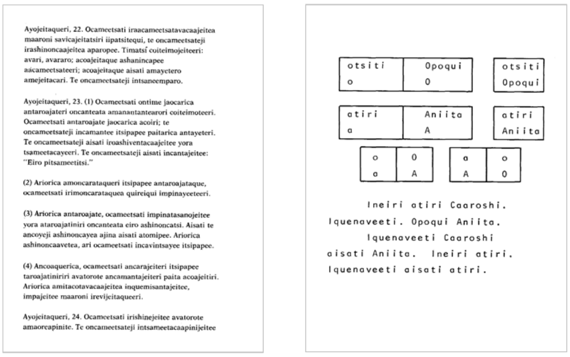


Figure 4: Number of pages available in the SIL repository for documents written in Peruvian languages. The selected categories exclude handwritten and unconstrained size texts.



Ayojeitaqueri, 22. Ocameetsati iraacameetsatavacaajeitea maaroni savicajeitatsiri iipatsitequi, te oncameetsateji irashinoncaajeitea aparopee. Timatsí coiteimojeiteeri: avari, avararo; acoajeitaque ashanincapee aácameetsateeri; acoajeitaque aisati amayetero amejeitacari. Te oncameetsateji intsaneemparo.

Ayojeitaqueri, 23. (1) Ocameetsati ontime jaocarica antaroajateri oncanteata amanantantearori coiteimoteeri. Ocameetsati antaroajate jaocarica acoiri; te oncameetsateji incamantee itsipapee paitarica antayeteri. Te oncameetsateji aisati iroashiventacaajeitee yora tsameetacayeeri. Te oncameetsateji aisati incantajeitee: "Eiro pitsameetitsi."

(2) Ariorica amoncarataqueri itsipapee antaroajataque, ocameetsati irimoncarataquea quireiqui impinayeeteeri.

(3) Ariorica antaroajate, ocameetsati impinatasanojeitee yora ataroajatiniri oncanteata eiro ashinoncatsi. Aisati te ancoyeji ashinoncayea ajina aisati atomipee. Ariorica ashinoncaavetea, ari ocameetsati incavintsayee itsipapee.

(4) Ancoaquerica, ocameetsati ancarajeiteri itsipapee taroajatiniriri avatorote ancamantajeiteri paita acoajeitiri. Ariorica amitacotavacaajeitea inquemisantajeitee, impajeitee maaroni irevijeitaqueeri.

Ayojeitaqueri, 24. Ocameetsati irishinejeitee avatorote amaoreapinite. Te oncameetsateji intsameetacaapinijeitee

(a) Group 1: Only text

otsiti	Opoqui
o	O

otsiti
Opoqui

atiri	Aniita
a	A

atiri
Aniita

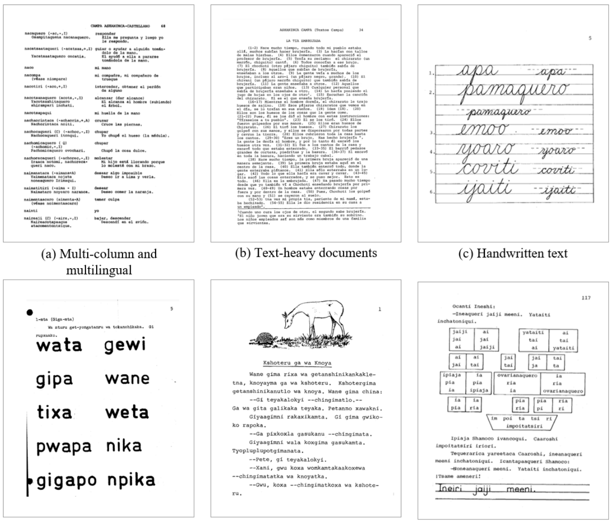
o	O
a	A

a	O
A	O

Ineiri atiri Caaroshi.
Iquenaveeti. Opoqui Aniita.
Iquenaveeti Caaroshi
aisati Aniita. Ineiri atiri.
Iquenaveeti aisati atiri.

(b) Group 2: Text delimited by boxes

Figure 5: Examples of pages from Group 1 and Group 2. Group 1 consists of text-heavy documents, whereas Group 2 presents either the entire text or portions of the text within tables.



(a) Multi-column and multilingual

(b) Text-heavy documents

5

1. apa --apa--
2. pamaquero --pamaquero--
--pamaquero--
3. emoo --emoo--
4. yoaro --yoaro--
5. coviti --coviti--
6. ijaiti --ijaiti--

(c) Handwritten text

5

1-mta (Giga-mta)

Wa mturu get-yongatanru wa tokanchikaka. Gi rupxaako.

wata	gewi
gipa	wane
tixa	weta
pwapa	nika
gigapo	npika

(d) Unconstrained size text

1

Kshoteru ga wa Knoya

Wane gima rixa wa getanshinikankakletna, knoyayma ga wa kshoteru. Kshotergima getanshinikanutlo wa knoya. Wane gima china:

--Gi teyakalokyi --chingimatlo.--
Ga wa gita galikaka teyaka. Petanno xawakni.

Giyaagimni rakaxikamta. Gi gima gwikoko rapoka.

--Ga pixkoxla gasukanu --chingimata.

Giyaagimni wala koxgima gasukamta. Tyopluplupotgimanata.

--Pete, gi teyakalokyi.

--Xani, gwu koxa womkamtakaakoxewa --chingimatatka wa knoyatka.

--Gwu, koxa --chingimatkoxa wa kshoteru.

(e) Text with images

117

Ocanti Ineshi:

—Ineaqueri jaiji meeni. Yataiti inchatoniqui.

jaiji jai ai	ai jai jaiji

yataiti tai ai	ai tai yataiti

ai jai	ai tai

jai tai

jai ja	tai ta

ipiaja pia ia	ia pia ipiaja

ovarianaquero ria ia	ia ria ovarianaquero

ia pia	ia ria

pia ria

pia pi	ria ri

im poi ta tsi ri
impoitatsiri

Ipiaja Shamoco ivancoqui. Caaroshi impoitatsiri iriori.

Tequerarica yareetaca Caaroshi, ineanaqueri meeni inchatoniqui. Icantapaaqueri Shamoco:

—Noneanaqueri meeni. Yataiti inchatoniqui. ¡Tsame ameneri!

Ineiri jaiji meeni.

(f) Text within tables and guidelines

Figure 6: Examples of document pages considering the classification made by Arrigo et al. (2022).

Language	Official characters	Unofficial characters	Punctuation marks	Digits
Asháninka	a, b, ch, e, i, j, k, m, n, ñ, o, p, r, s, sh, t, ts, y	á, c, d, é, f, g, í, l, ó, q, u, ú, v, x, z	!, ", (,), ,, -, ., /, :, ;, ?, —,¡, ¿	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
Shipibo-Konibo	a, b, ch, e, i, j, k, m, n, o, p, r, s, sh, t, ts, y	á, c, d, é, e , f, g, h, í, l, ñ, ó, q, s̈, u, ú, v, z	!, ", (,), ,, -, ., :, ;, =, ?, ¡, ¿, —	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
Yanesha	a, b, ch, e, ë, g, j, k, ll, m, n, ñ, o, p, r, rr, s, sh, t, ts, y	á, ä, b̃, c, c̈, d, f, h, i, í, l, m̃, ó, p̃, q, t̃, u, ú, v, z	", ', (,), ,, -, ., /, :, ?, ¿, —, '	0, 1, 2, 3, 4, 7, 8, 9
Yine	a, ch, e, g, i, j, k, l, m, n, o, p, r, s, sh, t, ts, u, w, x, y	á, b, c, d, é, f, í, ú, ü, v	", (,), *, ,, -, ., /, :, ;, ?, ¿, —, _, !, ¡	0, 1, 2, 3, 4, 5, 6, 7, 8, 9

Table 3: Characters present in the documents. Official characters: Belonging to the official alphabet of the language. Unofficial characters: Not belonging to the official alphabet of the language.

C Resource collection

Title	Language
Shitsa pajitachari anquilostoma aisati ameba	Asháninka
Campa 2	Asháninka
Ocantacota nonampi	Asháninka
Nantayetiri nonampiqui aisati noquemayetiri	Asháninka
Shiquiri	Asháninka
Naturaleza y vida social 1, 2	Asháninka
Naturaleza y vida social 1, 2	Asháninka
Quenquetsarentsi	Asháninka
Ina	Asháninka
Timayetatsiri quipatsiqui	Asháninka
Ompiquiri 6	Asháninka
Tsame aneanatacoteri Caaroshi (Vamos a leer sobre Carlos: Libro 3 para la lectura y escritura)	Asháninka
Tuberculosis (Libro de ciencias naturales 6)	Asháninka
Timatsi cameetsatatsiri acoajeitaqueri: Te oncameetsateji intsaneemparo	Asháninka
Jaoca ancantajeari antecatsijeitantajeari maaroni	Asháninka
Jaoca icanteetirori aamaacoventearo ajipee	Asháninka
Icantacota peeraniniri	Asháninka
Avatsa (El cuerpo humano: Libro de ciencias naturales 2)	Asháninka
Icantacota Shintsia	Asháninka
Jaoca acanteriri ameneri cameetsa vaca	Asháninka
Ameneri cameetsa aparoni jananequi	Asháninka
Tsame aneanatacoteri ompiquiri (Vamos a leer acerca de Ompíquiri)	Asháninka
Campa 3	Asháninka
Gigkanni Pirana	Yine
Gitaklu pirana ga wa prachi	Yine
Giyoliklu pirana	Yine
Gwacha Ginkakle	Yine
Jitomta 3	Yine
Lima pirana	Yine
Mgenoklumta	Yine
Muchikawa kewenni pirana ga wa pimri ginkaklukaka 10	Yine
Naturaleza y vida social 1: Manual para los cursos de naturaleza y vida social, y práctica de Castellano, para primer año	Yine
Naturaleza y vida social 2	Yine
Naturaleza y vida social 3	Yine
Nopra kina 2	Yine
Papa nikchi gijga	Yine
Papa-mta 1	Yine
Papisho 5	Yine
Pejri-mta 4	Yine
Walo-mta 7, 8	Yine
Yine ginkaklekaka 12 (Cartilla 12)	Yine
Yine sana kamruta	Yine
Yineru tokanu 3a	Yine
Yineru tokanu 3b	Yine
Yineru tokanu IX	Yine

Title	**Language**
Cuentos de la zorra y el zorro	Shipibo-Konibo
Ëa- tapaman caní 5	Shipibo-Konibo
Japari peoquin yoyo ati quirica	Shipibo-Konibo
Jascaaquin baqueshocobo coiranhanan, jahuequiamati yoii ica	Shipibo-Konibo
Jatibiainoa joni coshibaon, jascaas̈hon jacon jahuequi aresti jonibaon jahuequescamabi itiaquin shinana	Shipibo-Konibo
Moa peoquin yoyo ati: Quirica 4	Shipibo-Konibo
Moa peoquin yoyo ati: Quirica 5	Shipibo-Konibo
Moa peoquin yoyo ati: Quirica 6	Shipibo-Konibo
Nanbonyabi nacan, noa isin meniai yoia	Shipibo-Konibo
Naturaleza y vida social 2	Shipibo-Konibo
Naturaleza y vida social 3	Shipibo-Konibo
Non paron ja jahuequibo 1 (Nuestros recursos naturales: Guía didáctica 1 de ciencias naturales)	Shipibo-Konibo
Quimisha Incabo ini yoia (Leyendas de los shipibo-conibo sobre los tres Incas)	Shipibo-Konibo
Quirica 10 (Afianzamiento de lectura 10: Animales del mundo)	Shipibo-Konibo
Quirica 4	Shipibo-Konibo
Quirica 5	Shipibo-Konibo
Quirica 6	Shipibo-Konibo
Quirica 7	Shipibo-Konibo
Quirica 8	Shipibo-Konibo
Quirica 9 (Libro 9: Afianzamiento para la lectura)	Shipibo-Konibo
Ach	Yanesha
Ahuat̃ serraparñats at̃o eñalleta atsne'ñam̃a arrorr	Yanesha
Amuesha 7 - SHAñE'	Yanesha
Apa ñam̃a ach (Papá y mamá: Libro 3 para la lectura y escritura)	Yanesha
Atet̃cha'yecue'cheshat̃oll	Yanesha
At̃o'yepotamperra Meshtaso ñam̃a po'poñ serrparñats	Yanesha
Berroc̈hno ñeñt̃ Africo'marnesha'	Yanesha
Cartilla 9 (Besllom̃)	Yanesha
Chom - Amuesha 8	Yanesha
Homenaje a la Declaración Universal de Derechos Humanos en su 40 aniversario 1948-1988	Yanesha
Ma’yarr poyoc̈her ñam̃a po’poñec̈hno serrparñats	Yanesha
Manual de ganadería	Yanesha
Naturaleza y vida social 3	Yanesha
Nochcar (Mi perro: Libro 6 para la lectura y escritura)	Yanesha
Ot̃ec̈hno	Yanesha
Pa'namen alloch yechopene'champesyen	Yanesha
Pa'namen atsnañtsoc̈hno	Yanesha
Pepe	Yanesha
Pepe payara	Yanesha
Pepe ñam̃a ema'(Pepe y la niñita: Libro 4 para la lectura y escritura)	Yanesha
Posho'll (La ardilla: Libro 7 para la lectura y escritura)	Yanesha
Tempo pueserrpareñ	Yanesha
YANESHA'	Yanesha
Yehuom̃cheña	Yanesha

Table 4: Resources utilized from the SIL repository.