Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2914–2923
Marseille, 11–16 May 2020

No data to crawl? Monolingual corpus creation from PDF files of truly low-resource languages in Peru

Gina Bustamante[♠], **Arturo Oncevay**[♠◇] **and Roberto Zariquiey**[♣]
[♠]Artificial Intelligence Research Group, Pontificia Universidad Católica del Perú
[◇]School of Informatics, University of Edinburgh
[♣]Department of Humanities, Linguistics Unit, Pontificia Universidad Católica del Perú
{gina.bustamante,arturo.oncevay,rzariquiey}@pucp.edu.pe

Abstract

We introduce new monolingual corpora for four indigenous and endangered languages from Peru: Shipibo-konibo, Ashaninka, Yanesha and Yine. Given the total absence of these languages in the web, the extraction and processing of texts from PDF files is relevant in a truly low-resource language scenario. Our procedure for monolingual corpus creation considers language-specific and language-agnostic steps, and focuses on educational PDF files with multilingual sentences, noisy pages and low-structured content. Through an evaluation based on language modelling and character-level perplexity on a subset of manually extracted sentences, we determine that our method allows the creation of clean corpora for the four languages, a key resource for natural language processing tasks nowadays.

Keywords: Shipibo-Konibo, Ashaninka, Yanesha, Yine, endangered languages, indigenous languages, low-resource languages, pdf processing, monolingual corpus, corpus creation

1. Introduction

Several natural language processing (NLP) tasks are currently improving their performance based on deep learning models, which require large amounts of data to learn from. For this reason, unlabelled text corpora have become an elemental resource to improve performance in various NLP applications, such as in representation learning (Devlin et al., 2019), language modelling (Buck et al., 2014), and neural machine translation with back-translation (Sennrich et al., 2016) or in an unsupervised scenario (Artetxe et al., 2018; Lample et al., 2018). Other non-language generation tasks, such as text classification, can also benefit in a transfer learning setting, as the model can focus on the specific goal and not in learning the representation of words or subwords (Howard and Ruder, 2018).

Monolingual data is usually inexpensive to obtain. A popular and widely method is crawling different web pages. However, this is only possible under the assumption that there are sufficient web sites written in the target language. This is not the case for many endangered languages that do not have a consistent presence in the World Wide Web. In this way, the lack of web content limits the creation of language resources and developing further NLP applications for truly low-resource languages.

The aforementioned scenario is common for most of the indigenous and endangered languages in Peru, a multicultural and multilingual country. There is almost not web content of the Peruvian native languages, even in official Government sites. Nonetheless, recent efforts in bilingual education (Spanish – native language) for indigenous communities have endorsed the creation of digital educational resources written in local languages. These documents are stored in online repositories, mostly in Portable Document Format (PDF).

We could take advantage of the available PDF files to extract as much monolingual text as possible. However, to obtain plain text files with monolingual sentences of the target language, we need to overcome different issues caused by the unstructured nature of the documents. Educational-related content usually includes graphical guides or incomplete sentences/phrases as tasks for students. Bilingual content is also a challenge, as we cannot disregard all the sentences with the minimum presence of an external language. Most native languages in the country have adopted several loanwords given the extended contact with Spanish. Therefore, we need to carefully identify and clean the best sentences to build an NLP resource.

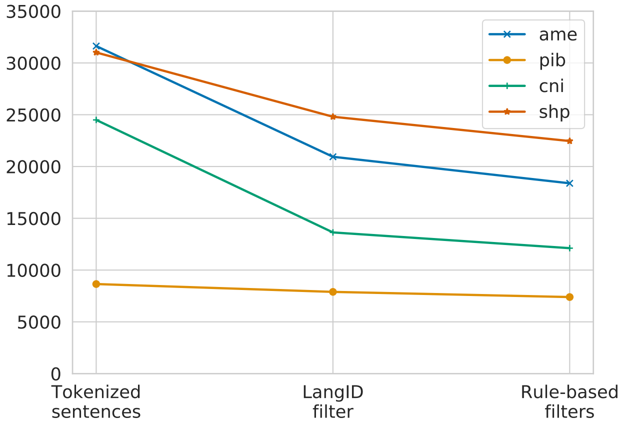


Figure 1: Summary of the number of sentences extracted and filtered per language in different stages

In this study, we focus on the extraction and processing of monolingual corpora for four indigenous Peruvian languages given noisy content from PDF files. Our contributions are:

- New monolingual corpora for four indigenous, endangered and truly low-resource languages from Peru: Shipibo-Konibo, Ashaninka, Yanesha and Yine. The four corpora include almost 60,000 sentences that have been automatically extracted from PDF files. Figure 1 summarises the extraction process.

- A procedure we followed to extract, clean and select the best sentences to build the monolingual corpora. The process includes language-agnostic (e.g. sentence tokenization) and language-specific (e.g. language identification) steps, where we also compare different rule-based and machine learning-based steps.
- Validation of our extracted data through language modelling and perplexity. We also created gold-standard subsets for each language, by manually sampling sentences from specific files.

2. Language specifics

In Peru, Castilian Spanish (*spa*) is the primary official language and is spoken by most of the population. Additionally, there are 48 indigenous languages, grouped in 19 language families, almost all (44) located in the Amazon region (Sullón Acosta et al., 2013; Zariquiey et al., 2019). Nonetheless, most of them are considered low-resource in the computational scope, since they have few available digital resources, such as computational linguistics corpora or NLP tools (Mager et al., 2018).

In this study, we focus on three languages from the Arawak family: Ashaninka (*cni*), Yanesha (*ame*), and Yine (*pib*); and one from the Pano language family: Shipibo-konibo (*shp*). All of them are Amazonian languages. To define our target languages, we considered the educational offer in one of the most important bilingual education institutions in the country[1], and chose the two most in-demand languages for native-spoken students (Asahaninka and Yanesha). In addition, we selected Yine given its poor status of language documentation (Zariquiey et al., 2019) and included Shipibo-Konibo to provide continuity to previous NLP research (Montoya et al., 2019).

Zariquiey et al. (2019) describe the vitality status of all Peruvian languages and dialects regarding the Agglomerated Endangerment Scale proposed in Glottolog (Hammarström et al., 2017) and GlottoScope (Hammarström et al., 2018), where they classify the following stages: not endangered, threatened, shifting, moribund, nearly extinct, extinct. The four languages in our scope are endangered in different levels: Shipibo-Konibo and Ashaninka are labelled as *threatened*, whereas Yanesha and Yine are in a *shifting* process.

3. Related work

Unstructured web corpus as Common Crawl[2] and crawled texts from Wikipedia have made a significant contribution providing large scale training data to learn word representations (Grave et al., 2018; Pennington et al., 2014), estimate n-gram language models (Buck et al., 2014) and apply transfer learning (Howard and Ruder, 2018). Furthermore, ParaCrawl (Esplà et al., 2019) is a parallel corpus for official languages in the European Union that aims to improve machine translation tasks with data retrieved from the web. The process is also feasible in some low-resource scenarios if there is data available. For instance, Sabeti et al. (2018) crawled over 250 websites to create a novel automatically generated corpus for Persian.

Besides, previous work on corpora creation from PDF documents has mostly focused on processing scientific articles (Goodman et al., 2018; Ferrés et al., 2018) to allow further scientific text mining. Additionally, Daudert and Ahmadi (2019) addressed the problem of sentence boundary detection on PDF files from the financial domain. Listed studies are remarkably precise in their domain, but they are strictly constrained to a specific PDF layout and fail to generalise to files from different sources.

Regarding monolingual corpora processing for Peruvian languages, the only precedent is the work of Espichán-Linares and Oncevay-Marcos (2017), where they extracted translations mostly from the Bible and linguistic material, for 29 indigenous language and 20 dialects, to build a language identification tool. With respect to our target languages, they worked with a considerable amount of sentences (almost 66,000 in total): 17,451 (*ame*), 11,997 (*shp*), 14,645 (*pib*), 22,057 (*cni*). We continue the effort and almost double the monolingual corpora for the four Peruvian languages with our two additional contributions. First, the educational sources we work with have been developed after an official revision of the language alphabets, in contrast with the old Bible translations, and thus, the texts are more suitable to support NLP applications targeting a modern style. Second, we perform a validation step to verify how much noise we are able to reduce for the final output.

4. Source of PDF documents

During the last decade, many native languages in Peru have been declared as official ones by the Peruvian Government (Sullón Acosta et al., 2013). However, the different official web sites of governmental entities and organisations have not been translated into the indigenous languages yet. Nonetheless, after a long process of documentation, the Government is publishing an increasing number of educational resources in PDF format to support bilingual education in indigenous communities. Therefore, we decided to exploit a digital repository[3] that includes books, guides and educational material for school teachers.

We downloaded all the documents with a label or keyword associated with any of the four target languages. Then, we classified the files in different types according to their content. Table 1 summarises the total number of documents per language and per type. For the four languages, Elementary School Workbook is the most frequent category. Conversely, document types of Dictionary and Community Traditions are rare. Furthermore, the languages with the highest number and variety of documents are Ashaninka (*cni*) and Shipibo-Konibo (*shp*), while Yanesha (*ame*) and Yine (*pib*) have small document type diversity. The mild case of Yine was expected, as it is the least documented language among the four in the linguistic domain (Zariquiey et al., 2019).

[1] UCSS, NOPOKI: `https://www.ucss.edu.pe/`

[2] https://commoncrawl.org

[3] Perueduca: `http://www.perueduca.pe`

Document type	ame	pib	cni	shp
Language Manual	3	3	1	3
Elementary School Workbook	12	6	18	9
Kindergarten Workbook	4	1	3	3
Dictionary	1	0	0	1
Tales	4	1	4	5
Community Tradition	0	0	1	6
Teaching Manual	0	0	6	5
Total	24	11	33	32

Table 1: Summary of the retrieved PDF documents and their categories per language

Method	ame	pib	cni	shp	Total
Rule-based	28,345	9,478	23,909	31,934	93,666
Unsupervised	31,630	8,651	24,498	31,012	95,791

Table 2: Comparison of rule-based and unsupervised sentence tokenizers by the number of split sentences

As most of the document types target a young audience (elementary school and kindergarten) or have a pedagogic goal, we expected to find visualisations (e.g. figures, plots, charts) intermixed with textual content. Figure 2 (and Figures 5 and 8 in the Appendix) presents some examples of the unstructured nature of the documents we need to deal with. For that reason, we analyse different approaches for corpus extraction. The procedure is detailed in the following section.

5. Methodology for corpus creation

To extract and create the corpora, we define three specific steps: (1) transformation of PDF into plain text files; (2) boundary detection of the sentences by comparing an unsupervised and a rule-based method; and (3) selection and filtering according to different criteria, such as language identification and specific heuristics.

5.1. PDF transformation

We take advantange of the open-source PDF-to-text converter PDFMiner[4] and generate a plain text file for each PDF. This step is not restricted to a specific document layout, source, or language.

Another possibility could be the transformation of the PDF documents into intermediate file formats with native markups, such as XML or HTML, and then to identify and extract the specific elements with raw text (e.g. headers, body). We tested a PDF to XML transformation, but the process was significantly longer in steps and time, and there was no noticeable improvement in the amount of extracted sentences with respect to a straightforward conversion to plain text format.

5.2. Sentence boundary detection

Automatically transformed text files present a vast diversity of content and layouts, and the identification of sentence boundaries is not a simple task. We can find different punctuation marks, bullet entries with or without enumeration, titles/subtitles/headers without any delimiter punctuation in the raw text, among others. Therefore, we must be able to handle different kinds of noise to obtain as many correct sentences as possible. For our study, we test an unsupervised sentence tokenizer[5] and manually designed regular expressions (a rule-based approach) for the following cases:

1. **Section titles** are found inside a capital letter and two line breaks.
2. A **standard sentence** is contained among a capital letter and a period.
3. **Questions** are texts between the two question marks ¿?. In this step, we only collect standalone questions but not the ones that are part of a previously extracted sentence.
4. Similarly, **exclamations** are texts inside two exclamation marks ¡!. We focus only on the unconnected exclamations.

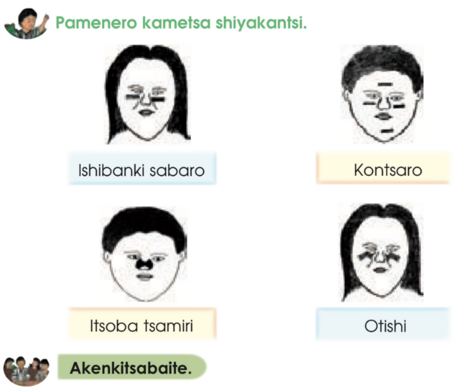


Figure 2: Example of an Ashaninka exercise where there are several blank spaces in between the different blocks of the layout (Perú. Ministerio de Educación. Dirección General de Educación Básica Alternativa, Intercultural Bilingüe y de Servicios Educativos en el Ámbito Rural, 2018a).

	Extracted sentences
Unsupervised	1. Pamenero kametsa shiyakantsi. 2. Ishibanki sabaro 3. Kontsaro 4. Itsoba tsamiri 5. Otishi 6. Akenkitsabaite.
Rule-based	1. Pamenero kametsa shiyakantsi. 2. Ishibanki sabaro Kontsaro Itsoba tsamiri Otishi 3. Akenkitsabaite.

Table 3: Split sentences from Figure 2.

[4] `https://github.com/euske/pdfminer`

[5] NLTK Punkt Sentence Tokenizer: `https://www.nltk.org/api/nltk.tokenize.html#module-nltk.tokenize.punkt`

5. For **quotes** identification, we look for the text within two quotation marks “ ”.

The unsupervised method splits more sentences than the rule-based one, as we can observe in Table 2. By manually inspecting the outputs, we found that most of the differences are generated from documents with specific layouts as in Figure 2 (mostly from children workbooks). Table 3 includes the output of the two methods for the aforementioned example. Since we expect to increase the reproducibility and generalise our procedure as much as possible, we decided to work with the unsupervised sentence tokenizer for the following steps.

5.3. Sentence filtering

Selection of the least noisy sentences is a crucial step in the overall procedure, as our long-term goal is to support the development of robust NLP applications for the target languages. For this purpose, we work with a language identification model and specific heuristics.

5.3.1. Language identification

Similar to the findings of Buck et al. (2014), we noticed that even when the expected language is known, some files contain instructions or entries written in a different language than our targets. Consequently, we use a Peruvian language identification tool, developed by Espichán-Linares and Oncevay-Marcos (2017), to label each sentence and drop the texts identified as written in languages out of our scope. They have reported highly accurate results even for some dialects withing a language, so we relied on the tool for the filtering process[6].

After the first filtering, the first entry of Table 5 shows that the language with the lowest number of ignored sentences is Yine (8.75%), whereas Ashaninka (44.34%) has the highest amount of dropped ones in proportion. The reported gap might be caused by the different types of retrieved documents per language. More than a half of the Ashaninka files are Elementary School Workbooks, which usually include texts in Spanish and other languages so that children can learn in a bilingual context. Besides, Ashaninka has a small number of Language Manual type of files, which are commonly the cleanest type of texts. More details about the filtering per type of document are included in Table 9 of the Appendix.

5.3.2. Rule-based heuristics

Following the language identification step, we perform a manual inspection of part of the output and identified specific issues. Thus, we propose different heuristics to clean and prepare a higher-quality corpus. Table 5 summarises the number of sentences removed at each step of the rule-based filtering process.

1. **Out-of-language characters (OOL):** Characters might be added or transformed incorrectly by the automatic PDF-to-text tool. Figure 5 in the Appendix presents an example where the original text includes images replacing some terms, and the transformation output adds random characters or symbols. In addition, we detect some code switched sentences with Spanish names (Figure 6 in the Appendix) that had been classified as part of the language in Section 5.3.1.[7]. As we cannot identify a specific set of characters or words to clean, we generalise a filtering rule considering the official alphabet of each language (Table 4) and exclude sentences containing words formed by graphemes outside the alphabet.

2. **Number of tokens:** Titles or Section headers that are usually composed by one single token are not too useful for language generation tasks such as language modelling or spell-checking. In this step, we remove sentences containing only one word. Results on Table 5, indicate that Yanesha (ame) loses more than a thousand sentences in this step (far more than the other languages). Examining possible explanations for this situation, we found out that in Yanesha it is common that only one word represents a whole phrase. However, we don’t have at this point an automatic way of differentiating which words represent a sentence and which do not, so we do not consider this phenomenon in the study.

3. **Token types per number of tokens ratio:** There are specific sentences with a large number of duplicated

	Alphabet
ame	a, b, bh, ch, xh, e, ë, g, j, k, kh, ll, m. mh, n, ñ, o, p, ph, r, rr, s, sh, t, th, ts, w, y
pib	a, ch, e, g, i, j, k, l, m, n, o, p, r, s, sh, t, ts, u, w, x, y
cni	a, b, ch, e, i, j, k, m, n, ñ, o, p, r, s, sh, t, ts, ty, y
shp	a, b, ch, e, i, j, k, m, n, o, p, r, s, sh, t, ts, w, x, y

Table 4: Official alphabet per language

Filter	ame	pib	cni	shp
Lang. identification	10,690	757	10,863	6,207
OOL characters	1,143	181	1,161	1,869
Number of tokens	1,140	108	77	44
V/N ratio	3	5	2	6
Token length	3	11	29	2
Split tokens	12	14	13	43
Math. expressions	259	170	237	381
Filtered sentences	41.89%	14.52%	50.54%	27.58%
Accepted sentences	18,380	7,395	12,116	22,460

Table 5: Number of filtered senteces in each step of the rule-based procedure.

[6] It is possible to update the language identification tool with the new monolingual texts for the four languages. Nevertheless, it is not the primary scope of this study.

[7] Although we are aware that this phenomenon is not necessarily incorrect, we remove the code-switched or language-mixed sentences to guarantee a clean corpus.

tokens (e.g. writing exercises for children in language guides or workbooks). Thus, we compute the ratio per sentence of token types *V* and number of tokens *N* to identify the duplication process. We define a 0.4 threshold for the *V/N* ratio, and dismiss all sentence below the value. We highlight that none of the targeted languages present re-duplication in their morpho-syntactic processes.

4. **Token length:** We observe some token length values above 40 characters. Even with the agglutinative nature of Peruvian native languages, those numbers seem unrealistic. Hence, we established a threshold and filtered sentences containing words with more than 40 characters. Furthermore, by analysing a potential source issue, we found that they are caused by specific document layouts. Figure 8 in the Appendix presents an example of an official alphabet that is transformed incorrectly in a very long word.

5. **Split tokens:** There are text orientations that are not entirely flat (see Figure 7 in the Appendix as an example). Thereafter, some tokens could be split in n-grams during the transformation process. Initially, we tried to reconstruct the entries, but we did not find a pattern. Thus, we employed regular expressions to detect and delete sentences holding this kind of issue. Specifically, we look for sentences with three or more sequential tokens composed of one or two characters at most.

6. **Mathematical expressions:** Elementary School Workbooks, a document type within our sources (Table 1), frequently contain maths concepts and exercises. During the plain text conversion, some of the captured mathematical expressions are incorrectly located within a sentence, which loose its original meaning (see Figure 9 in the Appendix). Therefore, we establish a rule where we look for sequences of numbers and operators inside a sentence and remove them from our final corpus.

There is room for improvement in the overall procedure, specifically in the OOL character filter. For instance, we are disregarding potential Spanish loanwords and Spanish-based lemmas in the filtered sentences. We plan to carefully process the language-mixed sentences with unsupervised word segmentation models (e.g. Byte Pair Encoding) in a joint training setting with concatenated Spanish texts.

6. Corpora description

We perform a large number or rare events (LNRE) modelling[8] to statistically analyse the processed corpora. The computed values are presented in Table 6.

We notice that both the vocabulary and hapax growth rate (V/N, $V1/N$) are similar for Yanesha, Yine, and Ashaninka (all from the Arawak family) despite the large gap in the number of sentences between them. Contrarily, even when having the largest number of tokens (N), Shipibo-Konibo or *shp* (from the Pano family) has the lowest growth rates. Language contact and the borrowing of words could help to explain the growth rate difference. For instance, the Arawak language family is spoken in a wider territorial extension and is influenced by several other languages and communities, whereas Pano languages persist almost exclusively to the central Amazonian region (Sullón Acosta et al., 2013). Furthermore, we could also assess the morphological complexity of the two families to explain the phenomena. For example, Arawak languages usually have a greater number of morphological feature values (e.g. cases) than Pano ones (Aikhenvald, 2012)[9].

	ame	**pib**	**cni**	**shp**
S	18,380	7,395	12,116	22,460
N	154,730	58,023	99,177	208,418
V	30,727	13,988	23,456	23,954
V1	4,626	1,902	3,257	4,062
V/N	0.198	0.241	0.236	0.114
V1/N	0.125	0.16	0.153	0.063
mean	6.035	4.148	4.228	8.700

Table 6: Corpora description: *S* = sentences collection size; *N* = number of tokens; *V* = vocabulary size; *V1* = number of tokens occurring once (hapax); *V/N* = vocabulary growth rate; *V1/N* = hapax growth rate; *mean* = word frequency mean.

Moreover, there is a high presence of tokens occurring once or hapax legomenon. In some cases, the large number of hapax is related to a poor quality of the corpus that might be caused by spelling errors or the presence of foreign words (Nagata et al., 2018). However, our scenario is expected given the agglutinative nature of the four target languages, so they might present a vast vocabulary diversity.

Finally, we also analyse the rich morphological nature of the indigenous Peruvian languages in terms of the number of characters per word and sentences. Firstly, Figure 3 describes the case of tokens, where we observe that the average length is in a range from 8 to 13 characters, and there are tokens with even more than 25 or 30 characters. Specifically, Ashaninka (*cni*) is the language with the largest token length average. We expected the scenario given that Ashaninka belongs to the Campa branch from the Awarak language family, which is known for having more complex languages than the other family-branches. Secondly, Figure 4 shows the sentence length distribution at character-level. We notice that the $\log_{10}$ length average for all languages is around 1.70, whereas Ashaninka is slightly higher with 1.80. The difference might not be related to the methodology or document types but due to the largest average word length of Ashaninka in contrast with the other three languages.

[8] We use the LNRE calculator developed by Kyle Gorman: `https://gist.github.com/kylebgorman`

[9] The measurement of morphological complexity is an open problem (Sagot, 2013) where counting methods are one of the most simple approaches.

	ame	**pib**	**cni**	**shp**
Non-filtered corpora	**3.42** ± 0.003	3.18 ± 0.001	3.09 ± 0.001	3.26 ± 0.001
Random sub-sampling	3.59 ± 0.005	3.19 ± 0.002	3.13 ± 0.002	3.25 ± 0.001
Our filtered corpora	3.43 ± 0.001	**3.15** ± 0.001	**3.06** ± 0.001	**3.18** ± 0.001

Table 7: Character-level perplexity scores (↓ lower is better). For our filtered corpora and the random sub-sampling set we perform ten different iterations and statistically compare the distributions (p-value). Non-filtered corpora presents the score using the entire corpora without any filter.

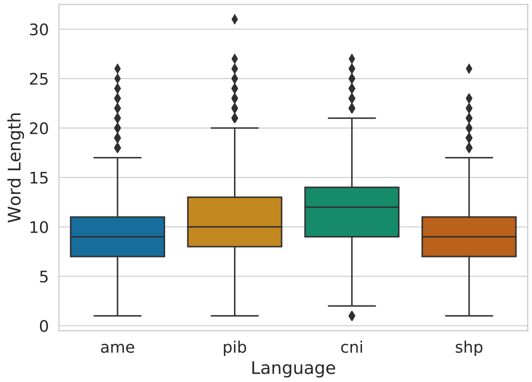


Figure 3: Word length distribution in number of characters per language

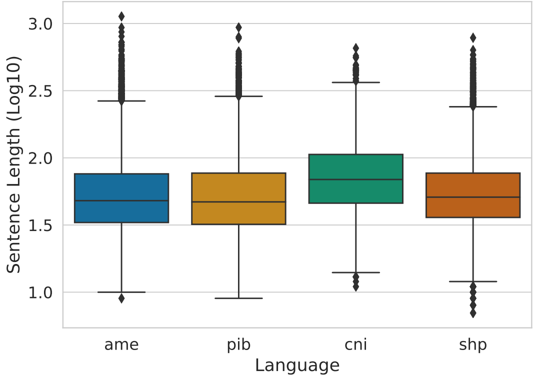


Figure 4: Sentence length distribution in number of characters per language. The vertical axis is in $\log_{10}$ scale.

7. Evaluation of the filtered corpus

We perform a downstream task to evaluate the quality of the corpora after the filtering process. Specifically, we train an open-vocabulary language model at character-level (Mielke et al., 2019) and measure the results with character-level perplexity (Mielke, 2019). For the experimentation, we use a low-compute version of a recurrent neural network, named Average SGD Weight-Dropped (Merity et al., 2018), with a smaller embedding size (300 units) for faster training. Concerning the evaluation, we introduce a new gold-standard set for each language. Thereafter, we explain our baselines for comparison and provide results with statistical significance.

7.1. Gold standard set

The evaluation sets are composed from two sources. First, we manually extracted sentences from additional Language Manuals documents provided by teachers of a bilingual educational programme (UCSS, NOPOKI). Second, we included extra sentences that were extracted from the most common types of documents in each language. We meticulously checked that none of the evaluation sentences was duplicated in the training set to avoid data leakage. More details about the composition of the test set are presented in Table 10 of the Appendix. Finally, we randomly split the evaluation set into development and test sets per language (see Table 8).

7.2. Evaluation and statistical comparison

The evaluation task is based on language modelling and perplexity. Given that the filtering process for obtaining quality corpus reduce the amount of sentences to work with, the most elemental baseline is to randomly select a subsample of sentences of the same size. Our second baseline is to use the whole non-filtered corpus, as there is evidence that the largest amount of data (and potential noise) could positively impact in perplexity scores (Prasad et al., 2018). Besides, we perform ten different experiments with each kind of filter by applying different random seeds, as we must make a valid statistical comparison to ensure that the results are not coincidental.

Table 7 presents the average results of the different runs per language for the two baselines and our filtering process. We observe that our filtering method completely outperforms the random sub-sampling baseline. Similarly, we obtain outstanding results in contrast with the baseline based on total amount of sentences, where our perplexity is better in all cases but for Yanesha (*ame*).

The case of Yanesha might be explained due to its large number of filtered sentences (41.89% dropped from 31,630) and most diverse alphabet among the four lan-

Lang	**Filtered Corpora**	**Random sub-sampling**	**Non-filtered corpora**	**Valid**	**Test**
ame	18,343	18,343	31,630	637	636
pib	7,347	7,347	8,651	614	614
cni	12,010	12,010	24,498	593	592
shp	22,035	22,035	31,012	780	780

Table 8: Corpus size information in number of sentences

guages (see Table 4). Therefore, our method had the challenge to train a language model with a large vocabulary but with a quarter less data than the baseline. Nevertheless, we observe a very narrow gap in the results, which suggests that the proposed filtering process is feasible to reduce the noise of the corpus.

8. Conclusions and future work

We described our method to extract new monolingual corpora for four indigenous and endangered languages from Peru. We perform a filtering process with language-specific and language-agnostic steps to select the least noisy candidates after, as our target is to support robust NLP applications in the native languages. By a validation with character-level perplexity in language modelling, we conclude that the filtered sentences obtain comparable or better results as with the whole non-filtered corpora or a random sample of the same size.

We plan to extend the monolingual corpus extraction for more Peruvian languages, as there are digital documents for several languages in the repositories we worked with. Similarly, there are additional digital sites with material written in native languages from public and private entities in the country. Thus, we expect to keep increasing the size and variety of the monolingual corpus for many Peruvian languages and update the language identification method with our new outputs, as it is a crucial tool for the purpose of new corpora extraction. Furthermore, there is a potential in extracting parallel texts from the PDF files, and thus, we are going to analyse the feasibility of building an automatic alignment tool for the unstructured layouts.

9. Acknowledgements

We are grateful to the computational linguistic team at PUCP: John Miller, Erasmo Gómez, Kervy Rivas, Gema Silva, Gildo Valero, Jaime Montoya and Gonzalo Acosta. Similarly, we thank the bilingual teachers from the UCSS who provided their own crafted material for the evaluation, and more specifically to Juan Rubén Ruiz for his support. Besides, we appreciate the comments of Fernando Alva-Manchego on a draft version and the feedback of our anonymous reviewers.

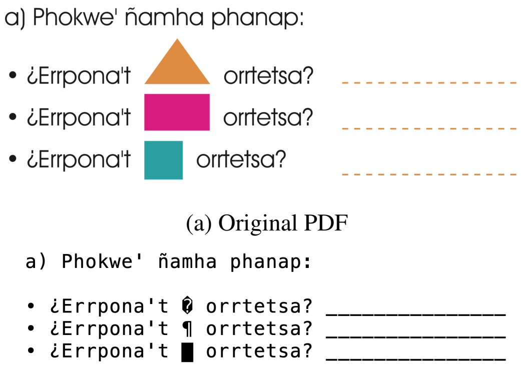
Finally, we acknowledge the research grant of the "Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica" (CONCYTEC, Peru) under the contract 183-2018-FONDECYT-BM-IADT-MU, and the support of NVIDIA Corporation with the donation of a Titan Xp GPU used for the study.

Appendix

Additional tables and figures with information about the corpora creation and final outcome.

	ame	**pib**	**cni**	**shp**
Language Manual	2.4	14.9	0.6	3.5
Elementary School Workbook	31.7	69.9	79.1	45.0
Kindergarten Workbook	3.2	4.4	2.2	4.5
Dictionary	59.0	0.0	0.0	16.2
Tales	3.7	10.8	6.4	6.4
Community Tradition	0.0	0.0	3.8	17.3
Teaching Manual	0.0	0.0	8.0	7.0

Table 9: Percentages of filtered sentences by document types per language



(a) Original PDF

```
a) Phokwe' ñamha phanap:

 • ¿Errpona't � orrtetsa? _______________
 • ¿Errpona't ¶ orrtetsa? _______________
 • ¿Errpona't █ orrtetsa? _______________
```

(b) Converted plain text file

Figure 5: Yanesha mixed figures and text example from (Perú. Ministerio de Educación. Dirección General de Educación Básica Alternativa, Intercultural Bilingüe y de Servicios Educativos en el Ámbito Rural, 2018b)

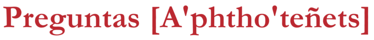
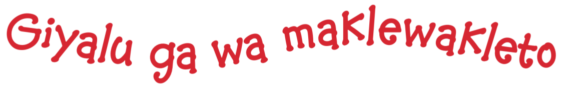

Preguntas [A'phtho'teñets]

Figure 6: Yanesha code switched sentence example (Perú. Ministerio de Educación. Dirección General de Educación Básica Alternativa, Intercultural Bilingüe y de Servicios Educativos en el Ámbito Rural, 2017)



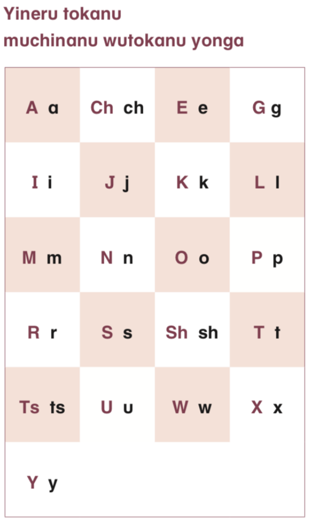
Giyalu ga wa maklewakleto

(a) Original PDF

```
Gi yalu ga  w a   m aklewakleto
```

(b) Converted plain text file

Figure 7: Example of Yine splited word during plain text conversion (Perú. Ministerio de Educación. Dirección General de Educación Básica Alternativa, Intercultural Bilingüe y de Servicios Educativos en el Ámbito Rural, 2018c)



Yineru tokanu
muchinanu wutokanu yonga

A a	Ch ch	E e	G g
I i	J j	K k	L l
M m	N n	O o	P p
R r	S s	Sh sh	T t
Ts ts	U u	W w	X x
Y y			

(a) Original PDF

Yineru tokanu
muchinanu wutokanu yonga

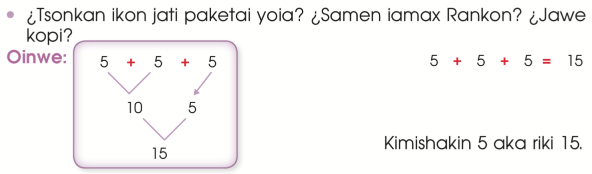
AaChchEeGgIiJjKkLlMmNnOoPpRrSsShshTtTstsUuWwXxYy

(b) Converted plain text file

Figure 8: Example of Yine alphabet automatic conversion from (Perú. Ministerio de Educación. Dirección General de Educación Básica Alternativa, Intercultural Bilingüe y de Servicios Educativos en el Ámbito Rural, 2012)

	ame	pib	cni	shp
Language Manual	608	590	621	641
Elementary School Workbook	473	423	564	424
Kindergarten Workbook	139	117	0	0
Dictionary	0	0	0	0
Tales	53	98	0	0
Community Tradition	0	0	0	495
Teaching Manual	0	0	0	0
Total	1273	1228	1185	1560

Table 10: Composition of the evaluation set in number of sentences by document type and language.



- ¿Tsonkan ikon jati paketai yoia? ¿Samen iamax Rankon? ¿Jawe kopi?

Oinwe:

5 + 5 + 5 = 15

Kimishakin 5 aka riki 15.

(a) Original PDF

```
¿Tsonkan ikon jati paketai yoia? 5  +  5  +  5
¿Samen iamax Rankon? ¿Jawe kopi? Oinwe:

5  +  5  +  5
5   +   5   +   5  =   15

10

5

15
```

(b) Converted plain text file

Figure 9: Example of math operation inside a Shipibo Konibo sentence (Perú. Ministerio de Educación. Dirección General de Educación Básica Alternativa, Intercultural Bilingüe y de Servicios Educativos en el Ámbito Rural, 2018d)

10. Bibliographical References

Aikhenvald, A. (2012). *The Languages of the Amazon*. OUP Oxford.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2018). Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*, April.

Buck, C., Heafield, K., and van Ooyen, B. (2014). N-gram counts and language models from the common crawl. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3579–3584, Reykjavik, Iceland, May. European Languages Resources Association (ELRA).

Daudert, T. and Ahmadi, S. (2019). NUIG at the FinSBD task: Sentence boundary detection for noisy financial PDFs in English and French. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, pages 108–114, Macao, China, 12 August.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Espichán-Linares, A. and Oncevay-Marcos, A. (2017). Language identification with scarce data: A case study from Peru. In *Annual International Symposium on Information Management and Big Data*, pages 90–105. Springer.

Esplà, M., Forcada, M., Ramírez-Sánchez, G., and Hoang, H. (2019). ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland, 19–23 August. European Association for Machine Translation.

Ferrés, D., Saggion, H., Ronzano, F., and Bravo, À. (2018). PDFdigest: an adaptable layout-aware PDF-to-XML textual content extractor for scientific articles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Goodman, M. W., Georgi, R., and Xia, F. (2018). PDF-to-text reanalysis for linguistic data mining. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Languages Resources Association (ELRA).

Hammarström, H., Forkel, R., and Haspelmath, M. (2017). Glottolog 3.0. *Max Planck Institute for the Science of Human History*.

Hammarström, H., Castermans, T., Forkel, R., Verbeek, K., Westenberg, M. A., and Speckmann, B. (2018). Simultaneous visualization of language endangerment and language description. *Language Documentation & Conservation*, 12:359–392.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July. Association for Computational Linguistics.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2018). Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations (ICLR)*.

Mager, M., Gutierrez-Vasques, X., Sierra, G., and Meza-Ruiz, I. (2018). Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Merity, S., Keskar, N. S., and Socher, R. (2018). Regularizing and optimizing LSTM language models. In *International Conference on Learning Representations*.

Mielke, S. J., Cotterell, R., Gorman, K., Roark, B., and Eisner, J. (2019). What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy, July. Association for Computational Linguistics.

Mielke, S. J. (2019). Can you compare perplexity across different segmentations? Available in: `http://sjmielke.com/comparing-perplexities.htm`.

Montoya, H. E. G., Rojas, K. D. R., and Oncevay, A. (2019). A continuous improvement framework of machine translation for Shipibo-konibo. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 17–23, Dublin, Ireland, 20 August. European Association for Machine Translation.

Nagata, R., Sato, T., and Takamura, H. (2018). Exploring the influence of spelling errors on lexical variation measures. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2391–2398, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Prasad, M., Breiner, T., and van Esch, D. (2018). Mining training data for language modeling across the world's languages. In *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-*

resourced Languages (SLTU 2018).

Sabeti, B., Abedi Firouzjaee, H., Janalizadeh Choobbasti, A., Mortazavi Najafabadi, S., and Vaheb, A. (2018). MirasText: An automatically generated text corpus for Persian. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Sagot, B. (2013). Comparing complexity measures.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August. Association for Computational Linguistics.

Karina Natalia Sullón Acosta, et al., editors. (2013). *Documento nacional de lenguas originarias del Perú*. Ministerio de Educación (MINEDU).

Zariquiey, R., Hammarström, H., Arakaki, M., Oncevay, A., Miller, J., García, A., and Ingunza, A. (2019). Obsolescencia lingüística, descripción gramatical y documentación de lenguas en el Perú: hacia un estado del arte. *Lexis (in-press)*, 44(2).

11. Language Resource References

Perú. Ministerio de Educación. Dirección General de Educación Básica Alternativa, Intercultural Bilingüe y de Servicios Educativos en el Ámbito Rural. (2012). *Retjemtanropa kiruka steno girixpoko giynumsatkota ga wa yonatkota yineru tokanu*. Ministerio de Educacioón.

Perú. Ministerio de Educación. Dirección General de Educación Básica Alternativa, Intercultural Bilingüe y de Servicios Educativos en el Ámbito Rural. (2017). *Ateth kowen e'morrtena ñamha kellkëna yanesha' po'ñoñ*. Ministerio de Educacioón.

Perú. Ministerio de Educación. Dirección General de Educación Básica Alternativa, Intercultural Bilingüe y de Servicios Educativos en el Ámbito Rural. (2018a). *Ayoyeteri añaantyari kametsa 1: amempori*. Ministerio de Educación.

Perú. Ministerio de Educación. Dirección General de Educación Básica Alternativa, Intercultural Bilingüe y de Servicios Educativos en el Ámbito Rural. (2018b). *E'ñot 1 : yeñóchepa'ch párro*. Ministerio de Educación.

Perú. Ministerio de Educación. Dirección General de Educación Básica Alternativa, Intercultural Bilingüe y de Servicios Educativos en el Ámbito Rural. (2018c). *Gigkaklu-mta*. Ministerio de Educacioón.

Perú. Ministerio de Educación. Dirección General de Educación Básica Alternativa, Intercultural Bilingüe y de Servicios Educativos en el Ámbito Rural. (2018d). *Toponti Xawe 3*. Ministerio de Educacioón.