Using Open-Source Tools to Digitise Lexical Resources for Low-Resource Languages

Ben Bongalon[1], Joel Ilao[2], Ethel Ong[3],
Rochelle Irene Lucas[4], Melvin Jabar[5]

[1] Independent Researcher, California, USA
[2,3] College of Computer Studies, De La Salle University, Manila, Philippines
[4] English and Applied Linguistics, De La Salle University
[5] Behavioral Sciences, De La Salle University
E-mail: ben@isawika.org, {joel.ilao, ethel.ong, rochelle.lucas, melvin.jabar}@dlsu.edu.ph

Abstract

Advances in open-source lexicography tools have made it more practical to digitise historical dictionaries and lexical resources. However, most retro-digitisation efforts have catered to dominant languages while ethnic minority and indigenous languages tend to be neglected. In countries with a large number of regional and local languages, such as the Philippines, retro-digitisation is a daunting challenge. Of its 186 languages and 500+ dialects, only a few are known to have e-dictionaries produced. The traditional "top-down" approach simply does not scale, since the community need for language documentation far outstrips the number of motivated linguists, lexicographers and funding entities available. This paper describes a complete tool chain and workflow that we used to digitise a Hanunoo-English dictionary originally published in the 1950s (Conklin, 1953). A trainable OCR engine, Tesseract (Smith, 2007), is used to handle the novel glyphs found in the dictionary. Post-edits were performed to fix OCR errors, extract lexical elements from the transcribed pages, and produce an XML-formatted electronic dictionary containing 5,779 entries. The Lexonomy dictionary editor (Měchura, 2017) was used to edit the entries and host the access-controlled electronic dictionary online.

Keywords: indigenous language; retro-digitisation; electronic lexicography; OCR; LSTM

1. Introduction

Starting with the publication of "Samuel Johnson: A Dictionary of the English Language" on CD–ROM in 1996 (Schneiker, 2009; McDermott 1996), a growing number of projects to digitise historical dictionaries have been launched. The reasons for undertaking these projects vary and include: disseminating resources of "great historical value for European lexicographical heritage" (Salgado, 2019), aiding research to trace "the history of the language" and understand "society's situation at the time of the publication" (Özcan, 2018), providing "valuable information on the first attestations of words, on their variants (ranging e.g. from formal to diachronic or diatopic kinds), on the authors who quote them, and on their etymologies" (Sassolini, 2019).

Having a dictionary in one's mother tongue confers many advantages (SIL, 2020) including:

- Validating the use of the vernacular language and boosting the community's self-esteem
- Promoting literacy and serving as a bridge to mainstream languages
- Helping mother-tongue writers record their oral traditions and author new material
- Helping in creating educational resources in the local language
- Facilitating translation of health bulletins, news and other informational materials

Moreover, when dictionaries are digitised and made available online or as mobile applications, they promote cultural identity and a sense of pride, foster language use in youth (who heavily use mobile apps), and encourage learners around the world to interact and use the language which helps in preserving it.

Despite the numerous benefits of having retro-digitised lexical resources, many speakers of minority and indigenous languages today do not have electronic dictionaries and grammar reference books for their own communities to use. Why is this so? We believe the overall cost of retro-digitisation projects in terms of the time, money and skills required are still too high, making them out of reach for marginalised language communities. Without adequate funding and institutional support, these communities often depend on external partners who happen to express interest in their mother tongue to initiate the projects on their behalf.

Creating dictionaries from scratch takes considerable time and resources. Not only is the initial word collection effort expensive, but even the subsequent phase of producing the dictionary typically requires two people working full time for 12 to 18 months (SIL, 2020). This is where historical dictionaries can play a vital role. Many dictionaries for languages of ethnic minority and indigenous groups have been published in the last 100 years. Often it took years to compile them given the language barriers and extreme difficulty in reaching the target communities, who often lived in remote locations. Thus they contain substantial linguistic and cultural knowledge, and while no doubt many words have shifted in meaning or are no longer used by today's native speakers, core vocabularies are surprisingly resilient to semantic shift and can be used to bootstrap or augment modern dictionary-building initiatives when desired by the community. In other words, retro-digitisation enables ethnic minority and indigenous communities to start building e-dictionaries for their language with less risk, cost and effort.

However, retro-digitisation presents a huge challenge for countries with a large number of minority and indigenous languages. The traditional "top-down" approach where language documentation projects typically require multi-year efforts and sizable budgets simply does not scale (i.e., the number of languages to be documented far outstrips the number of motivated linguists, lexicographers and funding entities available). The Philippines makes for a good example. With 186 languages (Eberhard et al., 2021) and 500+ dialects, it is the 25th most linguistically diverse country in the world (World Atlas, 2009), but almost half of these languages are considered endangered (Eberhard et al., 2021), and thus the need to produce more language resources to revitalise them.

In this paper, we describe our project to retro-digitise a historical dictionary developed for the Hanunoo Mangyan language. Hanunoo (IPA: [hanunuʔɔ]) is spoken by one of the eight Mangyan ethnic groups in Mindoro, an island in the southwestern part of the Philippines. Other languages include Alangan, Iraya, Buhid and Tadyawan (Zorc, 1974). It is classified as an Austronesian language, a sub-classification of Malayo-Polynesian, further sub-classified as a Greater Central Philippine language (South Mangyan) (Eberhard et al., 2021; Blust, 1991). There were approximately 25,100 speakers of Hanunoo Mangyan as of 2010 (Eberhard et al., 2021).

2. Related Work

The Hanunoo Mangyan is a unique ethnolinguistic group in the Philippines as it has its own indigenous system of writing, known as the Surat Mangyan. Their system of writing is said to have descended from the ancient Sanskrit alphabet. There are 18 characters in the syllabary, three of which are vowels; the remaining 15 are written in combination with the vowels (Conklin, 1953). However, the writing system is no longer used in the day-to-day encounters of the Hanunoo Mangyan population.

Prior works in documenting the Hanunoo language are found in literature. Studies on the Hanunoo vocabulary (Scannel, 2015) and Hanunoo and English (Conklin, 1953, 1955, 1962) have been conducted and dictionaries produced. Harold Conklin, an American anthropologist who studied the indigenous Hanunoo culture in the Philippines after serving in the US Army during WWII, authored the “Hanunoo-English Vocabulary” (Conklin, 1953) using field notes from his voluntary fieldwork in Mindoro. It is this dictionary that inspired our retro-digitisation project.

Digitising historical dictionaries has been carried out for various languages including English (Johnson, 1996), German (Christmann, 2003), Portuguese (Simões, 2016; Salgado, 2019b), Turkish (Özcan, 2018). Italian (Sassolini, 2019), French (Salgado, 2019b) and Spanish (Salgado, 2019b). Text capture, the process of converting print pages into text, can be grouped into three approaches. For digital-born dictionaries that were printed from LaTex or tagged PDF documents, the embedded markup in the typesetting files was used directly to create XML-formatted e-dictionaries with minimal processing (Simões, 2016; Salgado, 2019b). Some projects, including the Oxford English Dictionary 2nd Edition and the Deutsches Wörterbuch (Christmann, 2003) relied on brute force, employing typists to manually enter the entire text, in some cases double-keyed to achieve higher accuracy. The third and most common approach is to apply OCR technology to transcribe scanned page images to text (Sassolini, 2019).

The Text Encoding Initiative Guidelines (TEI Consortium, 2016) is a *de facto* standard for digitally encoding all types of written texts, ranging from novels and poetry to mathematical formulae or music notation (Salgado, 2019a). Its “Dictionaries” chapter provides guidelines for encoding human-oriented monolingual and multilingual dictionaries, glossaries and similar documents. TEI-Lex0 (Banski, 2017) is a proposed extension to address representational ambiguities in TEI with a stricter set of encoding rules. It has been used to construct the Nxaʔamxcín (Czaykowska-Higgins, 2014), Portuguese, Spanish, and French Academy Dictionaries. Salgado (2019b) proposed further enhancements to TEI-Lex0, most notably in terms of diatextual labels.

3. Materials and Methods

In this section, we discuss how the Hanunoo dictionary was digitised and published for our target audience. We use the workflow stages defined in the DariahTeach's "Digitizing Dictionaries" course (DariahTeach, 2020) to organise our presentation.

Several post-editing tasks were needed to convert the original book into a user-accessible digital resource. In this retro-digitisation project, we trained an OCR engine to recognise special characters used in the Hanunoo dictionary because out-of-the-box OCR engines did not perform well and thus were put aside. Proofreaders were employed to correct residual errors in the OCR output, and to format the content to conform to an XML schema we defined for semantic markup.

3.1 Planning

Planning was simple given that the project is a loose collaboration between the primary author (independent researcher) and faculty members of the De La Salle University's (Philippines) English and Applied Linguistics, Behavioral Science and Computer Technology departments. We aimed to explore innovative ways to leverage mutual interest in developing electronic lexical resources for the Philippines' indigenous languages.

The immediate goal was to produce a high-quality, digitised version of the Conklin dictionary which could serve as: 1) an accessible historical reference of the Hanunoo language, and 2) an auxiliary source of lexical data to augment recent Hanunoo language documentation projects. To make the e-dictionary accessible to our target users, we published it as a web-based application and shared the data for research and community use by providing the XML source. To ensure a high-quality final output, each page would be proofread. While we did not set a formal project schedule, we discussed a soft target of three to six months.

3.2 Image and Text Capture

Because the Conklin dictionary is out-of-print and rare, we sent our copy to a book-scanning service for non-destructive scanning in order to preserve it. We received an image scan of all the pages as a PDF file, as well as an OCR-ed version in Microsoft Word. However the OCR output had too many transcription errors which the company could not correct, so an alternate OCR solution was needed.

In analysing the transcription errors, we found a systematic pattern. Most were due to two special characters used in the Conklin dictionary that stumped out-of-the-box OCR engines: the ŋ (eng) letter and the ʔ glottal stop symbol. They were often mis-transcribed as ‘g’ and question mark ‘?’ characters, respectively. Another set of common errors were the sporadic omission of diacritical marks on vowels. The ŋ and diacritical mark errors were especially problematic, because being both pervasive and subtle, manually correcting them would have been very labour-intensive and so it is desirable to have them accurately transcribed.

To overcome these errors, we searched for OCR engines that can be trained to recognise new symbols. Of the two that we found, Tesseract (Smith, 2007) and OCRopus (Breuel, undated), we chose the former because it supports many more pre-trained language models[1] and is actively maintained. Moreover, starting with version 4, Tesseract employs Deep Learning technology (LSTM neural networks) for more accurate text recognition.

3.2.1 Training the OCR Engine

Training Tesseract began with finding a pre-trained language model that can recognise the most characters present in the source document’s character set. For Conklin’s dictionary, a reasonable assumption would be to use the Tagalog model (tgl.traineddata), since both Tagalog and Hanunoo are Philippine languages. However, our experiment showed that the Spanish model (spa.traineddata) was a better starting point because it recognised diacritical marks in vowels (á, é, í, ó, ú) more accurately than the Tagalog model.

Next, we strategised on how to handle the ŋ and ʔ special characters. The ŋ (eng) symbol, a ligature of the digraph “ng”, is pervasive in some Philippine languages. Thus we wanted the OCR to recognise ŋ accurately to avoid a massive number of post-corrections. On the other hand, question marks ‘?’ were seldomly used in the vocabulary pages so globally replacing them with a glottal stop symbol yielded very few errors which were easily corrected during proofreading. We will revisit the theme of minimising the production cost in the Discussion section. The key point is that by choosing a good starting language model and allowing for a small number of expected transcription errors, we reduced the OCR training task to recognising just one new character (ŋ).

The high-level steps are described in Table 1. We wrote scripts to execute each step as single-line commands. For reference, the scripts and the detailed steps are available on GitHub[2]. To create the training data, we chose 20 sample pages from the scanned dictionary, preferring pages with Hanunoo words containing ŋ in different positions (first, middle, last letter of the words).

[1] For a list of Tesseract language models, see https://github.com/tesseract-ocr/tessdata

[2] Our project repository can be found at https://github.com/isawika/retro-digitization

Step	Notes

1. Prepare the training data.
Split the PDF document into individual pages.
$ pdftk book.pdf burst
Convert the PDF pages to TIFF format.
$ pdf2tiff *.pdf
	Output: page-01.pdf, page-02.pdf, etc.
	We use TIFF image files because both Tesseract and jTessBoxEditor support it.
	Output: page-01.tiff, page-02.tiff, etc.

2. Create a Tesseract box file for each page. $ for i in *.tif; do ../tessbox.sh $i; done;	A box file contains Tesseract's predicted characters in the page. OCR is performed using a pre-trained Spanish language model.
3. Open each page in jTessBoxEditor, then find and correct the OCR errors.	jTessBoxEditor saves the edits in the box file.
4. Convert each box file into a plain text file. $ for i in *.box; do ../box2lines $i; done;	**Output:** page-01.txt, page-02.txt, etc.

5. Create the training text.
Combine the plain text files from Step 4.
$ cat page*.txt > hanunoo.txt
Prune the file and add to the Spanish training data.
$ cat hanunoo.txt >> spa.training_text
	Multiple experiment runs may be needed to determine the appropriate mix of new and original training data. See 3.2.2 for details.

6. Run the Tesseract fine-tuning procedure. $ tesstrain.sh; combine_tessdata; lstmtraining	For brevity, the full commands are not shown. They mimic the commands in the "Fine Tuning" section of the Tesseract tutorial. [3]

Table 1: Steps for fine-tuning the Tesseract OCR engine

3.2.2 Evaluating the models

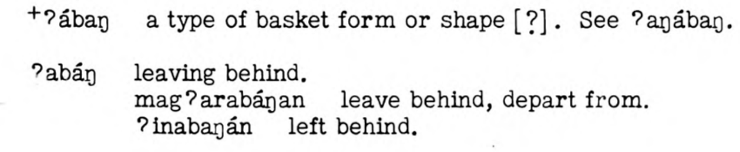
Only a small amount of sample text is needed to fine-tune the OCR engine. For the Conklin dictionary, we found that adding 40 lines of Hanunoo text to the original 68 lines of Spanish training data (*spa.training_text*) yielded the best results. In fact, including more Hanunoo text resulted in more OCR errors. Even more surprising, removing the Spanish text completely and replacing it with Hanunoo text produced a model that performed the worst and generated unknown words ("hallucinations" in Tesseract parlance). In the latter two cases, we believe the resulting neural net models

[3] Training, see https://tesseract-ocr.github.io/tessdoc/tess4/TrainingTesseract-4.00.html

were overfitted to the training data. We ran eight experiments in total, from which we selected the best performing model.

3.2.3 Using the trained model

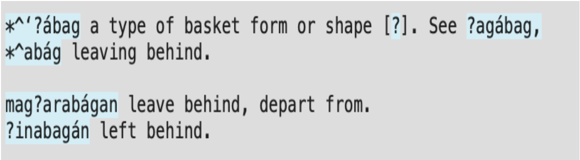
We transcribed the vocabulary pages (N=270) using the best re-trained language model “X3”, then replaced all occurrences of question marks ‘?’ with glottal stop ‘ʔ’ symbols. Figure 2 shows a sample result. All ŋ symbols were recognised. However the glottal stop substitution rule incorrectly replaced the question mark symbol “[ʔ]” in Line 1 (an infrequent error). These need to be fixed in the post-edit step.



```
+ʔábaŋ   a type of basket form or shape [?]. See ʔaŋábaŋ.

ʔabáŋ   leaving behind.
        magʔarabáŋan   leave behind, depart from.
        ʔinabaŋán   left behind.
```

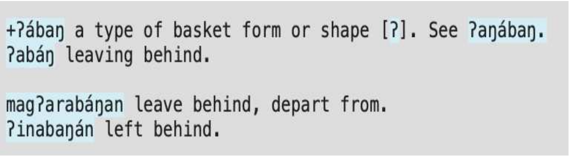
Figure 2a: Source PDF



```
*^‘?ábag a type of basket form or shape [?]. See ?agábag,
*^abág leaving behind.

mag?arabágan leave behind, depart from.
?inabagán left behind.
```

Figure 2b: Transcription before training



```
+ʔábaŋ a type of basket form or shape [ʔ]. See ʔaŋábaŋ.
ʔabáŋ leaving behind.

magʔarabáŋan leave behind, depart from.
ʔinabaŋán left behind.
```

Figure 2c: After OCR training & glottal stop replacement

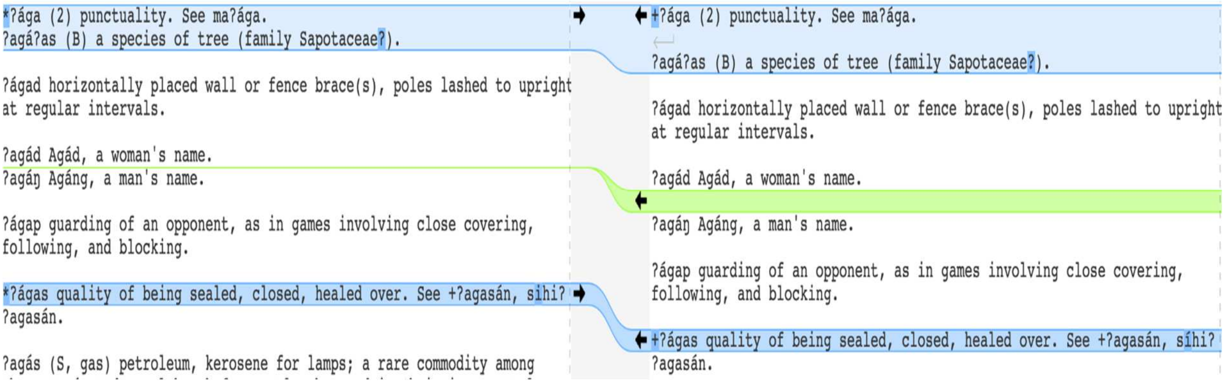
3.2.4 Post-Editing

The transcribed pages needed manual review to correct residual errors. We used UpWork[4] to find freelance proofreaders and had a positive experience. After posting the project for five days, we received 31 bids, screened applicants with a sample task, and hired two freelancers to work in parallel. While we initially planned to hire a third person to provide 2X coverage on 25% of the pages, this proved unnecessary as the quality of the two proofreaders’ work was excellent.

We spot-checked the pages on a MacBook computer using the open-source Meld tool[5], visually comparing the OCR transcript with the proofreaders’ edits and consulting the original PDF page as needed. Figure 3 shows an example output.

[4] see http://upwork.com

[5] see https://meldmerge.org



```
*?ága (2) punctuality. See ma?ága.
?agá?as (B) a species of tree (family Sapotaceae?).

?ágad horizontally placed wall or fence brace(s), poles lashed to upright
at regular intervals.

?agád Agád, a woman's name.
?agáŋ Agáng, a man's name.

?ágap guarding of an opponent, as in games involving close covering,
following, and blocking.

*?ágas quality of being sealed, closed, healed over. See +?agasán, sihi?
?agasán.

?agás (S, gas) petroleum, kerosene for lamps; a rare commodity among
```

```
+?ága (2) punctuality. See ma?ága.

?agá?as (B) a species of tree (family Sapotaceae?).

?ágad horizontally placed wall or fence brace(s), poles lashed to upright
at regular intervals.

?agád Agád, a woman's name.

?agáŋ Agáng, a man's name.

?ágap guarding of an opponent, as in games involving close covering,
following, and blocking.

+?ágas quality of being sealed, closed, healed over. See +?agasán, síhi?
?agasán.
```

Figure 3: Comparison of an OCR output (left) and the proofread page (right). Blue highlights denote modified lines, with the actual changes in dark blue. The green highlight denotes a blank line that was added to separate two dictionary entries.

3.3 Data Modelling and Enrichment

Data modeling and data enrichment were intricately enmeshed in our project and so we discuss them together. First, we analysed the dictionary entries to identify the various semantic elements present to design the encoding schema in Figure 4a. We followed the TEI-Lex0 standard (Banski, 2017) with some deviations for a simpler markup. For example we skipped the use of <form> elements, inserting the <headword> and <pronounce> elements directly under the <entry> node.

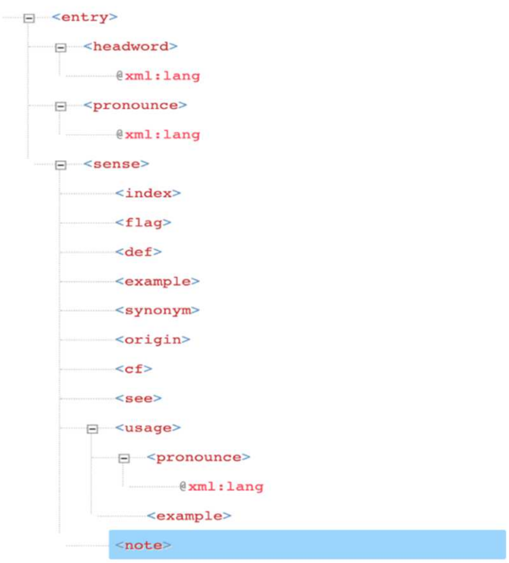
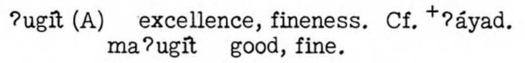
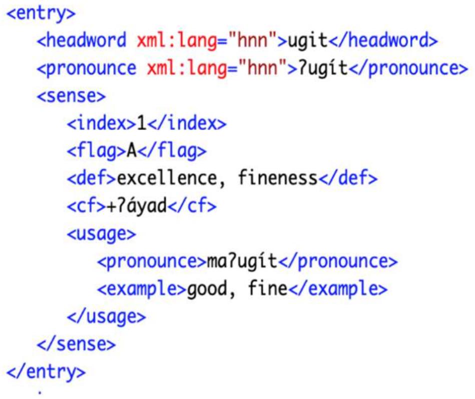


Figure 4a: Schema for Conklin dictionary



?ugı̂t (A) excellence, fineness. Cf. +?áyad.
ma?ugı̂t good, fine.

Figure 4b: An entry in an OCR-ed page



```
<entry>
   <headword xml:lang="hnn">ugit</headword>
   <pronounce xml:lang="hnn">?ugít</pronounce>
   <sense>
      <index>1</index>
      <flag>A</flag>
      <def>excellence, fineness</def>
      <cf>+?áyad</cf>
      <usage>
         <pronounce>ma?ugít</pronounce>
         <example>good, fine</example>
      </usage>
   </sense>
</entry>
```

Figure 4c: Entry formatted in XML

The entries in the Conklin dictionary intermixed references to synonyms, word origins, "c.f." / "see also" terms or other annotations with the definition body (Figure 4b). We wrote a Python script (*conklin2xml.py*) to unpack them into separate XML elements. The textual flow followed a fairly regular pattern, making it easy to define pattern-extraction rules.

To make the dictionary searchable, the script also created two XML elements for each headword. The <headword> field contained a Romanised form of the word with syllable hyphens and glottal stop symbols removed, and with "ŋ" symbols changed to "ng". The <pronounce> field retained the original orthography. For example:

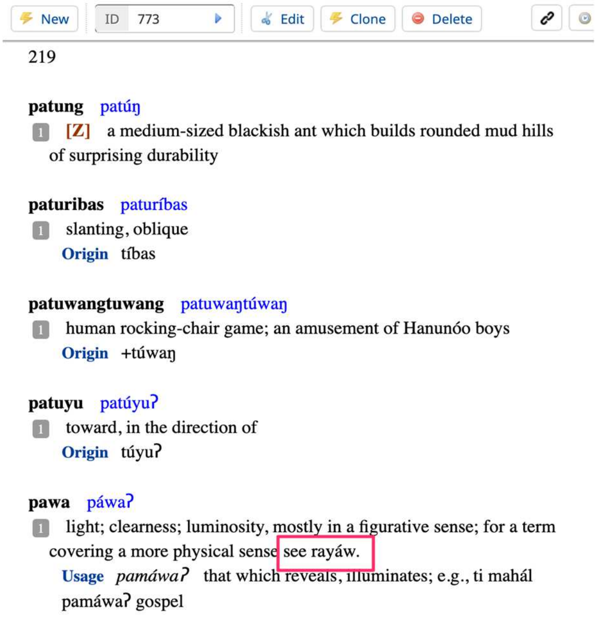
"ʔínaʔ ʔulúŋ" (stepmother) **becomes** <headword>ina ulung</headword>
<pronounce>ʔínaʔ ʔulúŋ</pronounce>

Calling the Python script with the OCR-ed text as input, as shown below, will produce a fully-formatted XML document (Figure 4c):

```
$ conklin2xml.py page021-ocr.txt > page021.xml
```

As in the OCR text capture, the output XML documents contained errors that needed manual correction. In addition, post-edits were needed to undo several "typographical and editorial conventions of the print medium" (Tasovac, 2010), specifically to merge lexical entries that spanned across two pages and to dehyphenate words that wrapped at the end of a text line.

We hired a third freelancer to perform the post-edits, a task that took 14 days to complete. To simplify the editing task, we loaded a specially formatted version of the XML documents into a self-hosted Lexonomy dictionary editing application (Měchura, 2017), where each "entry" embodied a page's worth of lexical entries. This "page view" format significantly aided proofreading because it was easier to visually compare a virtual page against its original PDF source (Figures 5a and 5b) and make corrections to the virtual page (Figure 5c).



219

patung patúŋ
1 [Z] a medium-sized blackish ant which builds rounded mud hills of surprising durability

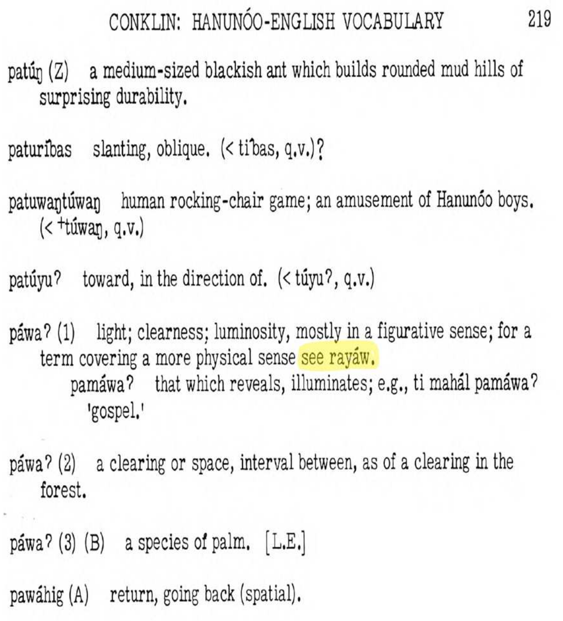
paturibas paturíbas
1 slanting, oblique
Origin tíbas

patuwangtuwang patuwaŋtúwaŋ
1 human rocking-chair game; an amusement of Hanunóo boys
Origin +túwaŋ

patuyu patúyuʔ
1 toward, in the direction of
Origin túyuʔ

pawa páwaʔ
1 light; clearness; luminosity, mostly in a figurative sense; for a term covering a more physical sense see rayáw.
Usage *pamáwaʔ* that which reveals, illuminates; e.g., ti mahál pamáwaʔ gospel

Fig 5a: "Page view" has an entry with dangling text caused by a run-on sentence


CONKLIN: HANUNÓO-ENGLISH VOCABULARY 219

patúŋ (Z) a medium-sized blackish ant which builds rounded mud hills of surprising durability.

paturíbas slanting, oblique. (< tíbas, q.v.)?

patuwaŋtúwaŋ human rocking-chair game; an amusement of Hanunóo boys. (< +túwaŋ, q.v.)

patúyu? toward, in the direction of. (< túyu?, q.v.)

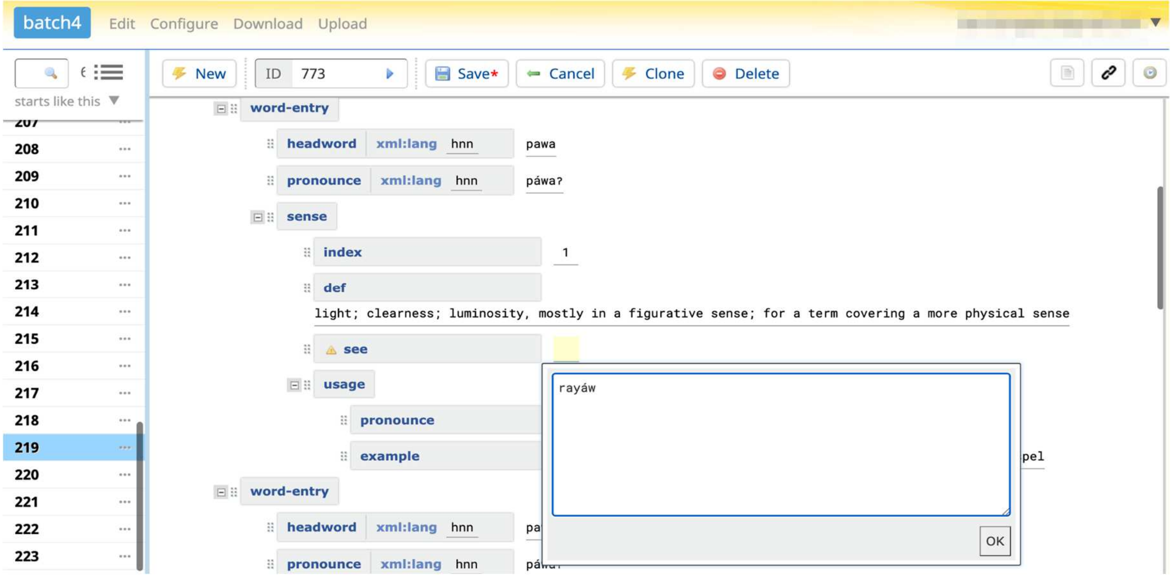
páwa? (1) light; clearness; luminosity, mostly in a figurative sense; for a term covering a more physical sense see rayáw.
pamáwa? that which reveals, illuminates; e.g., ti mahál pamáwa? 'gospel.'

páwa? (2) a clearing or space, interval between, as of a clearing in the forest.

páwa? (3) (B) a species of palm. [L.E.]

pawáhig (A) return, going back (spatial).

Figure 5b: Source PDF (error highlighted)



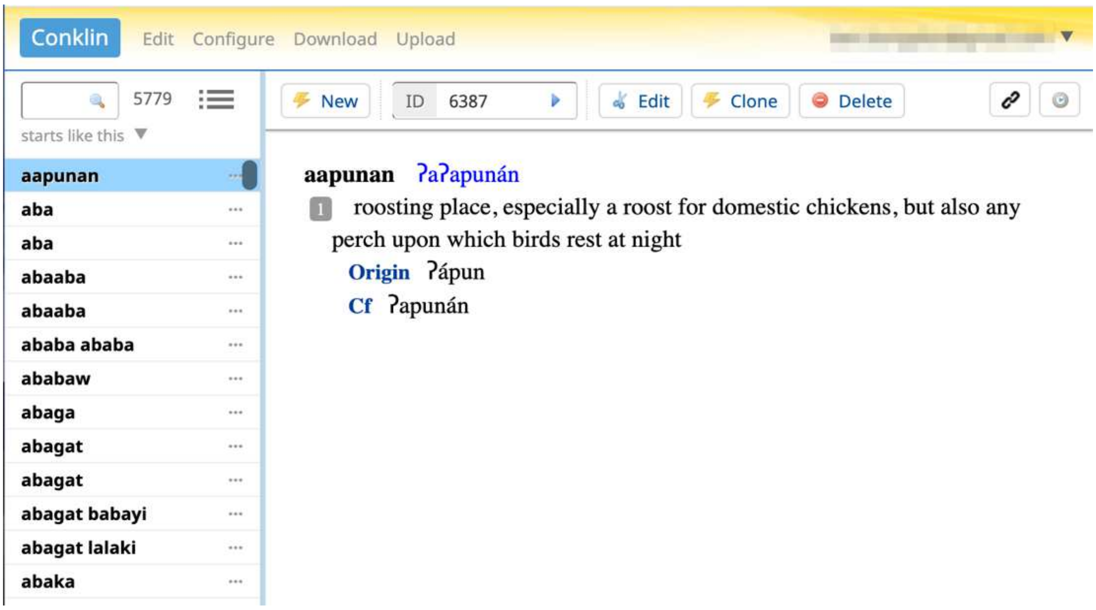
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223

word-entry
headword xml:lang hnn pawa
pronounce xml:lang hnn páwaʔ
sense
index 1
def
light; clearness; luminosity, mostly in a figurative sense; for a term covering a more physical sense
see
usage
pronounce
example
rayáw
word-entry
headword xml:lang hnn

Figure 5c: Entry is fixed by splitting the reference into a "see" XML element

3.4 Publishing

After the data enrichment edits were completed, the XML documents were downloaded from the Lexonomy[6] platform. The documents were reformatted to detach the individual dictionary entries from the page frames and were re-uploaded. The resulting e-dictionary contains a total of 5,779 headwords, as shown in Figure 6.



aapunan
aba
aba
abaaba
abaaba
ababa ababa
ababaw
abaga
abagat
abagat
abagat babayi
abagat lalaki
abaka

aapunan ʔaʔapunán
1 roosting place, especially a roost for domestic chickens, but also any perch upon which birds rest at night
Origin ʔápun
Cf ʔapunán

Figure 6: Format of the dictionary after the entries were detached from the page frames

The Hanunoo-English dictionary is online and access-controlled with individual permissions granted in consultation with representatives of the Mangyan community. The same Lexonomy platform used for editing is used to publish the e-dictionary.

[6] see https://www.lexonomy.eu

4. Discussion

Despite its introduction over 20 years ago, retro-digitisation technology is still immature. While numerous projects have documented their workflow and tools to share knowledge, there are no clear guidelines to help lexicographers figure out which solution is best for their needs. Often they must find out by trial and error. This is an inconvenience for larger and better-funded organisations but a barrier for the resource-constrained, many of whom represent or support ethnic minority and indigenous communities. We thus aimed to help address this issue by introducing a complete digitisation workflow that leverages open-source tools to eliminate or significantly reduce software expenses, and by sharing techniques that contribute to best practices for digitising lexical resources.

In implementing our project, we observed some limitations in the tools we used:

- The Tesseract training program (tesstrain.sh) randomly shuffles the input training data which unpredictably varies the performance of the trained model. To compensate, we ran experiments multiple times to obtain the best model for a given training setup.
- Lexonomy does not support limiting user access to a subset of a dictionary. To prevent proofreaders from accidentally overwriting others' work, we created separate dictionaries containing only the entries each one was responsible for.
- Lexonomy has no built-in support for the "page view" editing as described in Section 3.3. We jerry-rigged it by temporarily reformatting the XML document.
- There appears to be a lack of data interoperability among lexicography tools from different providers. For example, an organisation that wants to use SIL's Dictionary App Builder[7] to create a mobile version of their Lexonomy e-dictionary would first need to build a custom translator.

We admit that the workflow we propose still includes steps that may be challenging and intimidating to less technical users. Training the Tesseract OCR remains to be an art and needs to be simplified. Similarly converting the OCR-ed dictionary pages into XML documents requires someone skilled in writing Python scripts. For the latter, tools such as GROBID-dictionaries (Khemakhem, 2017) which allow users to specify the transformation rules by giving examples can enable laypeople to do the task.

There are also aspects of our method that require further exploration. While our solution worked well for digitising the Hanunoo-English dictionary, we do not know how generalisable it is. Questions include: *How likely will other projects be able to find a good OCR language model as a starting point? How does the number of unknown characters in the source's alphabet affect training complexity? What conditions make it possible to achieve high recognition accuracy on mixed-language text with a single*

[7] See https://software.sil.org/dictionaryappbuilder

language model? In our case, we obtained surprisingly excellent transcription quality for both Hanunoo and English text from a language model that we did not train with English text included.

Digitising the Hanunoo-English dictionary presented some ethical concerns. While the dictionary itself became public domain when its US copyright expired, the vocabulary it contains is considered property of the Mangyan people. Therefore publishing it online requires their Free, Prior and Informed Consent (FPIC) as mandated in the Philippine Indigenous Peoples' Rights Act of 1997 (IPRA, 1997) because "the copyright to their indigenous language has no expiration" (private communication). There is also the question of whether our team is guilty of treating "language as data" (Bird, 2020). In this regard, Bird seems to level criticism against researchers who employ "zero resource" techniques that automatically "discover the language" from audio recordings or transcriptions without further input from linguists, speakers or previously developed language resources. Our project takes a completely opposite approach, reusing and repurposing linguistic knowledge that Conklin and several members of the Hanunoo tribe meticulously documented 70 years ago. However, due to these concerns we took the measured approach of making the e-dictionary available only to the Mangyan community and for limited research. While the Mangyan people are reluctant to publicly share their vocabulary online for fear of cultural misappropriation, they supported and participated in building the vocabulary for an earlier e-dictionary project initiated by the De La Salle University research team (Uy, 2020). In that project the community acknowledged the importance of digitising their language for preservation purposes, affirming their openness to change.

5. Conclusion and Further Work

We presented a tool chain and detailed workflow for digitising a historical dictionary which required the use of a trainable OCR engine to recognise special characters. While the technique was successfully demonstrated in one dictionary, we believe it is applicable to other similar projects. In designing the workflow, we aimed to lower the bar to retro-digitisation in order to encourage more paper dictionaries for other languages to be digitised. We also hope to give minority and indigenous communities an easier way to build and shape their own language resources so help them become more active participants in the digital age.

We plan to host the Hanunoo e-dictionary online indefinitely given the modest cost of hosting (US$800 to $2,500 per year). We will seek volunteers and explore support options for maintaining the dictionary content and the website. Our group intends to expand the research to the other Mangyan languages, namely Buhid, Tawbuwid, Alangan, Iraya and Tadyawan, and possibly to other Philippine indigenous languages.

In doing so, we anticipate some challenges ahead. First, data availability is a concern because there may be fewer printed lexical materials and native speakers available to build a dictionary for the other indigenous languages. Related to this is the issue of combining digital resources for the same language. Various sources are likely to differ in levels of organisation, from unstructured (narratives, poems) to structured (dictionaries), and some materials may even incorporate the orthographies of the Mangyan indigenous writing scripts. These informational mismatches must be reconciled, with a suitable XML dictionary schema developed, so that content can be merged. Third, maintaining and growing the e-dictionaries will require more robust data management processes to enable faster, distributed content creation without sacrificing data quality. As an example, we would like to harness crowdsourcing to build dictionaries more rapidly but with appropriate submissions screening and review processes in place. Another issue is that when working with indigenous groups, securing the appropriate ethical approvals for research takes time and this can significantly delay or curtail the data gathering process. Finally, funding grants for language documentation is difficult in the Philippines given the limited government support for such research endeavours. Despite these challenges, we remain determined to pursue these projects and leverage the open-source, retro-digitisation solution we developed.

6. Acknowledgements

The research team wishes to express their gratitude to the following agencies: Department of Science and Technology (DOST) - National Research Council of the Philippines (NRCP), and Summer Institute of Linguistics (SIL) Manila, Philippines.

7. References

Banski, P., Bowers, J., & Erjavec, T. (2017). *TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms.* HAL Archives.

Bird, S. (2020). Decolonising Speech and Language Technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3504-3519. Barcelona, Spain.

Blust, R. (1991). The Greater Central Philippines Hypothesis. *Oceanic Linguistics*, 30(2), pp. 73-129. University of Hawai'i Press. https://doi.org/10.2307/3623084. Available at: https://www.jstor.org/stable/3623084 (22 March 2021)

Breuel, T. The OCRopus Open Source OCR System. Accessed at: https://github.com/ocropus/ocropus.github.io (22 March 2021)

Christmann, R. & Schares, T. (2003). Towards the User: The Digital Edition of the Deutsche Wörterbuch by Jacob and Wilhelm Grimm. *Literary and Linguistic Computing*, 18(1), pp. 11–22. https://doi.org/10.1093/llc/18.1.11

Conklin, H. (1953). *Hanunóo-English Vocabulary.* Berkeley: University of California Press.

Czaykowska-Higgins (2014). Using TEI for an Endangered Language Lexical Resource: The Nxaʔamxcín Database-Dictionary Project. Available at: https://scholarspace.manoa.hawaii.edu/bitstream/10125/4604/8/czaykowska.pdf

DariahTeach (2017). *Digitizing Dictionaries* course. Accessed at: https://teach.dariah.eu/mod/page/view.php?id=343 (22 March 2021)

Eberhard, D., Simons, G. & Fennig, C. (2021). *Ethnologue: Languages of the World. Twenty-fourth edition.* Dallas, Texas: SIL International. Available at: http://www.ethnologue.com

Harrison, K.D., Lillehaugen, B.D., Fahringer, J., & Lopez, F.H. (2019). Zapotec Language Activism and Talking Dictionaries. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & S. & C. Tiberius (eds.) *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2019 Conference*, pp. 31-50. Brno: Lexical Computing CZ, s.r.o. Available at: https://works.swarthmore.edu/fac-linguistics/252

IPRA (1997). "The Indigenous Peoples' Rights Act of 1997". Republic Act No. 8371. Philippine Official Gazette. October 29, 1997. Available at: https://www.officialgazette.gov.ph/1997/10/29/republic-act-no-8371/ (5 April 2021)

Jabar, M., Lucas, R., Collado, Z., & Regadio, C. (2019). An Ethnolinguistic Vitality Study of the Hanunoo Mangyan Language. *Terminal Report*, De La Salle University, Philippines.

Johnson, S. & McDermott, A. (1996). *A Dictionary of the English Language on CD–ROM.* Cambridge, England and NY, USA: Cambridge University Press.

Khemakhem, M., Foppiano, L., & Romary, L. (2017). Automatic Extraction of TEI Structures in Digitized Lexical Resources Using Conditional Random Fields. In I. Kosem et al. (eds.) *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference.* Leiden, Netherlands. Hal-01508868v2

Měchura, M. (2017). Introducing Lexonomy: An Open-source Dictionary Writing and Publishing System. In I. Kosem et al. (eds.) *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference.*

OED (2019). Digitizing the OED: the making of the Second Edition. OED Blog, 15 January 2019. Available at: https://public.oed.com/blog/digitizing-the-oed-the-making-of-the-second-edition/ (5 April 2021)

Özcan, E. (2018). Retro-digitizing Turkish Dictionaries Using GROBID-dictionaries. *Lexical Data Masterclass Symposium.* Berlin: Germany. (HAL-01969337)

Postma, A. (1986). *Primer to Mangyan Script (1st ed.).* Oriental Mindoro, Philippines: Mangyan Research Center.

Postma, A. (2002). *Primer to Mangyan Script (1st Rev. ed.).* Oriental Mindoro, Philippines: Mangyan Heritage Center.

Postma, A. (2013). *Primer to Mangyan script (2nd Rev. ed).* Oriental Mindoro, Philippines: Mangyan Heritage Center.

Salgado, A., Costa, R., & Tasovac, T. (2019a). Improving the Consistency of Usage Labelling in Dictionaries with TEI Lex-0. *Lexicography ASIALEX* 6, pp. 133–156. https://doi.org/10.1007/s40607-019-00061-x

Salgado, A., Costa, R., Tasovac, T., & Simões, A. (2019b). TEI Lex-0 In Action: Improving the Encoding of the Dictionary of the Academia das Ciências de Lisboa. In Kosem, I., Zingano Kuhn, T., Correia, M., Ferreria, J. P., Jansen, M., Pereira, I., Kallas, J., Jakubíček, M., Krek, S. & Tiberius, C. (eds.) *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2019 Conference.*

Sassolini, E., Khan, A.F., Biffi, M., Monachini, M., & Montemagni, S. (2019). Converting and Structuring a Digital Historical Dictionary of Italian: A Case Study. In In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & S. & C. Tiberius (eds.) *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2019 Conference.*

Schneiker, C., Seipel, D., & Wegstein, W. (2009). Schema and Variation: Digitizing Printed Dictionaries. In *Proceedings of the ACL-IJCNLP 2009 Third Linguistic Annotation Workshop (LAW 2009)*, pp. 82-89.

Schreibman, S., Agiatis, B., Clivaz, C., Ďurčo, M., Huang, M., Papaki, E., Scagliola, S., Tasovac, T. & Wissik, T. (2016). #dariahTeach: online teaching, MOOCs and beyond. *Digital Humanities 2016: Conference Abstracts.*

SIL. (2020-21). *Dictionary-Making and Lexicography Course.* Accessed at: https://sites.google.com/sil.org/dls-course (27 March 2021)

Simões, A., Almeida, J.J., & Salgado, A. (2016). Building a Dictionary using XML Technology. In *Proceedings of the 5th Symposium on Languages, Applications and Technologies* (SLATE'16). https://doi.org/10.4230/OASIcs.SLATE.2016.14

Smith, R. (2007). An Overview of the Tesseract OCR Engine. In *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR 2007)*, pp. 629-633. Curitiba, Brazil. https://doi.org/10.1109/ICDAR.2007.4376991.

Tasovac, T. (2010). Reimagining the dictionary, or why lexicography needs digital humanities. *Digital Humanities*, pp. 254–256. Center for Computing in the Humanities, Kings College London.

TEI Consortium, eds. (2016). TEI P5: Guidelines for Electronic Text Encoding and Interchange. TEI Consortium. Available at: http://www.tei-c.org/Guidelines/P5/ (13 February 2017)

Uy, D. (2020). Hanunoo Mangyan Project: Saving Languages through Technology, *The La Sallian.* Accessed at: https://thelasallian.com/2020/03/10/50467/ (22 March 2021)

World Atlas. (2009). Countries where the most languages are spoken. Accessed at: https://www.worldatlas.com/articles/the-most-linguistically-diverse-countries-in-the-world.html (4 April 2021)