(Re-)Digitizing 吳守禮 Ngôo Siú-lé's Mandarin – Taiwanese Dictionary

Pierre Magistry
INALCO, ERTIM
pierre.magistry@inalco.fr

Afala Phaxay
INALCO, ERTIM

Abstract

This paper presents the efforts conducted to obtain a usable and open digital version in XML-TEI of one of the major lexicographic work for Taiwanese bilingual dictionaries, namely the 《國臺對照活用辭典》 *Practical Mandarin-Taiwanese Dictionary*, (吳 , 2000). The original dictionary was published in 2000, after decades of work by Prof. 吳守禮 (Ngôo Siu-le/Wu Shouli).
We publish the resulting TEI files on Zenodo[1]

1 Introduction

This paper describes the efforts, the issues and some proposed solutions to conduct the re-digitization of the two volumes (over 2800 pages) of the 《國臺對照活用辭典》 *Practical Mandarin-Taiwanese Dictionary*, a masterpiece of Taiwanese lexicography authored by Prof. 吳守禮 (Ngôo Siu-le/Wu Shouli) in the last decades of the 20th century and published in 2000.

The author started its work before the Unicode was founded, at a time when commercial sinographic word processing softwares were designed mostly for Mandarin. He had to come up with workaround solutions to print sinograms and phonetic symbols specific to Taiwanese, involving the creation of thousands of glyphs.

As a result, our work was not strictly speaking a digitizing project, since the original document was already digital-born. But we had to face a number of challenges to turn the original files from their obscure proprietary format into a more modern, open and standard format (we choose XML-TEI) in order to make it easily usable in future projects.

1.1 Context

This project begun after 吳 Ngôo's family members, as right holders of the dictionary, decided to release the work of their ancestor under a permissive and open license. They were willing to provide a larger access to this work and ensure the continuity of this legacy. Their first choice was to turn to the Wikimedia Foundation and target WikiSource to host a public version of the dictionary. Volunteers from Wikimedia Taiwan worked out all the legal aspects of this project and enabled Ngôo's family to release the original data under a Creative Commons license, allowing us to conduct our work. Unfortunately, due to typographic and other technical difficulty which will be described below, the conversion to mediawiki was all but straightforward. Wikimedians turned to the g0v community (involving one author of this paper) for technical support and we finally took the decision to first convert the document into a more standard XML-TEI file so it would be easier to work on it and later to provide all sorts of browsing interfaces. This work was also slowed down by inevitable limitation of time available to volunteers from Wikimedia and g0v, and is now being conducted in a more academic environment. We hope it can return to the public and NGO sphere once we achieve a good level of felicity to the original author's work.

1.2 The author and the Dictionary

Prof. 吳 Ngôo (1909 – 2005) was born in Tainan and received a primary education in Taiwanese. He later graduated from the Taihoku Imperial University, continued his research and work as a translator and lexicographer in Japan where he learned Mandarin before this language was brought to Taiwan by the Kuomintang. He came back to Taiwan to conduct research and teach at the Taihoku Imperial University which later became National Taiwan University (NTU) where he became professor and dedicated is research to the study of Taiwanese.
The 《國臺對照活用辭典》 dictionary is only one of his numerous publications. He finished it

[1] https://doi.org/10.5281/zenodo.1308746)

The 2nd International Workshop on Natural Language Processing for Digital Humanities (NLP4DH), pages 174–178
November 20, 2022.

after he retired from NTU. It was published in 2000 and the next year it obtained the *presidential price for culture* 總統文化獎 .

2 Structure of the dictionary

This massive piece of work covers over 12,000 sinograms and 40,000 lexical entries, providing information including traditional phonology, syntax, meaning (bilingual) and semantic relations. In this section we describe how the dictionary was organized.

2.1 macrostructure

The macrostructure of the dictionary presents two layers of information, which is typical of sinitic languages dictionaries. The first layer focuses on Mandarin syllables, ordered according to the canonical order of the 注音符號 (*zhuyin fuhao*) transcription.

Under each syllable section, one can find the list of corresponding sinograms, ordered following the number of strokes in the sinograms. Under each sinogram entry, lexical entries (words) which include this sinogram, ordered by the number of sinograms in the word and following the order of *zhuyin fuhao* for words of the same length.

2.2 microstructure

Each sinogram entry comes with a description of its various readings. It includes the traditional phonological description (the 反切 *fanqie*), possible readings in Mandarin and possible readings from the *bân-lâm* group, drawing from many sourcing, notably William Campbell's dictionary (Campbell, 1913) for Taiwanese and the *Pumin* dictionary representative of Xiamen (Amoy) readings (University, 1982). These descriptions include the customary distinctions between literary and popular readings (a traditional distinction in ban-lam studies, 文白異讀).

Word entries represent actual lexical items. They can be numbered in cases of homographs. The author provides the part of speech, definition in Mandarin and translation or translation of the term and/or the definition in Taiwanese. Interestingly, all the text written in Taiwanese comes with phonetic transcription in the form of *zhuyin fuhao*, on the side of each character (just like the *furigana* in Japanese). Possible regional variations in pronunciation or orthography are indicated with a slash / character. See Figure 1 for an example.

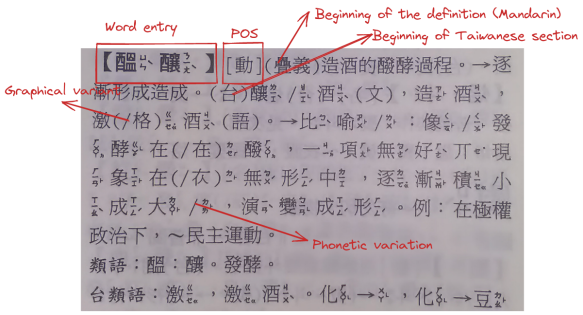

Figure 1: example of a word entry, with some annotation of the types of information provided

3 Digitization stages

The starting point was a set of floppy disks. It required a succession of steps involving various tools and strategies to obtain a usable dataset out of it.

One of the main issues we had to face was to understand how Prof. Ngô managed to include thousands of glyphs which were specific to Taiwanese texts and absent from the encodings of that time.[2]

3.1 Recoding

The first step was to figure out how to read the files, especially in terms of encoding. Prof. Ngô started his work on his dictionary before the Unicode even existed. Knowing this work had been done on a word-processing software for traditional Chinese in Taiwan, Big5 encoding was a safe guess. However, Big5 was only a *de facto* standard, with various vendors extensions, and the use of the Private Use Area (PUA) for new glyphs may differ from one vendor to another.

After a few trials, we wrote a small Perl script to perform the conversion from CP950 (Microsoft version of Big5-ETEN) to UTF-8, and at the same time converting Big5 PUA codes into Unicode PUA, keeping a simple mapping between the two so we can later go back to the original value.[3]

3.2 Reverse engineering the format

Once the text was in an easily readable UTF-8 encoding, we could investigate how to parse the syntax of the text formatting software. The file

[2] with the noticeable exception of the CCCII encoding, but it seems that text processing softwares of that time, being more focused on Mandarin, did not support it.

[3] The script can be read here, kudos to Audrey Tang https://github.com/g0v/koktai/blob/master/a-tsioh_sandbox/recode_utf8.pl

format relies on a set of control expressions written in plain text. We were able to guess the most important ones by comparing the content of the file to the printed version of the book.

One important command to find out was the switch between different fonts. The author had to add so many new glyphs that the PUA space available on Big5 was too small. To be able to print all the desired characters, he had to fill it twice, with two different font. As a result, some original Big5 PUA codes are ambiguous and we need to know the font intended for rendering to know which character it corresponds to.

To ease the parsing and XML output, we wrote this part in Scala, which provides convenient parser combinators[4] to describe the grammar of the file format and can natively and elegantly deal with XML as it is part of the language basic syntax.

3.3 Characters in the Private Use Area

At this stage we could face the issue of the thousands of characters encoded in the PUA.

We had the list of codes, associated with a font name and were also provided the font files from which we could extract the images of each glyph. This left us with thousands of images to map to their code.

There were two main cases for these glyphs. Some represented syllable transcriptions in *zhuyin fuhao*. These glyphs are actually the combination of smaller characters transcribing subcomponents of the syllable (typically initial and final, sometime a simple phoneme). We choose to transcribe these glyphs into their decomposed form, which is easy to type and process.

The other type of glyph is actual sinograms, which are specific to Taiwanese or too rare in Mandarin to have been included in Big5 encodings. For some of them, they were later included in Unicode, in which case we could recode them into the new non PUA Unicode code point. Some other were still missing and we could only provide the original image and the description of its composition as a Unicode Ideographic Description Sequence IDS.

3.4 Crowdsourcing the mapping

In order to obtain the mappings described in the previous subsection, we relied on the strong online community of g0v followers to launch a crowdsourcing campaign during a hackathon. The description of the syllable glyphs could simply be keyed in by Taiwanese netizen familiar with *zhuyin fuhao*. On the other hand, the missing sinograms were mostly obscure and often unknown characters for nonspecialist of Taiwanese script (Mandarin being the language of education in Taiwan).

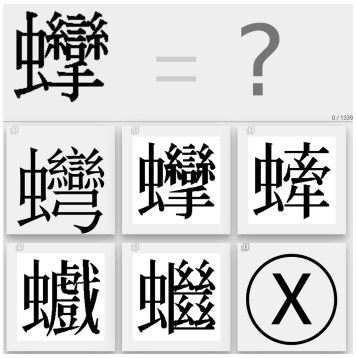

Figure 2: screenshot of the crowdsourcing interface. A simple click or tap on the match will take the user to another character

Figure 3: glyph for a Taiwanese syllable (typically romanized as *bih*)

To make it easier and faster for the netizen to help us, we pre-processed the data with an image similarity algorithm based on Python's OpenCV library. By doing so, we could provide a simple online interface presenting a targeted glyph and a few similar glyphs (extracted from a Unicode font of similar style). In that way, participants in the crowdsourcing just had to click on corresponding characters when present.

Same glyph was presented multiple times (at least 3 participants had to agree on the mapping), and we proceed in multiple waves, to remove already mapped characters and focus on missing data. We counted over $13,000$ user sessions in about 30h.

3.5 OCR for missing pages

Once we obtained a first usable version of this dataset and started working on a web interface, we noticed some discrepancies between our data and the printed version. We soon realized that missing entries were a result from corrupted files on one of the original floppy disks. Some hundred pages

[4] https://index.scala-lang.org/scala/scala-parser-combinators

of entries were therefore not available. In order to recover those data, we decided to use Optical Character Recognition (OCR) on the scanned version of the missing pages. Granted we had enough data in our hands to train a model. It would have been way too time-consuming and tedious to retrieve them manually rather than limiting the manual work to post-correction of the OCR.

From the numerous OCR tools available, we chose eScriptorium [5], which claim to "focus on pre-typographical and/or non-alphabetic cultures." It is based on Kraken [6] which is "optimized for historical and non-Latin script material." Our situation was a good test case to see how it would fare against sinograms.

eScriptorium's web interface was not too difficult to handle thanks to its tutorial. With eScriptorium, we had two major steps : document segmentation and characters' recognition. All the information is saved into a PAGE [7] schema file. At first, we directly performed the models' training on the original scanned page of the dictionary. The document segmentation gave quite good results with errors such as recognizing one long line of text instead of two lines from two separate columns or not recognizing the entirety of the vertical *zhuyin fuhao* as part of the text's region. On the other hand, the character recognition step would be the difficult part, especially with the specific typography of this dictionary, we would have to train a model from scratch. But the XML we already built could serve as the basis to create training data.

This led us to not only generate our own images of lines from the TEI output of the existing data, with corresponding PAGE files. Our images consist of one single line per images of similar length and police size compared to what was in the original data, however, the police and size used would randomly be chosen among several ones to ensure a more robust model. Sadly, eScriptorium could not correctly recognize and segment our generated images, perhaps it was because the model was trained on the original pages which structure was not the same as ours : a one line image. When the police size was smaller, we didn't come upon this issue.

The result on generated data reached over 95% of accuracy, But it did not perform as well on actual pages from the scanned version. At this point the model could rarely recognize one line entirely. We had to add some original pages during the model's training to introduce the structure of what we wanted the model to be used on, so that it would not be encountering it for the first time. That allowed the new model to now recognize several lines entirely.

[5] https://escriptorium.fr/

[6] https://kraken.re/master/index.html

[7] https://github.com/PRImA-Research-Lab/PAGE-XML/blob/master/documentation/XML%20File%20Structure.pdf

4 TEI output

To distribute the dataset in a format as standard as possible, we do our best to follow the TEI guidelines with specific sections for dictionaries. We use **entryFree** elements to describe sinograms and **entry** elements for words. Special characters mapped from the original PUA are described inside the **teiHeader** element as many **charDecl**.

For the moment, we count $6,900$ **charDecl**, $11,805$ **entryFree** and $43,926$ **entry** elements. These figures do not include data from the OCR which is still a work in progress. The exact XML schema is also subject to evolution as we plan to deeper analyze the content of the entries, so we invite the readers to refer to the current version released on the Zenodo repository.

https://doi.org/10.5281/zenodo.1308746

5 Conclusion and Future Work

We hope to achieve a high degree of felicity to the original work by Prof. 吳 Ngôo. We will then be able to provide easier and broader access to this valuable material, with full-text search and reverse index. Our experiment with eScriptorium also encourages us to address other volumes, such as the handwritten Taiwanese Dictionary published in 1986 by the same author(吳 , 1986).

Acknowledgments

Part of this work was funded by a *Taiwan Fellowship* grant from Taiwan's MOFA. We express our gratitude to prof. 吳 Ngôo's family who initiated this work by sharing the data openly. We also thanks all the participants from Wikimedia Taiwan and g0v at the initial hackathons, especially Chou Soichi and Audrey Tang for their contributions on legal and technical aspects of the work, as well as the numerous yet anonymous netizens who kindly contributed to the crowdsourcing.

References

William Campbell. 1913. *A Dictionary of the Amoy Vernacular spoken through out the prefectures of Chin-chiu, Chang-Chiu and Formosa*. Fukuin Printing Co., Yokohama.

Xiamen University. 1982. 普通話閩南方言詞典 . 福建人民出版社 *Fujian People's Publishing House*, Xiamen.

守禮 吳 . 1986. 綜合閩南臺灣語基本字典初稿 . 文史哲出版社 *The liberal arts press co., ltd.*, Taipei.

守禮 吳 . 2000. 國臺對照活用辭典 , *Practical Mandarin-Taiwanese Dictionary*. 遠流出版事業股份有限公司 Yuan-Liou Publishing Co., Ltd., Taipei.