Journal of Open Humanities Data

Printed Text Recognition for Lexical Lists in Chinese-International Phonetic Alphabet (IPA) Glossing

DATA PAPER

SHIHUA LI
NATHAN HILL
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

This study presents a dataset serving as a benchmark for the recognition of printed text in lexical lists using Chinese-IPA glossing. The paper provides an overview of the baseline model, transcription model, and PyLaia engines employed in the research. Furthermore, it elucidates the specific need for digitizing the aforementioned lexical lists, outlines the methodology employed for training the baseline model for layout analysis, and describes the training process of the transcription model using the ground truth data generated on Transkribus. This comprehensive approach encompasses both the images of the lexical list content and their corresponding transcriptions as input. Additionally, the study highlights the limitations of the model and identifies avenues for future development. By making this dataset openly accessible, it can be utilized by researchers seeking to digitize lexical lists using Chinese-IPA glossing. Moreover, since the model can recognize both Chinese characters and IPA symbols, it has the potential to contribute to linguistic analysis of languages documented in Chinese-IPA glossing.

CORRESPONDING AUTHOR:
Shihua Li
Trinity Centre for Asian Studies, Trinity College Dublin, Dublin, Ireland
sli7@tcd.ie

KEYWORDS:
printed text recognition; Chinese; IPA; Burmish and Tujia languages; lexical lists; baseline model; transcription model; Transkribus

TO CITE THIS ARTICLE:
Li, S., & Hill, N. (2023). Printed Text Recognition for Lexical Lists in Chinese-International Phonetic Alphabet (IPA) Glossing. *Journal of Open Humanities Data*, 9: 15, pp. 1–8. DOI: https://doi.org/10.5334/johd.119

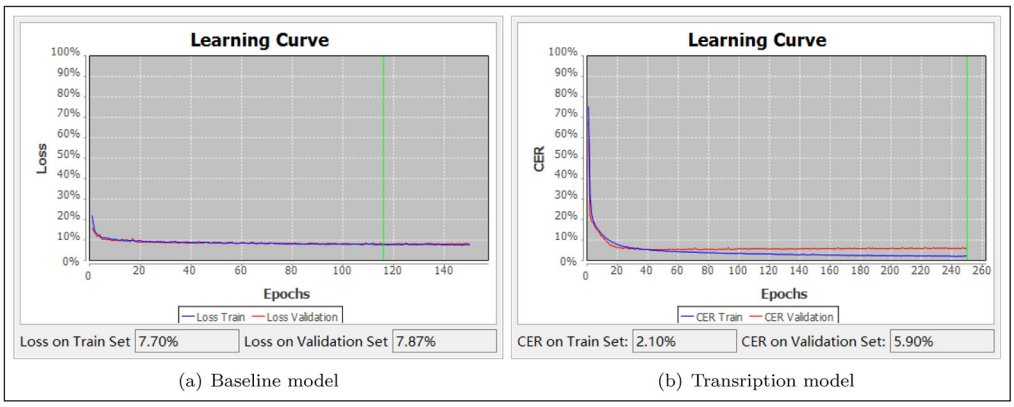
1 OVERVIEW

1.1 REPOSITORY LOCATION

https://doi.org/10.5281/zenodo.8325375

1.2 CONTEXT

Transkribus, a specialized tool developed for Handwritten Text Recognition (HTR) and powered by the PyLaia engine, currently lacks publicly available models for Chinese character and/or IPA (International Phonetic Association, 1999) symbol recognition. In the context of my own research on the Tujia language in China, I have encountered numerous related studies published predominantly between the 1980s and 2000s. The original materials are typically in an outdated print style, which makes it challenging to use the scanned copies for further analysis. Unlike other materials that have been readily digitized, the scanned versions of these Tujia materials do not offer the same ease of use. For instance, the search function cannot be applied to identify specific keywords, and it is not possible to convert the recorded lexical entries into convenient Excel or XML formats that are beneficial for our studies and research. Lexical lists or word lists have played a significant role in prior Tujia studies; however, the antiquated format and print style of these lexical lists poses challenges for further analysis. Given this context, I developed two models on Transkribus for the recognition of printed texts, as in Figure 1. The baseline model achieved a loss rate of 7.87% on the validation set, while the transcription model attained a character error rate (CER) of 5.90% on the validation set. The loss rate denotes the percentage of information that the baseline model failed to capture on a page level, and the CER of the transcription model indicates the rate of incorrect transcriptions generated at the character level. These models were trained using the digitized lexical lists of Burmish languages made available and openly accessible by Hill and Cooper (2020). These lexical lists pertain to Burmish languages, which belong to the Tibeto-Burman language family and are primarily spoken in the Republic of the Union of Myanmar and the neighboring country of China (Bradley 2012).



(a) Baseline model (b) Transription model

Figure 1 Learning curves of latest models on Transkribus.

2 METHOD

To achieve accurate printed text recognition using Transkribus, two distinct models need to be developed: the baseline model and the transcription model. These models are designed to cater to specific requirements and factors, including text content, input text alignment, document format, and desired results, among other considerations. While Transkribus provides default models for each purpose, the generated results may not be optimal due to variations in the input data. Therefore, it is crucial to develop models that align with individual needs and document types. The process of employing Transkribus for printed text recognition can be divided into two main stages. Firstly, the text is segmented at the page level using the trained baseline model, and subsequently, corresponding transcriptions are generated for the segmented text. The training of models was based upon lexical lists adapted from Hill and Cooper (2020), comprising a total of 345 pages of files. Within the training process, 311 pages were employed, with the remaining 34 pages serving as a validation set. The baseline model underwent training over the course of 100 epochs, while the transcription model was subjected to 250 epochs. The most recent model has acquired knowledge encompassing approximately 2,363 Chinese characters, 100 IPA symbols, 44 common symbols, and Arabic numerals.

This section will elaborate on both the training and usage of the baseline model, as well as the transcription model.

2.1 LAYOUT SEGMENTATION – BASELINE MODEL

The preliminary step in preparing for the transcription process involves layout segmentation. This entails dividing the content of each page into distinct text regions, text lines, baselines, and establishing the appropriate reading order. Prior to using Transkribus to generate transcriptions for the input data, it is essential to perform this segmentation. While Transkribus offers default baseline models, their effectiveness relies heavily on the alignment of the texts of the original document.

In the context of our study, our primary objective was to develop Transkribus models capable of effectively processing lexical lists as illustrated in Figure 2. These lexical lists typically comprise multiple columns, with one column presenting the lexical meanings in Chinese, while the remaining columns provide corresponding transcriptions in IPA for three distinct dialects. Figure 2 below illustrates this format. To segment the content in Figure 2, we have first employed the default baseline model provided by the Transkribus team, namely the 'Horizontal Text Line Orientation' (Transkribus, n. d.). According to the description provided by the Transkribus team on their desktop software, this default model is trained exclusively on the cBad dataset, which exhibits a similar layout. Consequently, this model primarily identifies horizontal and vertical lines.

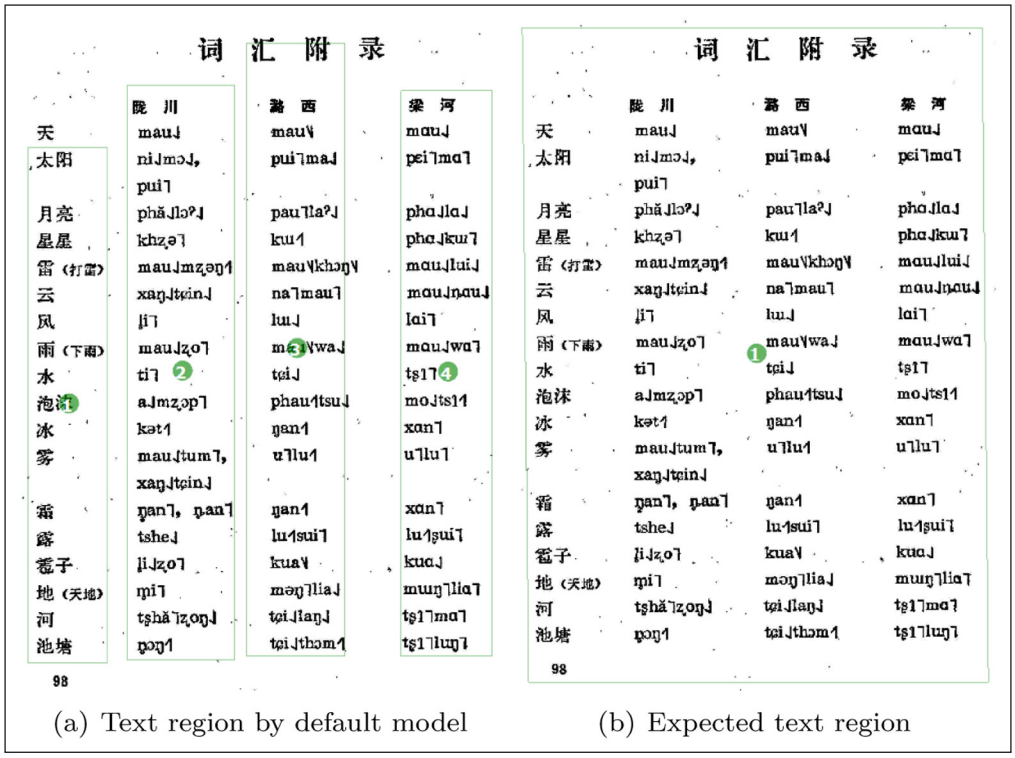


Figure 2 Text regions identified by Transkribus models: default and desired.

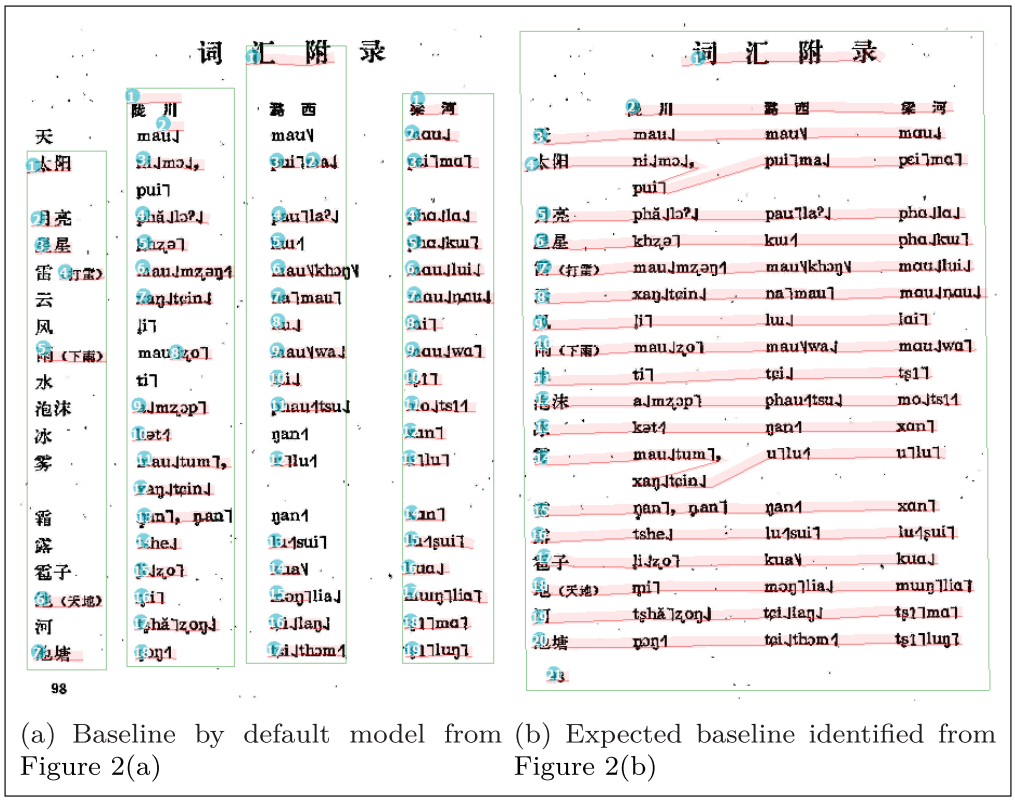
2.1.1 Text region recognition

As depicted in Figure 2(a), the content within the image has been divided into four text regions, labeled as 1, 2, 3, and 4, delineated by the green squares. While this segmentation identifies additional text regions, it is evident that certain content in the image, such as the page number '98', has been overlooked. For our purposes, it would be preferable to encompass all the content on the page within a single text region, as demonstrated in Figure 2(b). Figure 2(b) represents our desired outcome for the recognition of lexical list images at the text region level, achieved through manual correction.

2.1.2 Baseline recognition

In addition to text region recognition, the recognition of baselines provides further detail regarding the reading and transcribing order within blocks or paragraphs on the image. Figure 3(a) illustrates the baselines identified by the default Transkribus baseline model within each text region, represented by red lines. Upon reviewing the reading order, indicated by

numbers in a vertical view, no apparent mistakes are observed, except for some ignored text. However, when considering the horizontal order, the numbering scheme appears somewhat inconsistent. In Figure 3(a), the first baselines (2 and 3) in the third column from the left should be numbered in the opposite order. Furthermore, the two baselines identified should be recognized as a single baseline since they represent different parts of the same lexical word. To address these inconsistencies, Figure 3(b) displays the manually corrected baselines. In summary, our expectation for the baseline model is to recognize content on the same line as a single baseline. The workflow involves initially segmenting the texts using the default baseline model. Subsequently, we trained a new baseline model using the manually corrected results.



(a) Baseline by default model from Figure 2(a)

(b) Expected baseline identified from Figure 2(b)

Figure 3 Baselines identified by Transkribus models: default and desired.

2.2 TEXT TRANSCRIPTION – TRANSCRIPTION MODEL

The transcription process in Transkribus relies on the PyLaia engine (Mocholí Calvo et al., 2018), which is built using the PyTorch toolkit (Paszke, Gross, & Chintala, 2017). In general, we input transcriptions for the content previously identified by the baseline model. Specifically, we transcribe each baseline recognized, following the assigned baseline order. However, it is important to note that on Transkribus, we can only provide transcriptions for text that falls within the identified text regions and text lines. Therefore, if any text has been ignored by the model and not manually corrected, we will not be able to input transcriptions for that particular text.

The transcription model we developed was trained using a compilation of ten Burmish lexical lists, as shown in Table 1. These lexical lists originate from ten distinct sources, documenting various Burmese languages in China, specifically Achang, Bola, Chashan, Langsu, Leqi, and Zaiwa. They are predominantly composed of both Chinese characters and IPA symbols. As Table 1 shows, most of these works were published in the 2000s, with some dating back to the 1980s. Consequently, the quality of printing and alignment of the content may not be ideal in comparison to contemporary publications.

Table 1 Burmish lexical lists.

LANGUAGE RECORDED	SOURCE
Achang	Dai and Cui (1985)
Bola	He and Chen (2004) and Dai, Jiang, and Kong (2007)
Chashan	Dai (2010)
Langsu	He and Chen (2004), and Dai (2005)
Leqi	He and Chen (2004), Dai and Li (2006), and Dai and Li (2007)
Zaiwa	Xu and Xu (1984)

Since we have trained the transcription model using these lexical lists, its focus primarily lies in recognizing Chinese characters and IPA symbols. As a result, the challenges we encountered mainly revolve around variations in the representation of Chinese characters across different systems, as well as differences in the usage of IPA symbols.

In relation to Chinese characters, the model endeavors to identify analogous alternatives within its lexicon. For instance, the model employs the character '作' as a substitute for '昨', and '了' as a substitute for '子'. Additionally, it becomes apparent that the quality of the recognized baseline significantly impacts the error rate of the transcription model. If the text is inadequately covered by the baseline, the likelihood of incorrect transcription for a given character or symbol increases. In terms of IPA symbols, the error rate is generally low, and the overall transcription quality is better. However, there is a specific type of error observed in the transcriptions of IPA symbols generated by the trained model, particularly related to tone representation. The lexical lists we used employ two different tonal recording systems: Chao tone letter (Chao, 1930) and the corresponding representation by numbers. The model performs well with numbers for tone representation, but it tends to make numerous mistakes when transcribing tones with Chao tone letter. This indicates a specific challenge in accurately transcribing tones using Chao tone letter in our trained model.

In summary, we trained new models by using the default model as the base and incorporating manually corrected transcriptions for the recognized texts. This process helps improve the accuracy and quality of the transcriptions generated by Transkribus.

3 PERFORMANCE OF TRANSKRIBUS MODELS

To evaluate the efficiency of using Transkribus for transcribing lexical lists, a test was conducted involving three PhD students from Trinity College Dublin, all specializing in linguistics. Each participant was given one hour to transcribe a Tujia lexical list from Tian et al. (1986: 176–222). The task involved typing the content seen on each page into Word or Excel files using their preferred input method. All three participants commenced simultaneously within the same room. The lexical list consisted of three columns per page: one for the transcriptions of the northern Tujia variant, one for the transcriptions of the southern Tujia variant, and another for the corresponding meanings in both Chinese and English.

During the test, each line was considered as one entry, and the precise character count was also recorded. Table 2 present some details regarding the participants, the number of entries and characters they completed within the given one-hour time frame, and the CER they achieved.

Table 2 Participants' information and performance.

PARTICIPANT	INPUT METHOD	ENTRIES	CHARACTERS	CER (%)
1	TypeIt (Szynalski, 2023)	64	2027	1.62
2	Keyman (SIL International, 2023)	109	3424	0.32
3	Sougou (Sougou Corporation, 2023)	60	1952	1.22

To obtain the transcriptions of the same Tujia lexical list using Transkribus, we first utilized the baseline model to segment the document's layout. Since Participant B transcribed the most entries (109), we only needed to segment six pages containing a total of 131 lexical entries. However, the Transkribus model completed the segmentation for all 47 pages of the lexical list in just one minute. We then manually corrected the segmentation performed by the Transkribus baseline model, which took approximately six minutes. Next, we used the Transkribus transcription model we trained to transcribe the first six pages of the lexical list, which had a similar number of entries as Participant B. The model completed 131 lexical entries and 4,173 characters in one minute and 40 seconds, with a CER of 8.65%.[1]

Similarly, during the 28 minutes we spent correcting the model's transcriptions, most of the time was dedicated to correcting the English content. In summary, it took the model one minute for layout analysis, six minutes for manual correction, one minute and 40 seconds

1 The model was solely trained to recognize Chinese characters and IPA symbols, while the Tujia lexical list also encompasses word meanings in English. The reported CER of 8.65% specifically pertains to errors unrelated to English, whereas the overall error rate, inclusive of English errors, stands at 19.5%.

for transcription analysis, and 28 minutes for manual correction. Thus, the total time required for transcribing 109 lexical entries using Transkribus was 36 minutes and 40 seconds, while Participant B took one hour to complete the same task without even considering the error rate. In essence, for the most recent model we trained, manual correction proves essential for both layout analysis and transcription; however, it significantly reduces the overall time required compared to entirely manual transcription.

4 DATASET DESCRIPTION

- **Object name** – OCR model for lexical lists in Chinese-IPA Glossing, Ground Truth
- **Format names and versions** – jpg, xml, pdf and docx
- **Creation dates** – 2023-09-07
- **Data creators** – Shihua Li, Trinity Centre for Asian Studies, Trinity College Dublin, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization
- **Language** – Chinese, Burmish languages including Achang, Bola, Chashan, Langsu, Leqi, and Zaiwa
- **License** – Creative Commons Attribution 4.0 International
- **Repository name** – Zenodo
- **Publication date** – 2023-09-07

5 REUSE POTENTIAL AND FUTURE DEVELOPMENT

The primary objective of training the baseline and transcription models on Transkribus is to facilitate the digitization of Tujia lexical lists for future research. Nevertheless, these models can also be effectively employed for documents primarily written in Chinese and IPA, which aligns with the customary practice observed in studies conducted on national languages in China. Consequently, this project will not only contribute to the digitization of lexical lists using Chinese-IPA glossing, but also enable the digitization of materials following the same glossing style. Additionally, these efforts will make valuable contributions to the Lexibank project by expanding the collection of wordlists pertaining to national languages in China. There is still room for improvement in both of the two trained models, as their training was based on a limited amount of data.

5.1 LAYOUT ANALYSIS

The training of the baseline model for layout analysis is highly dependent on the specific characteristics of the documents being considered. The key factors that significantly influence the training process include the original alignment of content on each page and the desired format and structure of the transcription. In our case, we focused on training the model to accurately segment content into lines. However, for alternative purposes, it may be beneficial to refer to other publicly available baseline models or train new models for specific needs.

5.2 CHINESE TRANSCRIPTION

Contrary to our initial expectations, the transcription model exhibits impressive performance in recognizing and transcribing Chinese characters within untrained documents, particularly for those characters it has been previously trained on and can readily identify. However, when confronted with untrained characters, the model tends to generate new characters based on its existing inventory, leading to frequent inaccuracies. Additionally, the model is susceptible to errors when the original printing quality of the documents is not as high as anticipated. For example, tone '35' was frequently recognized as tone '25' due to the poor printing quality.

5.3 IPA TRANSCRIPTION

Contrary to our expectations, the transcription model's performance on IPA symbols falls short compared to its proficiency in recognizing Chinese characters. This discrepancy may be attributed to the fact that the ten Burmish lexical lists used for training cover a limited range of

IPA symbols. Surprisingly, even when the original documents have excellent printing quality, the model still exhibits a tendency to make mistakes with similar IPA symbols, such as confusing [ɛ] and [e], [l] and '1', and [ʔ] and '2'. Consequently, it becomes imperative to train the model using a more extensive dataset of IPA symbols to improve its accuracy and proficiency in identifying and recognizing such symbols.

5.4 FIELD MODELS

Transkribus is currently in the process of testing new layout analysis models known as Field Models. These models are specifically designed to analyze various types of contexts, such as newspapers that contain distinct text regions within the same page. Once the testing phase is complete and these models are officially released, we plan to proceed with the development of a new version of the layout analysis model based on the newly introduced Field models.

ACKNOWLEDGEMENTS

I would like to express my gratitude to Doug Cooper for his inspiration in motivating me to develop Transkribus models for the purpose of digitalizing Tujia lexical lists. Furthermore, I am deeply thankful to Franz Xaver Erhard for introducing me to Transkribus and its functionalities. Lastly, I am immensely grateful to the three participants who graciously offered their assistance during the testing phase of this study.

FUNDING INFORMATION

This work was funded by the Arts and Humanities Research Council (AHRC), UKRI, as part of the project "The Emergence of Egophoricity: a diachronic investigation into the marking of the conscious self." Project Reference: AH/V011235/1. Principal Investigator: Nathan Hill, SOAS University of London.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Shihua Li: Data curation, Formal Analysis Investigation, Methodology, Validation, Visualization, Writing – original draft.

Nathan Hill: Conceptualization, Methodology, Supervision.

AUTHOR AFFILIATIONS

Shihua Li orcid.org/0000-0002-6090-1383
Trinity Centre for Asian Studies, Trinity College Dublin, Dublin, Ireland

Nathan Hill orcid.org/0000-0001-6423-017X
Department of East Asian Languages and Cultures, SOAS University of London, London, UK; Trinity Centre for Asian Studies, Trinity College Dublin, Dublin, Ireland

REFERENCES

Bradley, D. (2012). The characteristics of the Burmic family of Tibeto-Burman. *Language and Linguistics*, *13*(1), 171–192.
Chao, Y.-R. (1930). A system of tone letters. *Le maître phonétique*, *30*, 24–27.
Dai, Q. (2005). 浪速语研究 *[Langsuyu Yanjiu]*. Beijing: Minzu Chubanshe.
Dai, Q. (Ed.) (2010). 片马茶山人及其语言 *[Pianmachashanren Ji Qi Yuyan]*. Shanghai: Shangwu Yinshuguan.
Dai, Q., & **Cui, Z.** (1985). 阿昌语简志 *[Achangyu Jianzhi]*. Beijing: Minzu Chubanshe.
Dai, Q., Jiang, Y., & **Kong, Z.** (2007). 波拉语研究 *[Bolayu Yanjiu]*. Beijing: Minzu Chubanshe.
Dai, Q., & **Li, J.** (2006). 勒期语概况 [Leqiyu Gaikuang]. 民族语文 *[Minzu Yuwen]*, *1*, 66–80.
Dai, Q., & **Li, J.** (2007). 勒期语研究 *[Leqiyu Yanjiu]*. Beijing: Zhongyang Minzu Daxue Chubanshe.
He, L., & **Chen, F.** (2004). 云南特殊语言研究 *[Yunnan Teshu Yuyan Yanjiu]*. Kunming: Yunnan Minzu Chubanshe.

Hill, N., & **Cooper, D.** (2020). A machine readable collection of lexical data on the Burmish languages [Data set]. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.3759030

International Phonetic Association. (1999). *Handbook of the international phonetic association: A guide to the use of the international phonetic alphabet*. Cambridge: Cambridge University Press.

Mocholí Calvo, C. (2018). *Development and experimentation of a deep learning system for convolutional and recurrent neural networks*. (Unpublished doctoral dissertation). Valencia: Universitat Politècnica de València.

Paszke, A., Gross, S., & **Chintala, S.** (2017). *Pytorch*. Retrieved from https://github.com/hughperkins/pytorch-pytorch

SIL International. (2023). *Keyman*. Retrieved from https://keyman.com

Sougou Corporation. (2023). *Sougou Keyboard*. Retrieved from https://pinyin.sogou.com (搜狗)

Szynalski, T. P. (2023). Typeit.org. Retrieved from https://www.typeit.org/

Tian, D., He, T., Chen, K., Li, J., Xie, Z., & **Peng, X.** (1986). 土家语简志 *[Tujiayu Jianzhi]* (Vol. 1). Beijing: Minzu Chubanshe.

Transkribus. (n. d.). *Advanced layout configuration settings*. Transkribus. Retrieved from https://help.transkribus.org/advanced-layout-configuration-settings

Xu, X., & **Xu, G.** (1984). 景颇族语言简志 (载瓦语) *[Jingpozu Yuyan Jianzhi (Zaiwayu)]*. Beijing: Minzu Chubanshe.

TO CITE THIS ARTICLE:
Li, S., & Hill, N. (2023). Printed Text Recognition for Lexical Lists in Chinese-International Phonetic Alphabet (IPA) Glossing. *Journal of Open Humanities Data*, 9: 15, pp. 1–8. DOI: https://doi.org/10.5334/johd.119

Submitted: 11 July 2023
Accepted: 12 September 2023
Published: 13 October 2023

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.