WikiHan: A New Comparative Dataset for Chinese Languages

Kalvin Chang and Chenzuan Cui and Youngmin Kim and David Mortensen
Language Technologies Institute, Carnegie Mellon University
{kalvinc, cxcui, youngmik, dmortens}@cs.cmu.edu

Abstract

Most comparative datasets of Chinese varieties are not digital; however, Wiktionary includes a wealth of transcriptions of words from these varieties. The usefulness of these data is limited by the fact that they use a wide range of variety-specific romanizations, making data difficult to compare. The current work collects this data into a single constituent (IPA, or International Phonetic Alphabet) and structured form (TSV) for use in comparative linguistics and Chinese NLP. At the time of writing, the dataset contains 67,943 entries across 8 varieties and Middle Chinese. The dataset is validated on a protoform reconstruction task using an encoder-decoder cross-attention architecture (Meloni et al., 2021), achieving an accuracy of 54.11%, a PER (phoneme error rate) of 17.69%, and a FER (feature error rate) of 6.60%.

1 Introduction

The Chinese language family consists of many mutually unintelligible languages, which linguists cluster into seven well-established subgroups (Mandarin, Yue, Wu, Min, Hakka, Gan, and Xiang) and a few more debated subgroups (Jin, Hui, and Pinghua; Handel, 2015). Each subgroup is made of varieties that may or may not be mutually unintelligible. Most scholars of Sinitic believe that all varieties except for Min languages descended from Middle Chinese (Handel, 2015), which is a set of old varieties documented in Qièyùn (切韻), a rhyme dictionary compiled in 601 CE.

Today, there are a total of 1.3 billion speakers of Sinitic varieties, making the family one of the largest in terms of speaker count (Eberhard et al., 2022). Eight of the subgroups have at least tens of millions of speakers. In all Chinese-speaking societies except for Hong Kong, Mandarin is the language of administration, meaning that other varieties are not commonly written, although it is possible to do so, either with Han characters (Chinese logographic characters) or romanized orthographies. Because Standard Written Chinese diverged from the spoken varieties over time (Chen, 2015), it is unclear which Han characters should be used for transcription, for many words in the spoken varieties. However, scholars can trace from which “original character” in Middle Chinese or Old Chinese the modern word descended (bènzì 本字) (Yang, 2000).

To determine the pronunciations of these original characters in Middle Chinese, historical linguists and dialectologists compare the pronunciations of words across modern Sinitic varieties. More broadly, protoform reconstruction is the inference of morphemes or words as they appeared in the ancestral languages of a set of daughter languages. Cognates—words deemed to descend from the same ancestral form—are inputs to the reconstruction, while this ancestral form is the output of the reconstruction. See Table 2 for an example of a cognate set, with Middle Chinese as the protoform and all columns to its right as the cognates in the daughter languages.

To enable fair comparison across the daughter languages, the pronunciations should be in a common phonetic or phonemic transcription system, as opposed to divergent romanizations. The problem is that large datasets that do offer phonetic transcriptions such as the Great Dictionary of Modern Chinese Dialects 現代漢語方言大詞典 (Li, 2002) are often in print form. Wiktionary, on the other hand, offers a digital compilation of pronunciation entries across several sources manually entered by users (Wiktionary contributors, 2022a). However, entries are stored in subgroup-specific romanizations on the back end. The IPA transcription...
tions are generated on the front end by Lua scripts that convert romanizations to IPA\(^2\) and as such are not available in data exports, which only store the romanized forms. To obtain a large dataset of pronunciations in IPA, one could scrape the front end of Wiktionary, but this would be inefficient. Instead, we write our own romanization to IPA conversion modules in Epitran (Mortensen et al., 2018) for each subgroup based on those from Wiktionary. Using the outputs of these Epitran modules, we compile a large dataset of pronunciation entries in IPA in TSV form. We then show that the entries we generate can be used in a computational protoform reconstruction of Middle Chinese, demonstrating that our dataset addresses the inadequacies of previous digital comparative Sinitic datasets with respect to computational reconstruction.

2 Related Work

Linguists have created digital comparative datasets since the birth of computers. Refer to Table 1 for a list of Sinitic datasets. Wang (1970) in particular features Middle Chinese, Sino-Xenic (Japanese Kan-on, Japanese Go-on, and Sino-Korean) loanwords, and Old Mandarin (from the Zhōngyuán Yīnyūn 中原音韻). Similar to ours, the main objective of the dataset was to provide data for computer-assisted reconstructions of Chinese phonology. The Multi-function Chinese Character Database is interesting in that it organizes characters by all the possible syllables in the dialect, including the tone. The Database is itself a compilation of multiple print sources. Wu and List (2021) stand out for their manual annotation of salient morphemes that contributed to cognacy judgments, which they use to convert partial cognates to full words in creating cognate sets and wordlists for phylogenetic inference.

One challenge with many of the aforementioned datasets is that they organize entries by words instead of characters. Across subgroups, synonymous compound words may use different characters to express the same meaning.\(^3\) Even though partial cognacy of such compounds is possible, the method by which partial cognates are coded into full cognates in word lists will affect the downstream task of phylogenetic inference, which assumes that words in the inputted word lists descend from the same proto-word (Wu and List, 2021). Many other sources exist and may have significantly more entries, such as the Great Dictionary of Modern Chinese Dialects. However, to our knowledge, they are not digitized, making computer-assisted reconstruction difficult.

3 Dataset and Methodology

We obtain Sinitic pronunciation entries using a CBOR snapshot of the zh-pron module on Wiktionary.\(^4\) For heteronyms, characters with multiple pronunciations within a variety, each pronunciation is stored as a separate entry in the snapshot and is grouped along with pronunciations in other varieties believed to be cognate with that particular pronunciation. The dataset spans eight subgroups: the seven conventionally recognized ones (Mandarin, Yue, Wu, Min, Xiang, Gan, and Hakka) and the proposed Jin subgroup, for which Wiktionary has entries. We chose the dialects with the most data to represent each subgroup. Refer to Table 3 for the full list of dialects.\(^5\) We restrict the dataset to single characters (morphemes), allowing us to establish cognacy between pronunciation entries across subgroups by assuming that readings grouped under the same character by Wiktionary are cognates descended from the same original character (本字).

To convert from romanization to IPA, we extend Epitran with additional modules for each subgroup (Mortensen et al., 2018). Refer to Table 3 to see which romanization system was used for each subgroup. Wiktionary supplies mapping tables for each variety\(^6\), and we compare the mapping tables with other sources such as Wiktionary’s Lua conversion scripts (see Appendix A for the full list).

We represent tones using IPA tone letters. We do not mark the neutral tone, as opposed to marking it as IPA tone 3. The transcription is relatively narrow, with diphthongs always represented as se-

\(^2\)https://tools-static.wmflabs.org/templatehoard/dump/latest/zh-pron.cbor
\(^3\)We acknowledge that including Taiwanese Hokkien entries may present a problem during reconstruction because its dialects are a mix of the Quanzhou and Zhangzhou dialects of Hokkien (a language within the Southern Min branch of the Min subgroup) to varying degrees.

\(^4\)See https://en.wiktionary.org/wiki/Module:nan-pron for an example of a conversion script.
\(^5\)See https://en.wiktionary.org/wiki/Wiktionary:About_Chinese/Xiang for an example
sequences of vowel symbols and glides marked by the IPA non-syllabic diacritic: [i̯a] rather than [ja]. Fewer symbols are preferred over many: [n] rather than [ɲ].

Middle Chinese (MC) transcriptions following Baxter and Sagart (2014) are derived programatically from fǎnqiè formulae7 from Qiéyùn (available for around 20,000 characters). Despite concerns raised by Norman and Coblin (1995) and others, we treat MC pronunciations as the gold standard protoform. We process the Qiéyùn descriptions in three stages. First, the Middle Chinese descriptions of initial, final, tone, division (等), and openness (合開) are converted to the ASCII romanization system in Baxter and Sagart (2014). The romanizations are then converted to IPA with Epitran, using a mapping table based on SinoPy (List, 2019). Finally, we rewrite some IPA phonemes to match the phonetic transcription convention used in this dataset (e.g. [ʨ] → [t͡ɕ], [ȵ] → [ɲ]). For the tones, we use superscripts 1 through 4 to indicate what would traditionally be denoted as 平上去入.

The final result is a list of nearly 20,000 characters, each with a reconstruction written in IPA symbols.

4 Experiments

We show that the dataset can be used for the protoform reconstruction task. Meloni et al. (2021) model the Latin protoform reconstruction task as a sequence to sequence transduction problem with a character-based encoder-decoder (Cho et al., 2014) with cross-attention (Bahdanau et al., 2015).8 We reimplement their architecture, originally written in DyNet (Neubig et al., 2017), in PyTorch. The architecture consists of a language and token embeddings, an encoder GRU (Cho et al., 2014), a decoder GRU, and a multi-layer perceptron.

All daughter forms within one cognate set are concatenated into one string before entering the encoder. To distinguish between each variety, a language code is first prepended before each pronunciation entry. In the encoder, a language embedding is learned for each dialect. The same is done in the decoder for the proto-language. Token embeddings are applied to individual characters (close to a phone) in the input and are shared across each language. There is a residual connection between the attention output and the decoder RNN output before entering the multi-layer perceptron. The objective function is cross-entropy loss between the protoform and the predicted protoform.

The only difference between our PyTorch version and their code is that we do not implement variational dropout in the encoder (Gal and Ghahramani, 2016), but DyNet comes with this flavor of dropout in its RNN modules. We do implement variational dropout for the decoder, though.9

---

Table 1: Comparative Sinitic datasets. “SIPA” refers to “Sinological IPA.” In our dataset, a heteronymic character can have multiple cognate sets, reflecting different sets of pronunciation variants that are only cognate with variants in the same set. Entries refers to the total number of pronunciation records across all varieties.

<table>
<thead>
<tr>
<th>Name</th>
<th>Citation</th>
<th>Varieties</th>
<th>Sets</th>
<th>Entries</th>
<th>Transcription</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Great Dictionary of Modern Chinese Dialects</td>
<td>Li (2002)</td>
<td>42</td>
<td>—</td>
<td>170,000+</td>
<td>SIPA</td>
<td>print</td>
</tr>
<tr>
<td>Dictionary on Computer</td>
<td>Wang (1970); Streeter (1972)</td>
<td>17</td>
<td>2,444</td>
<td>58,012</td>
<td>SIPA</td>
<td>tabular</td>
</tr>
<tr>
<td>Multi-Function Chinese Character Database</td>
<td>Research Centre for Humanities Computing, CUHK</td>
<td>20</td>
<td>7,554</td>
<td>—</td>
<td>SIPA</td>
<td>HTML</td>
</tr>
<tr>
<td>Peking University</td>
<td>Peking University (2021)</td>
<td>18</td>
<td>905</td>
<td>18,059</td>
<td>SIPA</td>
<td>CLDF</td>
</tr>
<tr>
<td>Wiktionary</td>
<td>Hou (2004)</td>
<td>39</td>
<td>1,023</td>
<td>&gt;39,000</td>
<td>IPA</td>
<td>XML</td>
</tr>
<tr>
<td>Phonological Database of Chinese dialects</td>
<td>List (2021)</td>
<td>15</td>
<td>140</td>
<td>2,789</td>
<td>SIPA</td>
<td>CLDF</td>
</tr>
<tr>
<td>Liu et al. (2007) annotated with salient morphemes</td>
<td>Wu and List (2021)</td>
<td>19</td>
<td>201</td>
<td>&gt;3,000</td>
<td>CLDF</td>
<td>CLDF</td>
</tr>
<tr>
<td>WikiHan (ours)</td>
<td>N/A</td>
<td>8</td>
<td>21,227</td>
<td>68,368</td>
<td>IPA</td>
<td>TSV</td>
</tr>
</tbody>
</table>

---

7 Fāngqié spelling provides equivalence classes for the pronunciation of a syllable by using one character with the same onset and another with the same rhyme.

8 Meloni et al. (2021)’s code is available at https://github.com/shauli-ravfogel/Latin_reconstruction.

9 PyTorch’s RNN, LSTM, and GRU modules do not come with variational dropout. It is possible to overwrite the respective classes with a version that implements variational dropout, though.
Table 2: Example of a complete cognate set in the dataset for the word 犬 (dog)

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Dialect Chosen</th>
<th>Romanization</th>
<th>犬 romanized</th>
<th>Number of entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mandarin</td>
<td>Beijing Pinyin</td>
<td>quán</td>
<td>20369</td>
<td></td>
</tr>
<tr>
<td>Yue</td>
<td>Cantonese Jyutping</td>
<td>hyun2</td>
<td>16727</td>
<td></td>
</tr>
<tr>
<td>Wu</td>
<td>Shanghainese Wiktionary’s romanization</td>
<td>2qyoe</td>
<td>2877</td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>Hokkien</td>
<td>khián</td>
<td>6145</td>
<td></td>
</tr>
<tr>
<td>Hakka</td>
<td>Sixian</td>
<td>khién</td>
<td>5215</td>
<td></td>
</tr>
<tr>
<td>Gan</td>
<td>Nanchang Wiktionary’s romanization</td>
<td>qyon3</td>
<td>1195</td>
<td></td>
</tr>
<tr>
<td>Xiang</td>
<td>Old Xiang Wiktionary’s romanization</td>
<td>qye3</td>
<td>1258</td>
<td></td>
</tr>
<tr>
<td>Jin</td>
<td>Taiyuan</td>
<td>qye1</td>
<td>1410</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: The dialect chosen for each subgroup and its romanization of the word 犬, in addition to a count of the number of pronunciation entries per subgroup. For Min, the pronunciations are a mix of the Xiamen, Quanzhou, Zhangzhou, and Taiwanese dialects of Hokkien. For Middle Chinese, we have 14653 entries.

We find that dropout makes a significant difference when trained on the small 39-variety dataset of 1,000 cognate sets from Hóu (2004) (which ended up being around 800 sets because not every entry in Hóu (2004) had an entry in Qièyùn). 10

We now discuss how we adapted our dataset for Meloni et al. (2021)’s model. In order for the model to learn correspondences between phonemes as linguists would, we tokenize by phonemes, for example /tʰ/ and /t͡ɕʰ/, instead of characters. These two example phonemes should each be treated as one consonant despite being represented with several Unicode characters. We treat diphthongs and triphthongs as one token because they constitute one syllable, phonetically speaking. We also restrict ourselves to cognates with at least 4 entries including Middle Chinese to avoid being biased to varieties with more entries, such as Mandarin. Another decision we made is to arbitrarily take the first pronunciation when multiple variants are included in the same entry on Wiktionary.

We compare Meloni et al. (2021) against two baselines. The random daughter baseline selects a daughter form at random and takes that as the reconstructed protoform. This assumes that no sound change occurred from the protoform to the daughter. The majority constituent baseline first separates daughter forms into onset, nucleus, and coda with the consonantal feature of the phoneme obtained using PanPhon (Mortensen et al., 2016), reflecting domain knowledge about the syllable structure of Chinese languages. This allows us to easily obtain sound correspondences across the daughter languages. Within each constituent (onset, nucleus, and coda), we take the most common phoneme sequence. This relies on the majority wins heuristic employed by historical linguists wherein the most frequent sound across the daughter languages is chosen as the proto-sound (Campbell, 2013).

Meloni et al. (2021) outperforms both baselines on 3 different metrics (see Table 4): (1) Accuracy, the rate at which hypothesis and reference match exactly, (2) Phoneme Error Rate (PER), the cumulative number of phoneme edits between the hypothesis and the reference normalized by the total length of the reference (in phonemes), and (3) Feature Error Rate (FER) the cumulative edit distance in terms of PanPhon (Mortensen et al., 2016) features (drawn from articulatory phonetics) between the hypothesis and the reference, normalized by the total number of features in the reference (the total length of the reference in phonemes multiplied by the number of features per phoneme).

PER is more suited for the protoform reconstruction task than character error rate or edit distance because many phonemes are written with more than 1 character in IPA, as shown in the examples from above. As for FER, its benefit lies in how it is able to assign partial credit to hypothesized phonemes that are more phonetically similar. Intu-
<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Phoneme error rate</th>
<th>Feature error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meloni et al. (2021)</td>
<td>0.5411</td>
<td>0.1769</td>
<td>0.0660</td>
</tr>
<tr>
<td>Random daughter</td>
<td>0.0290</td>
<td>0.7367</td>
<td>0.2600</td>
</tr>
<tr>
<td>Majority constituent</td>
<td>0.0271</td>
<td>0.7320</td>
<td>0.2209</td>
</tr>
</tbody>
</table>

Table 4: Evaluation of Meloni et al. (2021)’s model and 2 baselines on our dataset using 3 different evaluation criteria.

positively, we would want, for example, [t] to be penalized less than [x] in some scenario where the reference is [tʰ]. Both [t] and [x] differ from the reference by one character, but [t] and [tʰ] are both voiceless alveolar plosives that differ only in aspiration. Finally, we prefer error rates over edit distances because it is difficult to compare results across different language families, which differ in word lengths. Sinitic words in particular are often shorter than Romance words because the former is composed of monosyllabic characters.

5 Discussion and Future Work

Our dataset is intended for a computational reconstruction of Middle Chinese, as we have demonstrated in the experiments, but can be used to accomplish much more. It can also be used for cognate prediction and for dialectometry (quantifying relationships between linguistic varieties). Along the same lines, it can be used to build phylogenetic models of Sinitic that can shed light on the history of Chinese populations. Additionally, Chinese speech models could benefit from a phonetic language model (Dalmia et al., 2019) trained on our data or from estimations of phone distributions (Li et al., 2021) in low resource varieties present in our dataset.

In a future release, we will include other varieties available on Wiktionary (Taishanese, Southwest Mandarin, Teochew, Min Bei, and Min Dong). Wiktionary also contains pronunciations for Sino-Xenic loanwords in Korean, Japanese, and Vietnamese, which linguists often reference when creating Chinese reconstructions. The more languages we include, the fewer the number of sources that Chinese historical phonologists need to consult, reducing the tediousness of work in this field.

References


Peking University. 2021. CLDF dataset derived from Beijing University’s ”Chinese Dialect Vocabularies” from 1964.


Wiktionary contributors. 2022a. Wiktionary, the free dictionary. [Online; accessed 11-May-2022].

Wiktionary contributors. 2022b. 眼淚. Wiktionary, the free dictionary. [Online; accessed 20-Jul-2022].


A Sources

First, we must credit Wiktionary users for manually compiling pronunciation entries across different sources. Without them, this corpus would not be possible. While creating the Epitran romanization to IPA mapping tables and while building the Middle Chinese data, we also consulted their scripts and documentation, in addition to other sources listed below.

A.1 Middle Chinese
   • SinoPy (List, 2019)
     • https://github.com/ycm/cs221-proj/blob/master/preprocessing/dataset/pron/mc-pron.csv

A.2 Mandarin
   • https://en.wiktionary.org/wiki/Module:cmn-pron
   • https://en.wikipedia.org/wiki/Pinyin
   • https://en.wikipedia.org/wiki/Help:IPA/Mandarin

A.3 Cantonese (Yue)
   • https://en.wikipedia.org/wiki/Jyutping

A.4 Taiwanese Hokkien (Min)
   • https://en.wiktionary.org/wiki/Module:nan-pron
   • https://blgjts.moe.edu.tw/doc/tmt_compare.pdf
   • Ministry of Education, Taiwan (2008)
     • https://zh.wikipedia.org/wiki/Help:%E6%B9%98%E6%AF%AD

A.5 Xiang
   • https://en.wiktionary.org/wiki/Module:han-pron
   • https://zh.wikipedia.org/wiki/%E6%B9%98%E6%AF%AD

A.6 Jin
   • https://en.wiktionary.org/wiki/Module:cjy-pron
   • https://en.wiktionary.org/wiki/Wiktionary:About_Chinese/Jin

A.7 Gan
   • https://en.wiktionary.org/wiki/Module:gan-pron

A.8 Wu
   • https://en.wiktionary.org/wiki/Module:wuu-pron

A.9 Hakka
   • https://en.wiktionary.org/wiki/Sixian_dialect
   • https://en.wiktionary.org/wiki/Module:hak-pron