WikiHan: A New Comparative Dataset for Chinese Languages

Kalvin Chang and **Chenxuan 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.[1] 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 *un*intelligible 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 (*běnzì* 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 transcrip-

[1] The data is available at https://github.com/cmu-llab/wikihan and the code is available at https://github.com/cmu-llab/meloni-2021-reimplementation/.

Proceedings of the 29th International Conference on Computational Linguistics, pages 3563–3569
October 12–17, 2022.

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 Yīnyùn* 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, synonymic 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]See `https://en.wiktionary.org/wiki/Module:nan-pron` for an example of a conversion script.

[3]For instance, tears is 目屎 (lit. eye feces) in Hokkien, 目汁 (lit. eye juice) in Hakka, and 眼淚 (eye tears) in Mandarin (Wiktionary contributors, 2022b). Here 目屎 and 目汁 are partial cognates since not every character descends from the same protoform.

[4]`https://tools-static.wmflabs.org/templatehoard/dump/latest/zh-pron.cbor`

[5]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.

[6]See `https://en.wiktionary.org/wiki/Wiktionary:About_Chinese/Xiang` for an example

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

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.

quences of vowel symbols and glides marked by the IPA non-syllabic diacritic: [i̯a] rather than [ja]. Fewer symbols are preferred over many: [ɲ] rather than [n̠ʲ].

Middle Chinese (MC) transcriptions following Baxter and Sagart (2014) are derived programmatically from *fǎnqiè* formulae[7] from *Qièyùn* (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 *Qièyùn* 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. [tɕ] →[t͡ɕ], [n̠] →[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]

[7] *Fǎnqiè* 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.

Character	Middle Chinese	Yue	Gan	Hakka	Jin	Mandarin	Min	Wu	Xiang
犬	/kʰwen²/	[hyːn˧˥]	[t͡ɕʰy̯ɵn˨˩˧]	[kʰi̯en˧˩]	[t͡ɕʰy̯e˩]	[t͡ɕʰy̯ɛn˨˩˦]	[kʰi̯ɛn˥˧]	[t͡ɕʰy̯ø˧˦]	[t͡ɕʰy̯e̞˦˩]

Table 2: Example of a complete cognate set in the dataset for the word 犬 (*dog*)

Subgroup	Dialect Chosen	Romanization	犬 romanized	Number of entries
Mandarin	Beijing	Pinyin	quǎn	20369
Yue	Cantonese	Jyutping	hyun2	16727
Wu	Shanghainese	Wiktionary's romanization	2qyoe	2877
Min	Hokkien	Pe̍h-ōe-jī	khián	6145
Hakka	Sixian	Pha̍k-fa-sṳ	khién	5215
Gan	Nanchang	Wiktionary's romanization	qyon3	1195
Xiang	Old Xiang	Wiktionary's romanization	qye3	1258
Jin	Taiyuan	Wiktionary's romanization	qye1	1410

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 the *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 protosound (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-

[10]Available on Wiktionary at `https://en.wiktionary.org/wiki/Module:zh/data/dial-pron/documentation`.

Model	Accuracy	Phoneme error rate	Feature error rate
Meloni et al. (2021)	**0.5411**	**0.1769**	**0.0660**
Random daughter	0.0290	0.7367	0.2600
Majority constituent	0.0271	0.7320	0.2209

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

itively, we would want, for example, [t] to be penalized less than [x] in some scenario where the reference is [t^h]. Both [t] and [x] differ from the reference by one character, but [t] and [t^h] 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

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

William H Baxter and Laurent Sagart. 2014. *Old Chinese: A new reconstruction*. Oxford University Press.

Lyle Campbell. 2013. *Historical Linguistics: an Introduction*. Edinburgh University Press.

Ping Chen. 2015. *The Oxford Handbook of Chinese Linguistics*, chapter Language Reform in Modern China. Oxford University Press.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Siddharth Dalmia, Xinjian Li, Alan W Black, and Florian Metze. 2019. Phoneme level language models for sequence based low resource ASR. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6091–6095.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2022. Ethnologue: Languages of the world. twenty-fifth edition. Dallas, Texas: SIL International. Online version: `http://www.ethnologue.com`. Accessed on 11 May 2022 at `https://web.archive.org/web/20190605194504/https://www.ethnologue.com/language/zho`.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Zev Handel. 2015. *The Oxford Handbook of Chinese Linguistics*, chapter The Classification of Chinese. Oxford University Press.

侯精一 Jīngyī Hóu, editor. 2004. *Xiàndài Hànyǔ fāngyán yīnkù 现代汉语方言音库 [Phonological database of Chinese dialects]*. Shànghǎi Jiàoyù 上海教育, Shànghǎi 上海.

Xinjian Li, Juncheng Li, Jiali Yao, Alan W Black, and Florian Metze. 2021. Phone distribution estimation for low resource languages. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7233–7237.

李榮 Rong Li. 2002. 現代漢語方言大詞典 *[Great Dictionary of Modern Chinese Dialects]*. 現代漢語方言大詞典. 江蘇教育出版社.

Johann-Mattis List. 2019. lingpy/sinopy: Python library for quantitative tasks in Chinese historical linguistics. Version 0.3.4.

Johann-Mattis List. 2021. CLDF dataset derived from Hóu's "Phonological Database of Chinese Dialects" from 2004.

刘俐李 Lili Liu, 王洪钟 Hongzhong Wang, and 柏莹 Bai Ying. 2007. 現代漢語方言核心詞·特徵詞集. *[Collection of basic vocabulary words and characteristic dialect words in modern Chinese dialects]*. 鳳凰出版社.

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. 2021. Ab antiquo: Neural proto-language reconstruction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4460–4473, Online. Association for Computational Linguistics.

Ministry of Education, Taiwan. 2008. 臺灣閩南語羅馬字拼音方案使用手冊 [Taiwanese Hokkien Romanization Handbook]. `https://ws.moe.edu.tw/001/Upload/FileUpload/3677-15601/Documents/tshiutsheh.pdf`.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori S. Levin. 2016. Panphon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. DyNet: The dynamic neural network toolkit.

Jerry L. Norman and W. South Coblin. 1995. A new approach to Chinese historical linguistics. *Journal of the American Oriental Society*, 115(4):576–584.

Peking University. 2021. CLDF dataset derived from Beijing University's "Chinese Dialect Vocabularies" from 1964.

Research Centre for Humanities Computing, CUHK. 漢語多功能字庫 Multi-function Chinese Character Database. `https://humanum.arts.cuhk.edu.hk/Lexis/lexi-mf/dialect.php`. Accessed: 2022-05-08.

Mary L. Streeter. 1972. Doc, 1971: A Chinese dialect dictionary on computer. *Computers and the Humanities*, 6(5):259–270.

William S.-Y. Wang. 1970. Project DOC: Its methodological basis. *Journal of the American Oriental Society*, 90(1):57–66.

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].

Mei-Shin Wu and Johann-Mattis List. 2021. Annotating cognates in phylogenetic studies of South-East Asian languages. *Language Dynamics and Change*. `https://doi.org/10.17613/rabq-7z45`.

楊秀芳 Hsiu-fang Yang. 2000. Concepts and methods in the study of original characters of dialects 方言本字研究的觀念與方法. 漢學研究 *[Chinese Studies], Vol. 18*, pages 111–146.

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
- https://en.wiktionary.org/wiki/Wiktionary:About_Chinese/Cantonese
- https://en.wikipedia.org/wiki/Help:IPA/Cantonese

A.4 Taiwanese Hokkien (Min)

- https://en.wiktionary.org/wiki/Module:nan-pron
- https://en.wiktionary.org/wiki/Wiktionary:About_Chinese/Hakka
- https://blgjts.moe.edu.tw/doc/tmt_compare.pdf
- Ministry of Education, Taiwan (2008)
- https://zh.wikipedia.org/wiki/Help:%E8%87%BA%E7%81%A3%E8%A9%B1%E5%9C%8B%E9%9A%9B%E9%9F%B3%E6%A8%99

A.5 Xiang

- https://en.wiktionary.org/wiki/Module:hsn-pron
- https://en.wiktionary.org/wiki/Wiktionary:About_Chinese/Xiang
- https://zh.wikipedia.org/wiki/%E6%B9%98%E8%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/Wiktionary:About_Chinese/Gan
- https://en.wiktionary.org/wiki/Module:gan-pron

A.8 Wu

- https://en.wiktionary.org/wiki/Wiktionary:About_Chinese/Wu
- https://en.wiktionary.org/wiki/Module:wuu-pron

A.9 Hakka

- https://en.wiktionary.org/wiki/Wiktionary:About_Chinese/Hakka
- https://en.wikipedia.org/wiki/Sixian_dialect
- https://en.wiktionary.org/wiki/Module:hak-pron