

Report on the Iu Mien—Chinese—English Dictionary Project

Greg Aumann and Pan Chengqian

Abstract:

This dictionary is a dictionary of Iu Mien as spoken in Laibin County, Guangxi Zhuang Autonomous Region, China. The intended audience includes Chinese and English speaking linguists. But the main audience is Iu Mien speakers. Thus the dictionary has some unusual features designed to make it simpler and more useful for Iu Mien speakers. The first is that it has definitions in Iu Mien. Hence it is like a combined monolingual and bilingual dictionary. The second is that the Chinese part of each entry includes a Chinese gloss, pinyin of the gloss and a Chinese definition. It is hoped that the dictionary will be useful in helping Iu Mien speakers to learn to read their own language and also to learn Chinese characters and pinyin.

This dictionary will have reverse Chinese—Iu Mien and English—Iu Mien indexes. Generating the reverse Chinese—Iu Mien index requires being able to sort Chinese in a standard order. The order most frequently used in modern dictionaries in China sorts characters first by pronunciation, then by stroke count and thirdly by stroke categories. This sort order is easy to use for Chinese speakers but difficult to implement. It requires disambiguation for characters with multiple pronunciations and the stroke counts and types for each character in the reverse index. The available databases of character information were generally not accurate and didn't contain all the necessary data so character data suitable for sorting Modern Simplified Chinese was developed.

Introduction to Iu Mien

Iu Mien belongs to the Mienic branch of the Hmong-Mien language family. It has more speakers than any other Mienic language. The bulk of the speakers are found in China and Vietnam. But there are significant numbers also in Laos, Thailand, the United States and Europe. In Asia Iu Mien speakers are grouped with speakers of other, mostly related languages, and called Yao. Thus the term Yao is not a very precise term when discussing linguistics.

There are three orthographies in use. A Thai based orthography and an Old Roman orthography are used in Thailand. The New Roman orthography is used outside of Thailand, especially in the United States and in Europe. The New Roman orthography is the preferred orthography also in China but its use is not widespread because of the low literacy rate in Iu Mien. Chinese and American Iu Mien cooperated in the design of the New Roman orthography (Purnell, 1987). It was designed to have transfer value to both English and Chinese Pinyin. The New Roman orthography is used slightly differently in China. The differences are relatively minor and are convertible in either direction, much like the differences between written British and American English.

Orthography chart

Consonants

/p ^h /	p	/t ^h /	t	/ts ^h /	c	/c ^h /	q-	/k ^h /	k	/ʔ/	-q
/p/	b	/t/	d	/ts/	z	/c/	j	/k/	g		
/b/	mb	/d/	nd	/dz/	nz	/j/	nj	/g/	nq		
/f/	f	/s/	s							/h/	h
/m/	m	/n/	n			/ɲ/	ny	/ŋ/	ng		
/m̥/	hm	/n̥/	hn			/ɲ̥/	hny	/ŋ̥/	hng		
		/l/	l								
		/l̥/	hl								
/w/	w					/j/	y				

The letter q represents two different phonemes. In syllable initial position /c^h/ and in syllable final position /ʔ/. The letters p, t, and k when used syllable finally are realised without aspiration. Note that the use of nasal symbols with the voiced stops indicates voicing not prenasalisation.

Vowels

/i/	i					/u/	u
/e/	e	/ə/	er			/o/	o
/æ/	ae	/a/	a	/a:/	aa	/ɔ/	or

Tones

33	-	53	v	24	x
31	h	231	z	11	c

The letters h, z and c signify consonants when used in syllable initial position and tones in syllable final position. There is a short high tone and a short low tone that occur only in closed syllables. These are considered allotones of the v and c tones respectively. In this dictionary surface tone and not underlying tone is written in the orthography. Tone sandhi is represented in the orthography by the resultant tone and a hyphen between the syllables. In this orthography word boundaries are not marked. Spaces indicate syllable boundaries where there is no tone sandhi.

The Dictionary Team

This dictionary is a joint project between the Central University for Nationalities, Beijing, SIL East Asia Group and Biola University. The personnel compiling the dictionary are: Prof. Pan Chengqian, Prof. Deng Fanggui, and Zhuo Xiaoqing from CUN, Prof. Herbert Purnell from Biola and Greg and Misako Aumann from SIL. Initial work began on the dictionary in 2001. We hope that it might be completed in approximately another two years. However, it is very difficult to be certain where dictionaries are concerned. It is expected that the dictionary will have roughly seven thousand entries when completed.

Previous Dictionaries

This dictionary is by no means the first dictionary of Iu Mien. Savina's (1926) was the first to include a variety of Iu Mien from Vietnam. Lombard and Purnell (1968) is of the Thailand variety of Iu Mien. Mao et. al. (1992) is a Chinese—Iu Mien dictionary rather than a Iu Mien—Chinese dictionary. Thus it doesn't really describe the Iu Mien language in any depth. It is from Longsheng county in northeast Guangxi, China. Smith Panh's dictionaries (1995, 2002) are of Iu Mien spoken by immigrants to the United States from Laos. Our dictionary is the first dictionary that describes a Chinese variety of Iu Mien.

Target Audience

The target audience for any dictionary needs to be considered quite carefully so as to decide appropriately what should be included and how the dictionary should be arranged. For this dictionary the main audience is Iu Mien speakers themselves. More precisely it is speakers of Iu Mien living in China. Some features will be included to increase its usefulness to linguists and learners of Iu Mien, both Chinese and non-Chinese, but they will not be the main consideration. It is particularly hoped that this dictionary might be a useful tool in helping Iu Mien to learn to read both Iu Mien and Chinese. In recent times all Chinese school children are required to learn some English and the English aspects of the dictionary will certainly increase its acceptability. The English definitions and especially the English reverse index may also prove useful for this even though they are primarily for the benefit of non-Chinese linguists and learners of Iu Mien.

Dictionary Format

Preliminary nature

This dictionary project is far from being finished. So far only the first draft has been completed but not even all of that has even been keyboarded. So much of what is described in this paper is fairly preliminary in nature and certainly some will be changed before the dictionary is finally published.

Overall Structure

The bulk of the dictionary will consist of three sections. These will be the dictionary proper, a reverse Chinese index and a reverse English Index. The dictionary proper will contain definitions of Iu Mien headwords in Iu Mien, Chinese and English. The reverse indexes will be fairly simple and refer back to the relevant Iu Mien headword. There will of course be some front matter but the content of that is yet to be decided.

Iu Mien Definitions

The inclusion of Iu Mien definitions is the main innovation of this dictionary. This means that it combines the functions of a monolingual and a bilingual dictionary. The motivation for including definitions in Iu Mien comes directly from consideration of the target audience. Including them makes the dictionary much more useful to the Iu Mien themselves.

The Iu Mien definitions also increase the usefulness of the dictionary to linguists. This is because, together with the example sentences, they provide a small corpus of Iu Mien text, which can be analysed for various linguistic features. We expect they will also prove useful to learners of Iu Mien. They have certainly already proved useful to us in compiling the dictionary. We have used computer programs to compare the words that are in the definitions and examples with those used as headwords in the dictionary. This has enabled us to find spelling errors and also words that were forgotten. A particularly horrifying example of a word that was forgotten in the first draft was *bun* ‘give’. This word is very common and has important grammatical functions in Iu Mien, as it does in many other languages of the area.

Chinese Models

As much of the target audience for this dictionary live in China we are intending for the dictionary to resemble as much as possible Chinese dictionaries that are in widespread use or held in high regard in China. The *Dictionary of Modern Chinese* is certainly the most highly regarded dictionary of the modern language in China. This dictionary has sold more than 40 million copies. The most widely used dictionary seems to be the *Xinhua Character Dictionary* which has sold more than 380 million copies. This dictionary is commonly used by Chinese school children and thus is also a good model to follow. The value of using influential Chinese dictionaries as models is that it makes it easier for users to transition between dictionaries. A hypothetical example might be Iu Mien children in a bilingual education programme. They might begin their education in Iu Mien and use the Iu Mien—Chinese—English dictionary. Later they would transition to education in Chinese and move to using a Chinese dictionary. Of course the reverse transition could occur if you have Iu Mien educated in Chinese and coming across our dictionary and trying to use it.

There are of course some significant differences between our dictionary and the dictionaries we have chosen as models so there needs to be some discernment used. Not every aspect should be used as a model. Many features are purely cosmetic and there is no reason not to adopt them. For example our model dictionaries use two columns and use a line to separate the columns and another line across the top of the columns. Pinyin in these dictionaries is typeset in a sans serif typeface. Sense numbers are white in a black circle in sans serif type e.g. **①②③④⑤**. Examples are separated from definitions by a colon and multiple examples are separated by a vertical line in the *Dictionary of Modern Chinese* and by a slanted vertical line in the *Xinhua Dictionary*. We plan to use Chinese style brackets where suitable, e.g. [] . Initially we were using the Multidictionary Formatting (MDF) approach of consistently using bold type for all Iu Mien and normal type for English. This was not a very helpful arrangement in our dictionary. Having the Iu Mien definitions in bold made it difficult to distinguish the headwords from the definitions as the Iu Mien definition came near the front of the entry. As the dictionary also includes quite a few Iu Mien example sentences that were also in bold there was a lot of bold text that took a lot of space, overwhelmed the eye and, worst of all, was difficult to read. This was a particular concern as the majority Iu Mien users are expected to be new readers and more interested in reading the Iu Mien than the Chinese or the English. Thus we decided to use regular type for Iu Mien (except for headwords) and italics for English. This results in a small decrease in readability but it was not felt to be a concern as readers interested in the English probably have enough reading experience for it not to matter. Most dictionaries use italics for parts of speech but we are using Chinese abbreviations for parts of speech which are set off in parentheses.

Dictionaries of Chinese almost always put compounds as subentries of the first character in the word. Iu Mien has a similar compounding structure. Our dictionary team has not yet decided whether to list compounds as subentries of the first syllable or not. If we use subentries then that resembles Chinese dictionaries. However, there are features of the Iu Mien language that make the use of subentries less useful than they are in Chinese. In Iu Mien the first syllable of a word fairly frequently occurs with a reduced form. There are many words where the reduced form is so prevalent that it is not possible to determine what the full form of the compound is. The reduced forms make it difficult for inexperienced users to look up words in the dictionary because the reduction often extends to changing the first letter of the word. Thus *biouh-gomh* ‘pomelo’ has the following forms that occur in free variation: *buh-gom*, *gerh-gomh* and *beih-gomh*. We are still undecided about the best way to help the user to find the definition of the the word he or she is looking for despite this variation.

Entry Structure

Each lexical entry begins with the Iu Mien headword and then a list of senses. Senses generally begin with a part of speech in Chinese, a Iu Mien definition, a Chinese gloss followed by the pinyin of the Chinese gloss and a Chinese definition. After the definitions there may be some example sentences with translations into Chinese and English. Some entries also have some lexical functions that link to classifiers, synonyms, antonyms etc. The three different languages that are used in this dictionary consistently occur in the order Iu Mien, Chinese and then English. The Iu Mien comes first because it is the language the dictionary is describing. We felt that it was much better to put the Chinese before the English. The reason is that Iu Mien and English both use a Roman script and putting Chinese between them with its markedly different script separates them very clearly. This is particularly helpful to Chinese who are not used to Roman scripts but is also helpful for other users. Some example entries are included below.

Selected Dictionary Entries

caangv ① (动) *i bung nzaeng*. 抢 *qiǎng*: 双方争夺。 *two parties fight each other*: ~ *jouh* 抢球 *snatch the ball from someone's hands*. ② (动) *nzaeng ndaangc*. 抢 *qiǎng*: 争先。 *compete for*: ~ *jenv gorngv wac* 争着说话 *try to speak first* ③ *siepv*. 抢 *qiǎng*: 快。 *fast*: ~ *zuangx* 抢种 *plant in a rush*.

caangv daapv (动) *baeqc laanh naaic nyei sic*, *caangv ndaangc daapv cuotv daaih*. 抢答 *qiǎngdá*: 他人提问的问题, 抢先把它回答出来。 *hurry to answer a question before others*.

ceuv ① *cuh haic*. 吵 *chǎo*: 很嘈杂。 *very noisy*: *Zorqv yie ~ nyei av*. 把我吵醒了。 *The noise*

woke me up. ② (动宾) *nzaeng*. 吵 *chǎo*: 争吵。 *quarrel*: *Ninh mbuo i dauh ~ jiax*. 他们两人吵架。 *Two people are quarreling*.

ceuv jiax (动宾) *nzaeng jiax*. 吵架 *chǎojià*: 争吵。 *quarrel*: *Ninh mbuo i hmuangv mv ~ jiex jiax*. 他们夫妻俩从来不吵架。 | *Gorngv duh leiz mv duqv ~*. 讲道理, 别吵架。 *Pay attention to common sense, do not quarrel*.

ciex *mv ziangx*. 斜 *xié*: 不正。 *Naaiv diuh ndiouh mv ziangx, ~ nyei*. 这根柱子不正, 斜的。 *This pillar is not straight*. 〔反〕 *ziangx*.

Selected Reverse Chinese Index Entries

扒拉 *bāla*: — *biuih*;
(名) *pan*;
(动) *biah*.

拔 *bá*: (动) *baeng₂* ①;
(动) *cun₂*.

拔草 *bácǎo*: (动宾) *baeng miev*.

白菜 *báicài*: (名) *laih-baeqc*.

百 *bǎi*: (数) *baeqv₁*.
百叶 *bǎiyè*: — *ngong ziepc nyeic pin*;
(名) *baeqc-ipc*.

摆 *bǎi*: (动) *baaiv* ①;
(动) *baaiv* ②;
(动) *baaiz*.

摆动 *bǎidòng*: (动) *baaiv dongz*.

摆架子 bǎijiàzi: (动宾) baaiv jiax zeiv.
摆设 bǎishè: — baaiv ㉔;
(动) baaiv ㉕.
摆摊 bǎitān: — baaiv taan.
柏油 bǎiyóu: (名) baeq̄c youh.

败 bài: (动) baaic ㉑.
败兵 bàibīng: (名) baaic baeng.
败国 bàiguó: (名) baaic guoqv.

Dictionary Software and Procedure

We are using Shoebox/Toolbox¹ for data entry, Concurrent Versions System² (CVS) for revision tracking and synchronisation of data files, and custom software written in Python³ for dictionary consistency checks and generating formatted output.

Entries are generated by working through the list of initials in the language and testing which finals and finals go with them. The above mentioned dictionaries of Iu Mien are also consulted as a source of entries. The dictionary entries are first written on paper approximately 10 cm by 15 cm. They are then reviewed with corrections made on the paper. Then the entries are keyboarded and a draft English translation made. Entries that are not clear or are problematic in some way have questions attached to them. These entries are then printed out with the questions. This output leaves half the page blank for space for written corrections. We haven't worked our way through a whole revision cycle yet so the revision procedures are not yet worked out in detail.

We are in the process of classifying the entries into semantic domains. We hope that this will help the development of better definitions and also help us to discover missed headwords. We think that semantic domains can help us develop more consistent definitions. Semantically related words have parts of their meaning the same and parts that are different. Ideally they should be defined as a set so that both the similarities and differences are brought out, as much as possible, by the definition. Currently we plan to use semantic domains as a tool to improve the quality of the dictionary and have no plans to include them in the final dictionary.

The reason for writing custom software for generating the formatted output is that the Multidictionary Formatting (MDF) software (Coward and Grimes, 1995) that is part of Shoebox/Toolbox is problematic for our dictionary. It was designed for Austronesian languages and has significant limitations when dealing with East Asian languages. Some of these limitations are not severe and can be worked around. For us the critical limitation was that it doesn't cope very well with Chinese. In particular it cannot generate reverse Chinese indexes. But also its method of including subentries is very awkward to use if you want more than a few subentries under a lexeme. Furthermore it cannot sort subentries. Also its conceptual model of subentries is quite different to the model used in Chinese dictionaries. At the beginning of this project we were planning to make extensive use of subentries so this was an important consideration. Another consideration is that its customisability is limited and it didn't seem to offer us the flexibility that we needed. We wanted to make the dictionary look as Chinese as possible and to include unusual features such as including a Chinese gloss, pinyin of the gloss as well as a Chinese definition.

A further limitation of Shoebox/Toolbox and hence also of the MDF software is that it doesn't support the sort order used for Iu Mien. The sort order for Iu Mien sorts syllables which are identical except for tone next to each other, which is clearly a desirable property, e.g. *ba*, *bac*,

¹Available from <http://www.sil.org/computing/toolbox/>

²See <http://www.cvshome.org>

³See <http://www.python.org>

bah, bav, bax, baz, baaic, baaiv, etc. The difficulty is caused by the fact that some letters in the orthography are used to represent both a segment and a tone. There is no ambiguity as in syllable initial position they represent a segment and syllable final position a tone. However, the Shoebox/Toolbox collation algorithms cannot distinguish these two uses and so sort incorrectly. This is not a serious concern for data entry but it is for printed output. The sorting problem can be worked around by using a sort string in the the \lx field and putting the head word in the \lc citation form field. This work around causes some other difficulties implementing this kind of sorting in the custom software is helpful. The development of this custom software has been quite time consuming.

Chinese Reverse Index

The inclusion of a reverse Chinese—Iu Mien index is quite an important feature of the dictionary. It is, however, also the most problematic feature. The reason is because sorting Chinese is quite a bit more complex than sorting most other languages and the Shoebox/Toolbox Dictionary Formatting software that we are using for data entry doesn't support the creation of reverse indexes in Chinese.

Sorting Chinese

There are numerous ways of sorting Chinese. The method used in the *Dictionary of Modern Chinese* and in many other modern dictionaries is to sort words by characters. Characters are sorted first by their pronunciation, second by the number of strokes in the character and third by the categories of the strokes.

The pronunciation of Modern Standard Chinese is written using the Pinyin Romanisation system. Ignoring the tone marks and other diacritics pinyin is sorted using English alphabetical order. There are two letters with diacritics that are not tone marks: u-diaeresis (ü) and e-circumflex (ê). Ignoring tones, syllables with a diacritic sort after the syllable that is identical except for the diacritic. This is a common ordering principle for languages that have diacritics. Tones are less significant than the diaeresis or circumflex in the sort order. Neutral tone is sorted after tones 1 to 4 rather than before. Thus the following are in correct sort order: lou, lū, lú, lǔ, lù, lu, lū, lú, lǔ, lù, lū, luán, luǎn, luàn, lǔě, lǔè, lun.

If two characters have the same pronunciation, including the same tone then they are sorted in order of increasing number of strokes in the character. Thus 艾 ài is listed before 爱 ài because 艾 has five strokes as compared to 爱's ten strokes. But there are also quite a few instances where two characters have the same pronunciation and the same number of strokes. In that case they are ordered using the third aspect which is to sort according to the types of strokes. In this aspect each stroke is given a number 1 to 5: a horizontal stroke or 横 (e.g. 一) is 1, a vertical stroke or 竖 (e.g. 丨) is 2, a left falling stroke or 捺 (e.g. 丿) is 3, a right falling stroke or 撇 or 点 (e.g. 丶) is 4 and all other strokes are 5 (e.g. ㇇ 乙 乚). Then the stroke numbers are written in the order in which the strokes are written. Thus 艾 ài is 12234. Strokes are generally written in a top to bottom, left to right order. Thus for example 璦 ài, 14 strokes is 11213443451354 and sorts before 霏 ài, 14 strokes 11543443451354, which sorts before 暧 ài, 14 strokes, 25113443451354.

There are two strokes for which the category is not obvious. The first is a left to right rising stroke or 提. It is considered the same as a horizontal stroke and thus is category 1. The

second more difficult case is the 竖钩 stroke (J). This stroke is a vertical line with a small hook at the bottom. The only dictionary that we have found with an explanation of whether this is a category 2 or category 5 stroke is the *Dictionary of Modern Chinese*. Its explanation of the order of the radical index puts it in category 5. However, the explanation of the ordering of the main body of the dictionary, like other dictionaries, is ambiguous about this point. In contrast, most dictionaries explicitly state that a horizontal stroke with a hook or 横钩 (冫) is a category 5 stroke. Examination of the order of the characters in this dictionary is more consistent with treating it as a category 5 stroke rather than a category 2 stroke. Determination of which categorisation of the 竖钩 stroke a dictionary uses is not straightforward because there are often small ordering errors and so a large number of cases need to be examined until a clear pattern emerges.

There are a few instances of different characters that have the same pronunciation, same number of strokes and the same categories for all their strokes. In that case there is no standard order and in our dictionary we are putting them in Unicode code point order. The Unicode code point order is not meaningful to users but this is not really a concern. The whole reason for having a sort order is to make the sought after character or word easy to find in a list. There are very few instances where the Unicode code point order is used to sort characters that are the same in the other three respects. When the Unicode code point is used there will be the character being looked up will in the worst case be only one position removed from where it might have otherwise been. Thus uncertainty at this level of the sorting will have virtually no impact on the ease of finding the character in the list.

Sorting Data

The sorting method as described above is fairly well understood and on the whole is not difficult to implement. There are two problem areas however. The first is caused by characters that have multiple pronunciations. The second difficulty is obtaining accurate data to use for sorting the characters, i.e. pronunciation, stroke count and stroke categories data.

First, we will consider the problems of characters with multiple pronunciations. Approximately 5 % of characters have more than one pronunciation. Some have three or four different pronunciations. The character 欸 has six different pronunciations listed in the *Dictionary of Modern Chinese*. It should be noted that we are discussing here the pronunciations of characters that are used for sorting in a standard order. Thus we are only considering pronunciations that are recognised as standard, i.e. are listed in the *Dictionary of Modern Chinese*. We choose this dictionary as defining the standard pronunciation because one of its main purposes is language standardisation, it is held in high regard and it pays particular attention to pronunciation. This approach of considering only pronunciations recognised as standard is different from the usual descriptive linguistic approach of recognising the pronunciations that people actually use because our purpose is sorting and the reason for sorting is to make it easy to find a word which is only achievable if there is a generally agreed upon standard.

When we say that characters have multiple pronunciations that means that their pronunciation will be different in different contexts. In any given context only one pronunciation will be used. Other pronunciations frequently, but not always, signal a difference in meaning or function. For example: 着 zhāo ‘a move (as in chess)’, 着 zháo ‘touch, feel’, 着 zhe ‘(particle marking continuous aspect)’, 着 zhuó ‘wear (clothes)’. The end result of characters having pronunciations which are different in different contexts means that a computer cannot

correctly sort a list of words without additional information over and above the pronunciations for each character, their stroke counts and stroke categories. It must also know which pronunciation is correct in each context. This could be done by using a Chinese lexicon that contained all the contexts. This would not be completely reliable and would also require the compilation of a large Chinese computer lexicon. A better alternative for the purposes of this dictionary is to include the pronunciation with each entry in the Chinese reverse index. This is not too much of an additional burden as the bulk of the reverse index will come from the Chinese glosses and we already plan to include pinyin for them to aid Iu Mien trying to learn Chinese and also any foreign users whose Chinese might be limited. Also we are using the Chinese lexical data that we have to guess the pronunciations and then checking them manually. An important consequence of the multiple pronunciations is that it is not possible, in general, to say whether one character sorts before another unless specific contexts are also given. Thus characters can occur in multiple places in a sorted list.

The compilation of the data necessary for sorting was not straightforward. The main reason is that although there is some data on the internet for pronunciations and stroke counts, most of it is lacking in accuracy and there was no data available for download for the stroke categorisations. To compile the sorting data we first cross-matched different sources to find errors. Once the pronunciations and stroke counts were mostly correct, the characters were sorted and cross-matched against the *Dictionary of Modern Chinese*. This provided accurate pronunciations. We used the radical indexes in the dictionary to divide the characters into their subcomponents in a recursive fashion. We could then categorise the strokes of the simpler components and calculate the stroke categories of the whole characters by combining the stroke categories of their respective subcomponents. See the table below, which shows this for the character 癌 *ái* ‘cancer’:

Character	Components	Stroke Categories
癌	疒 品 山	41341251251251252
疒		41341
品	日 品 山	251251251252
日	日 口 口	251251251
口	口 口	251251
口		251
山		252

Unicode and the Dictionary of Modern Chinese

The following table gives numbers of characters in the Dictionary of Modern Chinese and which part of Unicode they are included in. This is relevant because different sections of Unicode have different levels of support in computer systems. Support for the first unified ideographs section is quite good, though some less common characters are not in the pinyin input methods. It is not easy to find fonts which include the characters in extension A and they are very unlikely to be included in the input methods. However, things are much worse when one considers extension B. There is not much software at all that supports these characters as yet. Fortunately not many of them are in current use. Worst of all, of course, is the nearly four hundred characters in the dictionary that are not in Unicode at all. A small number of these may actually be in extension B. Hopefully the remaining characters will be in the forthcoming set of characters.

Section of Unicode 4.0	Total Characters in Unicode Section	Characters in <i>Dictionary of Modern Chinese</i>
CJK Unified Ideographs	20,902	8,268
CJK Unified Ideographs Extension A	6,582	142
CJK Unified Ideographs Extension B	42,711	13
forthcoming	40,000–60,000	
not found		372
other (○)		1
Total	70,195 (+40,000–60,000)	8,796

Unicode 4.0 and Modern Chinese Characters.

Conclusion

We have described the structure and format of the Iu Mien—Chinese—English Dictionary. We have concentrated on some of its more unusual features that result from the primary target audience being speakers of Iu Mien. Making dictionaries with Chinese as one of the languages involved causes many complications for lexicography software such as Shoebox/Toolbox. Thus much of our work on this project has been to resolve these complications, and we have described this at some length.

Bibliography

- No author. 2004. *Xinhua Character Dictionary* 《新华字典》 10th edition. Beijing: Commercial Press.
- Coward, David F. and Charles E. Grimes. 1995. *Making dictionaries: A Guide to Lexicography and the Multi-Dictionary Formatter*. Version 1.0. Waxhaw, North Carolina: Summer Institute of Linguistics.
- Dictionary Compiling Unit, Language Research Institute, Chinese Academy of Social Sciences. 2002. *Dictionary of Modern Chinese: 2002 enlarged edition* 《现代汉语词典：2002年增补本》 Beijing: Commercial Press.
- Jenkins, John H. 1999. “New Ideographs in Unicode 3.0 and Beyond.” Paper presented at the 15th International Unicode Conference, San Jose.
<http://developer.apple.com/fonts/WhitePapers/IUC15Han.pdf>
- Lombard, Sylvia J. (compiler) and Purnell, Herbert C., Jr. (ed.). 1968. *Yao-English Dictionary*. Ithaca, New York: Department of Asian Studies, Cornell University.
- Máo Zōngwǔ 毛宗武 (ed.), Zhào Xūn 赵勋, Zhèng Zōngzé 郑宗泽 and Méng Cháojí 蒙朝吉. 1992. 《汉瑶简明分类词典(勉语)》 [Concise Chinese-Yao Topical Dictionary: Mien Language]. Chengdu: Sichuan Nationalities Press.
- National Language and Writing System Committee, China News Publisher. 1988. 《现代汉语通用字表》 [Chart of Commonly used Characters in Modern Chinese]. In Language

- and Literature Press (ed.). 1997. 《语言文字规范手册》 [Handbook of Language and Writing Standards]. Beijing: Language and Literature Press.
- Panh, Smith. 1995. *Mienh In-Wuonh Dimv Nzangc Sou: Mien-English Everyday Language Dictionary*. Visalia, California: Smith and Jenny Panh.
- . 2002. *Modern English-Mienh and Mienh-English Dictionary*. Victoria, British Columbia: Trafford.
- Purnell, H. C. 1987. “Developing Practical Orthographies for the Iu Mien (Yao) 1932–1986: A Case Study.” *Linguistics of the Tibeto-Burman Area* 10(2):128–141.
- Savina, F.M. 1926. Dictionnaire Français-Mán: Précédé d'une note sur les Mán Kim-Đi-Mun et leur langue. *Bulletin École Française d'Extrême-Orient*, 26.11–255.