Proposal of Taiwanese CJK Unified Ideographs

Doc Type: Proposal

Title: Proposal of Taiwanese CJK Unified Ideographs

Source: Peter Burkimsher

Status: Individual Contribution

Action: For independent feedback before consideration by the UTC & IRG

Date: 2018-07-21

Abstract

The Taiwanese/Minnan/Hokkien/閩南語 dialect of Chinese is spoken by 68 million people, particularly in Taiwan and Fujian. There is no standardized written language, but religious texts and song lyrics do exist, using Hanzi or mixed orthography. When studying the Bible, I discovered 7 missing CJK ideographs. In this proposal, I present scans of printed examples of these characters, frequency counts, and verse references.

1   Introduction

Sinitic languages form a dialect continuum, and different varieties are often not mutually intelligible. Mandarin Chinese and other dialects can differ as much as Romance languages in Europe[1]. The number of native speakers of Chinese dialects is comparable[2] to other major world languages.

The Taiwanese/Minnan/Hokkien/閩南語/台語 dialect is prevalent in southern China and Taiwan[3]. The first Bible translation was made in 1884, and introduced the POJ Romanization method[4]. A second edition was released in 1933. The first transliteration into Chinese Hanzi was made in 1996.

The education systems in Taiwan and China encourage the use of Chinese characters, and therefore many native speakers are unfamiliar with the POJ Romanization method. Mixed orthography[5] texts, with Chinese and POJ Romanization, are used when the appropriate Chinese Hanzi is not known. The translators of the 1996 Bible and 2009 PCT hymnal applied some rare characters in order to express the entire lexicon using Hanzi. Those are the characters I present in this proposal.

---

[1] Norman, Jerry (1988), Chinese, Cambridge: Cambridge University Press, ISBN 978-0-521-29653-3.

[2] http://peterburk.github.io/i2018n/#chinese

[3] https://joshuaproject.net/languages/nan

[4] https://amoytainan.wordpress.com/2010/08/21/minnan-hokkien-amoy-taiwanese-tainanese/

[5] http://bible.fhl.net/new/read.php?VERSION7=bklcl&VERSION8=bklhl&strongflag=1

2   Suitability for Inclusion

The Unicode Glossary[6] defines a character (4) as "The English name for the ideographic written elements of Chinese origin." The proposed characters are not currently included in the Unicode Pipeline Table or BETA. Including these characters into Unicode would allow plain-text representation of religious texts that are actively used by churches in Taiwan.

3   Evidence of Use in Running Text

Four of the missing characters were discovered when trying to parse an electronic version of the 1996 Bible available from the Lingshyang[7] website. The text is interspersed with inline JPG images representing 52 different characters. Of those, 47 characters are already supported in Unicode, and 1 is in the pipeline. See Appendix 1 for a table with all the unrepresented characters from this data source.

The fifth missing character was found from a collection of character JPGs used by Dexian Church in Appendix B. It is also found in the 2009 hymn book character references[8] published by the PCT (Presbyterian Church of Taiwan). There are 238,372 members among 1205 congregations[9] – the largest Christian denomination in Taiwan.

The sixth and seventh characters were found in the PCT2009 character references, tabulated in Appendix C. In order for the PCT authors to prepare that document, Private Use Area codes were used. When possible, I replaced the characters with the corresponding Unicode equivalents.

There is a lack of awareness of existing characters. The DOC versions of the PCT2009 hymn book obtained from Dexian Church use an image to represent 愞 (tok) in hymn number 534. More work must be done to provide input methods for these characters. The Taigi-Hakka IME software[10] is usually used in Taiwan, and that program lacks the 愞 (tok) character.

As a workaround for the text encoding problem, PPT files used in church use an image to represent each entire slide. The folder of all 657 hymns is 898 MB in size, which is far larger than the 44 MB required for a comparable English-language collection of PPT formatted song lyrics. Including the proposed characters in Unicode will allow plain-text software for managing songs, chords, and Bible reading to handle the Taiwanese language effectively.

Photographs in this document are taken from the from the 1996 Taiwanese Han Character Bible, ISBN 957-99771-4-3, and the PCT Hymnal Sèng-si Seventh Edition, 4th printing April 2016, ISBN 978-986-7134-30-1. Regarding copyright of these pages, fair use is asserted for the small portions sampled in this research.

---

[6] http://unicode.org/glossary/#character

[7] http://www.lingshyang.com/bible/taiwan_Bible/2ch/2ch14.htm

[8] http://hymn.pct.org.tw/modulecontrol/GetModule.aspx?fid=U2006121900001_20110414103952_3

[9] http://english.pct.org.tw/enWho_int_Sta.htm

[10] http://taigi.fhl.net/TaigiIME/

4   Character Properties

The proposed characters are as follows:

Ex. 1	Ex. 2	Ex. 3	Ex. 4	IDS	PUA	Name	References
鬠	髟意	髟意		⿰髟意	U+F7F2	chhang	Job 39:19, Job 4:15
瘄	瘄	瘄	瘄	⿸⿰粒	U+F7F3	liap	1 Sa 5:6, 1 Sa 5:9, 1 Sa 5:12, 1 Sa 6:4, 1 Sa 6:5, 1 Sa 6:11, 1 Sa 6:17, 2 Ki 20:7, Exo 9:9, Exo 9:10, Exo 9:11, Exo 9:11, Isa 38:21, Lev 13:18, Lev 13:19, Lev 13:20, Lev 13:23
躉	躉	躉		⿱⿳亠口冖足	U+F7F4	lo-	Deu 1:28, Deu 2:10, Deu 9:2
犆	犆	犆		⿰周	U+F7F6	tiau	1 Ch 17:7, 1 Sa 24:3, 2 Ch 14:15, 2 Ch 32:28, 2 Sa 7:8, Amo 6:4, Est 8:10, Eze 25:5, Eze 34:14, Eze 34:14, Hab 3:17, Jhn 10:1, Jhn 10:7, Jhn 10:16, Jug 5:16, Mal 4:2, Num 32:16, Num 32:24, Num 32:36, Pro 14:4, Psm 50:9, Psm 68:13, Psm 78:70, Zep 2:6
呣	呣	呣	呣	⿰口毋	U+F1DC	m̄	PCT2009 Hymnal 127, 145, 213, 243, 245, 253, 352, 467, 530, 557, 634
萏	萏	萏	萏	⿱帕	U+F1E7	phè	PCT2009 Hymnal 70, 95, 588, 604, 614, 638
莊	莊	莊	莊	⿱吐	U+F1E8	puh	PCT2009 Hymnal 7, 76, 195, 438, 578

The IDS column refers to the Ideographic Description Sequence, which is the standard way to represent characters that are not available in Unicode.

The PUA column refers to the Private Use Area codes in the BabelStone Han font, available here:

http://www.babelstone.co.uk/Fonts/PUA.html

High-resolution photographs of the full pages, so that the characters can be viewed in context, are available on Flickr:

https://www.flickr.com/photos/150180606@N08/albums/72157693398863955/with/28657064117/

4.1   Character in Pipeline

One of the missing characters is already in the Unicode inclusion pipeline. It is listed here along with its references to reiterate the urgency of adding it in the next revision of the CJK ideographs.

瘑	ko.jpg	UTC-02663	UK-2015	104.1	0769.011	□疒哥	UTCDoc L2/15-260 1351

There are 62 references for the ko character in the 1996 Bible. These are:

2 Ch 26:19, 2 Ch 26:20, 2 Ch 26:21, 2 Ch 26:21, 2 Ch 26:23, 2 Ki 15:5, 2 Ki 5:1, 2 Ki 5:3, 2 Ki 5:6, 2 Ki 5:11, 2 Ki 5:27, 2 Ki 5:27, 2 Ki 7:3, 2 Sa 3:29, Deu 24:8, Exo 4:6, Gen 34:30, Lev 13:2, Lev 13:3, Lev 13:8, Lev 13:11, Lev 13:12, Lev 13:13, Lev 13:15, Lev 13:20, Lev 13:25, Lev 13:25, Lev 13:27, Lev 13:30, Lev 13:42, Lev 13:43, Lev 13:44, Lev 13:45, Lev 13:47, Lev 13:49, Lev 13:51, Lev 13:52, Lev 14:2, Lev 14:3, Lev 14:3, Lev 14:7, Lev 14:32, Lev 14:44, Lev 14:54, Lev 14:55, Lev 22:4, Luk 17:12, Luk 4:27, Luk 5:12, Luk 5:14, Luk 7:22, Mak 1:40, Mak 1:42, Mak 14:3, Mat 10:8, Mat 11:5, Mat 26:6, Mat 8:2, Mat 8:3, Num 12:10, Num 12:10, Num 5:2

5   Example Font

The BabelStone Han font includes the first 4 proposed characters.

http://www.babelstone.co.uk/Fonts/Han.html

For the PCT2009 Hymnal, I need to find the original font, or redraw the characters.

6   Anticipated Objections

It is possible that some of the characters may already exist in the CJK Unified Ideographs, but escaped the scrutiny of my research. The IDS data used to search for existing characters was sourced from kawabata on Github[11], and queried using Pingtype[12], software developed by the author for entering Chinese characters.

Further discussion is recommended regarding the correct construction of the chhang character. Is the correct IDS □□髟煮 (Lingshyang) or □髟煮 (paper)? This is probably merely a difference in font.

---

[11] https://github.com/cjkvi/cjkvi-ids

[12] https://pingtype.github.io

7 Drawing the characters

The proposed characters are CJK Ideographs used in Chinese dialects, most notably Taiwanese. It is suggested that the ROC Ministry of Education[13] be consulted regarding the correct construction of these characters.

8 Authors

The address for correspondence is:

Peter Burkimsher
94, Route du Pont de la London
F-01170 Chevry
France

Email: peterburk@gmail.com

Tel: +1 210 369 8741

Other contributions include:

Searching for characters: Richard Cook, rscook@unicode.org

Proofreading: Grant McLean, grant@catalyst.net.nz

9 Summary and Conclusion

The Bible is the most translated book in the world, and comprises the largest written corpus for some languages[14]. It is important to allow the Presbyterian Church of Taiwan to continue conducting religious services in the Taiwanese language without resorting to inline images or mixed orthography. For believers to quote certain passages using social media, it is essential to express those in plain text. Therefore I implore the Unicode Consortium to include these CJK Ideographs as soon as possible.

References

(should references be kept as footnotes or listed at the end?)

A. Appendix 1 – Table of Lingshyang Characters

---

[13] http://dict.revised.moe.edu.tw/cbdic/

[14] https://en.wikipedia.org/wiki/Hakka_Chinese#Writing_systems

These characters are used in the 1996 Taiwanese Han Character Bible, available at
http://www.lingshyang.com/bible/taiwan_Bible/2ch/2ch14.htm.

JPG	Character	Hex	Filename
	儣	23350	boe.jpg
	(proposed)		chhang.jpg
	跙	47D9	chhih.jpg
	趱	27F75	chong.jpg
	勢	20895	gau.jpg
	脑	26706	ham.jpg
	扠	39D2	hiat.jpg
		2A736	in.jpg
	揪	22BE9	jih.jpg
	扮	39CE	khoe.jpg
	咱	20CBC	khu.jpg
	噡	20F3B	ki.jpg
	(pipeline)	UTC-02663	ko.jpg
	襈	27794	koa.jpg
	焒	24259	kun.jpg
	跜	47DC	lap.jpg
	(proposed)		liap.jpg
	蹽	8E7D	liau.jpg
	摺	3A28	liu.jpg
	(proposed)		lo-.jpg
	喀	21014	lo.jpg
	抙	22B33	lut.jpg
	�	22AEB	lut2.jpg
	襛	2771E	moa.jpg
	瞱	25349	nih.jpg
	囕	2330D	nit.jpg
	囕	2330D	nith.jpg
		2CFEA	oh.jpg
		2C9B0	peh.jpg
	箠	25C14	phin.jpg
	酾	2BC54	phoe.jpg
	坡	24FCE	pit.jpg
	殳	47DD	poa.jpg
	猯	2479A	siau.jpg
	淌	23EF8	siong.jpg
	瘗	24EC4	sui.jpg
	潜	23E04	tam.jpg

	啲	5572	teh.jpg
	冇	20547	teng.jpg
	扐	2BF18	tha.jpg
	甯	25981	thang.jpg
	迌	28468	thit.jpg
	蟶	274B4	thoa.jpg
	跈	27FEC	thun.jpg
	胼	3B39	ti.jpg
		2B8B6	tiam.jpg
	(proposed)		tiau.jpg
	扡	62D5	tioh.jpg
	虰	2B2A5	to.jpg
	煍	243F8	toh-.jpg
	愗	2272C	tok.jpg
	廐	22285	tui.jpg

B. Appendix 2 – Table of Dexian Church Characters

These characters were found in the Dexian Church data source, obtained along with lyrics for the hymns from the projector computer at 德賢教會 in Kaohsiung. The filenames may be useful for providing alternative pronunciations.

JPG	Character	Hex	Filename
		2B8B6	diam.jpg
	煍	243F8	do.jpg
	扤	39D2	hiat.jpg
		2A736	in.jpg
	蹴	280DF	pwang.jpg
	啲	5572	teh.jpg
	(proposed)		不.jpg
	(proposed)		不要.jpg
	犦	23350	未.jpg
	甲	7532	甲.jpg
	矖	2330D	聶.jpg
	坙	3639	追.jpg

C.   Appendix 3 – Table of PCT2009 Characters

This table may be downloaded from:
http://hymn.pct.org.tw/modulecontrol/GetModule.aspx?fid=U2006121900001_20110414103952_3

ct	標楷	新細明	白話字	片語	Character	Hex
1	儎	儎	bōe		儎	23350
2	勢	勢	gâu		勢	20895
3	抾	抾	hiat		抾	39D2
4	佪	佪	in			2A736
5	嘰	嘰	ki	Ki-lō͘-peng	嘰	20F3B
6	嚕	嚕	lō͘	Ki-lō͘-peng	嚕	21014
7	呣	呣	m̄	m̄-sī		(proposed)
8	襪	襪	moa		襪	2771E
9	娜	娜	nà	kng-nà-nà		2DD34
10	瞴	瞴	nih		瞴	25349
11	傲	傲	oh	oh-tit		2CFEA
12	跖	跖	peh	peh lâu-thui		2C9B0
13	萜	萜	phè	chhì-phè		(proposed)
14	蹔	蹔	pôaⁿ	pôaⁿ-kè	蹔	280DF
15	荘	荘	puh	puh-íⁿ		(proposed)
16	迌	迌	thit	thit-thô	迌	28468
17	蟶	蟶	thōa	thâng-thōa	蟶	274B4
18	跉	跉	thún	thún-tảh	跉	27FEC
19	佔	佔	tiàm	tiàm-tī		2B8B6
20	煯	煯	tỏh	tỏh-hé	煯	243F8
21	愣	愣	tỏk	chhek-tỏk	愣	2272C
22	偲	偲	ên			2B8C6
23	嬌	嬌	kiung			2BC54
24	拷	拷	lâu		拷	39EF
25	偃	偃	ngài		偃	2028E