L2/19-177

Proposal to encode Cantonese Bopomofo Characters

Ben Yang
Director of Technology
PanLex — The Long Now Foundation
ben@panlex.org

Eiso Chan (陈永聪)

April 29, 02019

1. Introduction

This proposal is a followup to L2/19-100 *Preliminary proposal to encode four extended Bopomofo letters for Cantonese in BMP* by Eiso Chan (陈永聪).

Cantonese (廣東話, 广东话, Gwong2 Dung1 Waa2) is a variety of Chinese spoken by approximately 80 million native speakers. Between the 01930s and 01950s, Bopomofo (or Zhuyin Fuhao) was adapted to transcribe Cantonese by adding 4 additional characters, currently unencoded in Unicode. These additional characters are proposed below.

2. Request

- This proposal requests the addition of 4 new characters in the Bopomofo block with the following names and code points:
  - U+3100 ㄆ BOPOMOFO LETTER GW
  - U+3101 ㄇ BOPOMOFO LETTER KW
  - U+3102 ㄈ BOPOMOFO LETTER OE
  - U+3103 ㄉ BOPOMOFO LETTER AH

  Note: The preliminary proposal (L2/19-100) for these characters suggested code points in the Bopomofo Extended block. However in the course of our research, we have discovered a large number (possibly dozens) of potentially encodable Bopomofo characters for non-Mandarin Chinese varieties. Due to this, we believe the best course of action would be to fill all remaining space in the Bopomofo and Bopomofo Extended blocks, to reduce the size a potential new Bopomofo Supplement block.
- Additionally, this proposal requests the following changes to the Unicode Core Spec, section 18.3 Bopomofo (under Extended Bopomofo):
  - Remove the following line from the first paragraph:

    “There are no standard Bopomofo letters for the phonetics of Cantonese or several other Southern Chinese dialects.”

- Add the following to the end of the first paragraph:

  “The four characters encoded at U+3100..U+3103 were designed to cover additional sounds found in Cantonese.”

- Add the following subsection after the second paragraph:

  “In Cantonese, final consonants not covered by the set of standard Bopomofo with final “N” and “NG” are marked with a standard-sized character (ㆷ, ㆶ, ㆻ, ㆹ, ㆸ, ㆵ). They are occasionally also marked with a following middle dot, represented by U+00B7 MIDDLE DOT.”

3. Justification

- The images below, from 全國主要方言區方音對照表 (one of the initial documents proposing these characters) show first the two initials (Figure 1), BOPOMOFO LETTER GW (here marked as kw) and BOPOMOFO LETTER KW (here marked as kʻw)[1], then the two finals (Figure 2), BOPOMOFO LETTER OE (here marked as œ) and BOPOMOFO LETTER AH (here marked as ɐ):

Figure 1: 全國主要方言區方音對照表 pg. 6

本表注音字母音值表

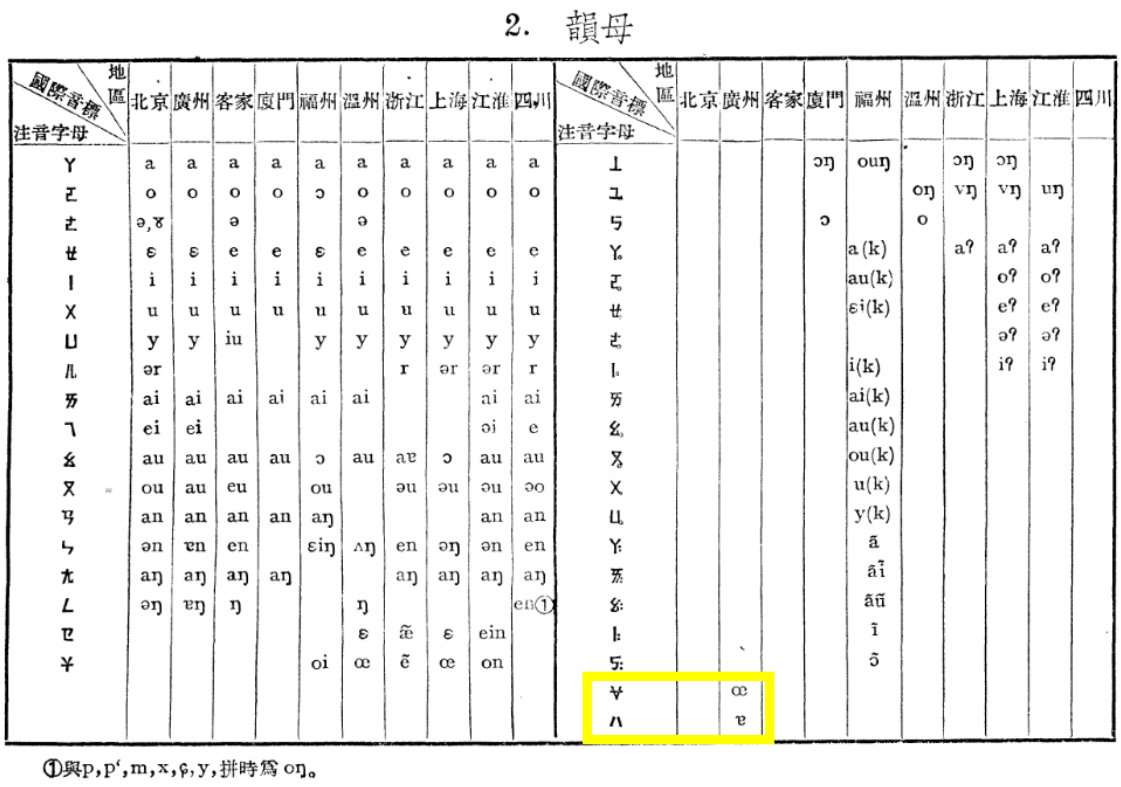
1. 聲母

國際音標 / 地區 / 注音字母	北京	廣州	客家	廈門	福州	温州	浙江	上海	江淮	四川
ㄅ	p	p	p	p	p	p	p	p	p	p
ㄆ	pʻ	pʻ	pʻ	pʻ	pʻ	pʻ	pʻ	pʻ	pʻ	pʻ
ㄇ	m	m	m		m	m	m	m	m	m
ㄈ	f	f	f		f		f	f	f	f
ㄪ					v		v	v		
ㄉ	t	t	t	t	t	t	t	t	t	t
ㄊ	tʻ	tʻ	tʻ	tʻ	tʻ	tʻ	tʻ	tʻ	tʻ	tʻ
ㄋ	n	n	n		n	n	n	n	n	
ㄌ	l	l	l	l	l	l	l	l	l	l
ㄍ	k	k	k	k	k	k	k	k	k	k
ㄎ	kʻ	kʻ	kʻ	kʻ	kʻ	kʻ	kʻ	kʻ	kʻ	kʻ
ㄫ		ŋ	ŋ		ŋ	ŋ	ŋ	ŋ		ŋ
ㄏ	x	h	h	h	h	h	h	h	x	x
ㄐ	tɕ	tɕ	①		③	tɕ	tɕ	tɕ	tɕ	tɕ
ㄑ	tɕʻ	tɕʻ	②		④	tɕʻ	tɕʻ	tɕʻ	tɕʻ	tɕʻ
ㄬ			ɲ			ɲ	ɲ	ɲ		ɲ
ㄒ	ɕ	s	s⑤	h⑥		ɕ	ɕ	ɕ	ɕ	ɕ
ㄓ	tʂ		č⑦							

國際音標 / 地區 / 注音字母	北京	廣州	客家	廈門	福州	温州	浙江	上海	江淮	四川
ㄔ	tʂʻ		čʻ⑧							
ㄕ	ʂ		š⑨							
ㄖ	ʐ								ʑ	ʑ
ㄗ	ts		ts	ts⑩		ts	ts	ts	ts	ts
ㄘ	tsʻ		tsʻ	tsʻ⑪		tsʻ	tsʻ	tsʻ	tsʻ	tsʻ
ㄙ	s		s	s	s	s	s	s	s	s
ㄚ				bʻ		bʻ	bʻ	bʻ	⑫	
ㄞ						dʻ	dʻ	dʻ		
ㄟ				gʻ			gʻ	gʻ		
ㄠ						ɦ	ɦ	ɦ		
ㄡ						dʑʻ	dʑʻ	dʑʻ		
ㄢ						dzʻ	dzʻ			
ㄣ						z	z	z		
ㄤ						ɦ(i)	ɦ(i)	ɦ(i)		
ㄥ						ɦ(u)	ɦ(u)	ɦ(u)		
ㄦ						ɦ(y)	ɦ(y)	ɦ(y)		
ㄨ		kw								
ㄩ		kʻw								

①尖音時爲 ts，圓音時爲 k。 ②尖音時爲 tsʻ，匿音時爲 kʻ。 ③ ts，tɕ 之間。 ④ tsʻ，tɕʻ 之間。 ⑤ ʂ，ɕ 之間。 ⑥在 i 音前。 ⑦ tʂ，tɕ 之間。 ⑧ tsʻ，tɕʻ 之間。 ⑨寬式的 ʂ。 ⑩在 i 前讀 tɕ。 ⑪在 i 前讀 tɕʻ。 ⑫半子音，帶音。 ⑬溫州、浙江、上海的濁音送氣記號係根據趙元任"現代吳語的研究"註的。按實際上這些音的送氣是很輕微的。

Figure 2: 全國主要方言區方音對照表 pg. 7



2. 韻母

國際音標 / 地區 / 注音字母	北京	廣州	客家	廈門	福州	溫州	浙江	上海	江淮	四川
ㄚ	a	a	a	a	a	a	a	a	a	a
ㄛ	o	o	o	o	ɔ	o	o	o	o	o
ㄜ	ə,ɤ		ə			ə				
ㄝ	ɛ	ɛ	e	e	ɛ	e	e	e	e	e
ㄧ	i	i	i	i	i	i	i	i	i	i
ㄨ	u	u	u	u	u	u	u	u	u	u
ㄩ	y	y	iu		y	y	y	y	y	y
ㄦ	ər						r	ər	ər	r
ㄞ	ai	ai	ai	ai	ai	ai			ai	ai
ㄟ	ei	ei							əi	e
ㄠ	au	au	au	au	ɔ	au	aʊ	ɔ	au	au
ㄡ	ou	au	eu		ou		əu	əu	əu	əo
ㄢ	an	an	an	an	aŋ				an	an
ㄣ	ən	ɐn	en		ɛiŋ	ʌŋ	en	əŋ	ən	en
ㄤ	aŋ	aŋ	aŋ	aŋ			aŋ	aŋ	aŋ	aŋ
ㄥ	əŋ	ɐŋ	ŋ			ŋ				en①
ㄭ						ɛ	ɛ̃	ɛ	ein	
ㄮ					oi	œ	œ̃	œ	on	

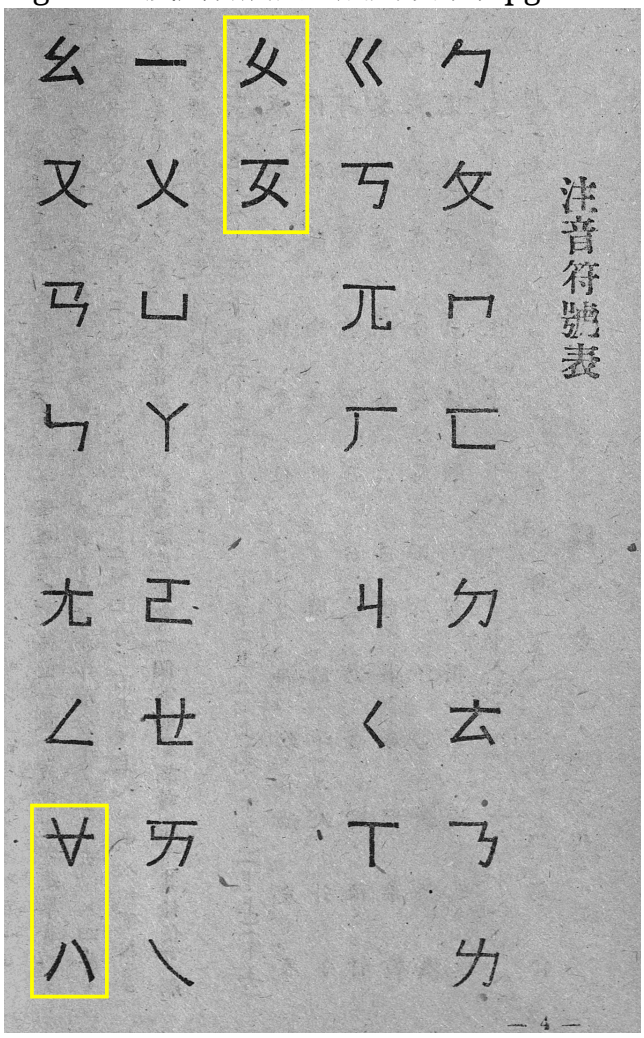
國際音標 / 地區 / 注音字母	北京	廣州	客家	廈門	福州	溫州	浙江	上海	江淮	四川
ㄫ				ɔŋ	ouŋ		ɔŋ	ɔŋ		
ㄬ						oŋ	vŋ	vŋ	uŋ	
ㄭ				ɔ		o				
ㄱ					a(k)		a?	a?	a?	
ㄲ					au(k)			o?	o?	
ㄳ					ɛi(k)			e?	e?	
ㄴ								ə?	ə?	
ㄵ					i(k)			i?	i?	
ㄶ					ai(k)					
ㄷ					au(k)					
ㄸ					ou(k)					
ㄹ					u(k)					
ㄺ					y(k)					
ㄻ					ā					
ㄼ					āĩ					
ㄽ					āũ					
ㄾ					ĩ					
ㄿ					ɔ̃					
ㅀ		œ								
ㅁ		ɐ								

①與p,pʻ,m,x,ɸ,y,拼時爲oŋ。

[1]Both Cantonese and Mandarin do not distinguish between voiced and voiceless consonants, but rather voiceless unaspirated and voiceless aspirated consonants. In older transliteration schemes, this distinction is reflected as without and with an apostrophe (ta vs. t'a, ka vs. k'a, etc.). However, most modern transliteration schemes including Yale, Jyutping, and Pinyin reflect this distinction as voiced and voiceless letters (da vs. ta, ga vs. ka, etc.). The proposed Unicode names reflect this modern transliteration style.

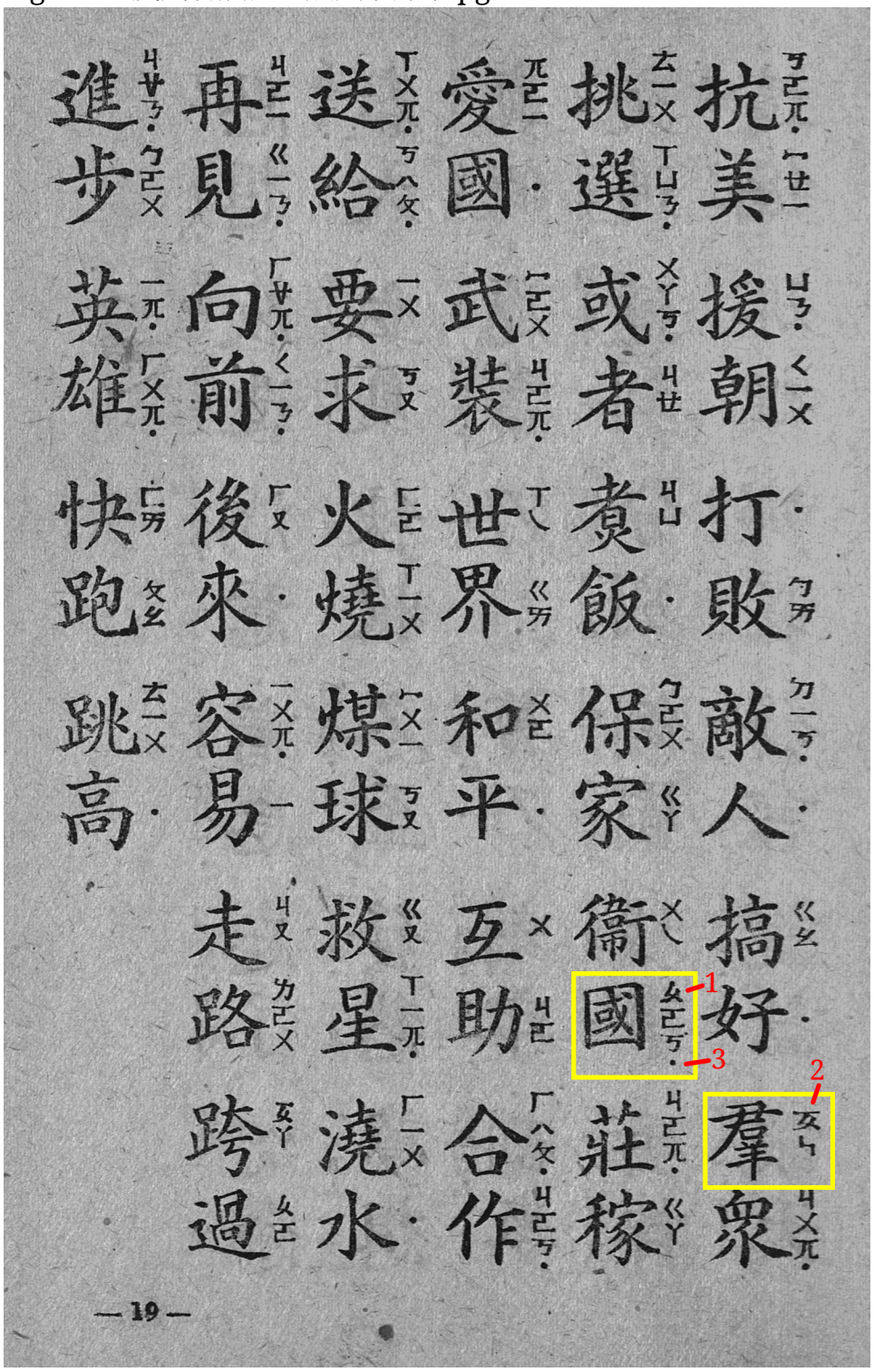
- Figure 3 below (from 廣州音農民速成識字課本, a literacy textbook for farmers) shows clear representations of the glyph shapes and the position of the new characters in the standard Cantonese Bopomofo ordering. Note that the reading direction is top-to-bottom, right-to-left.

Figure 3: 廣州音農民速成識字課本 pg. 4



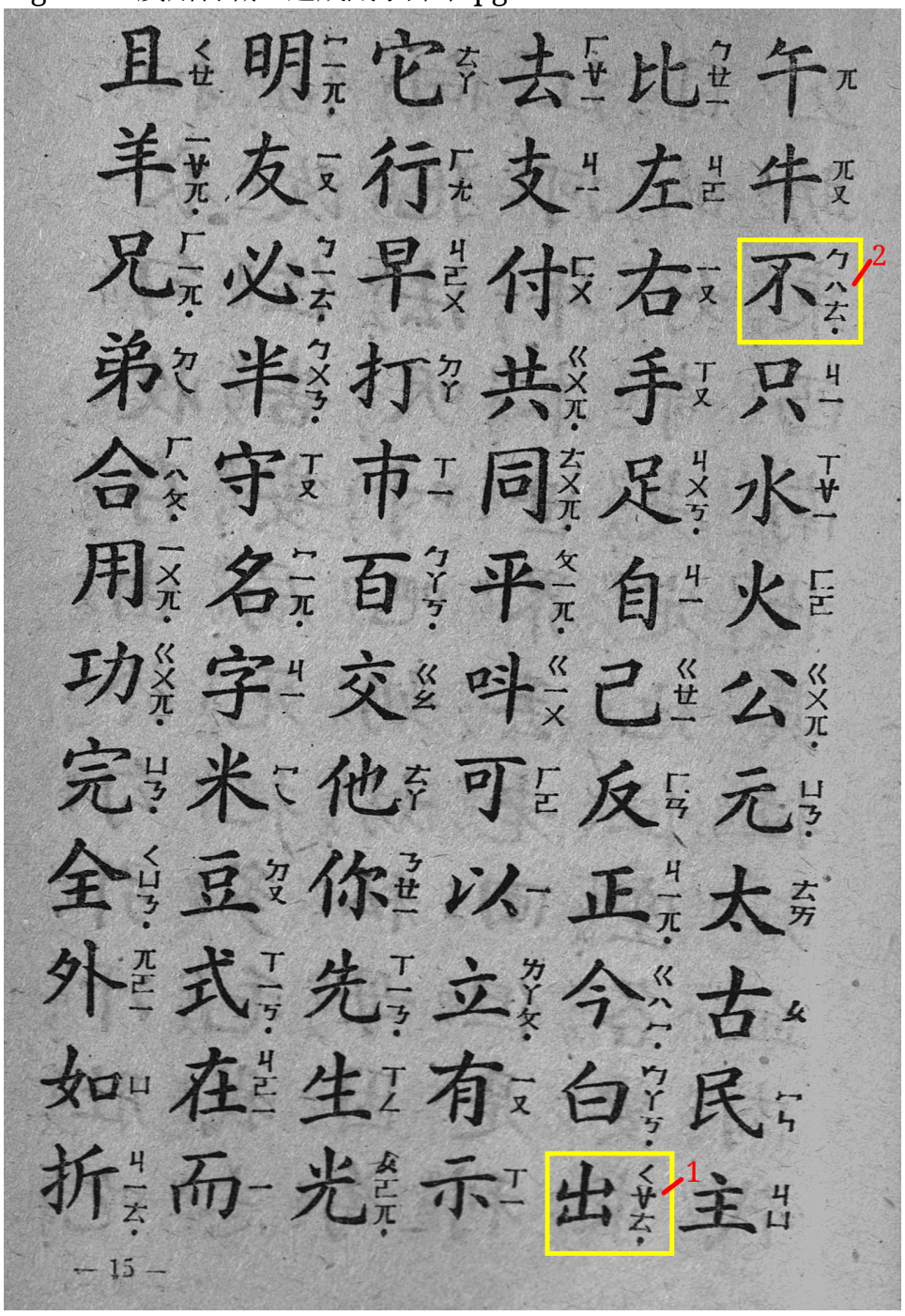
- Figure 4 below (also from 廣州音農民速成識字課本) shows usage of BOPOMOFO LETTER GW (1) and BOPOMOFO LETTER KW (2). These two characters are transcribed as *gwok* and *kwan* (in Jyutping) respectively. Also note the use of the final middle dot (3) to mark the final -k in *gwok*.

Figure 4: 廣州音農民速成識字課本 pg. 19



- Figure 5 below (from 廣州音職工速成識字課本, a literacy textbook for laborers) shows usage of BOPOMOFO LETTER OE (1) and BOPOMOFO LETTER AH (2). These two characters are transcribed as *ceot* and *bat* respectively.

Figure 5: 廣州音職工速成識字課本 pg. 15



- Figures 6 and 7 shows the use of all four Cantonese-specific Bopomofo characters in a horizontal context. Note in Figure 8 that in this source, the middle dot is not used after final consonants.

Figure 6: 全國主要方言區方音對照表 pg. 61

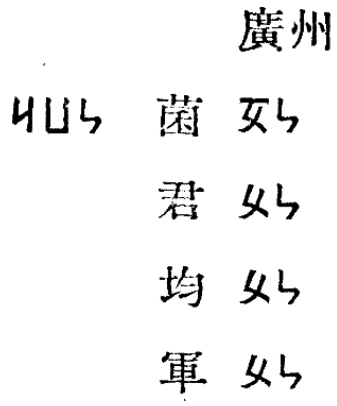


Figure 7: 全國主要方言區方音對照表 pg. 63

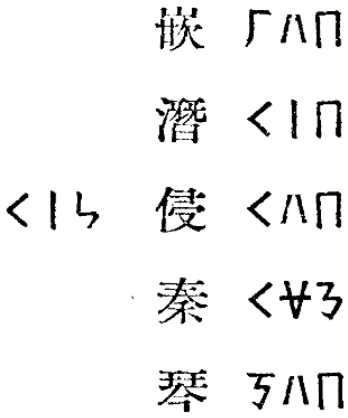
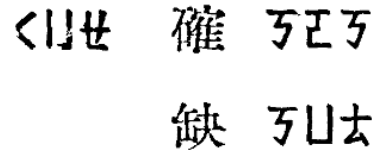


Figure 8: 全國主要方言區方音對照表 pg. 64



- Note: Tones are not marked in any of sources of Cantonese Bopomofo, and thus no additional tone marks are proposed at this time

4. Alternative encoding possibilities

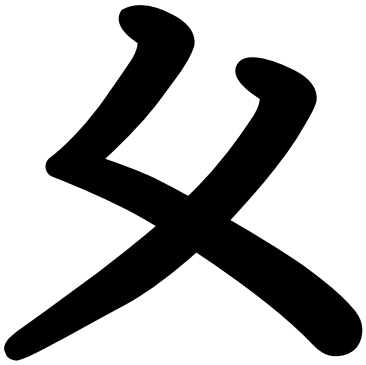
The four Cantonese Bopomofo characters cannot be depicted using the currently encoded set of Bopomofo characters, and are showed to be semantically contrasted with the other Bopomofo glyphs.

The only possibility for unification is BOPOMOFO LETTER AH superficially resembles U+30CF KATAKANA LETTER HA ハ. However, the drastically different use case of the Bopomofo character necessitates that it not be unified with this Katakana character.

5. Character Data

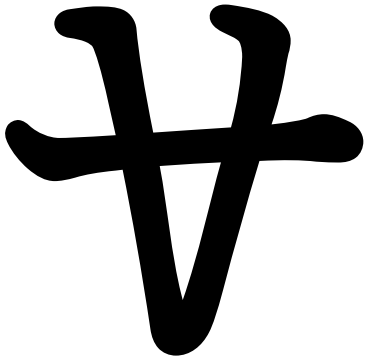
5.1 Glyphs

- U+3100 BOPOMOFO LETTER GW

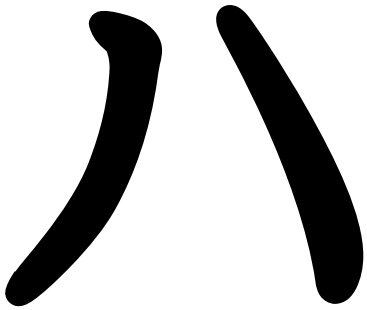


- U+3101 BOPOMOFO LETTER KW

- U+3102 BOPOMOFO LETTER OE



- U+3103 BOPOMOFO LETTER AH



5.2 Character properties

In `UnicodeData.txt` format:

- `3100;BOPOMOFO LETTER GW;Lo;0;L;;;;;N;;;;;`
- `3101;BOPOMOFO LETTER KW;Lo;0;L;;;;;N;;;;;`
- `3102;BOPOMOFO LETTER OE;Lo;0;L;;;;;N;;;;;`
- `3103;BOPOMOFO LETTER AH;Lo;0;L;;;;;N;;;;;`

All other properties are identical to U+3105 BOPOMOFO LETTER B

5.3 Collation order

As indicated by Figure 3 above, the collation order for Cantonese Bopomofo is:

ㄅ<ㄆ<ㄇ<ㄈ<ㄉ<ㄊ<ㄋ<ㄌ<ㄍ<ㄎ<ㄏ<ㄐ<ㄑ<ㄒ<**ㄠ**<**ㄡ**<ㄧ<ㄨ<ㄩ<ㄚ<ㄛ<ㄜ<ㄞ<ㄟ<ㄢ<ㄣ<ㄤ<ㄥ<ㄦ<ㄧㄥ<**ㄭ**<**ㄮ**

However, as the other Cantonese Bopomofo characters have already been encoded, the ordering of the four proposed characters should be:

ㄠ<ㄡ<ㄭ<ㄮ

6. Sources

- 全國主要方言區方音對照表, Chinese Character Reform Commission, Beijing: ZhonghuaBook Company, 1954
- 廣州音農民速成識字課本, South China Peoples Publishing House, Guangzhou: Xinhua Bookstore, 1952
- 廣州音職工速成識字課本, South China Peoples Publishing House, Guangzhou: Xinhua Bookstore, 1952

7. Acknowledgements

We would like to thank Alex DelPriore, Jerry You, Hinata Syaoran, and Debbie Anderson for their help in obtaining materials for evidence of these characters and for assistance with editing this proposal.

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646

Please fill all the sections A, B and C below.

Please read Principles and Procedures Document (P & P) from http://std.dkuug.dk/JTC1/SC2/WG2/docs/principles.html for guidelines and details before filling this form.

Please ensure you are using the latest Form from http://std.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html.

See also http://std.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html for latest *Roadmaps*.

Form number: N4502-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05, 2009-11, 2011-03, 2012-01)

A. Administrative

1.**Title:** ***Proposal to encode Cantonese Bopomofo Characters***
2. Requester's name: ***Ben Yang and Eiso Chan***
3. Requester type (Member body/Liaison/Individual contribution): ***Individual contribution***
4. Submission date: ***02019-04-29***
5. Requester's reference (if applicable):
6. Choose one of the following:
   This is a complete proposal: ***YES***
   (or) More information will be provided later:

B. Technical - General

1. Choose one of the following:
   a. This proposal is for a new script (set of characters): ***NO***
      Proposed name of script:
   b. The proposal is for addition of character(s) to an existing block: ***YES***
      Name of the existing block: ***Bopomofo***
2. Number of characters in proposal: ***4***
3. Proposed category (select one from below - see section 2.2 of P&P document):

A-Contemporary	B.1-Specialized (small collection)	***X*** B.2-Specialized (large collection)	
C-Major extinct	D-Attested extinct	E-Minor extinct	
F-Archaic Hieroglyphic or Ideographic		G-Obscure or questionable usage symbols	

4. Is a repertoire including character names provided? ***YES***
   a. If YES, are the names in accordance with the "character naming guidelines" ***YES***
   b. Are the character shapes attached in a legible form suitable for review? ***YES***
5. Fonts related:
   a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?
   Ben Yang
   b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):
   Ben Yang, OFL
6. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? ***YES***
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? ***YES***
7. Special encoding issue

Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)? ***YES***

see proposal

8. Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see UAX#44: http://www.unicode.org/reports/tr44/ and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? ***YES***

If YES explain ***Preliminary proposal was made as L2/19-100***

2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? ***NO***

If YES, available relevant documents:

3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? ***YES***

Reference: ***see proposal***

4. The context of use for the proposed characters type of use; common or rare) ***common***

Reference: ***Used in all transcriptions of Cantonese using Bopomofo***

5. Are the proposed characters in current use by the user community? ***No***

If YES, where? Reference:

6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? ***NO***

If YES, is a rationale provided?

If Yes, reference:

7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? ***YES***

8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? ***NO***

If YES, is a rationale for its inclusion provided?

If Yes, reference:

9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? ***NO***

If YES, is a rationale for its inclusion provided?

If Yes, reference:

10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to, or could be confused with, an existing character? ***YES***

If YES, is a rationale for its inclusion provided? ***YES***

If Yes, reference: ***see proposal***

11. Does the proposal include use of combining characters and/or use of composite sequences? ***NO***

If YES, is a rationale for such use provided?

If Yes, reference:

Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?

If Yes, reference:

12. Does the proposal contain characters with any special properties such as

control function or similar semantics? *NO*

If YES, describe in detail (include attachment if necessary)

13. Does the proposal contain any Ideographic compatibility characters? *NO*

If YES, are the equivalent corresponding unified ideographic characters identified?

If Yes, reference: