The influence of orthographic repertoire on Written Cantonese across different jurisdictions

Chaak Ming LAU – The Education University of Hong Kong
lchaakming@eduhk.hk

Hilário DE SOUSA – CRLAO, École des hautes études en sciences sociales
hilario@bambooradical.com

Cantonese is a Sinitic language.[1] Amongst the modern Sinitic languages, there are two extremes in terms of their relationship with writing. At one extreme is Standard Mandarin, a standardised language supported by several governments, and an official language in a number of important international organisations. At the other extreme are the vast majority of other Sinitic languages, which are not written, or scarcely written. Then there is Cantonese, one of the very few Sinitic languages that are somewhat in between these two extremes. In this talk we will have a look at some issues surrounding how Cantonese is written, especially how being spoken in a number of jurisdictions caused differences in the way Cantonese is written.

"Cantonese"
Cantonese is the language of Canton, i.e. *Gwong2 zau1* / *Guǎngzhōu* 廣州.[2] Standard Cantonese is the representative variety of the Yuè 粵 (*Jyut6*) dialect group within the Sinitic language family.[3] When Mandarin speakers talk of *Yuèyǔ* 粵語, it means Cantonese by default.

The linguistic distance between Standard Cantonese and the other Yue dialects varies; some Yue dialects are very close to Standard Cantonese, while some are not intelligible to speakers of Standard Cantonese. People use the term *Cantonese* differently: for some authors, the term *Cantonese* only refers to the speech of Canton; for some, *Cantonese* refers to the entire Yue dialect group. Most usually, *Cantonese* is used to cover a range of Yue dialects somewhere between these two extremes. In this paper, a rather narrow definition of Cantonese is used: Cantonese includes the language of Canton, plus the language maintained by their migrant communities who left the Canton area[4] within the last 150 years or so. (See

---

[1] The modern members of the Sinitic language family are the descendents of Old Chinese. In Mandarin, there are the terms *yǔyán* 語言 and *fāngyán* 方言: Cantonese is considered a *fāngyán* of the Chinese *yǔyán*. *Fāngyán* is often translated as "dialect". Nonetheless, the semantics of *fāngyán* is wider than that of *dialect*, and calling Cantonese a dialect of Chinese often conjures the wrong impression in terms of the linguistic distance between these "dialects" of Chinese. When compared with the Western linguistic concept of *language* versus *dialect*, which is based on intelligibility, *fāngyán* is a dialect or a regional language. Cantonese and other Sinitic languages like Mandarin, Teochew, and Shanghainese are separate languages from a Western linguistic point of view, as they are not mutually intelligible. The reader need not be alarmed at why something is called a language, dialect, or speech variety; both the Western model of *language* versus *dialect*, and the Chinese model of *yǔyán* versus *fāngyán*, are problematic in their own ways anyway. See, e.g., Mair (1991).

[2] For Cantonese, the romanisation used is Jyutping (jyut6 ping3 [jyt˨ pʰɪŋ˧], https://en.wikipedia.org/wiki/Jyutping). Beyond normal Jyutping, <sl> is used to represent /ɬ/ which found in some non-standard Cantonese varieties. For Mandarin, the romanisation used is Hànyǔ Pīnyīn. The Chinese characters used are by default Traditional Chinese characters. Simplified Chinese characters are used in linguistic examples from jurisdictions where Simplified Chinese dominates, and when quoting publications in Simplified Chinese.

[3] The Sinitic languages are classified into dialect groups. Within the Yuè dialect group, there are many speech varieties that are not intelligible to speakers of Cantonese. In other words, a dialect group is not a single language from a Western linguistic point of view.

[4] The traditional counties of *Naam4 Hoi2* / *Nánhǎi* 南海 (which included the modern day city centre of *Fat6 Saan1* / *Fóshān* 佛山), *Pun1 Jyu4* / *Pānyú* 番禺 (which included the modern day city centre of Canton), and *Seon6 Dak1* / *Shùndé* 順德.

discussions in de Sousa 2021.) After the First Opium War (1839–1842), a large number of people emigrated from the Canton area.[5] Many Cantonese-speaking enclaves were formed in Far Southern China and overseas. Cantonese came to dominate many Chinatowns overseas, e.g. Saigon, Kuala Lumpur. In a number of cities in Far Southern China, e.g. *Zaam3gong1 / Zhànjiāng* 湛江 (Fort-Bayard), *Bak1hoi2 / Běihǎi* 北海, and *Naam4ning4 / Nánníng* 南寧, the city centre became Cantonese-dominant, and the pre-established Sinitic languages were pushed out into the suburbs and rural areas. In Hong Kong and Macau, Cantonese has caused the pre-established Sinitic languages (other Yue varieties, Hakka, Min) to become moriband. The most striking similarity that these Cantonese varieties share is their tones: while some minor segmental differences are easily observable amongst the various Cantonese varieties, the tones have remained remarkably similar. This contrasts with some Yue dialects that are spoken very close to Canton, e.g. *Dung1gun2* / Dōngguǎn 東莞, *Zung1saan1* / Zhōngshān 中山, which have vastly different tones (and segments) from those in Cantonese.

Different jurisdictions and their orthographic repertoires
The prosperity of Guangdong attracted European colonists. The trading opportunities in the then European colonies of Hong Kong, Macau, and Zhanjiang attracted a large number of Cantonese migrants, causing these places to become Cantonese-dominant. Currently some level of autonomy still exists in Hong Kong and Macau, and their governments still largely function in Cantonese, at least when spoken language is concerned. Nonetheless, due to the dominance of Modern Standard Written Chinese (contracted below as Modern Written Chinese below), i.e. written Mandarin, and the attitude that Cantonese is "just at *fong1 jin4 / fāngyán* 方言", there is a medium to high level of resistance towards Cantonese as a written language in the government and the education system. Cantonese is a language that is widely spoken, in a sense enjoys governmental support, but has no written standard (e.g. David Li 2000; Cheung & Bauer 2002; Snow 2004; Bauer 2018).

This is essentially a diglossic situation (Ferguson 1959): a H variety which is normally used in formal writing, and a L variety which is used in day-to-day conversation. The majority of the schools in Hong Kong and Macau have maintained the tradition of teaching the Chinese language subject in Cantonese (totally, or partially in Mandarin, instead of totally in Mandarin): the texts that the students read and compose are in Modern Written Chinese, i.e. Mandarin, but the characters are read out in Cantonese (in most cases the Cantonese cognates of the Mandarin morphemes involved). The writing of actual Cantonese is penalised by the education system. The following are two pairs of examples demonstrating some lexical and morphosyntactic differences between the Modern Written Chinese used in Cantonese societies and actual Cantonese. (There are many other lexical and morphosyntactic differences.)

---

[5] As a consequence of the First Opium War, China was forced to give up all forms of maritime prohibitions; there were no longer restrictions on civilian watercraft ownership. The heart of the Pearl River Delta was relatively prosperous (despite the war), and overpopulated. Many Cantonese people left, via the rivers or the sea, in order to seek new commercial opportunities elsewhere. Through their economic prowess, Cantonese came to dominate many city centres.

Modern Written Chinese

1(a) 他　　今天　　沒　　　　上課。
　　　tā　　jīntiān　　méi　　　　shàngkè　　　　(< Mandarin)
　　　taa1　gam1tin1　mut6　　　soeng5fo3　　(< Cantonese cognates; how it is read out)
　　　3SG　today　　NEG.PFV　attend:class
　　　'S/he did not go to school today.'

Cantonese

1(b) 佢　　今日　　冇　　　　　返學。
　　　keoi5　gam1jat6　mou5　　　faan1hok6
　　　3SG　　today　　NEG.PFV　　return:learn
　　　'S/he did not go to school today.'

Modern Written Chinese

2(a) 我　　給　　他　　錢。
　　　wǒ　　gěi　　tā　　qián　　　　(< Mandarin)
　　　ngo5　kap1　taa1　cin4/2　　(< Cantonese cognates; how it is read out)
　　　1SG　give　3SG　money
　　　'I give him/her money.'

Cantonese

2(b) 我　　俾　　錢　　佢。
　　　ngo5　bei2　cin2　keoi5
　　　1SG　give　money　3SG
　　　'I give him/her money.'

The various Cantonese varieties are understandably influenced by the (major) languages spoken in their local environment. For instance, in Hong Kong there is English; in Malaysia there is Malay, English, and various other Sinitic languages; in Guangzhou there is more Mandarin influence (in comparison with Hong Kong and Macau); in Nanning there is Zhuang, Pinghua, and Mandarin. How Cantonese is written there is also influenced by how the languages that they are in contact with are written. Below we shall first discuss some general principles of Cantonese writing, before talking about some regionally specific cases.

How Cantonese is written

In the history of Chinese writing, there have always been writings with dialectal elements. The earliest literary works with features that can be clearly identified as Cantonese were the *Muk6 jyu4 syu1* / *Mùyúshū* 木魚書, which started to appear towards the end of the Ming Dynasty (17th century). While most texts in the Canton area were written in Classical Chinese, and later on Modern Written Chinese, for centuries there have been writings that were partially or entirely in Cantonese. Particularly popular in Hong Kong newspapers in the 1940s to 1960s was the *Saam1 Kap6 Dai2* 三及第 writing style, which is a mixture of Written Cantonese, Modern Written Chinese, and Classical Chinese (e.g. Snow 2004: 127). Nowadays there is a spectrum of writings from Written Cantonese to Modern Written Chinese.

There are several principles governing the choice of graphemes used to represent Cantonese. (Given the lack of standardisation, there can be more than one grapheme commonly used for the same morpheme.) The principles are similar to those found in Chinese writings in general, and also in the other (larger) Chinese-derived scripts in the region like Saw-ndip (Zhuàng) and Chữ Nôm (Vietnamese). However, given that Cantonese is already a Sinitic language, the amount of "novel" characters needed for writing Cantonese is fewer than that in Saw-ndip and Chữ Nôm, of which the host languages are non-Sinitic. (The shapes of the Cantonese-specific characters are also much less deviant from the Chinese prototype than many Saw-ndip and Chữ Nôm characters.)

Most morphemes in Cantonese are cognates of the morphemes that people are familiar with from their education of Modern Written Chinese (i.e. Written Mandarin) and Classical Chinese. In these cases, usually the same characters are used, e.g. *ngo5* / *wǒ* 我 'I', *faat3 leot6* / *fǎ lǜ* 法律 'law'. The morphemes are not necessarily equally commonly used in Cantonese and Mandarin. For instance, the usual word for "eat" is *sik6* 食 in Cantonese and *chī* 吃 in Mandarin. Nonetheless, people are very familiar with the character 食, as *shí* 食 is used to a degree in Mandarin, e.g. *shípǐn* 食品 'foodstuff', and 食 is a common word in Classical Chinese.

Solutions are needed for other morphemes. The strategies are:

- creation of novel characters; most commonly created are phono-semantic compounds, and there are also some novel ideograms
- using homophonous or near-homophonous characters
- using synonymous characters (in Written Cantonese this is rare, unlike, e.g., Min languages)
- reviving rarer characters found in earlier stages of Chinese

There are a few categories of Cantonese morphemes where these solutions are required. Firstly there are the non-Sinitic morphemes (that are not found in Mandarin or Classical Chinese). The majority are words from the Kra-Dai substratum (e.g. Bauer 1996; Lǐ Jǐnfāng 李锦芳 2002), or European loanwords. For instance, the English loanword *lip1* 'elevator' (< *lift*) is commonly written 軩: this phono-semantic compound has a vehicle radical 車 on the left as its semantic component, and *laap6 ~ lap6* 立 on the right as its phonetic component. A commonly used radical is the mouth radical 口: usually this is not used to indicate that the semantics has to do with mouth or speech directly; the mouth radical is used to indicate that this character represents a morpheme that is commonly found in colloquial speech (*hau2 jyu5* / *kǒuyǔ* 口語), unlike the character without a mouth radical, which is used for a "proper" word. For instance, *nei4* 尼 is a "buddhist nun", whereas with a mouth radical, 呢 represents a similar-sounding morpheme used in the colloquial register, in this case the Cantonese demonstrative *ni1 ~ nei1* 'this' (of Kra-Dai origin, cf. Zhuang *neix* 'this', Thai นี้ *níi* 'this'). The morpheme *di1*, which is a mass classifier (indicating non-singularity), or a comparative marker for adjectives, is commonly written with a mouth radical character 啲, with the phonetic component *dik1* 的. (Another commonly used grapheme for *di1* is the Roman alphabet D (homophonous to Cantonese speakers), e.g. *faai3 di1* 'faster' is written 快D or 快啲. This morpheme is of Hmong-Mien origin (Yue-Hashimoto 1991).) Sometimes instead of creating a novel character, a homophonous character is used. For instance, there is the Kra-Dai word *han4* 'itch' (cf. Zhuang *haenz*, Thai คัน *khān*), written with the homophonous *han4* 痕, a Sinitic word meaning "scar". Using homophonous or near-homophonous characters (instead of creating novel characters) is the norm with European loanwords, e.g. *zyu1gu1lik1* 朱古力 'chocolate' (from *zyu1* 朱, *gu2* 古, and *lik6* 力). Something that happens less often in Cantonese is using an etymologically unrelated character of the same meaning to represent a morpheme (*fan3 duk6* / *xùndú* 訓讀). An example is *nap1* 'concave' (cf.

Zhuang *mboep* 'concave'), written with the synonymous 凹 (the regular Sinitic pronunciation of this character, *aau3*, is not well known).

There are also Sinitic morphemes where "solutions" are required. There are three categories of these. Firstly there are the colloquial doublets of literary morphemes. In Sinitic languages there is the phenomenon of *man4 baak6 ji6 duk6* / *wénbái yìdú* 文白異讀: due to historic influences from other varieties of Chinese, especially from the national or regional standard of the time, certain characters have more than one pronunciation, one used in literary contexts, and one used in colloquial contexts. (There can be more than two readings.) This phenomenon occurs to a small degree in Cantonese. However, Cantonese is not as tolerant at having multiple readings for one character (unlike, e.g. Hokkien), and sometimes a new character is created for the colloquial pronunciation (while the etymologically correct character is left for the literary pronunciation). For instance, there is the doublet of *loi4* and *lai4~lei4*, meaning "come". The etymologically correct character 來 is usually only used for the literary pronunciation *loi4* (e.g. *loi4 din6* 來電 'incoming call'). A separate character has been created for the colloquial pronunciation *lai4~lei4*: 嚟 'come' (e.g. *keoi5 lai4 zo2 laa3* 佢嚟咗嘛 's/he has come'). The character 嚟 is formed with a mouth radical (for "colloquial") and a phonetic component 黎 *lai4* 'Hlai'.

The second category involves grammatical morphemes which are of Sinitic origin, but their origins are not generally known. For example, the third person pronoun *keoi5* is usually written 佢, with a person radical 亻 on the left, and a phonetic component *geoi6* 巨 ('huge') on the right. (This morpheme comes from *keoi4* 渠, which is a Southern dialectal form of the third person pronoun in Classical Chinese. This fact is not widely known. Even if it is widely known, a novel character would likely still be invented, as *keoi5* 佢 's/he', in Standard Cantonese at least, is not homophonous with *keoi4* 渠 'drain'.) There is also the example of the ideogram 冇: from the normal Chinese character of *jau5* 有 'have/exist', the ideographic *mou5* 冇 'not have/ not exist' was created.

The third category involves morphemes from earlier stages of Chinese. There is an active culture of *haau2 bun2 zi6* / *kǎo běnzì* 考本字 'searching for original characters', i.e. the investigation into the true etyma of morphemes (e.g. Zhōu Shìmǐn 周仕敏 2015). Nonetheless, the quality of such research varies, and the level of influence that these scholars (and their followers) have also varies. The "original character" movements have injected even more variations into the (at times chaotic) landscape of Cantonese morpheme representation. A more-succesful example is the character for *bei2* 'give'. The etymologically correct character 畀 is now commonly used. (Also very commonly used to write *bei2* 'give' is the homophonous *bei2* 俾 'cause'.)

<u>Observations from the data</u>
The principles that outlined above apply to Written Cantonese in multiple jurisdictions, but this does not imply that all users will naturally choose the same written form for a word. Variation across regions can be described in terms of (i) glyph-level script choice issues, (ii) general preference on the default strategy, (iii) the reference point for the phonetic component for new words, (iv) the role of external systems (such as the orthography or a transliteration scheme of another language) and (v) the use of certain characters to highlight regional differences.

i. Glyph-level script choice issues

The Han script is the default script for Cantonese in multiple jurisdictions[6]. The Traditional Han script is by nature a subset of the Han characters that are considered unsimplified, and there is room for multiple flavours of Traditional Han with minor differences in codepoints and glyph selection.

Taiwan vs Hong Kong flavour of Traditional Han: In the earlier days Hong Kong relied heavily on Taiwanese software and the Big-5 encoding for Chinese processing. Certain important distinctions made in Hong Kong Chinese are not supported by Big-5 and compromises were made to coerce Hong Kong writing practice to the Big-5 standard. Up till today, glyph selection is still biassed towards Big-5 compliant characters. Examples include (i) the use of 溫 (*wan1*, *wēn*, *warm*) with "囚", which is commonly found in Taiwanese materials, over the use of the glyph variant 溫 with "日", which is the character used in the Hong Kong education system; (ii) the use of 戶 (*wu6*, *hù*, *family*) with a "丿"instead of 户 with a "丶". The pair 着 (*zoek3*, *zhuó*, to wear; zoek6, zháo/zhe to suffer, progressive) and 著 (*zyu3*, *zhù*, to author) is an exception. Although the two characters are merged (into 著) in the Taiwanese standard, they are used distinctively in Hong Kong materials. Governmental materials are more consistent in upholding the Hong Kong preferred characters, e.g. the use of 衞 (*wai6*, *wèi*, to protect) (with 帀) as opposed to 衛. Hong Kong users do not seem to make a conscious decision to maintain a different script than Taiwan, but in educational materials,

Japanese standard characters: There is no sign of Kana's or Japanese Kanji systematically adopted for the writing of Cantonese in any region, not even in Hong Kong where high popularity of Japanese culture is observed[7]. The use of Japanese Kanji is marked, and is limited to proper names from or related to Japan. An example of intended use of Japanese Kanji is the character 駅 (Japanese: *eki*), which can be found in shops and building names, and are pronounced *jik6*. Occasionally pronounced as 尺 (*cek3*, ruler) or 站 (*zaam6*, station) by the locals. Other instances are mostly unintentional. For example, the character 戸, the Japanese standard Kanji for 户, is attested (in proper names).

ii. General preferences on strategy to write out unknown words

If there is a need to write an unknown character and no existing acceptable characters already exist, Hong Kong users tend to resort to either homophones or English orthography, or avoiding the expressions. In a digital context, end users do not have the flexibility to test out newly created characters. Since there is

---

[6] The heritage speaker community may be an exception.
[7] Hong Kong Census 2021 shows that 2.0% of Hong Kong residents speak Japanese as an additional language

close to no way for an end-user to use a character that is not already present in existing fonts, the use of novel characters is limited to characters that have been encoded in Unicode and have been included by common fonts. The choice of Cantonese characters is nothing more than the set of collected Cantonese characters used in publication before the 2000s, collected by and submitted to the Unicode consortium by Cheung and Bauer (2002). Of course one can also choose one of the phono-semantic characters in the vast Chinese character set that has reasonable phonetic and semantic components for the morpheme. Semantic writing is an alternative to writing with homophones, but this strategy is not used systematically.

iii. Reference point for the phonetic component

Traditionally, the sound component in a phono-semantic compound is based on the character's Cantonese pronunciation. This is still being done in all Cantonese regions, and it is a strategy that works because this strategy allows other users to guess what the character denotes by reading out the phonetic component.

In Mainland China, one sees a strong shift towards the writing of Cantonese based on Mandarin phonetic principles. Among the push factors is the overwhelming dominance of Hanyu-Pinyin-based input methods (i.e. people choose characters that are easy to type in Mandarin). For instance, the modifier marker 嘅 *ge3* is nowadays usually written 噶 (from Mandarin 葛 *gé* 'poplin') in Mainland China. Other Mainland Chinese examples include the universal-quantitative marker 嘥 *saai3* (from Mandarin 塞 *sài*), sentence final particle 嘎 *gaa3* (戛 *gā*), and the expletive 黑/嘿 *hai1* (黑 *hēi*). None of these are formed from Cantonese phonetic principles; cf. Cantonese 葛 *got3*, 塞 *sak1*, 戛 *gaat3*, 黑 *hak1*.

iv. The role of external systems

Romanisation is sometimes used among strings of Han characters. Common examples in Hong Kong and Macau include *hea ~ hae* [*he3*] 'lazy', *chok* [*cok3*] 'shake' or 'pretentious', which are largely based on the English orthography. Other examples include AA制 (*ei1 ei1 zai3*, to go dutch), BB (*bi4 bi1*, baby), jer (*zoe1*, male genital), call (*ko1*, to call), chur (*coe2*, demanding), D (*di1*, plural classifier). Speakers in Hong Kong and Macau, and heritage speakers elsewhere, may occasionally write Cantonese entirely in romanisation for obfuscation, or as a technical workaround (from the inconveniences associated with computing in Han characters). Due to the lack of education on Cantonese phonology and romanisation, the lack of consistency in the romanisation used is easily observable. These words are not necessarily English loanwords, e.g. "D" (*di1*) is native to Cantonese, as discussed above. It may appear that the Latin script is more common in Hong Kong, but one should be aware that acronyms of romanisation are not popular in Hong Kong, while acronym usage (to replace sensitive words) is commonplace In Mainland China. The romanisation used is often based on Hanyu Pinyin, despite the fact that Hanyu Pinyin is designed for Mandarin and ill-suited for Cantonese. One example is the representation of the expletive 屌 *diu2* by Hanyu Pinyin *diao*, which violates Cantonese phonotactics. Mandarin acronyms can also be inserted into Cantonese text, e.g. YYDS (永遠的神, god-level player or person).

Hong Kong users occasionally use a Cantonese romanisation that is loosely based on English orthography, which has been reported as "Kong Girl Phonetics" (Wetters, 2021). The use of a loose Cantonese romanisation has been around not later than the 2000s, and the rules that govern its spelling are

never formally taught or documented. Some users are happy to use both English grapheme-to-phoneme rules as well as some form of Hanyu Pinyin.

v. Character choice that reflects regional features

The selected regions are linguistically similar, but certain regional morphemes made their way to the writing system. Examples are:
(3a)   戈 *go1* 'this' (Macau, from 戈 *go1* 'dagger')
(3b)   起 *hei2* 'at' (Hong Kong, from 起 *hei2* 'rise', reflecting a new reduced pronunciation)
(3c)   嘥 *slak1* 'one or so' (Nanning, from Nanning Cantonese 塞 *slak1* 'block (v.)').

In order to determine how popular it is to insert regional features to writing. Data from two Macau newspapers online (*Macao Daily* from 1st July 2020 to 31st March 2022, and a full-text searchable interface from *Exmoo news* (力報) since inception).

Written Cantonese in Macao Daily is negligible. Only 1027 out of 127925 pages (0.8%) contain the Cantonese perfective marker 咗 *zo2*. Most of them are direct quotes from Hong Kong entertainment news. All Macao official facebook pages were written in Standard Written Chinese. The word 戈 *go1* appears 1649 times in Macao Daily, and none of them are the expected demonstrative (most of them are foreign name transliteration). There are two occurrences of 戈 from Exmoo news that is used in this sense:

<u>More recent developments</u>

Texting variation that is only found in Hong Kong: Users on online forums use of numerals with the Latin script as a shorthand to longer terms, e.g. c9 (< 師奶 *si1 naai1*, middle-aged women), c6 (< 死佬 *sei2 lou2*, husband), 99 (< 奶奶 *naai4 naai2*, mother-in-law), 62 (< 老爺 *lou5 je4*, father-in-law); on9 (< 戀鳩 *ngong6 gau1*, f--ing stupid); siu4 (< 笑死 *siu3 sei2*, laugh-die), sor9 (< sor鳩, f--ing sorry), jm9 (< 做乜 鳩 *zou6 mat1 gau1*, what the f--k).

Hong Kong users also try to innovate as much as possible, by promoting innovative ways to distinguish key grammatical differences. Luke (2007) discusses a case of ambiguity caused by orthographic variation in sentence final particles found in the news, which can cause confusion to the readers.

(4)   開船   啦
       hoi1 syun4   *laa1/3*
       set sail   SFP

This line can mean either "Please set sail" (if the intended SFP is *laa1* with a high tone) or "It's already time to set sail" (if the intended sound is *laa3*). As at 2022, most users of Written Cantonese from Hong Kong distinguish the two by using separate characters for each morpheme, i.e. 喇 for *laa3*, as in 喇沙 (*laa3 saa1*, from La Salle, the name of a local school) and 啦 for *laa1*, as in 啦啦隊 (*laa1 laa1 deoi2*, cheerleaders). Luke (2007) in fact suggests that different characters should be used to distinguish tonal differences in SFPs, but his proposal uses 喇 for *laa1* as in 喇嘛 (*laa1 maa4*, Tibetan Lama), and 嘑 for

laa3 with the phonetic radical 罅 (*laa3*, gap). Luke's notation, which is also used in HKCanCor and 19th century Cantonese textbooks, is not known to the general public.

Some users also advocate distinguishing 咁 (*gam3*, so) and 噉 (*gam2*, in this way), which would otherwise be written with the first character, but this distinction is only found in some publication.

Discussion
A description of the orthographic repertoire on Written Cantonese provides insight to the mechanism of Written Cantonese development. Due to Hong Kong's cultural output and the status of Cantonese in the territory, one may argue that the use of Written Cantonese is a result of other regions following Hong Kong usage. From the examples above we see that local Macao and Nanning morphemes can show up in the text. It is more reasonable to assume that Cantonese regions develop their own ways to represent Written Chinese, and therefore be affected by what the users' education background, e.g. the set of characters, romanisation, and other writing systems that they are familiar with.

Although a list of strategies were described, we may need to know whether the speakers apply general strategies all the time when they need to write in Cantonese, or would it be the case that some users may have acquired the written form from the environment and pass down this usage through frequent usage. This is indeed the basis for distinguishing non-synonymous near-homophones. If only general strategies are available to the community, then the written forms for near-homophones will always be told apart by the context, and they would not be assigned different characters.

At the time of writing, linguistic exchange between different regions may be not as popular/common, due to the fact that Hong Kong and Macau fall under a different internet ecosystem and are less affected by Mainland Chinese trends. The pandemic which cut off Hong Kong from other Cantonese regions for over 2 years, and the border between Macau and mainland Chinese intermittently, may also have reduced interaction between these regions.

References

Bauer, Robert S. 1996. Identifying the Tai substratum in Cantonese. In *Pan-Asiatic linguistics – Proceedings of the Fourth International Symposium on Languages and Linguistics*, vol. 5, 1806–1844. Salaya: Institute of Language and Culture for Rural Development, Mahidol University.

Bauer, Robert S. 2018. Cantonese as written language in Hong Kong. *Global Chinese*, 4 (1). 103–142.

Cheung, Kwan-hin & Robert S Bauer. 2002. *The representation of Cantonese with Chinese characters*. Journal of Chinese Linguistics Monograph Series 18. Hong Kong: The Chinese University of Hong Kong Press.

Ferguson, Charles A. 1959. Diglossia. *Word* 15. 325–340.

Hóu, Xīngquán 侯兴泉 & Nánkāi Wú 吴南开. 2017. *Xìnxī chùlǐyòng Yuè fāngyán zìcí guīfàn yánjiū* 信息处理用粤方言字词规范研究 [A study of the standardisation of characters and words in the Yue dialect for information processing]. Guangzhou: Guangdong People's Publishing House 广东人民出版社**.**

Li, David C.S. 2000. Phonetic borrowing: key to the vitality of Written Cantonese in Hong Kong. *Written Language and Literacy* 3 (2). 199–233.

Lǐ, Jǐnfāng 李锦芳. 2002. *Dòng Tái yǔyán yǔ wénhuà* 侗台语言与文化 Tai-Kadai: language and culture. Beijing: The Ethnic Publishing House 民族出版社.

Luke, Kang Kwong. 2007. 粤語語氣詞的書寫方式 [Written forms of Cantonese sentence final particles]. In Joanna Ut-Seong Sio and Sze-Wing Tang (eds.), *Studies in Cantonese Linguistics 2*. Hong Kong: The Linguistic Society of Hong Kong.

Mair, Victor H. 1991. What is a Chinese "dialect/topolect"? Reflections on some key Sino-English linguistic terms. *Sino-Platonic Papers* 29. 1–31.

Snow, Don. 2004. *Cantonese as written language: The growth of a written Chinese vernacular*. Hong Kong: Hong Kong University Press.

Sousa, Hilário de. 2021. The expansion of Cantonese over the last two centuries. In Zhengdao Ye (ed.), *The Palgrave handbook of Chinese language studies*, 1–32. London: Palgrave Macmillan. https://doi.org/10.1007/978-981-13-6844-8_35-2.

Wetters, Ruth. 2021. 'Kong Girl Phonetics': Loose Cantonese Romanization in the 2019 Hong Kong Protest Movement. *Sino-Platonic Papers* (no. 317 [August, 2021]).

Yue-Hashimoto, Anne Oi-kan. 1991. The Yue dialect. In William S. Y. Wang (ed.), *Languages and dialects of China*, 292–322. Journal of Chinese Linguistics Monograph Series Number 3.

Zhān, Bóhuì 詹伯慧, Xiǎoyàn Fāng 方小燕, Yū'ēn Gān 甘于恩, Xuéqiáng Qiū 丘学强, Choi Lan Tong 汤翠兰, Jiànshè Wáng 王建设 & Qí Zhōng 钟奇 (eds.). 2002. *Guǎngdōng Yuè fāngyán gàiyào* 广东粤方言概要 [An outline of Yue dialects in Guangdong]. Guangzhou: Jinan University Press 暨南大学出版社.

Zhōu, Shìmǐn 周仕敏. 2015. Yuèyǔ běnzì kǎodìng zhī guānniàn yǔ fāngfǎ wèntí chúyì 粤语本字考订之观念与方法问题刍议 [Conceptual and methodological issues in textual researches on etymological characters of Cantonese]. *Journal of Guangdong Polytechnic Normal University* 广东技术师范学院学报 2015 (9). 48–55.