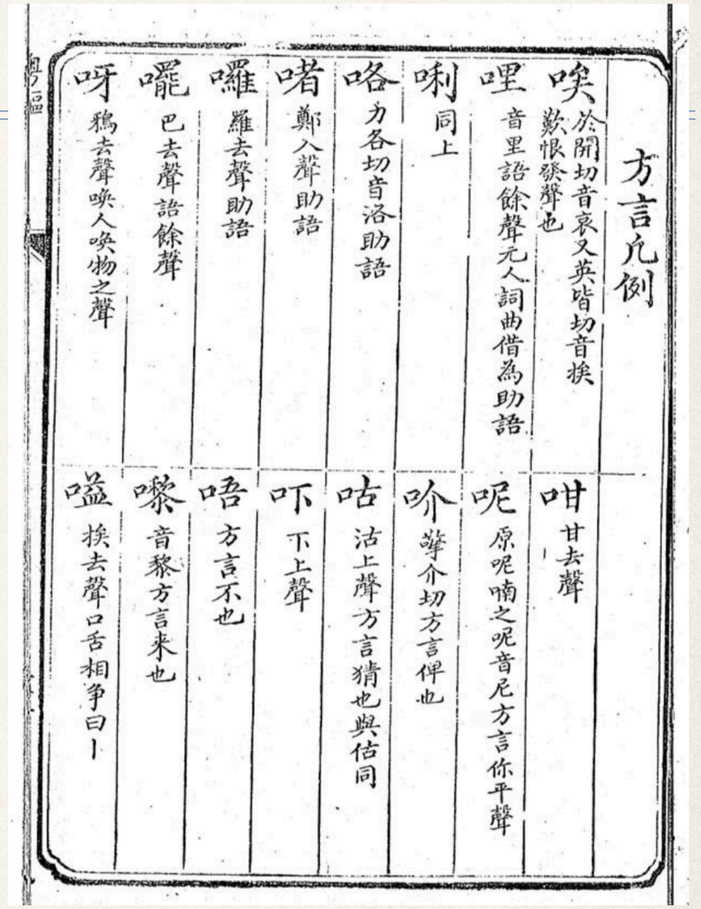
Decentralised standardisation of **Written Cantonese** in **Hong Kong**: implications to dictionary compilation and language testing

香港粵文嘅「『冇大台』標準化」：對詞典編寫、語言測試嘅意義

Chaak-ming Lau（劉擇明）<chaakming@gmail.com>
The Chinese University of Hong Kong

Society for Hong Kong Studies First Annual Conference 2019 *22 June 2019 @The University of Hong Kong*

What happened


方言凡例

唉 於開切音哀又英皆切音挨 歎恨聲聲也
哩 音里語餘聲无人詞曲借為助語
唎 同上
咯 力各切音洛助語
啫 鄭入聲助語
囉 羅去聲助語
嘥 巴去聲語餘聲
呀 鴉去聲喚人喚物之聲

咁 甘去聲
呢 原呢喃之呢音尼方言你平聲
吟 聲介切方言俾也
咕 沽上聲方言猜也與估同
吓 下上聲
唔 方言不也
嚟 音黎方言來也
嗌 挨去聲口舌相爭曰～

1. Cantonese has always had a prestigious/standard variety
   - 廣州話 is used as a lingua franca between speakers of other varieties of *Yue*
2. It did not have an authoritative written standard
3. Gradually standardised after repeated usage

- **Questions:** **1) How is that possible**
  2) What dictionaries and institutions should do to facilitate this process

Cantonese as a Written Language

- See Snow (2004, et. seq.) and Li (2011/2017) for its historical development
- Development was driven by the need to represent the native language in an "as-is" manner
- Gradually spread from one register to another register
- Fuelled by the availability of new communication means (e.g. forums, social media, PM, etc.), which gives rise to new norms

Snow, D. 2004. Cantonese as Written Language: The Growth of a Written Chinese Vernacular. Vol. 1. Hong Kong University Press
———. 2008. 'Cantonese as Written Standard?' Journal of Asian Pacific Communication.
———. 2013. 'Towards a Theory of Vernacularisation: Insights from Written Chinese Vernaculars'. Journal of Multilingual and Multicultural Development.
李婉薇。2011。《清末民初的粵語書寫》。香港：三聯書店（香港）有限公司。

How is it possible?

- **Question**
  How is it possible to write a language without a standard.

- **Answer**

  - Individuals acquire Han characters through school education, and learn Cantonese-specific usage through repeated exposure, and use innovative way to represent the language when necessary

  - The Community collectively reinforce the use of existing norms, spread accepted innovations, and suppress unwelcome written forms

How is it possible?

- Educated Speakers of Cantonese
  (before the introduction of Putonghua)
  - speak the vernacular (i.e. Cantonese)
  - taught to read the “common language” in local pronunciation
  - map each graph to a Cantonese syllable
- The nature of Han characters
  - loosely represents meaning
  - phono-semantic characters (形聲字) retain some information about the sound of the characters

“Common language” (通語) refers to either Classical Chinese or Standard Written Chinese (in essence, Written Mandarin).

The average educated Cantonese speaker can

- read and write 書面語 syu1 min2 jyu5 (Written Mandarin)
  - 我 今天 不 吃飯
    ngo5 gam1tin1 bat1 hek3faan6
- understand basic 文言文 man4 jin4 man2 (Classical Chinese)
  - 子曰：「學而時習之，不亦說乎？」
    zi2 joek6 hok6 ji4 si4 zaap6 zi1 bat1jik6 jyut6 fu4
- All these are done in Cantonese, without the need of any Putonghua knowledge (before the introduction of Putonghua in Chinese lessons)

Luckily

- Han character knowledge allows Cantonese speakers to write out nearly everything in Written Cantonese
- Compare
  - 我 今天 不 吃飯
    ngo5 gam1tin1 bat1 hek3faan6

    Standard Written Chinese
  - 我 今日 唔 食飯
    ngo5 gam1jat6 m4 sik6faan6

    Written Cantonese

日 - a common character that is taught in kindergarten
食 - a common radical, and frequently used in Classical Chinese
唔 - not found, requires exposure

Descriptive approach

- Cheung & Bauer (2001)
  - Acknowledges the fact that there is no rigid standardisation
  - Collected characters used in true Written Cantonese texts
  - These characters have been accepted into Unicode and provides a large pool of available forms for future innovation

030 / 08	FB7E	啲	gau6	lump; clsfr. for magnet, stone	一~石；一~攝石	*jat1~sek6* a stone; *jat1~sip3sek6* a magnet	JDB 1908:129 130
038 / 07	FB70	娞	gei2	see ex.	佢有身~	*keoi5jau5san1~* she's pregnant	HPP 1970:455
030 / 20	FBD6	嚤	geng6	guard against; take precautions	~住	*~zyu6* take precautions	HKJ 1998a:1
064 / 13	FCE3	撳	geng6	guard against; take precautions; take care doing sth.	响病房行路都~住	*hoeng2 beng6 fong4/2 haang4 lou6dou1~zyu6* take care when walking in sick room so as not to make noise	ZBH 2002:228
037 / 17	FA5F	煠	gip6	narrow; petty; crowded	窄~	*zaak3~* narrow; crowded	YMX 1999:111
085 / 10	FA60	滐	git6	dense; thick; great	個湯好~嘅	*go3tong1hou2~ge3* soup is very thick	YMX 1999:112
030 / 08	FA41	哿	go2	that	~個人	*~go3jan4* that person	MKL 1998:119
124 / 10	FA5D	猗	go6	spoil; dote on; shield	佢好~佢細佬	*keoi5hou2~keoi5sai3lou2* he very much	YMX 1999:84

Writing Strategies

Note1: Rare characters can be recycled in the process of character creation.

Note2: Not necessarily English loans

Note3: and all shared items with Mandarin / Classical Chinese

Use homophones	琴日	kam4 jat6	yesterday	啤啤 晒 緊 啦嚨 騎呢 升呢 比 左 乜水 仲 聽日
Create new characters	揮	bat1	to scoop	𨋢 噏 揼 揦 跣 瀡 𡳞 癐 [illegible] 扱 轆 叻 諗 攞 [1]
English orthography	hea	he3	lazy	BB, mum mum, pat pat, jetso, keep, D [2]
Etymological roots	攰	gui6	tired	重 滴 井 乜誰 畀 奇離 天光日 [3]

- Multiple forms may exist for the same lexical items. E.g. 啤啤 - BB, 左-咗, 晒-嗮, 癐-攰.
- Etymological roots (本字) are not necessarily the preferred form for various reasons, e.g. **disambiguation**, but other strategies are sometimes considered inferior.

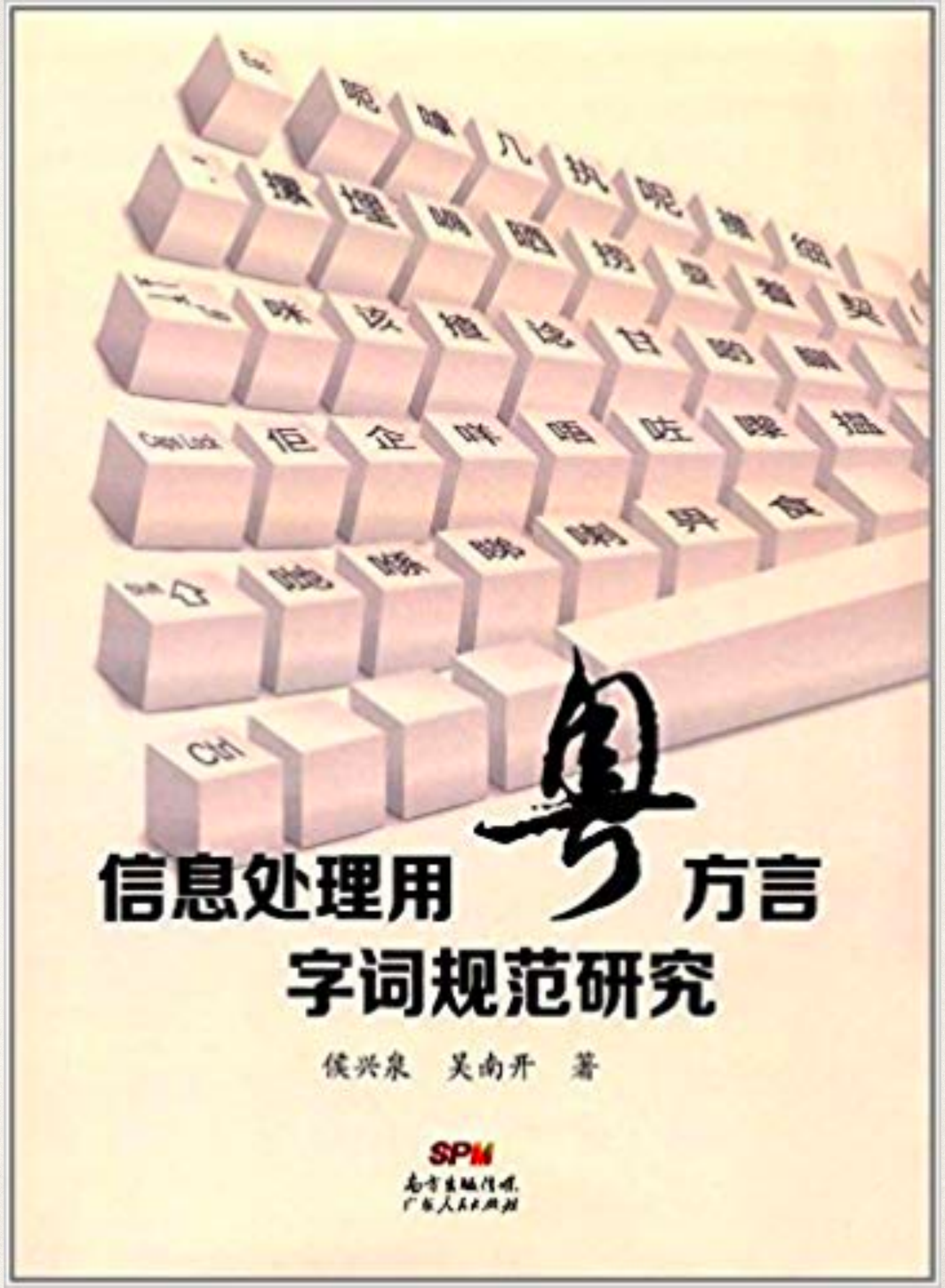
Etymological claims

- Etymological roots preference attracts dubious, controversial claims from both amateurs and experts. For example,
- Claimed etymology:
  - 尐人匸爾道等亙渠等
- Commonly accepted form:
  - 啲人喺呢度等緊佢哋
- The top row has been rejected by the community.

Standardisation Attempts in Mainland China

- 侯興泉、彭志峰、鍾奇、彭小川。2014。面向中文信息處理的粵方言字規範雛議。《語言教學與研究》4，107-112。

- 侯興泉、吳南開。2017。《信息處理用粵方言字詞規範研究》。廣東人民出版社。



2	通用粵字	學會粵拼	直音反切(註)	意思及例子	可能本字	疑問或建議/註解
597	摺	zip3	接	折、疊：～紙、～被		
598	澀	zit1	擸	液體濺出或噴出：畀啲水～到週身濕晒（被水濺到滿身濕透）、佢用水槍～我（他用水槍射我）		
599	擸/唧	zit1	澀	通過擠壓以把內部東西逼出：～牙膏、～暗瘡		
600	照	ziu3	詔	照顧，關照：放心啦，入面有熟人～住（放心吧，裏面有熟人關照）、我～你（我關照着你）		
601	嗤	ziu6	趙	嚼、咀嚼：～香口膠、～完鬆（比喻佔了便宜就跑）	噍	
602	撏/趙	ziu6	趙	打：～佢一鑊（打他一頓）	撏/肇	
603	咗	zo2	阻	表示動作完成，相當於「了」：完成～任務（完成了任務）、走～（走了）	徂/著	原爲"咀"
604	眦/裝	zong1	莊	窺視，偷窺：～人沖涼	瞛/覬	
605	盅	zung1	忠	一種體積較少的器皿：煙灰～、口～、茶～		
606	舂	zung1	忠	撞向或用拳擊打：～個頭埋去，一拳～過去	撞/捶	
607	仲/重	zung6	仲	還，再，更：～有（還有）、～係（還是）、～好（更好）		
608	啜	zyut3	至血切	用嘴吮吸，吻：～田螺、～一啖（吻一下）		

- 粵語協會/通用粵字表
  - Online effort to promote a standardised form for each morpheme
  - The list grew to 608 characters as of Jun 2019

But…

- Top-down approach has limited impact
  - It is important for dictionary editors and institutions to understand their role as descriptive facilitators, instead of authorities who dictate how things are done.
  - Standardisation in a self-organised act by the community
- **Core principle:**
  Individual chooses a character that maximises readability

From Reading Strategies to Writing Practice

- Recall that except for well-accepted written forms and characters acquired in school, users are not taught how Cantonese-specific lexical-items are written.
- Homophones and English orthography can be understood straightaway (but aesthetically not preferred, and affects the readability of the text as a whole).
- If a less-common character is used (whether it is newly-created or etymologically related), speakers would need to guess from the context.
- Writers have the incentive to use characters that readers can make a correct guess.

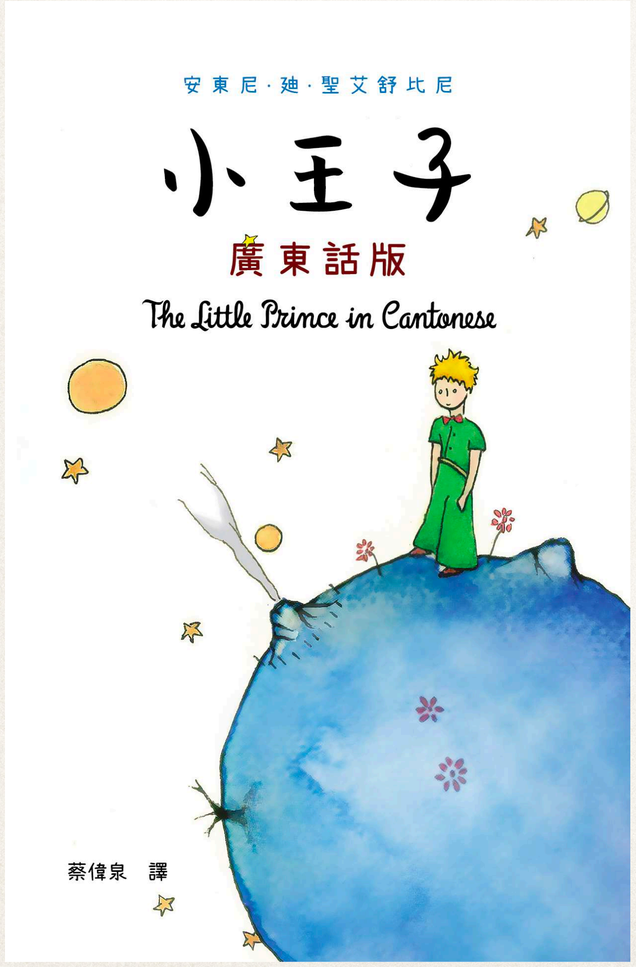
Positive Loop

- **Step 1: Morpheme X does not have a universally accepted character**
  e.g. the word for slippery: sin3
- **Step 2: Multiple candidates appear (innovation)**
  e.g. (a) 扇 (b) 鱔 (c) 跣 (d) 蹝 (e) sin (f) seen (g) 偃
- **Step 3: Effective forms stay, bad forms disappear**
- **Step 4: Best practice is copied by other writers**
- **Step 5: By consistently using the same character C to represent X, C becomes the universally accepted character for X**

A case of disambiguation

- Both 啦 and 喇 were used for sentence particle *laa*, which comes in four different tones: *laa1, laa3, laa4, laa5*.
- Homophone association:
  - 啦 is used in 啦啦隊 (laa1 laa1 deoi2) cheerleader
    喇 is used in 喇沙 (laa3 saa1). La Salle (proper name)
- Writing characters with the correct tone association improves readability. Writers prefer using 啦 instead of 喇 to represent laa1, creating a new norm
- 個 go3 vs. 嗰 go2, 係 hai6 vs. 喺 hai2 may have gone through a similar process

- 蔡偉泉譯。2017。《小王子廣東話版》。香港：藍出版。



凡例

廣東話至少有兩項特徵，令其書寫有一定挑戰。第一，廣東話中有不少字沒有標準寫法，甚或有音無字。有一些字雖然可寫，但寫法生僻，不為大部份人所認識。第二，廣東話中有不少句末助詞，當中有部份聲母韻母俱同，只透過不同聲調來區分微妙的語義差異。然而，這些聲調不同的助詞往往使用同一個漢字表達，容易混淆讀者。

對於第一個問題，本譯本以盡量便利溝通為原則。假如個別廣東話本字過於生僻，而該字有另一通用寫法，則使用通用寫法，並附加註腳補充。在有需要之處，亦會在註腳以粵拼系統標音。如有興趣對粵拼作進一步了解，可參考以下網站：

粵拼網上教室
http://www.iso10646hk.net/jp/learning/index.jsp

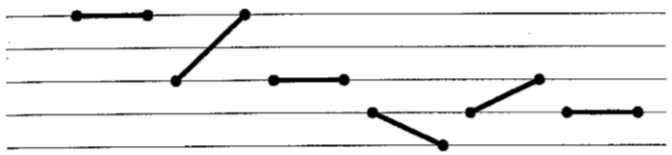
此外，在此特別針對「俾、畀」及「咁、噉」兩對字作補充說明。「俾」和「畀」本來有不同意思，「畀」本意指「給予」或被動的「容許」，而「俾」則有「令到」的意思。不過近年在香港，不少人一概以「俾」字表達兩個意思。同樣地，不少香港人以「咁」表示「咁」和「噉」兩個本來意義不同、甚至讀音也不

一樣的字。「咁」本唸「禁」，用作修飾形容詞，如「咁大」。「噉」唸「敢」，可作指示代詞，如「噉就啱嘞」；也可作結構助詞，用以標示副詞，如「慢慢噉食」。在本譯本中，兩對字皆沿用本來用法。

對於第二個問題，本書使用以下規則，區分部份常用句末助詞的讀音：

啦 laa1		喇 laa3	嗱 laa4
		嘞 laak3	
喺 gaa1	嘅 gaa2	㗎 gaa3	嘎 gaa4
		啫 gaak3	

以下提供廣東話的六個聲調比較作參考。



聲調	1	2	3	4	5	6
調類	陰平	陰上	陰去	陽平	陽上	陽去
調值	55	35	33	21	23	22
字例	詩	史	試	時	市	是

另外有三個入聲聲調以-p、-t或-k收尾，如「嘞（laak3）」。陰入、中入及陽入三個聲調調值分別為5、3及2。

Dictionary: words.hk

- **ROLE 1 descriptive documenter:**
  ensure existing acceptable forms are included (and outdated forms are excluded) in a timely manner

- **ROLE 2 innovator:**
  propose a written form if none exists

- **ROLE 3 gatekeeper:**
  reject dubious etymological claims, e.g. 𠂤, 匚

- **ROLE 4 facilitator:**
  take a side if necessary: set one as the default

Cantonese tests by LSHK

- Cantonese read-aloud test (https://lshk.org/crat)
- **Decision to make:** when there are two competing written forms, choose the one that increases readability, i.e. the character with only one pronunciation
  - 仲 (homophone) vs. 重 (etymologically better justified) for zung6 “still”
    - 重 has two other readings: cung4 (redo), cung5 (heavy) — **ambiguity**, e.g. 我重(cung4/zung6)讀緊中五. (I am re-studying / still studying Form five.)
    - 仲 is used in all our materials.

#01

有個喺深水埗賣遮嘅先生，叫威哥。佢做生意嘅手法，喺今時今日，完全係不可思議。	jau5 go3 hai2 sam1seoi2bou2 maai6 ze1 ge3 sin1saang1, giu3 wai1go1. keoi5 zou6 saang1ji3 ge3 sau2faat3, hai2 gam1si4gam1jat6, jyun4cyun4 hai6 bat1ho2si1ji5.
我唔知佢幾歲，應該都六十幾七十咁上下，一面白鬚，着住件短袖恤衫加西褲，聲如洪鐘，目光如炬。佢話祖上自香港割讓俾英國嗰陣已經開始遮呢行生意。唔係咩大企業，但係做咗兩百幾年都仲屹立不倒。港交所有冧把未？唔好意思，佢似乎滿足喺自己間小店嘅天地，而且淨係有興趣教人點樣開遮、收遮、摺遮，多過賣遮。有人拎把遮俾佢修理，佢一概照收；但因為得佢一個人做，依家囤積咗千幾把遮未整。佢強調話死之前都一定整唔晒添喎。	ngo5 m4 zi1 keoi5 gei2 seoi3, jing1goi1 dou1 luk6sap6gei2 cat1sap6 gam3soeng6haa2, jat1 min6 baak6sou1, zoek3zyu6 gin6 dyun2zau6 seot1saam1 gaa1 sai1fu3, sing1jyu4hung4zung1, muk6gwong1jyu4geoi6. keoi5 waa6 zou2soeng6 zi6 hoeng1gong2 got3joeng6 bei2 jing1gwok3 go2zan6 ji5ging1 hoi1ci2 ze1 ni1 hong4 saang1ji3. m4 hai6 me1 daai6 kei5jip6, daan6hai6 zou6zo2 loeng5baak3gei2 nin4 dou1 zung6 ngat6laap6bat1dou2. gong2gaau1so2 jau5 lam1baa2 mei6? m4hou2ji3si1, keoi5 ci5fu4 mun5zuk1 hai2 zi6gei2 gaan1 siu2dim3 ge3 tin1dei6, ji4ce2 zing6hai6 jau5 hing3ceoi3 gaau3 jan4 dim2joeng2 hoi1 ze1, sau1 ze1, zip3 ze1, do1gwo3 maai6 ze1. jau5 jan4 ling1 baa2 ze1 bei2 keoi5 sau1lei5, keoi5 jat1koi3 ziu3 sau1; daan6 jan1wai6 dak1 keoi5 jat1 go3 jan4 zou6, ji1gaa1 tyun4zik1zo2 cin1gei2 baa2 ze1 mei6 zing2. keoi5 koeng4diu6 waa6 sei2 zi1cin4 dou1 jat1ding6 zing2 m4 saai3 tim1 wo5.

塔巴〈賣遮先生〉（節錄改寫）

Cantonese tests by LSHK

- **Decision to make:** when there are two competing written forms, choose the one that increases readability, i.e. the character with only one pronunciation
  - 仲 (homophone) vs. 重 (etymologically better justified) for zung6 “still”
    - 重 has two other readings: cung4 (redo), cung5 (heavy) — **ambiguity**, e.g.
      我重(cung4/zung6)讀緊中五. (I am re-studying / still studying Form five.)
    - 仲 is used in all our materials.

Conclusion

- How? Cantonese speakers use characters from their Han character repertoire to represent Cantonese as they see fit. Exposure to Written Cantonese + Reader's strategy improve consistency across speakers.

- Writing Strategies: **(A)** Use etymologically related characters (for items shared with Standard Chinese or Classical Chinese), **(B)** If it does not exist, write with a homophone, English orthography, or create a new character. Good forms spread, bad forms are rejected.

- Top-down approaches do not work. Dictionaries and institutions as a relatively more powerful user. Two major roles: **(1)** descriptive documenter, most of the time, **(2)** innovator, when no good alternative exists. Additional roles which can facilitate the process (by words.hk and LSHK): (3) gatekeeper, who can reject dubious etymological claims, (4) deliberately choose a form that increases readability in case of a conflict.

References

- Lau, C. (2019). Building Cantonese Dictionaries Using Crowdsourcing Strategies: The words.hk Project. In: Tso A. (eds) Digital Humanities and New Ways of Teaching. Digital Culture and Humanities (Challenges and Developments in a Globalized Asia), vol 1. Singapore: Springer.

- Snow, D. (2013). Towards a theory of vernacularisation: Insights from written Chinese vernaculars. Journal of Multilingual and Multicultural Development, 34(6), 597-610.