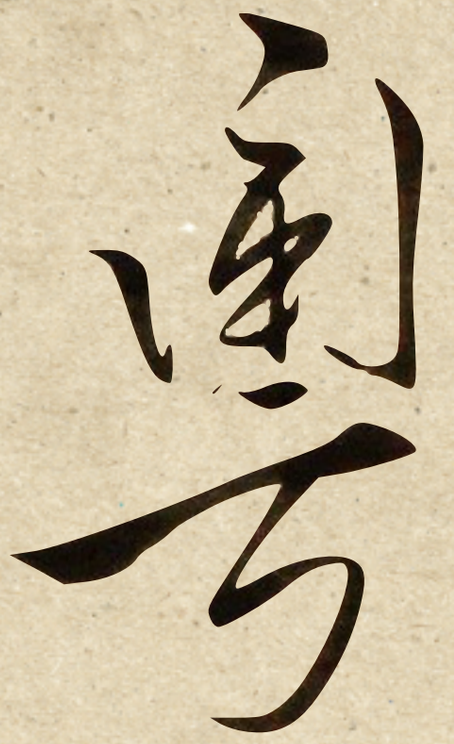


words.hk

Building Cantonese Dictionaries Using Crowd-sourcing Strategies: The words.hk Project

Chaak-ming Lau
words.hk / CUHK

2017.03.17

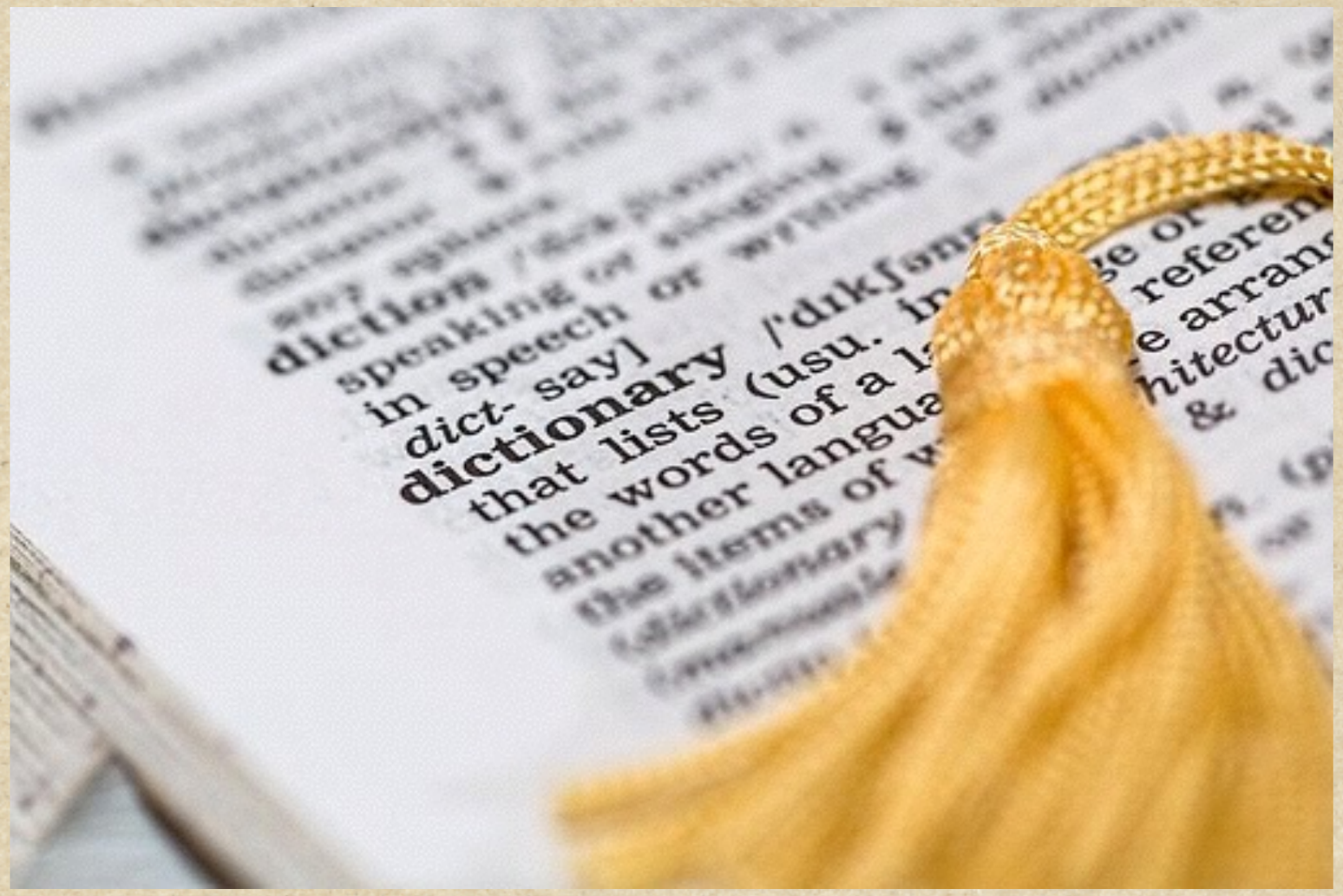
Cantonese

Is it a language?

- No → End of story
- Yes → Continue

What is a dictionary?

- Length
  - ✤ 50 words?
  - ✤ 1000 pages?
- Content
  - ✤ complete list?



Group Discussion

- in groups of four
- discuss (for 5 minutes)
  - ✤ What should be included in a Cantonese dictionary
  - ✤ length, what kinds of words, what should be excluded
  - ✤ Write a sample entry for your dictionary
  - ✤ Use 1 minute to present your ideas

Group Presentation

Let's turn to English ...

Which of these words...

- do you expect in an English dictionary?
  - the
  - man
  - attention
  - bling
  - dim-sum
  - pragmatics
  - Justin Bieber

Found one...

the

determiner · **UK** STRONG /ðiː/ WEAK /ðə/ **US** STRONG /ðiː/ WEAK /ðə/

the *determiner* **(PARTICULAR)**

A1 used before nouns to refer to particular things or people that have already been talked about or are already known or that are in a situation where it is clear what is happening:

I just bought a new shirt and some new shoes. The shirt was pretty expensive, but the shoes weren't.

Please would you pass the salt.

I'll pick you up at the airport.

A1 used before some nouns that refer to place when you want to mention that type of place, without showing exactly which example of the place you mean:

We spent all day at the beach.

Let's go to the movies this evening.

I have to go to the bank and get some Euros.

We have ...

Jesus Christ

noun • **UK** /ˌdʒiː.zəs ˈkraɪst/ **US** /ˌdʒiː.zəs ˈkraɪst/ (ALSO Christ,)(ALSO Jesus)

the man believed by his religious followers to be the son of God. Christianity is based on his life and teachings.

but not ...

Search suggestions for Justin Bieber

We have these words with similar spellings or pronunciations:

1	just in time
2	just in case
3	just-in-time
4	justifiable
5	justifiably
6	justified
7	justifies

man

noun · UK /mæn/ US /mæn/ PLURAL **men** UK /men/ US

man *noun* **(MALE)**

A1 [C] **an adult male human being**

dim sum

noun [U] · /ˌdɪm ˈsʌm/ /ˌdɪm ˈsʌm/

a Chinese meal or snack of small dishes including different steamed or fried foods:

pragmatics

noun [U] · /prægˈmæt.ɪks/ /prægˈmæţ.ɪks/ SPECIALIZED

the study of how language is affected by the situation in which it is

- **Found**
  - the (grammatical words)
  - man (basic words)
  - attention (harder words)
  - bling (slangs)
  - dim-sum (recent loanwords)
  - pragmatics (specialised words)
  - Jesus Christ (important proper names)
- **Not Found**
  - Justin Bieber

Look at another entry

What language is this?

bling

noun [U] · UK /blɪŋ/ US /blɪŋ/ INFORMAL

jewellery or decoration that attracts attention because it is very noticeable and looks expensive:

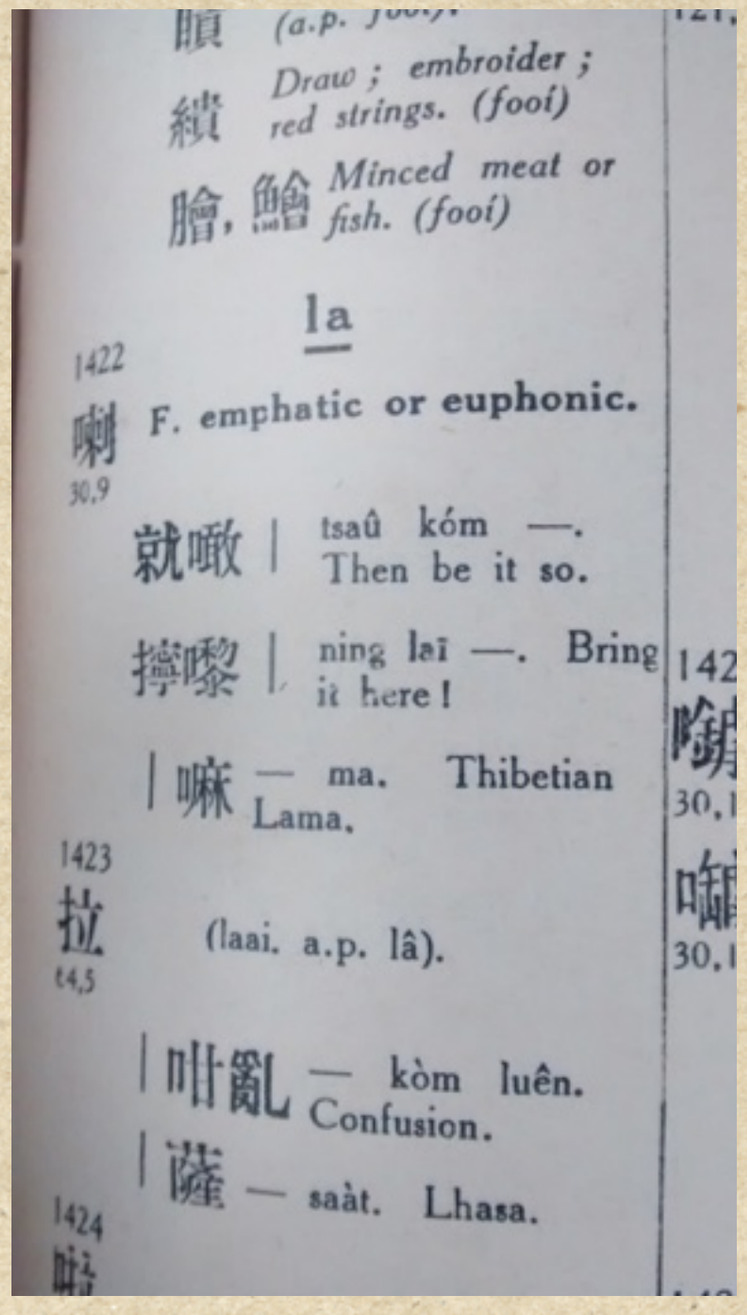
She wore a fake-fur coat, big sunglasses and lots of bling.

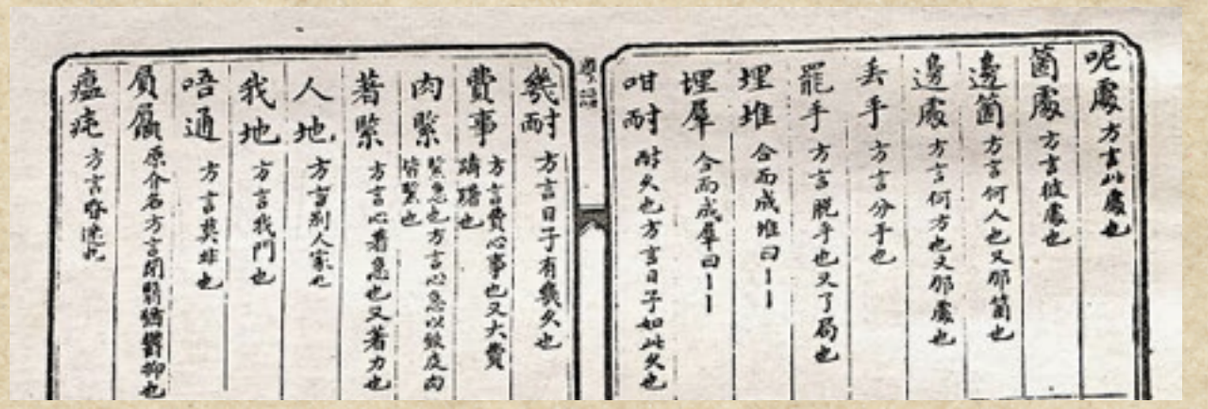
and what language is this?

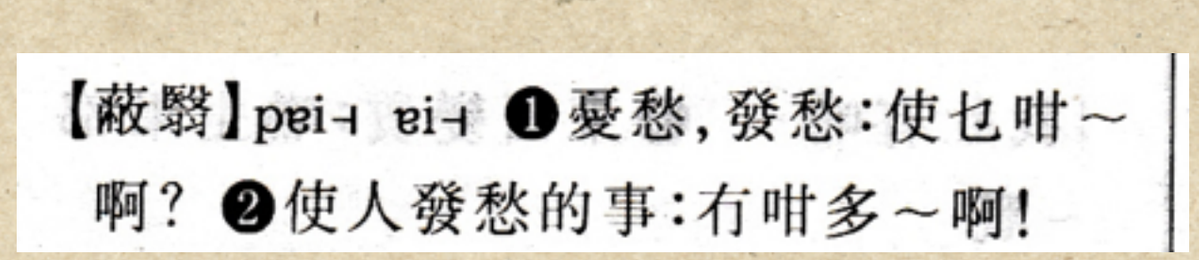
Expectation

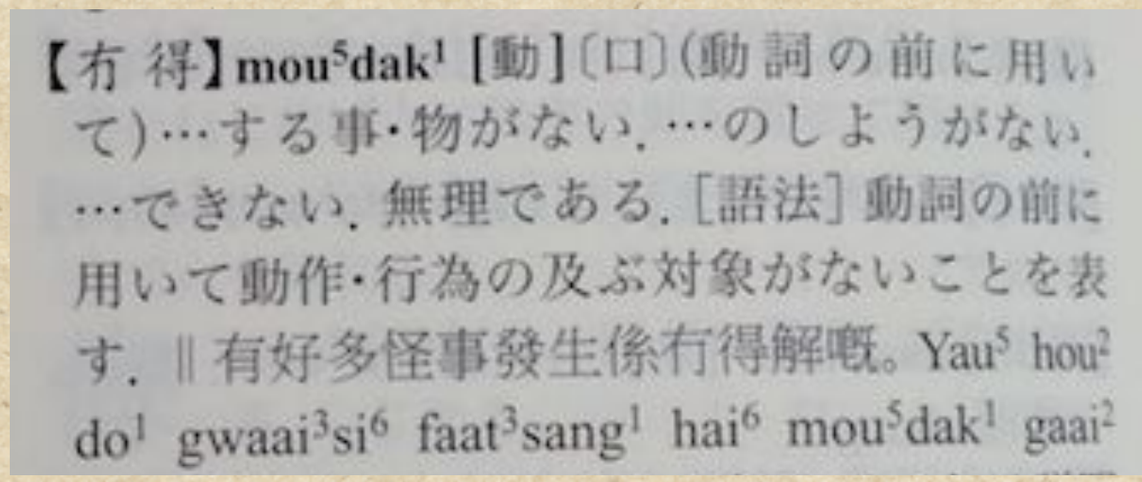
- English dictionaries
  - ✤ Comprehensive
  - ✤ Written in English
- Cantonese dictionaries
  - ✤ Same principles?
- Question: What should a Cantonese dictionary look like?

Existing Cantonese Dictionaries






【蔽翳】pɐi⊣ ɐi⊣ ❶憂愁，發愁：使乜咁～啊？ ❷使人發愁的事：冇咁多～啊！


【冇得】mou⁵dak¹ [動]〔口〕(動詞の前に用いて)…する事・物がない，…のしようがない，…できない，無理である．[語法] 動詞の前に用いて動作・行為の及ぶ対象がないことを表す．‖有好多怪事發生係冇得解嘅。Yau⁵ hou² do¹ gwaai³si⁶ faat³sang¹ hai⁶ mou⁵dak¹ gaai²

粵典

Monolingual dictionaries?

- Issues
  - ✤ not comprehensive, only contain:
    - ✦ words that do not overlap with "mainstream Chinese", i.e. Mandarin
    - ✦ or harder words

> ***objection***
>
> ✦ Sorry we don't have that word,
> have you tried a French dictionary

  - ✤ Descriptions written for Chinese / Mandarin users

Online resources

- There is a nice one, don't waste your time!
- Character-based, not word-based

words.hk I Cantonese Dictionary I Cantonese Corpus
brought to you by Hong Kong Lexicography Limited

The Project

Goal
Compilation
The Role of Social Networks

words.hk

- **Build a true Cantonese dictionary**
  - C1: Written in Cantonese ;
  - C2: Comprehensive, i.e. all common Cantonese words, spoken or written / unique to Cantonese or shared with Mandarin, should be included; and
  - C3: Words (rather than characters) will be the default unit

Compilation

- **NOT corpus-based compilation**
- **Existing bilingual dictionaries**
  - **HKUDict** (Luke et al., unpublished dictionary data)
  - **A Dictionary of Cantonese Slang** (Hutton & Bolton, 2005)
  - **Lexical Lists for Chinese Learning in Hong Kong** (Education Bureau, 2007, 2009)
- **Existing word-lists**
  - **現代標準漢語與粵語對照資料庫** (CUHK, 2001)
  - **Word lists from various Input Method Engines**

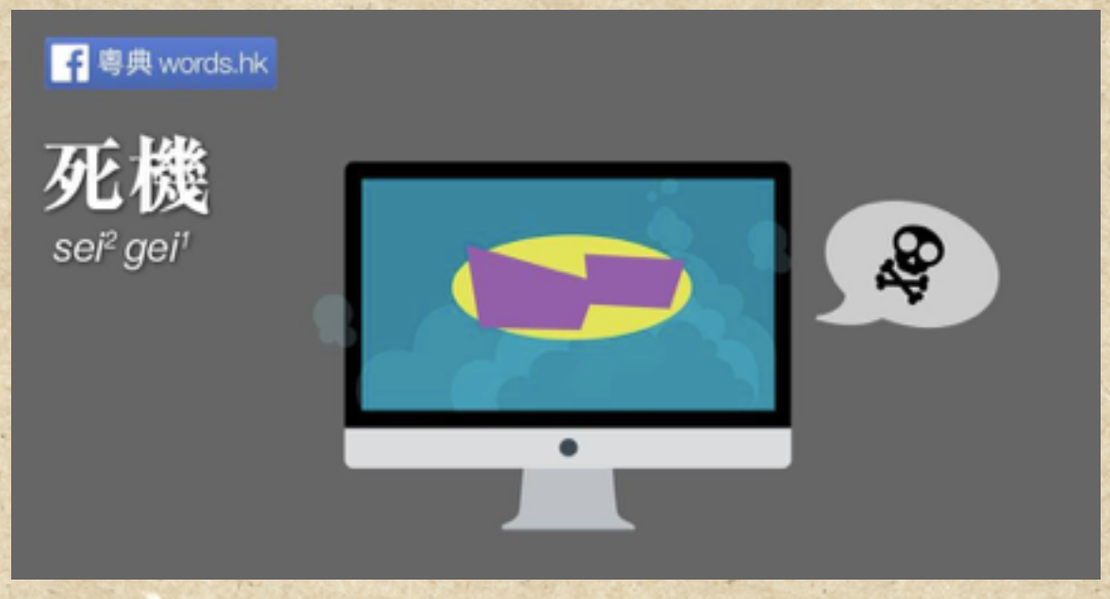
Compilation (con't)

- Imported entries are
  - ✤ used as reference only
  - ✤ rewritten or restructured manually by editors

The role of social networks

- **Facebook**
  - ✤ Page (words.hk)  words.hk
    - ✦ 10,000+ likes as of today
  - ✤ Member recruitment
    - ✦ Almost all members are recruited on Facebook
  - ✤ A word a day
- **Wordathon meetings**
  - ✤ Next: 18 Mar 2017 (Sat.) 3pm @Mong Kok

A word a day - 每日一字



Illustrator: Can's Illustrations

A word a day - 每日一字

Illustrator:P of KatLau

Sample Entries

食
sik^{6}

【動】1. 經食道吞下 to eat 我未食過臭豆腐，好想試下。*I've never eaten stinky tofu. I really want to give it a try.* 你食咗晏未？ *Have you had lunch yet?* 你食咗藥未？ *Have you taken your medicine yet?* **2. 經氣管或其他方法吸入或注入身體 to inhale, inject, or otherwise absorb into the body** 食煙 *to smoke (cigarettes)* 食白粉 *to shoot heroin* **3 .承受痛苦 to suffer; to be on the receiving end of bad things** 我食咗佢好多拳。 *I took a lot of punches from him* 業主又要加租，小商戶冇議價能力，唯有硬食。 *The landlord is increasing the rent again. With no bargaining power, the tenants have no choice but to accept it.* **4. 消耗、耗用 to consume, to expend** 呢隻app好食電 *This app consumes a lot of power* 食晒我啲時間 *taking up all my time* **5. 征服、吞併、控制、抽取 to conquer; to embrace, extend and extinguish; to control; to extract something from** 呢個明星恃住自己有錢又靚仔，食咗好多條女 *Using his*

「食」

出fb post版本 | 收返埋 | 修改 | 刪除

解釋 #1

讀音： *sik⁶*

詞性： 動詞

解釋：

1. *(yue)* 將一啲嘢擺入口到咬，然後經食道吞入肚裏面
   (eng) to eat

   配詞 / 用法：

   (yue) 食飯 *(sik6 faan6)*
   (eng) to have a meal

   (yue) 食晏 *(sik6 aan3)*
   (eng) to have lunch

   例句：

   (yue) 我未食過臭豆腐，好想試下。 *(ngo5 mei6 sik6 gwo3 cau3 dau6 fu6, hou2 soeng2 si3 haa5.)*
   (eng) I've never eaten stinky tofu. I really want to give it a try.

2. *(yue)* 服用藥物
   (eng) to take (medicine)

   例句：

   (yue) 你食咗藥未？ *(nei5 sik6 zo2 joek6 mei6?)*
   (eng) Have you taken your medicine yet?

3. *(yue)* 經氣管或者其他方法吸入或注入身體
   (eng) to inhale, inject, or otherwise absorb into the body

   配詞 / 用法：

Online Interface

- http://beta.words.hk/食

Status of the project

- On-going project
  - ✤ Active discussion every day
  - ✤ Bi-weekly wordathon meetings
- ~500 volunteers
- 42000+ entries
  - ✤ 34000+ edited w/ Cantonese definitions

Methodology

Agile vs. waterfall

Core Principles

Usage-over-etymology

Decision Problem Avoidance

Methodology

- Not executed by step-by-step plan
- Agile vs. waterfall
  - ✤ Concept borrowed from software engineering and lean start-ups
  - ✤ Ensure maximum flexibility
- Adhere to Core Principles

Core Principles

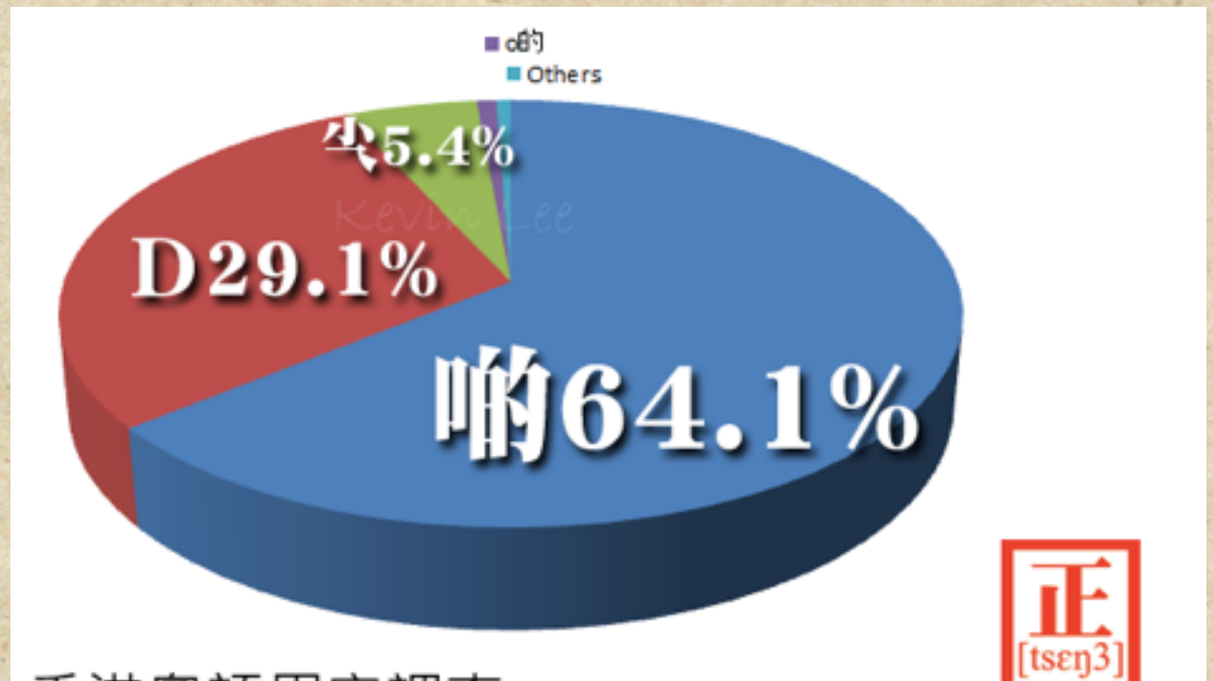
- **Usage-over-etymology**
  - ignore etymology
  - have everything based on usage
  - explaining actual usage is preferred over existing dictionaries or etymologies
- **Decision Problem Avoidance**
  - Decision problems:
    - something that requires a judgement
    - e.g. classification
  - Avoid these problems

Usage-over-etymology

- e.g. Which is the correct character?
  - ✤ The case of 啲 di1
    - ✦ Possible characters
      - ❖ 啲、D are commonly used
      - ❖ 尗 is the proposed character by amateur etymologists
      - ❖ 的 is used in older texts
      - ❖ 滴 is probably the true etymology, but no longer used / recognised by native speakers

Usage-over-etymology (con't)

- 啲／D／尐 are listed, ordered by frequency
  - based on editors' subjective judgement
  - agrees with Cantonese character usage survey conducted by 港語學



香港粵語用字調查

正 [tseŋ3] 港語學

			年齡					性別	
	全部	%	<19	20~29	30~39	40~49	>50	男	女
啲	2138	64.1%	+0.9%	-2.3%	+1.6%	+13.1%	-4.8%	-0.3%	+0.6%
D	972	29.1%	+1.8%	+4.5%	-7.3%	-19.4%	-9.4%	-1.1%	+1.7%
尐	181	5.4%	-2.3%	-2.1%	+5.4%	+6.7%	+8.2%	+1.3%	-2.2%
o的	27	0.8%	-0.4%	+0.0%	+0.4%	-0.3%	+1.7%	-0.1%	+0.2%

			本土派			無特定政治傾向	泛民主派	
	全部	%	民族論	城邦派	歸英			左翼
啲	2138	64.1%	+8.8%	+1.6%	-3.2%	-1.3%	-4.0%	+0.1%
D	972	29.1%	-7.5%	-5.3%	+2.8%	+2.8%	+5.6%	-0.8%
尐	181	5.4%	-1.1%	+3.9%	+0.3%	-1.3%	-1.8%	-1.6%
o的	27	0.8%	+0.1%	+0.0%	+0.3%	-0.0%	+0.0%	+1.1%

Decision Problem Avoidance

- Cantonese vs. Non-Cantonese
  - Major problem faced by previous Cantonese dictionaries
    - Three-way distinction (鄧思穎 2015, Ch. 3)
      - 通用詞 shared vocabulary
      - 社區詞 community vocabulary
      - 方言詞 topolectal vocabulary
- Issues
  - implies a decision problem for every entry created
  - no easy test: prone to dispute
- Solution
  - Only label items not used at all in spoken Cantonese

Decision Problem Avoidance (con't)

- Other problems
  - word-class classification
    - noun or verb? verb or adjective?
  - headword problem
    - should variants, e.g. 擦膠 (caat3 gaau1, eraser), 膠擦 (gaau1 caat3), 擦紙膠 (caat3 zi2 gaau1), listed separately?
    - should K (kei1, karaoke) listed separately?
      - or just under K房 (kei1 fong2, ~room), 唱K (coeng3 kei1, to sing~), 劈K (pek3 kei1, to get drunk at ~)?

Decision Problem Avoidance (con't)

- proper nouns
  - 周公 zau1 gung1
  - 青山 cing1 saan1
- jargons
  - e.g. jargons within political circles?
    - e.g. 左膠 zo2 gaau1
      - (Look it up on words.hk)

Conclusion

Issues
Future of Cantonese Lexicography
On Open Data

Issues

- Lexicography
  - Over-emphasis on Etymology / Characters
  - Prescriptivism
  - Fail to catch up with latest language developments
- Lack of Available resources
  - Lock-ups
  - Paper-based

Future of Cantonese Lexicography?

- Word-based dictionaries
- Use both word-list compilation and corpus-based approach
- Crowd-sourcing to generate content
- Let the users decide the preferred written form

Open Data Policy

- words.hk License Agreement
  - http://beta.words.hk/base/hoifong/
  - dictionary data are
    - freely accessible to the public
    - can be freely modified, adopted, and published for non-commercial and certain small-business purposes
- avoid lock-ups

Appeal to the community

- Locked-up resources
  - never been published, protected by copyright
  - unusable
- Say goodbye to the old “proprietary” age!

Finally

- Get the priorities right
  - ✤ w.r.t. Cantonese lexicography
  - ✤ it is language description
    - ✦ i.e. keep a record of:
      - ❖ usage
      - ❖ diversity
  - ✤ NOT decision problems
    - ✦ folk etymology (like where do the ?
    - ✦ existing本字考?

Reference

- Education Bureau (2007, 2009). *Lexical Lists for Chinese Learning in Hong Kong.* Available on http://www.edbchinese.hk/lexlist/
- Ferguson, C. (1959). *Diglossia.*
- Hutton, C., & Bolton, K. (2005). *A dictionary of Cantonese slang: the language of Hong Kong movies, street gangs and city life.*
- Snow, D. B. (2004). *Cantonese as written language the growth of a written Chinese vernacular.*
- 招子庸. (1821). 粵謳.
- 鄧思穎. (2015). 粵語語法講義. 香港: 商務印書館.