# Forms and Meanings of Lexical Reduplications in Cantonese: a corpus study

Charles Lam
charleslam@hsu.edu.hk
Department of English, The Hang Seng University of Hong Kong

Oct 25, 2020

Workshop on Multiword Expressions in Asian languages, PACLIC-34
VNU University of Science, Hanoi

http://charles-lam.net/presentations/

# Overview

- Reduplications in Cantonese: Productive vs. Non-productive
- Lexical reduplications (LR): Multiword Expressions (MWEs) that are frozen, unproductive, non-compositional
- Data extracted from two dictionaries as a lexical resource
- Potential use (novel term detection; POS-tagging)

# Reduplication in Cantonese

- Productive
- Predictable input-output relation
- Not multiword expressions

(1)

a. zek3 zek3 gau2 隻隻狗
CL CL dog
'every dog' — N → every N

b. haang6 haang6 haa5 行行吓
walk walk PRT
'while walking' — V → durative event

c. mong6 jat1 mong6 望一望
look one look
'to take a look' — V → brief occurrence

d. coeng4 coeng2 dei2 長長地
long long DEI
'long-ish; fairly long' — Adj → diminution; hedging

# What is Lexical Reduplication?

(2) haang4 haang4 kei5 kei5 行行企企
walk walk stand stand
'being idle and aimless'

(3) tiu3 tiu3 zaat3 跳跳紮
jump jump tie
'bouncy and active'

- Unproductive; frozen forms
- Found in all syntactic categories
- The meaning is often non-literal (but literal meaning is also possible for some items)
- Often unique in Cantonese and no direct equivalent in Mandarin

The orthography follows 《現代標準漢語與粵語對照資料庫》

# What is Lexical Reduplication?

- Most LRs are fixed and frozen
- Some items have multiple variations with no obvious differences in meaning or distribution
- No corresponding 'base form', i.e. neither **tiu3 zaat3* nor **zaat3 tiu3* exists as an lexical item

(4)
a. tiu3 tiu3 zaat3 跳跳紮
jump jump tie
'bouncy and active' = example (3)
b. zaat3 zaat3 tiu3 紮紮跳
tie tie jump
c. tiu3 tiu3 zaat3 zaat3 跳跳紮紮
jump jump tie tie
d. *zaat3 zaat3 tiu3 tiu3 * 紮紮跳跳
tie tie jump jump

# Color terms

(5)
a. hung4 bok1 bok1 紅卜卜
red BOK BOK
'bright red'
b. wong4 gam4 gam4 黃黔黔
yellow GAM GAM
'bright yellow'
c. *hung4 gam4 gam4 * 紅黔黔
red GAM GAM
d. *wong4 bok1 bok1 * 黃卜卜
yellow BOK BOK

- LRs of color terms are mostly intensification; sometimes 'excessive'
- The reduplicated syllables are often non-words
- Fixed matching between the color and the reduplicated syllables (see (5c) and (5d))

# Related Works

- Vast majority of the literature focuses on phonology/morphology of productive reduplications across languages (Wilbur, 1973; Botha, 2006; Frampton, 2009; Inkelas & Zoll, 2005; Hurch, 2005; Francis et al., 2011; Štekauer et al., 2012)
- Syntax-semantics of reduplications in Sinitic / Cantonese (Cheng, 2012; Lee, 2020; Lam, 2013; Basciano & Melloni, 2017)
- Idioms in Mandarin as a resource for learners (Chung & Hsieh, 2017; Wang et al., 2013)

**This study: Lexical Reduplications in Cantonese**

- Many LRs are not found in Mandarin
- Lexical resource for word segmentation / parsing

# Two data sets

1. Cantonese Dictionary by Cheung, Ngai and Poon (2018)
   張勵妍、倪列懷、潘禮美《香港粵語大詞典》(752 entries manually extracted, out of 12,000 entries)
2. Online dictionary `Words.hk` 《粵典》
   # of Types: 30,281; # of Tokens: 2,938,248

LRs represent 6.3% and 2.8% of the types in data sets #1 and #2, respectively.

# Distribution of LR

**CNP data set (Cheung, Ngai & Poon)**

| Length | Types | % of LR |
|---|---|---|
| 2 char | 23 | 3.06% |
| 3 char | 268 | 35.64% |
| 4 char | 298 | 39.63% |
| 5 char | 27 | 3.59% |
| 6 char | 23 | 3.06% |
| 7 char | 33 | 4.39% |
| 8 char | 9 | 1.20% |
| $\geq$ 9 char | 71 | 9.44% |
| *Total* | 752 | 100.00% |

Table 1: Summary of LR in the CNP data set

**WHK data set (Words.hk)**

| Length | Types | Tokens |
|---|---|---|
| 2 char | 138 (16.01%) | 9,870 (69.74%) |
| 3 char | 233 (27.03%) | 1,942 (13.72%) |
| 4 char | 491 (56.96%) | 2,341 (16.54%) |
| *Total* | 862 (100%) | 14,153(100%) |

Table 2: Summary of LR from Words.hk

# LRs with 3 characters

- All possible combinations are attested
- False positives are excluded (e.g. productive reduplication; items that do not involve word formation, e.g. the emergency number '999')

| | Types | |
|---|---|---|
| **Category** | **Words.hk** | **CNP Dictionary** |
| AAB | 88 (37.77%) | 77 (28.73%) |
| ABA | 27 (11.59%) | 11 (4.10%) |
| ABB | 117 (50.21%) | 180 (67.16%) |
| AAA | 1 (0.43%) | 0 (0%) |
| *Total* | 233 (100.00%) | 268 (100.00%) |

Table 3: 3-character LRs by types

Some items in CNP are productive, e.g. 估估下 'by guessing', and therefore excluded

# Examples of AAB

(6) AAB-template:

a. laap3 laap3 ling3 立立令
LAAP LAAP shiny
'shiny'

b. cyun3 cyun3 gung3 串串貢
sacarstic sacarstic GUNG
'sacarstic'

- Some reduplicated elements are meaningless, e.g. *laap3* in (6)
- The meaning might come from the reduplicated or the unreduplicated element

# Examples of ABA

(7) ABA-template:

a. gau2 m4 gau2 久唔久
long.time not long.time
'once in a while'

b. daap3 soeng6 daap3 搭上搭
contact over contact
'to liaise through a third party'

- Opaque, non-compositional meanings

# Examples of ABB

(8) ABB-template:

a. baak6 syut1 syut1 白雪雪
white snow snow
'very white'

b. jin6 dau1 dau1 現兜兜
now DAU DAU
'in cash'

- Many adjectives, but it remains to be confirmed whether ABB is more adjectival than other types by proportion (aside from the theoretical question whether Cantonese has a separate Adj category)
- All color-related LRs are in this format

# LRs with 4 characters

- 'A', 'B' represent reduplicated elements
- 'X', 'Y' are other random non-repeating elements
- The distribution are similar across the two data sets

| | Types | |
|---|---|---|
| **Category** | **Words.hk** | **CNP Dictionary** |
| AABB | 82 (16.70%) | 62(20.81%) |
| ABAB | 7 (1.43%) | 2 (0.67%) |
| AAXY | 68 (13.85%) | 6 (2.01%) |
| XYAA | 39 (7.94%) | 20 (6.71%) |
| AXAY | 259 (52.75%) | 182 (61.07%) |
| XAYA | 36 (7.33%) | 23 (7.72%) |
| AXYA | 0 (0%) | 3 (1.01%) |
| *Total* | 491 (100.00%) | 298 (100.00%) |

Table 4: Subcategories of 4-character LRs by types

# Examples of LR

- Many of the ‘XY’ strings are words by themselves, making these examples more prone to be misrecognized as separated from the 4-character string (e.g. 西裝骨骨、袋袋平安)
- Most logical possibilities are attested, making it difficult to predict when LR appears
- No ‘AAAX’ or ‘XAAA’ forms are attested

# Examples of AABB & ABAB

(9) loi4 loi4 heoi3 heoi3 來來去去
come come go go
'always' AABB-template

(10) bei2ci2 bei2ci2 彼此彼此
each.other each.other
'same to you / each other' ABAB-template

- Opaque meaning
- Elements can often be words (*loi4* 'come', *heoi3* 'go', *bei2ci2* 'each other')

# Examples of AAXY & XYAA

(11) AAXY-template

a. sei2 sei2 dei6 hei3 死死地氣
die die ground air
'reluctantly'

b. doi6 doi6 ping4on1 袋袋平安
pocket pocket peace
'to pocket something while it is available'

(12) sai1zong1 gwat1 gwat1 西裝骨骨
suits bone bone
'being dressed up' XYAA-template

- In (11a), *dei6 hei3* does not seem to contribute to the meaning compositionally
- In (12), *sai1zong1* 'suit' as a noun is clearly related to the meaning of the whole expression

# Examples of AXAY, XAYA & AXYA

(13) mou5 jan4 mou5 mat6 有人有物
no person no thing
'having nothing at all' AXAY-template

(14) sau2 ting4 hau2 ting4 手停口停
hand stop mouth stop
'living from hand to mouth' XAYA-template

(15) seng1 dou1 m4 seng1 聲都唔聲
voice even not voice
'does not even make a noise' AXYA-template

- Alternating pattern
- Parallel between the first and second halves

# Potential use as a lexical resource

```
import pycantonese as pc
pc.segment("好似有黃咁咁黑黵黵紅ㄊㄊ綠油油青BB 白雪雪咁藍色係咩？")
```

```
[ 好似',
 '有',
 黃',
 '咁',
 '咁',
 黑',
 黵',
 黵',
 紅',
 ㄊ',
 ㄊ',
 綠',
 油',
 油',
 '青',
 'BB',
 ' ',
 '白',
 '雪',
 '雪',
 '咁',
 藍色',
 '係',
 '咩',
```

- Improve word segmentation and POS-tagging (e.g. color terms in ABB form) (Lee et al., 2016)
- Novel term detection; flag for idioms (+ collocation of these flagged items)
- Future development: closer look at these forms may help predict their meanings

**Better recognition and comprehension for these idiomatic expressions**

# Conclusion

- Need for resources on the lexical reduplications (LR), which are idiomatic and unproductive
- Great variety in form and meaning in LRs
- Significant presence in the vocabulary, based on data sets
- Potential use for identification of novel / undiscovered LRs in NLP

Thank you!

Comments and questions are welcome!

charleslam@hsu.edu.hk

# References


Basciano, Bianca & Chiara Melloni. 2017. Event delimitation in Mandarin: The case of diminishing reduplication. *Italian Journal of Linguistics / Rivista di linguistica* 29(1). 147–170.

Botha, Rudolf P. 2006. *Form and meaning in word formation: A study of Afrikaans reduplication.* Cambridge University Press.

Cheng, Lisa Lai-Shen. 2012. Counting and classifiers. In Diane Massam (ed.), *Count and mass across languages*, 199–219.

Chung, Liang-Yi & Sheng-Min Hsieh. 2017. Using graphic digital materials in language learning. In *2017 international conference on applied system innovation (ICASI)*, 295–298. IEEE.

Department of Chinese Language and Literature, Chinese University of Hong Kong. 2011. A comparative study of modern Chinese and Cantonese in the development of teaching resources 現代標準漢語與粵語對照資料庫. `http://apps.itsc.cuhk.edu.hk/hanyu/Page/Cover.aspx`. Accessed: 2020-10-18.

Frampton, John. 2009. *Distributed reduplication*, vol. 52. MIT Press.

Francis, Elaine J, Stephen Matthews, Reace Wing Yan Wong & Stella Wing Man Kwan. 2011. Effects of weight and syntactic priming on the production of Cantonese verb-doubling. *Journal of psycholinguistic research* 40(1). 1–28.

Hurch, Bernhard. 2005. *Studies on reduplication.* Walter de Gruyter.

Inkelas, Sharon & Cheryl Zoll. 2005. *Reduplication: Doubling in morphology.* Cambridge University Press.

Lam, Charles. 2013. Reduplication across categories in Cantonese. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation*, 277–286. `http://aclweb.org/anthology/Y/Y13/Y13-1027.pdf`.

Lee, Jackson L, Litong Chen & Tsz-Him Tsui. 2016. PyCantonese: Developing computational tools for Cantonese linguistics .

Lee, Peppina Po-Lun. 2020. On the semantics of classifier reduplication in Cantonese. *Journa of Linguistics* (online first).

Štekauer, Pavol, Salvador Valera & Lívia Kőrtvélyessy. 2012. *Word-formation in the world's languages: a typological survey.* Cambridge University Press.

Wang, Zhimin, Li He & Yanqiu Shao. 2013. The idiom investigation of Chinese undergraduate textbook and the extraction of common used idioms. In *2013 IEEE/WIC/ACM international joint conferences on web intelligence (WI) and intelligent agent technologies (IAT)*, vol. 3, 208–212. IEEE.

Wilbur, Ronnie. 1973. *The phonology of reduplication.* Indiana University Linguistics Club Bloomington.