Lexical Reduplications in Cantonese

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

The Sixth Study Days on Chinese Linguistics
University of Bologna, Forlì
23-25 June 2021

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

Overview

- Cantonese Reduplications can be productive or non-productive
- Lexical reduplications (LR): frozen, unproductive, non-compositional expressions (ongoing work on Multiword Expressions)
- Data extracted from two dictionaries as a lexical resource
- Significance for L2 learners and Natural Language Processing (NLP)

Part of the quantitative study has been presented in MWE Workshop at PACLIC-34 in Nov 2020

Background: Many Reduplications 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 ZAAT
'bouncy and active'
b. zaat3 zaat3 tiu3 紮紮跳
ZAAT ZAAT jump
c. tiu3 tiu3 zaat3 zaat3 跳跳紮紮
jump jump ZAAT ZAAT
d. *zaat3 zaat3 tiu3 tiu3 * 紮紮跳跳
ZAAT ZAAT jump jump

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)
  1. Total vs. partial reduplication
  2. Identity / mapping between base and reduplicant
     These lines of research do not quite address the lexical reduplication in Cantonese
- Syntax-semantics of reduplications in Cantonese
  1. Nominal and quantification: (Cheng, 2012; Lee, 2020)
  2. Verbal, pluractionality and event modification: (Lam, 2013; Basciano & Melloni, 2017; Lam, 2020)

Related Works (continued)

Idiom list as a learning resource

- Mandarin:
  -
  - The need for graphical materials in language learning (Chung & Hsieh, 2017)
  - Common idioms can be extracted from textbooks as a source (Wang et al., 2013)
- Cantonese:
  Quadra-syllabic Idiomatic Expressions (QIEs) 粵語四字格成語語料及音檔庫 `http://www.livac.org/yueqie/` (T'sou, 2017)

This study: Lexical Reduplications in Cantonese

- Many LRs are not found in Mandarin
- Lexical resources for learning and NLP (esp. for low-resource languages)

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

(5) AAB-template:

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

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

- Some reduplicated elements are meaningless, e.g. *laap3* in (5)
- The meaning might come from the reduplicated or the unreduplicated element

Examples of ABA

(6) 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

(7) 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. 西装骨骨、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

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

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

- Opaque meaning → Challenge for learners!
- Elements are often words (*loi4* 'come', *heoi3* 'go', *bei2ci2* 'each other')

Examples of AAXY & XYAA

(10) 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'

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

- In (10a), *dei6 hei3* does not seem to contribute to the meaning compositionally
- In (11), *sai1zong1* 'suit' as a noun is clearly related to the meaning of the whole expression

Examples of AXAY, XAYA & AXYA

(12) mou5 jan4 mou5 mat6 冇人冇物
no person no thing
'having nothing at all' AXAY-template

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

(14) 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

Discussion 1: Body parts

LRs involving body parts:

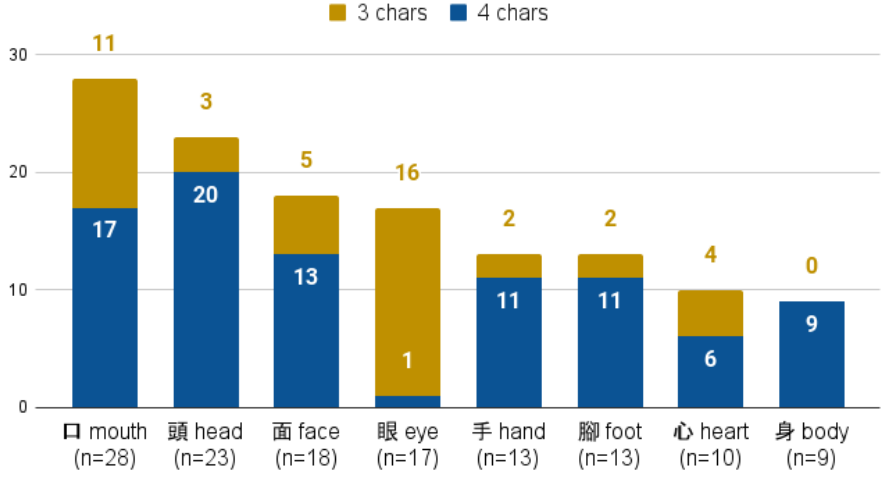

Figure 1: LRs with Body Parts (3- & 4-characters)

Discussion 1: Body parts

LRs involving body parts (words.hk):

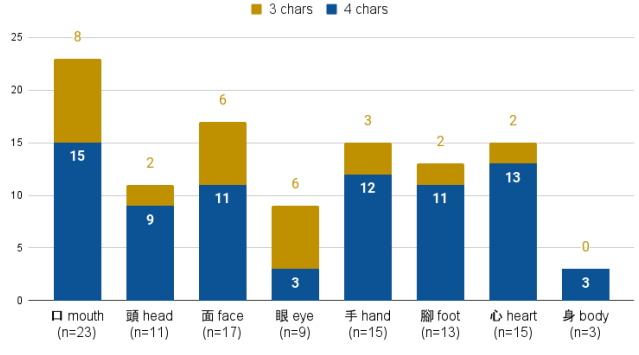

Figure 2: LRs with Body Parts (3- & 4-characters) in Words.hk dataset

Similar patterns with CNP dictionary data!

Discussion 1: Body parts

Body Part	Example	Meaning	JyutPing Romanization
口 mouth	口花花	being a sweet talker	hau2 faa1 faa1
	兜口兜面	right in the face	dau1 hau2 dau1 min6
頭 head	頭耷耷	looking downward and upset	tau4 dap1 dap1
	有頭有面	being famous and respected	jau5 tau4 jau5 min6
面 face	面青青	looking scared and pale	min6 ceng1 ceng1
	熟口熟面	looking familiar	suk6 hau2 suk6 min6
眼 eye	眼甘甘	staring	ngaan5 gam1 gam1
	篤眼篤鼻	looking annoying	duk1 ngaan5 duk1 bei6
手 hand	手多多	touching unnecessarily; handsy	sau2 do1 do1
	有手有腳	being capable and self-sufficient	jau5 sau2 jau5 goek3
腳 foot	急急腳	in a hurry	gap1 gap1 goek3
	落手落腳	getting hands-on	lok6 sau2 lok6 goek3
心 heart	心嘟嘟	tempted	sam1 juk1 juk1
	抌心抌肺	kicking oneself	dam2 sam1 dam2 fai3
身 body	嘟身嘟勢	moving around	juk1 san1 juk1 sai3
	身水身汗	sweaty	san1 seoi2 san1 hon6

- Co-occurrence of mouth and face; hand and foot
- Many of the items would be opaque to Mandarin speakers too

Discussion 2: Correlating Forms and Parts-of-Speech

Many ABB-forms appear to contain an adjective and two reduplicated syllables with no apparent meaning, especially for color terms:

(15) hung4 bok1 bok1 紅卜卜
red BOK BOK
'bright red'

(16) wong4 gam4 gam4 黃黔黔
yellow GAM GAM
'bright yellow'

(17) fei4 tan4 tan4 肥腩腩
fat TAN TAN
'chubby; fatty'

(18) je6 maa1 maa1 夜麻麻
late MAA MAA
'very late'

Discussion 2: Correlating Forms and Parts-of-Speech

Can we effectively predict the POS of the whole expression based on the parts?

- Descriptive adequacy for explanatory adequacy
- Recognition of novel terms and estimating their meaning

What are the useful indicators?

- reduplicated element?
  if the reduplicated element has no meaning, then the third element must carry the meaning?
- POS of the first element?
  true for many items, but many counterexamples too

More annotation required!

Lexical resource for learners & language preservation

- T'sou (2017) comments that younger speakers (below 50 years of age) no longer understand idiomatic expressions
- Increasing interests in Cantonese outside of homeland Hong Kong
- Rarely included formal education, or considered important
- Recent cancellation of oral examination in HKDSE Chinese (conducted in Cantonese) and promotion of Mandarin in Hong Kong

Lexical resource

- Improve word segmentation and POS-tagging
- Future development: closer look at these forms may help predict their meanings
  1. Novel term detection;
  2. Flagging idioms
  3. Collocation of these flagged items

  $\rightarrow$ Practical application for texting and word suggestion!
- Resources for content creators (story books, flash cards, crossword puzzles)

Better recognition and comprehension for these idiomatic expressions, by NLP systems or human learners

Conclusion

- Need for resources on the lexical reduplications (LR), which are idiomatic and unproductive
- Significant presence in the vocabulary, based on two data recent sets
- Emerging patterns; great variety in form and meaning in LRs (base-reduplication mapping; predicting novel terms)
- Potential use in NLP (research and daily application) and learner support
- Next step: Clean up and release of the data set

多謝各位！
do1 ze6 gok3 wai2

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.

Lam, Charles. 2020. The V-one-V construction in Cantonese. *Studies in Chinese Linguistics* 41(2). 161–184.

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.

T'sou, Benjamin Ka Yin. 2017. Cantonese 4-word idiomatic expressions excerpts and audio recording database.

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.