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

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

Abstract

Lexical reduplications (LR) in Cantonese are multiword expressions (MWEs) that are frozen and unproductive. The meanings of LR are often non-compositional, such as *hung4 bok1 bok1* 'bright red', where *bok1 bok1* does not contribute any meaning to *hung4* 'red'. This can pose a challenge in parsing and natural language understanding. LRs can be misinterpreted to bear literal meaning, if they are mistakenly treated as decomposable chunks. Identifying LRs is therefore an important step to reduce errors in word segmentation and natural language understanding. This study discusses a collection of LRs extracted from two data sets. Some common patterns are also identified in this study, which may guide parsers to automatically identify items that are novel to the system.

1 Introduction

Reduplication in general is a ubiquitous phenomenon in Cantonese. There are many types of reduplication and they are mostly productive. This means that reduplications are templatic, and their compatibility with the lexical items is rule-governed. For example, non-stative verbs (i.e., verbs involving actions) may be reduplicated to show durative events, as in (1a). Reduplication of the classifier (or measure words) denotes the meaning of 'every', as shown in (1b). To express the short duration of an event, one may use the 'V+one+V' reduplication in (1c). Since these reduplication types are rule-governed, their meaning and structure are predictable and cause no problem in NLP.

(1) a. haang6 haang6 haa5
walk walk PRT
'while walking'

b. zek3 zek3 gau2
CL CL dog
'every dog'

c. mong6 jat1 mong6
look one look
'to take a look'

d. coeng4 coeng2 dei2
long long DEI
'long-ish; fairly long'

Unlike the productive reduplications described above, lexical reduplications (LR) in Cantonese are multiword expressions (MWEs) that cannot be formed productively. LRs are considered frozen and are not coined by individual users. Many LRs are idiomatic in that their meanings are non-transparent. In (2a), the single use of *dai2* can only mean 'under / beneath', but not undergarment. Example (2b) does not have a corresponding base form (i.e. non-reduplicated) *tiu3 zaat3* either. When the item is used, *tiu3* must be reduplicated.

(2) a. dai2 dai2
under under
'undies' (casual term for undergarment)

b. tiu3 tiu3 zaat3
jump jump tie
'bouncy and active'
(Attested example from words.hk)

What makes LR tricky is that they may carry literal meaning in some cases. This can cause parsers to an-

alyze them as compositional and miss the intended idiomatic use. Example (3) contains the parallel of *haang4* 'walk' and *kei2* 'stand':

(3) haang4 haang4 kei5 kei5
walk walk stand stand
'being idle and aimless'
(From Cheung, Ngai and Poon (2018))

Taken literally, the phrase could mean walking and pausing intermittently. However, it actually means being idle and not doing anything and is often used to describe laziness. Its meaning is therefore unpredictable and must be learned by heart for speakers.

Some LRs contain parts that do not contribute meanings at all, most notably in some color terms. The reduplicated syllables *bok1* in (4b) and *gam4* in (5a) do not appear to bear any meaning[1], and there is no non-reduplicated 'base form', i.e., **hung4 bok1* alone is an illicit form.

(4) a. hung4 'red'
b. hung4 bok1 bok1
red BOK BOK
'bright red'

(5) a. wong4 gam4 gam4
yellow GAM GAM
'bright yellow'
b. *hung4 gam4 gam4
red GAM GAM
Intended: 'bright red'
c. *wong4 bok1 bok1
yellow BOK BOK
Intended: 'bright yellow'

In addition, examples (5b) and (5c) show that the combinations are fixed and cannot be changed. The same also applies to several other colors or adjectives. As the examples above have demonstrated, LRs are unproductive and frozen forms. The data presented in this study will potentially be useful for error / grammar detection (Jiang et al., 2012), or facilitate other studies on idiomatic expressions (Wang et al., 2019).

[1]Throughout this paper, these meaningless elements, reduplicated or not, are glossed with the romanization in all capital letters.

2 Related Works

Studies on lexical reduplication in Cantonese are limited. In the linguistics literature, most studies lie in the areas of phonology and phonetics, which deals with the relation between the underlying form of the reduplication and its realied pronunciation.

Since the seminal study by Wilbur (1973), there has been great progress in the investigation of sound systems (Botha, 2006; Frampton, 2009; Inkelas and Zoll, 2005). However, sound patterns might not be directly useful for LRs in the present study. For the morphosyntax and semantics of reduplication, previous studies focused on productive and predictable forms (Hurch, 2005; Francis et al., 2011; Štekauer et al., 2012), which cannot cover the LRs in the present study either.

Specific to Sinitic languages, Cheng (2012) and Lee (2020) provide thorough explanations on the mechanism of classifier reduplication in example (1b). Lam (2013) and Basciano and Melloni (2017) both discussed verbal reduplication in Cantonese and Mandarin. It seems that unproductive and frozen forms like LR have not received much attention in linguistics. This is probably because they do not display particular patterns, and therefore are not seen as theoretically important.

For idioms in Sinitic languages, the focus is often placed in learning enhancement, either with digital materials (Chung and Hsieh, 2017), or extracting a list of idioms from corpus data (Wang et al., 2013). The present study falls under the latter type. However, given the paucity of Cantonese teaching materials for children (which typically contain more idioms than materials for adults), this study uses dictionaries, both online and print, as the sources of data. As our data show, Cantonese LRs typically would not be considered a part of Chinese idioms or four-character expressions. It is therefore necessary to investigate LRs as a separate category from idioms.

Corpus resources for Cantonese are scarce, both for LRs specifically and for idiomatic expressions at large. While there are several Cantonese corpora that are widely used (Luke and Wong, 2015; Lee and Wong, 1998; Leung and Law, 2001), no previous works have been conducted on LR and their distribution. The present study is an attempt to in-

vestigate LRs through a comparison across corpora. As Cantonese is known to have a rich inventory of idiomatic expressions, many of the LRs identified in this study are not found in Mandarin. It is therefore not practical to assume Mandarin resources can be borrowed to handle Cantonese data. Therefore, this study aims to provide a constructed data set specific to LRs with the description of the data. The next section provides the statistics of the data sets, and the section after discusses their significance and potential uses.

3 Data sets

3.1 Cheung, Ngai and Poon (2018)

This study extracted two sources for the data set of LR. The first source of data was the Cantonese dictionary (Cheung et al., 2018), henceforth CNP. The CNP data set was extracted and digitized manually from the print dictionary. Out of the total of 12,000 entries in the dictionary, 756 entries contain reduplicated elements, which makes approximately 6.30% of the data set. Table 1 summarizes the data set.

Length of LR	Types	% of LR
2 characters	23	3.04%
3 characters	268	35.45%
4 characters	302	39.95%
5 characters	27	3.57%
6 characters	23	3.04%
7 characters	33	4.37%
8 characters	9	1.19%
$\geq$ 9 characters	71	9.39%
Total	756	100.00%

Table 1: Summary of LR in the CNP data set

Since the source is a dictionary, each entry represents a unique type and there is no number of tokens available.

3.2 Words.hk

The second source is the site `https://words.hk`, which is a crowd-sourced effort to create an online dictionary established in 2014. This study uses its data set of Cantonese articles, which includes lexical items that are attested in written articles but excludes the items that are only found in the constructed dictionary section of the site. This approach provides the number of tokens used in the texts, which can better reflect the use of LR, rather than the constructed list of types in the dictionary format. In the `words.hk` data (henceforth WHK), there are 30,821 unique types and 2,938,248 tokens from this data set. The longest lexical items have 4 characters. More details are listed in table 2:

Length	Types	Tokens
1 char	3,878 (12.59%)	2,253,458 (76.69%)
2 char	20,592 (66.88%)	636,566 (21.66%)
3 char	3,116 (10.12%)	32,408 (1.10%)
4 char	3,205 (10.41%)	15,419 (0.52%)

Table 2: Summary of all types in the Words.hk data set

Among these unique types, 881 are found to contain reduplicated elements. Since reduplicated forms entails more than one character, table 3 excludes one-character types. The percentages in table 3 are based on the types / tokens of LR.

Length	Types	Tokens
2 characters	138 (15.66%)	9,870 (67.92%)
3 characters	242 (27.47%)	2,302 (15.84%)
4 characters	501 (56.87%)	2,359 (16.23%)
Total	881 (100%)	14,351(100%)

Table 3: Summary of LR in the Words.hk data set

4 Distribution and Patterns of LRs

This section describes the patterns of the attested LRs in the two data sources. The reduplicated elements can form several patterns that are logically possible, but they are not equally distributed. While these sources are not meant to be exhaustive, the proportion of the different LR categories are similar, indicating that they are representative of LRs in the Cantonese language.

LRs with two characters cannot display any variation in pattern, due to their lengths. More fine-grained analysis on their parts of speech, meanings or syntactic distribution will require further investigation. The patterns of 3- and 4-character LRs will be discussed below.

4.1 3-character LR

LRs with three characters are attested in both sources and can be found in three templates.

(6) AAB-template:

a. laap6 laap6 ling3
LAAP LAAP shiny
'shiny'

b. cyun3 cyun3 gung3
sacarstic sacarstic GUNG
'sacarstic'

Similar to the examples of *hung4 bok1 bok1* 'bright red' in (4b) and *wong4 gam4 gam4* 'bright yellow' in (5a), some reduplicated elements are meaningless, as in (6). However, it is not always the case that the unreduplicated element denotes the meaning of the whole LR. Example (6b) shows that in some cases, it is the reduplicated element that indicates the meaning of the whole phrase, and the unreduplicated element is meaningless.

In the ABA-template, while examples like (7) and (7b) contain meaningful elements, the meaning of the entire phrase is not always compositional:

(7) ABA-template:

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

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

The ABB template includes many adjectives. As shown in the first section, color terms like examples (4b), (5a) and (8) often appear in the ABB template. In example (8b), *jin6* is likely a truncation of *jin6gam1* 'cash'. While the term is somewhat transparent, the whole phrase is used as an adverb that modifies paying events, not as a noun (as 'cash' is). Therefore, the function of the full phrase is not completely predictable by its elements.

(8) ABB-template:

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

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

Table 4 below shows the identified LRs and their distribution in the two data sets:

	Types	
Category	**Words.hk**	**CNP Dictionary**
AAB	88 (36.36%)	89 (31.79%)
ABA	31 (12.81%)	11 (3.93%)
ABB	122 (50.41%)	180 (64.29%)
AAA	1 (0.41%)	0 (0%)
Total	242 (100.00%)	280 (100.00%)

Table 4: Subcategories of 3-character LRs by types

More than half of the unique types belong to the ABB pattern across the two sources; and the second largest group is the AAB pattern, followed by the alternating ABA pattern. The distribution is the same across the two data sources, suggesting that it reflects the general pattern in the language as well.

On a side note, the AAA pattern is extremely rare in Cantonese. There are more than one attested entries from the WHK data, but a few of them were removed from LR, as they do not reflect idiomatic use of the language. For example, the entries of '999' (in arabic number) and *gau2 gau2 gau2* 'nine nine nine' (in Chinese character) are the emergency number in Hong Kong, so they do not reflect on the word formation in Cantonese. The only AAA item that ends up in the data set is *paak1 paak1 paak1*, which is onomatopoeic for sexual activities. The two data sets contain duplicated entries, such as:

(9) a. zing6 zing6 gai1
quiet chicken chicken
'very quiet'

b. suk6 hau2 suk6 min6
familiar mouth familiar face
'very familiar'

These duplicates were not removed, so that the two data sources are accurately represented. Researchers who want to make use of these data should remove the duplicated items. With the numbers indicating the unique types, table 4 provides a direct comparison between the two sources[2].

[2]Since the CNP dictionary data do not include naturalistic use in prose, there is no statistics on tokens. The numbers

4.2 Patterns of 4-character LR

LRs with four characters can be found in the following six templates.

(10) loi4 loi4 heoi3 heoi3
come come go go
'always' AABB-template

(11) sei2 sei2 dei6 hei3
die die ground air
'reluctantly' AAXY-template

(12) sai1zong1 gwat1 gwat1
suits bone bone
'being dressed up' XYBB-template

(13) bei2ci2 bei2ci2
each.other each.other
'same to you' ABAB-template

(14) mou5 jan4 mou5 mat6
no person no thing
'having nothing at all' AXAY-template

(15) sau2 ting4 hau2 ting4
hand stop mouth stop
'living from hand to mouth' XBYB-template

Similar to LRs with three characters, these items are idiomatic and cannot be modified. Some examples may seem to be transparent. However, the word order cannot be changed, even for the parallel ones. For example, (10) cannot be rephrased as **heoi3 heoi3 loi4 loi4*, and (14) cannot be **mou5 mat6 mou5 jan4*. This shows that these items should be considered frozen multiword expressions.

Table 5 shows the distribution of the types of LRs with four characters in the two data sets[3]. In both data sets, the AXAY template appears most frequently, followed by the AABB form. The two data sets diverge in the less frequent categories. For the two data sets, it is not clear why such difference exists, or whether such difference is representative.

	Types	
Category	**Words.hk**	**CNP Dictionary**
AABB	83 (16.57%)	61(20.07%)
AAXY	68 (13.57%)	7 (2.30%)
XYBB	39 (7.78%)	22 (7.24%)
ABAB	8 (1.60%)	5 (1.64%)
AXAY	267 (53.29%)	181 (61.18%)
XBYB	36 (7.19%)	23 (7.57%)
Total	501 (100.00%)	304 (100%)

Table 5: Subcategories of 4-character LRs by types

It is worth pointing out that the ABAB form results in rather low frequency, when compared to what one might expect from Mandarin. In Mandarin, the ABAB form is a productive morphological process to show tentative events. For instance, *chángshì* 'to try' can be reduplicated as *chángshì chángshì* 'to give it a try'. The process cannot be applied to Cantonese: While the cognate *soeng4 si3* 'to try' exists as a verb, the form **soeng4 si3 soeng4 si3* is not acceptable at all and therefore cannot be used to mean 'to give it a try'. This difference between Mandarin and Cantonese may explain the low frequency of ABAB reduplications. In all the attested Cantonese examples, the ABAB reduplications are not formed productively and therefore genuine multiword expressions.

5 Potential use of the data

It is possible that these forms of LRs can be used as part of novel term detection, or as a flag for idioms. From the examples above, it is clear that LRs are idiomatic and do not denote compositional meanings. Productive reduplications in Cantonese often come with specific morphemes, such as *dei2* for adjectival reduplication denoting diminution in example (1d), or *haa5* in V-one-V reduplication in example (1c). Reduplications that do not come with any particular marker can be a sign for LRs. At this point, only color terms in the ABB template form an observable pattern. Further studies on the form-meaning correlation in LRs will facilitate better recognition and comprehension for these idiomatic expressions. It is possible to identify and predict the meanings of based on the individual components. For example, LRs in ABB format with color terms and adjectives

of tokens in the Words.hk data set are as follows: AAB: 618 (26.85%); ABA: 778 (33.80%); ABB: 884 (38.40%); AAA: 22 (0.96%). In total, there are 2,302 tokens of LR with 3 characters.

[3]The tokens of 4-character LRs in the Words.hk data set are as follows: AABB: 472 (20.01%); AAXY: 274 (11.62%); XYBB: 149 (6.32%); ABAB: 16 (0.68%); AXAY: 1,265 (53.62%); XBYB: 183 (7.76%). In total, there are 2,359 tokens of LR with 4 characters.

as the first character are highly likely to be intensification of the denoted color or attribute.

6 Conclusion

This paper has highlighted the need for resources on the lexicial reduplication phenomenon (LR) and shown that the LRs are idiomatic and unproductive, which can be an issue for parsing or natural language understanding. Especially because many of them are unique to Cantonese, but not Mandarin, existing resources from Mandarin cannot be borrowed to recognize LRs, despite the abundance of Mandarin data.

The data show that LRs constitute a significant amount in the vocabulary. The reduplicated element may occur in various places within a Cantonese LR. Such unpredictability of LRs makes the phenomenon a challenge for natural language understanding. While some reduplication templates appear more frequently, it remains unclear how exactly the templates or forms correlate with the meanings.

Given that the two data sets are not exhaustive in listing the LRs in the Cantonese language, there will be novel LRs in NLP tasks in the real world. Therefore, the distribution from the presented data can potentially be used for identification of novel, undiscovered LRs.

References

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

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

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

Lai Yin Cheung, Lit Wai Ngai, and Lai Mei Poon. 2018. *The Dictionary of Hong Kong Cantonese*. Hong Kong: Cosmo Books.

Liang-Yi Chung and Sheng-Min Hsieh. 2017. Using graphic digital materials in language learning. In *2017 International Conference on Applied System Innovation (ICASI)*, pages 295–298. IEEE.

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

Elaine J Francis, Stephen Matthews, Reace Wing Yan Wong, and 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.

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

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

Ying Jiang, Tong Wang, Tao Lin, Fangjie Wang, Wenting Cheng, Xiaofei Liu, Chenghui Wang, and Weijian Zhang. 2012. A rule based Chinese spelling and grammar detection system utility. In *2012 International Conference on System Science and Engineering (ICSSE)*, pages 437–440. IEEE.

Charles Lam. 2013. Reduplication across categories in Cantonese. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation*, pages 277–286.

Thomas Lee and Colleen Wong. 1998. CANCORP: The Hong Kong Cantonese child language corpus. *Cahiers de Linguistique Asie Orientale*, 27(2):211–228.

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

Man-Tak Leung and Sam-Po Law. 2001. HKCAC: the Hong Kong Cantonese adult language corpus. *International journal of corpus linguistics*, 6(2):305–325.

Kang Kwong Luke and May Lai-Yin Wong. 2015. The Hong Kong Cantonese corpus: Design and uses. *Journal of Chinese Linguistics*, 25:309–330.

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

Zhimin Wang, Li He, and 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)*, volume 3, pages 208–212. IEEE.

Chengyu Wang, Yan Fan, Xiaofeng He, Hongyuan Zha, and Aoying Zhou. 2019. Idiomaticity prediction of Chinese noun compounds and its applications. *IEEE Access*, 7:142866–142878.

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