Exploring the Chinese Mental Lexicon with Word Association Norms

Oi Yee Kwong
Department of Chinese, Translation and Linguistics
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
Olivia.Kwong@cityu.edu.hk

Abstract

Our internal repository of words, often known as the mental lexicon, has primarily been modelled by psychologists as some kind of network. One way to probe its organisation and access mechanisms is by means of word association techniques, which have rarely been applied on Chinese. This paper reports on the design and implementation of a pilot word association test on native Hong Kong Cantonese speakers. The test contains 500 stimulus words, carefully selected and controlled on important factors including word frequency, part-of-speech, syllabicity, concreteness and vocabulary type. The resulting association norms based on 58 participants reveal interesting properties of the Chinese mental lexicon, such as the dominance of disyllabic and nominal concepts, and collocational associations. Despite its current small scale, the word association norms obtained from this study do not only offer first-hand psycholinguistic evidence for investigating the Chinese mental lexicon but also provide a useful resource to inform future studies in Chinese lexical access, lexical semantics and lexicography.

1 Introduction

Humans know tens of thousands of words, and their language behaviour suggests that the words are systematically organised and efficiently accessed in their internal word repository, often known as the mental lexicon. For a long time, psychologists have hypothesised the mental lexicon as a massive network of inter-connected nodes. The organisation and access mechanisms of the mental lexicon have primarily been studied with experimental approaches, using a variety of tasks like lexical decision, semantic verification, word association, etc. Despite some intrinsic weaknesses, word association techniques offer first-hand psycholinguistic evidence of our mental lexicon especially for revealing the extensiveness of the network and the different kinds of semantic associations among concepts in it. However, large-scale association norms obtained from native Chinese speakers comparable to those in English and other languages are lacking. This study, as a pilot attempt, intends to produce a set of word association norms from native Hong Kong Cantonese speakers, to complement other experimental methods and offer more insight for further studies on the Chinese mental lexicon.

Hong Kong Chinese is our focus in this study. Written vocabularies from Modern Standard Chinese and spoken vocabularies from the Cantonese dialect are supposed to co-exist in the mental lexicon of its speakers. It is therefore theoretically and practically interesting to explore how such mixed forms, together with other Chinese-specific factors like logographic meanings, syllabicity and the fuzziness of the word notion, as well as other general factors like word frequency, part-of-speech, concreteness and polysemy, shape the organisation of the Chinese mental lexicon.

Stimulus words were thus chosen carefully and systematically, with due consideration for the various important factors. The resulting word association norms are expected to shed light on a wide range of research questions on the Chinese mental lexicon, some of which are particularly addressed in the current study:

- What is the basic unit in the semantic memory of Chinese speakers?
- What kinds of semantic associations are found, and which kind (taxonomic, thematic, or otherwise) dominates?

27th Pacific Asia Conference on Language, Information, and Computation pages 153－162

- How do the associations differ for words of different frequency, part-of-speech, and concreteness?
- Are written and spoken vocabulary items stored together or separately?

In Section 2, we briefly review related work. In Section 3, the design and implementation of our word association test, and the compilation of word association norms, will be described. Preliminary analysis of the data will be discussed in Section 4, followed by a conclusion with further research agenda in Section 5.

2 Studies on the Mental Lexicon

As Aitchison (2003) pointed out, the general picture of the mental lexicon so far is one in which there are a variety of links between words, some strong, some weak. Strictly speaking, our knowledge of words includes phonological, morphological, syntactic, and even other lower level features like radicals, shapes and strokes. Hence network models in different forms and complexity have been proposed under the connectionist roof. For example, McClelland and Rumelhart's (1981) interactive activation model assumes three levels of processing (feature, letter, and word) which occur simultaneously with excitatory or inhibitory interactions. Others (e.g. Bock and Levelt, 1994; Caramazza, 1997) suggested that a lexical network should also connect a lemma level and a lexeme level for syntactic and phonological properties respectively, in addition to the conceptual level for semantic relations. Models for Chinese need to account for the role of radicals and their positions in a character, as well as the combination of characters to form words (e.g. Taft, 2006). The focus of this study is primarily on the semantic connections among words, or semantic memory.

The psychological reality of the network models receives support from experimental psychology, which often employs methods like lexical decision, semantic verification, etc. The spreading activation model of lexical access suggested in Collins and Loftus' (1975) classic study has been most influential especially with its account for the associative priming effects. Frequency effect and concreteness effect have been observed regarding lexical organisation (e.g. Kroll and Merves, 1986; Bleasdale, 1987). The role of polysemy with respect to lexical representation and word recognition, in isolation or in context, has also been widely studied (e.g. Rayner and Frazier, 1989; Klein and Murphy, 2002; Rodd et al., 2002). Similar studies on Chinese, though not as many, have identified some Chinese-specific properties of the mental lexicon, including the fuzziness of the word notion (e.g. Hoosain, 1992), word frequency (but not character frequency) effects and the competition between homophonic morphemes (e.g. Zhou and Marslen-Wilson, 1994), relations between morphemes and words (e.g. Myers, 2006), polysemy and meaning representation in the mental lexicon (e.g. Lin and Ahrens, 2010). Most of these studies on Chinese, however, were based on Mandarin Chinese, and the effect of concreteness has not been widely studied.

Word association techniques have also been a traditional approach to probe the mental lexicon. They can be used within experimental approaches, but more often, the resulting word association norms offer another useful resource for us to explore the mental lexicon from a broader perspective, which may in turn inform and complement experimental studies. Word association tests ask human subjects for the first word they can think of upon seeing or hearing a stimulus word and the percentage of subjects producing each response is computed from a large group of subjects to give the word association norms, which are useful in many respects (e.g. de Groot, 1989; Hirsh and Tree, 2001; Guida and Lenci, 2007). Examples of famous English word association norms include the 1952 Minnesota word association norms (Jenkins, 1970) and the Birkbeck word association norms (Moss and Older, 1996). Large-scale word association data are also available for other languages like Japanese (Joyce, 2005) and German (Schulte im Walde et al., 2008). However, comparable large-scale association norms obtained from native Chinese speakers are lacking. The pilot study reported in this paper thus intends to fill this gap with a focus on the mental lexicon of native Hong Kong Chinese, and to capitalise on the range of association data thus produced for qualitative and quantitative analyses with respect to various important factors.

3 Word Association Test

In this study, human subjects were asked to give the first word which occurs to them upon seeing a certain stimulus word, and thus it is a *discrete* word association test. The responses need not be

in any particular part-of-speech, and thus they lead to a set of *free* association norms. In the following, we describe the design and implementation of the test. Chinese examples are listed with Cantonese transcription in Jyutping (in italics) and an English gloss (in quotes) alongside.

3.1 Test Platform

Online version of the association test was developed. Registered participants were given instructions in English and Chinese (Figure 1), and asked to input their response on the web interface (Figure 2). The English and Chinese instructions are more or less equivalent. It was additionally specified in the Chinese instructions that there was no restriction on the part-of-speech and syllabicity for the responses. There is also slight difference in the example given in the instructions. In English, possible responses to "butter" are exemplified with "bread", "milk", "fat" and "spread"; while in Chinese, for the equivalent stimulus word 牛油 *ngau4-jau4* 'butter', the example responses given are 麵包 *min6-baau1* 'bread', 牛油果 *ngau4-jau4-gwo2* 'avocado', 牛奶 *ngau4-naai5* 'milk', 肥膩 *fei4-nei6* 'fatty', 黃色 *wong4-sik1* 'yellow', and 搽 *caa4* '(to) spread'. The inclusion of "avocado" and "milk" for the Chinese examples was to demonstrate the Chinese-specific cases as they share morphemes with the stimulus word.

3.2 Participants

All 58 participants (20 males and 38 females) were undergraduate students of the City University of Hong Kong. They were recruited from the Department of Chinese, Translation and Linguistics, and the Department of Computer Science. All of them are native Hong Kong Cantonese speakers. Each participant was rewarded with a shopping voucher upon completion of all test sessions.

3.3 Selection of Stimulus

To maximise the usefulness of the pilot collection of data and the resulting association norms, a balanced and representative sample of stimulus words was selected. The selection was done with systematic control on various factors deemed important in human lexical processing, which include word frequency (High, Mid, Low)[1], part-of-speech (Noun, Verb, Adjective, Fluid), syllabicity (Monosyllabic, Disyllabic, Trisyllabic), concreteness (Concrete, Abstract), and vocabulary type (Written, Spoken)[2]. As far as polysemy is concerned, most of the disyllabic and trisyllabic words are unambiguous. The categorially fluid ones are by default ambiguous, and the monosyllabic words are mostly ambiguous in Chinese. They were not particularly controlled for the number of meanings they possess. Altogether 500 stimulus words were selected. Their distribution with respect to the various factors and some examples for each category are shown in Table 1. All stimulus words were randomly divided into five test sessions with 100 words each.

3.4 Association Norm Preparation

The initially collected responses were checked for obvious typos (e.g. changing 普通詰 *pou2-tung1-kit3* to 普通話 *pou2-tung1-waa2* for 'Putonghua') and correcting homophones (e.g. changing 快子 *faai3-zi2* to 筷子 *faai3-zi2* for 'chopsticks'). For specific responses, we had to double check with participants and asked them to clarify, to help our subsequent classification of the responses. Similar responses (e.g. 蛋 *daan2* 'egg' and 雞蛋 *gai1-daan2* 'egg') were marked. In a word association test, the general patterns of responses from large groups of subjects are compiled into a set of word association norms. The percentage occupied by a certain response is assumed to indicate the associative strength between the stimulus and that response. The percentage of individual response types for each stimulus word was thus computed and the association norms were listed accordingly. Two examples are shown in Figure 3.

[1] The frequency distinction for the current study was based on a 2.6-million-character Chinese corpus collected over the web, which is composed of texts from three newspapers and five magazines covering a variety of topics. The corpus was word segmented. Two word lists were compiled from the news subcorpus and the magazine subcorpus respectively. The word-frequency list from the news subcorpus was divided into three frequency bands according to the cumulative percentage: Hi (below 80%), Mid (80-90%) and Low (above 90%). Only words appearing in both subcorpora were included as candidates.

[2] By written vocabularies, we mean those lexical items which can be acceptably used in standard written Chinese texts. In fact most of these items are also used in Cantonese speech. The spoken vocabularies, on the other hand, are often considered inappropriate to be used in formal written contexts.

Please read each word displayed on screen and then input the first word it brings to mind. Type your response in the textbox provided. For example, if the given word is "butter", the first word you think of might be "bread" or "milk" or "fat" or "spread" or something else. Please note that we are interested in the word that comes to mind immediately, and there is no right or wrong "answer", so you only need to give your immediate response, not after thinking about it for a while. Do not go back and change your mind after you have given your first response. Your responses should primarily be in the same language as the given words, but if it really happens that you immediately think of a word in another language, you may just input that word. Each test session contains 100 words. Try to finish one whole session without interruption at a time. Now please click "Start" to begin.

請觀看螢光幕顯示的詞語，然後盡快在空格內輸入第一個您即時聯想到的詞語，音節及詞性不限。 例如：當螢光幕顯示「牛油」一詞，大家可能會聯想到「麵包」、「牛油果」、「牛奶」、「肥膩」 、「黃色」、「搽」等等，您可以使用任何一種自己慣用的輸入法，盡快輸入您想到的第一個詞語。 我們旨在觀察一般人聯想的特點和規律，沒有對錯之分，所以您只需靠即時反應，不用特別思考，也不要更改已輸入的詞語。您應盡量以測試詞語的語言回應，但若您確實只能聯想到另一語言的詞語，也可輸入該詞。每一節測試包含 100 個詞語。每次請一氣呵成完成整節測試。現在請點擊 Start 開始測試。

Figure 1. Instructions Given to Participants

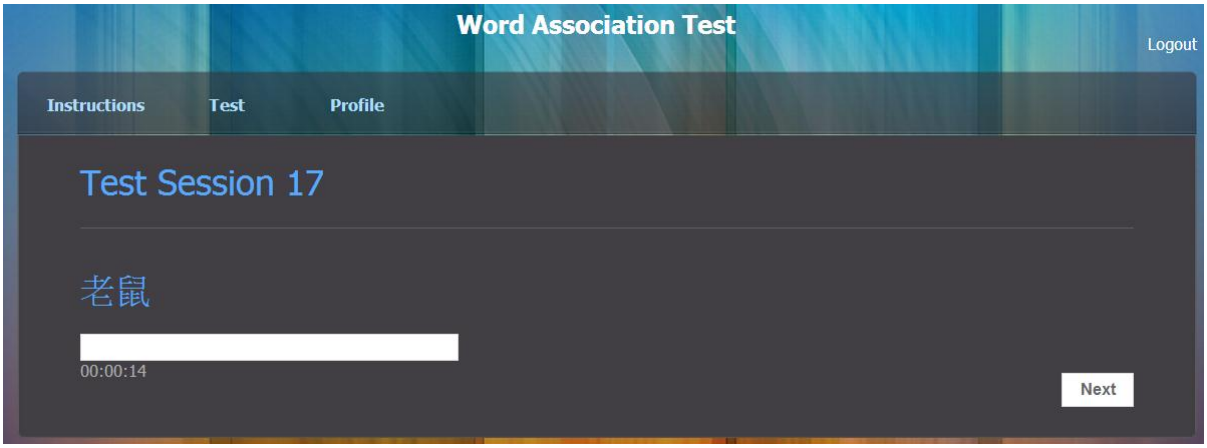

Figure 2. Online Platform for Submitting Association Responses

戒指 'ring'

結婚	'marriage'	41.4%
承諾	'promise'	8.6%
訂婚	'engagement'	8.6%
鑽石	'diamond'	8.6%
求婚	'propose'	6.9%
婚姻	'marriage'	5.2%
無名指	'ring finger'	3.4%
戴	'wear'	3.4%
決定	'decision'	1.7%
所羅門	'Solomon'	1.7%
金色	'gold (colour)'	1.7%
很貴	'very expensive'	1.7%
閃	'shiny'	1.7%
情人	'lover'	1.7%
責任	'responsibility'	1.7%
電影	'movie'	1.7%

閱讀 '(to) read'

書本	'book'	29.3%
書	'book'	20.7%
書籍	'book'	8.6%
小說	'fiction'	5.2%
理解	'comprehension'	5.2%
圖書	'(picture) book'	5.2%
計劃	'plan'	3.4%
習慣	'habit'	3.4%
文章	'article'	1.7%
全文	'full article'	1.7%
有益	'beneficial'	1.7%
困難	'difficult'	1.7%
知識	'knowledge'	1.7%
眼鏡	'spectacles'	1.7%
喜好	'preference'	1.7%
報告	'report'	1.7%
報章	'newspaper'	1.7%
廣泛	'widely'	1.7%
課外書	'leisure book'	1.7%

Figure 3. Examples of Association Norms

VOC	SYL	POS	CON	FRQ	Examples	
Wrt	Di	Noun	Abs	Hi	文化 *man4-faa3* 'culture'	事業 *si6-jip6* 'career'
				Mid	性格 *sing3-gaak3* 'personality'	治安 *zi6-ngon1* 'public order'
				Lo	良知 *loeng4-zi1* 'conscience'	潛質 *cim4-zat1* 'potential'
			Con	Hi	電腦 *din6-nou5* 'computer'	手袋 *sau2-doi2* 'handbag'
				Mid	畫家 *waa2-gaa1* 'painter'	菜刀 *coi3-dou1* 'chopper'
				Lo	花生 *faa1-sang1* 'peanut'	藥膏 *joek6-gou1* 'ointment'
		Verb	Abs	Hi	明白 *ming4-baak6* 'understand'	應付 *jing3-fu6* 'handle'
				Mid	珍惜 *zan1-sik1* 'cherish'	抗拒 *kong3-kui5* 'resist'
				Lo	猶豫 *jau4-ji4* 'hesitate'	承繼 *sing4-gai3* 'inherit'
			Con	Hi	畢業 *bat1-jip6* 'graduate'	參考 *caam1-haau2* 'refer'
				Mid	散步 *saan3-bou6* 'stroll'	挑選 *tiu1-syun2* 'select'
				Lo	徘徊 *pui4-wui4* 'linger'	溜冰 *lau4-bing1* 'ice-skate'
		Adj	--	Hi	明顯 *ming4-hin2* 'obvious'	重要 *zung6-jiu3* 'important'
				Mid	低調 *dai1-diu6* 'low-key'	慷慨 *hong2-koi3* 'generous'
				Lo	幼稚 *jau3-zi6* 'naive'	脆弱 *ceoi3-joek6* 'fragile'
		Fluid	--	Hi	服務 *fuk6-mou6* 'serve/service'	計劃 *gai3-waak6* 'plan'
				Mid	指揮 *zi2-fai1* 'conduct(or)'	練習 *lin6-zaap6* 'exercise'
				Lo	推斷 *teoi1-dyun3* 'infer(ence)'	青春 *cing1-ceon1* 'young/youth'
	Mono	Noun	Con	Hi	手 *sau2* 'hand'	海 *hoi2* 'sea'
				Mid	菜 *coi3* 'vegetable'	碳 *taan3* 'carbon'
				Lo	氧 *joeng5* 'oxygen'	虎 *fu2* 'tiger'
	Tri	Noun	Con	Hi	身分證 *san1-fan2-zing3* 'ID card'	辦公室 *baan6-gung1-sat1* 'office'
				Mid	小朋友 *siu2-pang4-jau5* 'child'	牛仔褲 *ngau4-zai2-fu3* 'jeans'
				Lo	天花板 *tin1-faa1-baan2* 'ceiling'	伺服器 *si6-fuk6-hei3* 'server'
Spk	--	--	--	--	回水 *wui4-seoi2* '(to) refund'	屋企 *uk1-kei2* 'home'

Table 1. Distribution and Examples of Stimulus Words by Various Factors
(Each group contains 20 stimulus words, amounting to 500 words altogether.)

3.5 Classification of Responses

All responses were then classified according to their syntactic and semantic nature. The syntactic classification, as shown in Table 2, categorises each response with respect to its constituent unit (WRD for words, PHR for phrases, SEN for sentences and INC for incomplete), vocabulary type (WRT for written words, SPK for spoken words, ENG for English or foreign words, and MIX for code-mixed items), and part-of-speech (NN for common nouns, VB for verbs, AJ for adjectives, PN for proper nouns, and OT for all other categories). For example, one of the responses to 三文治 *saam1-man4-zi6* 'sandwich' is 好食 *hou2-sik6* 'delicious', and this response is classified as WRD (word), SPK (spoken), and AJ (adjective). The semantic classification, on the other hand, had to consider the relation between individual stimulus words and responses. The granularity of classification may vary according to different purposes of analysis (e.g. Guida and Lenci, 2007; McRae et al., 2012). In this study, four main types of relations, from narrow to broad, are considered, namely taxonomic, collocational, thematic, and other relations. As exemplified in Table 3, taxonomic relations[3] comprises ANT for antonymy, HYP for hypernymy/hyponymy, MER for meronymy/holonymy, PRP for properties/attributes, SBL for siblings or coordinate terms, and SYN for (near-)synonymy. Collocations (COL) cover strongly collocated stimulus-response pairs or common syntagmatic patterns. Thematic relations (THM) include all broad and contextual relations between the stimulus and response, which can usually be connected within a given theme or context. Situational and personal associations, and those

[3] For simplicity, properties/attributes and part-of relations are grouped with other conventional taxonomic relations in this study.

involving subjective perception and value judgement, are grouped into the last category (OTH), which accommodates all other cases that cannot fit into any of the above types.

Category	Examples
Constituent Unit	
WRD	互聯網 *wu6-lyun4-mong5* 'Internet'
PHR	賣豬肉 *maai6-zyu1-juk6* 'sell pork'
SEN	知識就是力量 *zi1-sik1-zau6-si6-lik6-loeng6* 'Knowledge is power'
INC	其實我 *kei4-sat6-ngo5* 'actually I'
Vocabulary Type	
WRT	乞丐 *hat1-koi3* 'beggar'
SPK	老豆 *lou5-dau6* 'dad (colloquial)'
ENG	iPhone
MIX	做 GYM *zou6-zim1* 'work out in gym'
Part-of-Speech	
NN	大提琴 *daai6-tai4-kam4* 'cello'
VB	打架 *daa2-gaa3* '(to) fight'
AJ	宏偉 *wang4-wai5* 'grand'
PN	陳奕迅 *can4-jik6-seon3* 'Eason Chan'
OT	若然 *joek6-jin4* 'if'

Table 2. Syntactic Classification of Responses

Type		Examples	
		Stimulus	Response
TAXONOMIC	ANT	樂觀 *lok6-gun1* 'optimistic'	悲觀 *bei1-gun1* 'pessimistic'
	HYP	老鼠 *lou5-syu2* 'mouse'	動物 *dung6-mat6* 'animal'
	MER	引擎 *jan5-king4* 'engine'	飛機 *fei1-gei1* 'airplane'
	PRP	石頭 *sek6-tau4* 'stone'	堅硬 *gin1-ngaang6* 'hard'
	SBL	長褲 *coeng4-fu3* 'trousers'	短褲 *dyun2-fu3* 'shorts'
	SYN	開心 *hoi1-sam1* 'happy'	快樂 *faai3-lok6* 'happy'
COL		解決 *gaai2-kyut3* 'solve'	問題 *man6-tai4* 'problem'
THM		機場 *gei1-coeng4* 'airport'	旅遊 *leoi5-jau4* 'travel'
OTH		頸鏈 *geng2-lin2* 'necklace'	討厭 *tou2-jim3* 'detest'

Table 3. Semantic Classification of Responses

4 Preliminary Analysis and Discussion

4.1 Basic Profile of Data

Among the 29,000 responses to 500 stimulus words from 58 participants, there are about 16,000 distinct stimulus-response pairs. The response types elicited by individual stimuli range from 7 to 49, averaging at 32.

Among all response tokens, the majority falls under written vocabularies. Less than 1% of the responses are non-Chinese (English or numeric) or code-mixed items. Spoken vocabularies occupy about 4% of all responses. Nevertheless, for the 20 spoken stimulus words, slightly more than 13% of the responses are spoken items. This suggests that while spoken and written items co-exist in the mental lexicon of Hong Kong Chinese, spoken items remain the minority but they are more closely linked together and more readily activate one another.

Excluding non-Chinese and code-mixed responses, the majority of the response tokens (about 74%) are disyllabic, followed by about 14% of monosyllabic responses, 8% of trisyllabic ones, and the remaining (less than 4%) are quadrisyllabic or longer. For monosyllabic stimuli alone, 40% of the responses are also monosyllabic, and disyllabic responses occupy only 50%. On the other hand, for trisyllabic stimuli alone, monosyllabic responses fall back to a general 13% whereas trisyllabic responses occupy about 11%. In general, about 5% of the responses are obviously non-words. Most of these are phrases, and there are some sentences, and only a few incomplete constituents.

For the part-of-speech (POS) and semantic nature of the responses, we consider only those which appeared twice or more. Among them, nominal responses occupy about 59%, verbal responses 19%, and adjectival responses 17%. Proper nouns account for less than 4% of the responses. It is interesting to note that the proportion of nominal responses stays the largest regardless of the POS of the stimulus words (54% for nominal stimuli, 70% for verbal stimuli, and 61% for adjectival stimuli). Nevertheless, associations between verbs and adjectives are apparently weaker as compared to verb-verb and adjective-adjective associations respectively. For instance, for verbal stimuli, there are 20% verbal responses but only 7% adjectival responses. For adjectival stimuli, there are 23% adjectival responses but only 15% verbal responses. This suggests the central role of nominal nodes in the

mental lexicon. Verbs and adjectives have their own mass of associations, but most of the time they are more strongly associated to nominal concepts. In other words, paradigmatic relations may be more salient for nominal concepts, whereas syntagmatic relations, especially noun-verb and noun-adjective associations, are more significant for the learning and memory of verbs and adjectives.

Regarding the semantic associations between the stimuli and responses, taxonomic relations account for about 20%, collocational relations 42%, thematic relations 21%, and the remaining 17% (others) are mostly non-linguistic associations which often involve personal experience and judgement. Previous analysis of word association norms reveals that there are several common relations found between the responses and the stimuli, including coordination (e.g. salt to pepper), superordination (e.g. colour to red), synonymy (e.g. hungry to starved), collocation (e.g. net to butterfly), attributes (e.g. comfortable to sofa) and functions (e.g. rest to chair) (e.g. Aitchison, 2003), though their distribution is not clear. It appears that for Chinese, at least from the pilot data in this study, narrow taxonomic relations (including attributive and part-of relations) are relatively minor, whereas collocational and thematic associations have the largest share.

4.2 Frequency, POS and Concreteness

Figure 4 shows the distribution of the various kinds of semantic associations for concrete stimuli, including nouns and verbs of high and low frequency. Figure 5 shows similar results for abstract stimuli.

Some interesting facts are observed on the semantic associations with respect to the frequency, POS and concreteness of the stimulus words. First, collocational associations are particularly prominent for abstract stimuli. For nouns and verbs alike, abstract words were mostly responded with collocational items. Second, while collocational responses still occupy a large share for concrete verb stimuli, their importance for concrete noun stimuli is taken over by thematic associations. In other words, concrete words tend to elicit more thematic responses, and this is particularly true for nouns. Third, in general nouns tend to elicit more taxonomic relations than verbs, except for high frequency abstract nouns and verbs. Fourth, the role of the rather constant proportion of other non-linguistic associations in the semantic memory should by no means be ignored.

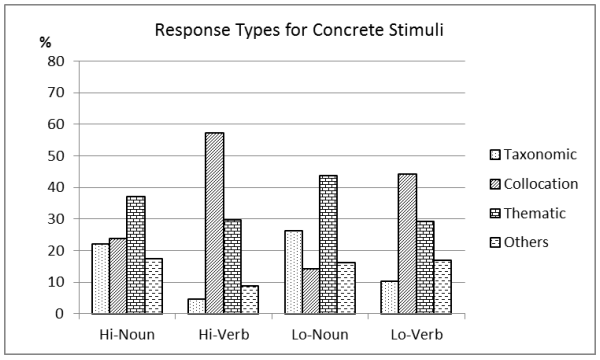

Figure 4. Response Types for Concrete Words

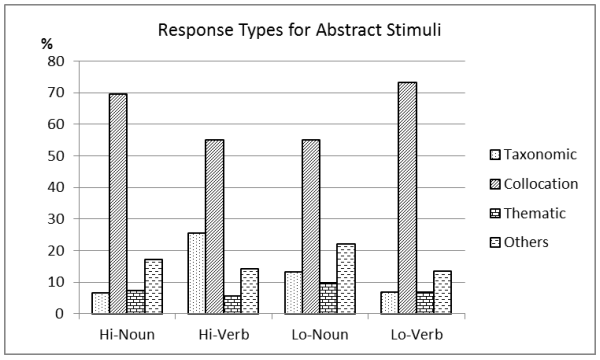

Figure 5. Response Types for Abstract Words

4.3 Some Implications

Referring to the questions raised in Section 1, the preliminary observations and analysis of the pilot association norms above thus suggest certain possibilities for the Chinese mental lexicon for further investigation.

Regarding the basic units, disyllabic words apparently dominate the Chinese mental lexicon. This is not only evident from the overall 74% disyllabic responses. In fact, the large number of monosyllabic items especially in response to monosyllabic stimulus words also points to the significance of disyllabic concepts. Many of the monosyllabic responses are not really elicited because of their morphemic meaning, but are intended to be combined with the stimulus to form a disyllabic word. For example, the stimulus 字 *zi6* 'character' elicited responses like 打 *daa2* '(to) hit' and 體 *tai2* 'body', probably because they form common words 打字 *daa2-zi6* '(to) type' and 字體 *zi6-tai2* 'font' respectively, and these disyllabic words are also among the responses for the same stimulus. There also remains the philosophical distinction between

concepts and lexical items in the mental lexicon. To a certain extent, the mental lexicon is not like a dictionary of words, but rather a repository of words superimposed with a semantic network of concepts. Hence different responses may simultaneously point to the same concept. For instance, among the responses for 閱讀 *jyut6-duk6* '(to) read' in Figure 3, the top three responses (書本 *syu1-bun2* 'book', 書 *syu1* 'book', and 書籍 *syu1-zik6* 'book') all refer to the same concept "book" except that they might be more conventionally used in different contexts or registers. This is quite specific to Chinese given its word formation mechanisms and may suggest a slightly different organisation of the mental lexicon for Chinese from that for English which is worth further investigation.

Regarding the semantic associations, although previous studies have identified several prominent kinds of responses particularly under narrow taxonomic relations, in this study we nevertheless found that the broader but less mentioned collocational relations and thematic relations made up the majority of the responses. This is in fact quite an unexpected finding, especially for nouns, though less so for verbs and adjectives. While taxonomic relations are by definition confined to stimulus and response under the same part-of-speech, collocational and thematic relations may apply to intra- or inter-POS pairs. Past studies on the semantic memory tend to focus on the connection of nominal concepts, and semantic lexicons often emphasise their taxonomic connections. When it comes to free association, however, it happens that even for nominal stimuli, taxonomic associations are not always more readily activated than collocational and thematic associations. This suggests that the associative strengths for pure ontological relations may not be as strong as others which are probably continuously reinforced through personal and media contact. The diversity of relations exhibited in word association norms should not be under-estimated, especially in view of the "Others" category which covers situational associations and personal judgement. For instance, one response for 漢堡包 *hon3-bou2-baau1* 'hamburger' was 好吃 *hou2-hek3* 'delicious' which appears to be a property of hamburgers, but it is nevertheless too subjective to be taken as a genuine semantic association as being delicious is hardly an intrinsic attribute for any food. This suggests that while the organisation of the mental lexicon as a network may be something universal, the semantic associations and associative strengths in individuals' semantic memory may be a personal copy containing the universal structure and ontological connections, but to a large extent augmented and shaped by one's experience, perception, and exposure to interpersonal as well as cultural and media influence. We must therefore be cautious in interpreting the association norms. The top responses for individual stimulus words may be considered more universal and they are expected to inform and correspond to what one designs for semantic lexicons. The infrequent responses, which make up the majority, constitute only personal mental pictures, and may not even be properly considered very weak associations in general.

Regarding the role of frequency, POS and concreteness, taxonomic relations as well as thematic relations are apparently more accessible for nouns than verbs, especially concrete nouns. The concreteness effect is also exhibited in the dominance of collocational responses among abstract stimuli. The frequency effect does not seem to be significant, except that high frequency abstract verbs were found to elicit unexpectedly many taxonomic associations. The dominance of nominal concepts in the mental lexicon is nevertheless obvious as nominal responses made up the majority regardless of the POS of the stimuli.

Regarding the organisation of written and spoken vocabulary items, the above preliminary analysis reveals that they co-exist in the mental lexicon of Hong Kong Chinese. While most lexical items can be used both in written and spoken form (and hence all considered written items), spoken items are more or less confined to specific contexts and registers, and are more readily activated when the stimulus word is also a spoken item. It seems that when prompted with a spoken item, participants tend to feel greater acceptability of colloquial spoken items to be given as responses.

4.4 Potential Applications

Results from word association norms have a direct role for advancing our understanding of the mental lexicon and the development of psycholinguistic models for human lexical processing. Upon quantitative and qualitative analyses of the norms, we can draw up hypotheses on the nature and strengths of semantic associations, the degree of isolation or

distinctiveness for particular concepts, the dispersion or diversity of the network, etc. for further investigation. One possible direction of study concerns the asymmetry of associative strength and thus readiness of activation between concepts. The traditional spreading activation model suggests that activation strengths depend, to a certain extent, on the number of connections (or density of nodes) from a concept and the distance in the activation path. With real data from the association norms (for example, the top response for the stimulus 衣櫃 *ji1-gwai6* 'wardrobe' is 衣服 *ji1-fuk6* 'clothing' (39.7%), but conversely, when 衣服 *ji1-fuk6* 'clothing' is the stimulus, 衣櫃 *ji1-gwai6* 'wardrobe' only accounts for 5.2% of the responses), this notion can be subject to more systematic investigation.

The association norms can also provide support for other areas including natural language processing, computational lexicography, and language pedagogy.

For natural language processing, the analysis of word association norms may reveal various kinds of semantic associations in relation to different salient factors including frequency, polysemy and concreteness, to inform the design of word sense disambiguation systems to better address the issue of lexical sensitivity, allowing a better exploitation of various knowledge sources with respect to different kinds of target words. For instance, Kwong (2012) suggested that concrete and abstract senses may be best distinguished and thus disambiguated by different knowledge sources.

For computational lexicography, the very notion of association plays an important role, and has thus significantly influenced the design of many lexical resources. WordNet (Miller et al., 1990) is perhaps the most typical example in this regard, as it started out as a psycholinguistic project on network models for the mental lexicon, but turned out to be one of the most popular semantic lexicons in computational linguistics. With the availability of large corpora, many studies have subsequently tried to simulate the observations from word association norms statistically from large corpora, and such statistical simulation provides concrete and scalable data to enhance lexicography (e.g. Church and Hanks, 1990; Wettler and Rapp, 1993; Ferret and Zock, 2006). Having a set of Chinese word association norms available will certainly enhance work in this regard.

Despite that participants were clearly asked to give the first "word" which the stimulus word brings to mind, responses other than words are still seen. This echoes that the concept of "word" is not really a clear and unambiguous one even for native speakers of Chinese. Monosyllabic responses comprise morphemes and words, and multi-syllabic responses contain phrases, sentences, and even non-constituent in addition to words. The many responses in the "Others" category involving extra-linguistic associations and personal mental pictures do not only reveal the organisation of the mental lexicon (especially of the younger generation) but also the influence of media and culture, which should be informative to educators to reflect on language pedagogy, even for L1 teaching.

5 Future Work and Conclusion

In this paper, we have reported on the design and implementation of a pilot word association test for Hong Kong Chinese. Preliminary analysis of the resulting association norms has revealed the dominance of disyllabic and nominal concepts, and collocational associations in the Hong Kong Chinese mental lexicon. The concreteness effect was also observed. In addition to traditional linguistic relations, semantic associations based on subjective experience and value judgement were also constantly found. More detailed quantitative and qualitative analysis of the responses is underway, and a larger-scale data collection will be planned. Notwithstanding the limitation of discrete word association, a benchmarking set of word association norms, which is currently lacking for Hong Kong Chinese, can offer a snapshot of the mental lexicon to inform and complement experimental studies. Further investigation on how the conventional spreading activation model may account for the asymmetry of associations will be done. The association norms will also be useful to other related areas like natural language processing, computational lexicography, and language pedagogy.

Acknowledgements

The work described in this paper was supported by grants from the City University of Hong Kong (Project No. 7002798 and funding from the Department of Chinese, Translation and Linguistics).

References

Aitchison, J. (2003) *Words in the Mind: An Introduction to the Mental Lexicon*. Blackwell Publishers.

Bleasdale, F.A. (1987) Concreteness dependent associative priming: Separate lexical organization for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 582-594.

Bock, K. and Levelt, W. (1994) Language production: Grammatical encoding. In M.A. Gernsbacher (Ed.), *Handbook of Psycholinguistics*. San Diego: Academic Press.

Caramazza, A. (1997) How many levels of processing are there in lexical access? *Cognitive Neuropsychology, 14*: 177-208.

Church, K.W. and Hanks, P. (1990) Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics, 16(1)*:22-29.

Collins, A.M. and Loftus, E.F. (1975) A spreading-activation theory of semantic processing. *Psychological Review, 82(6)*:407-428.

Ferret, O. and Zock, M. (2006) Enhancing electronic dictionaries with an index based on associations. In *Proceedings of COLING-ACL 2006*, Sydney, Australia, pp.281-288.

Groot, A.M.B. de (1989) Representational Aspects of Word Imageability and Word Frequency as Assessed Through Word Association. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15(5)*:824-845.

Guida, A. and Lenci, A. (2009) Semantic Properties of Word Associations to Italian Verbs. *Italian Journal of Linguistics, 19(2)*: 293-326.

Hirsh, K.W. and Tree, J.J. (2001) Word association norms for two cohorts of British adults. *Journal of Neurolinguistics, 14*:1-44.

Hoosain, R. (1992) Psychological reality of the word in Chinese. In H-C. Chen and O.J.L. Tzeng (Eds.), *Language Processing in Chinese*. Amsterdam: Elsevier Science Publishers, pp.111-130.

Jenkins, J.J. (1970) The 1952 Minnesota word association norms. In L. Postman and G. Keppel (Eds.), *Norms of Word Association*. New York: Academic Press, pp.1-38.

Joyce, T. (2005) Constructing a Large-Scale Database of Japanese Word Associations. (Special issue: Kanji corpus research, edited by Katsuo Tamaoka), *Glottometrics, 10*.

Klein, D.E. and Murphy, G.L. (2002) Paper has been my ruin: conceptual relations of polysemous senses. *Journal of Memory and Language, 47*: 548-570.

Kroll, J.F. and Merves, J.S. (1986) Lexical access for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*:92-107.

Kwong, O.Y. (2012) Psycholinguistics, Lexicography, and Word Sense Disambiguation. In *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation (PACLIC 26*), Bali, Indonesia, pp.408-417.

Lin, C-J. C. and Ahrens, K. (2010) Ambiguity Advantage Revisited: Two Meanings are Better than One When Accessing Chinese Nouns. *Journal of Psycholinguistic Research, 39*:1-19.

McClelland, J.L. and Rumelhart, D.E. (1981) An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88*: 375-407.

McRae, K., Khalkhali, S. and Hare, M. (2012) Semantic and Associative Relations in Adolescents and Young Adults: Examining a Tenuous Dichotomy. In V.F. Reyna, S.B. Chapman, M.R. Dougherty and J. Confrey (Eds.), *The Adolescent Brain: Learning, Reasoning, and Decision Making*. American Psychological Association.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K.J. (1990) Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography, 3(4)*:235-244.

Moss, H. and Older, L. (1996) *Birkbeck Word Association Norms*. Hove, U.K.: Psychology Press.

Myers, J. (2006) Processing Chinese compounds: A survey of the literature. In G. Libben and G. Jarema (Eds.), *The Representation and Processing of Compound Words*. Oxford: Oxford University Press, pp.169-196.

Rayner, K. and Frazier, L. (1989) Selection mechanisms in reading lexically ambiguous words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15(5)*, 779-790.

Rodd, J., Gaskell, G. and Marslen-Wilson, W. (2002) Making Sense of Semantic Ambiguity: Semantic Competition in Lexical Access. *Journal of Memory and Language, 46*:245-266.

Schulte im Walde, S., Melinger, A., Roth, M. and Weber, A. (2008) An Empirical Characterisation of Response Types in German Association Norms. *Research on Language and Computation, 6(2)*: 205-238.

Taft, M. (2006) Processing of characters by native Chinese speakers. In P. Li, L.H. Tan, E. Bates and O.J.L. Tzeng (Eds.), *The Handbook of East Asian Psycholinguistics, Volume 1: Chinese*. New York: Cambridge University Press.

Wettler, M. and Rapp, R. (1993) Computation of word associations based on the co-occurrences of words in large corpora. In *Proceedings of the 1st Workshop on Very Large Corpora: Academic and Industrial Perspectives*, Columbus, Ohio, pp.84-93.

Zhou, X. and Marslen-Wilson, W. (1994) Words, morphemes and syllables in the Chinese mental lexicon. *Language and Cognitive Processes, 9(3)*:393-422.