# Classifiers of Mandarin Alphabetical Words with Character-alphabet Structure

Xinlan Zhao[1], Yu-Yin Hsu[1], and Chu-Ren Huang[1*]

[1] Department of Chinese and Bilingual Studies, Hong Kong Polytechnic University, Hong Kong, People's Republic of China

**Abstract.** Mandarin alphabetical words (MAWs) refer to the code-mixing of Romanized letters and characters such as X 光 'X-ray' in the Mandarin lexicon. Previous studies have mainly focused on MAWs' formation but lacked empirical evidence regarding their morpho-syntactic behaviours. Classifiers have been used to infer nominals' semantic properties and characteristics. An intriguing yet less explored issue is the classifier-selection pattern of MAWs and MAWs' morpho-syntactic idiosyncrasies. We adopt a corpus-based approach to handle this issue. Assuming that a MAW's classifier is motivated by the head of that MAW, we hypothesize that when a MAW is integrated into the Mandarin lexicon, its dominant classifiers will be the semantically more specific ones and not the neutral classifier. We show that MAWs share a dominant compounding structure in Mandarin and that MAW's classifier is decided by that head even when the head is represented by alphabets.

**Keywords:** Mandarin Alphabetical Words, Classifiers, Code-mixing, Corpus-based approach

## 1 Introduction

In recent years, Chinese-English code-mixing has become increasingly common in Chinese society with frequent foreign exchanges. Code-mixing is the mixed usage of different languages, varieties, or orthographic systems, typically within the same sentence [1]. With frequent language contacts, people code switches/mixes to find proper expressions when there is no appropriate translation for the language being used or when communicators generate code-mixing language; social factors such as education, religion, gender, and age also have influences on the degree of this phenomenon [2, 3]. Mandarin Alphabetical Words (MAWs), as a result of language contact, mainly refer to the code-mixed words in Chinese, which usually consist of Chinese Characters and alphabetical letters (e.g., BB 霜 'BB cream', and 维 C 'vitamin C') [4].

In the literature, MAWs are further divided into three major subcategories based on the positions of the alphabetic letters and characters: A-C (alphabet + character), C-A (character + alphabet), and C-A-C (alphabet in the middle) [4]. Typical MAWs follow the Chinese modifier-modified (head)[1] Morphological rule. The morphological head of a MAW can be a Chinese character, as in BB霜 'blemish balm' with the character 霜 'cream' as its head, and can also be alphabets, as in 气垫BB 'air cushion blemish balm', which has the alphabets "BB" as its head.

MAWs have attracted much attention from scholars because of their wide use in natural settings and the complicated linguistic idiosyncrasies they demonstrate concerning the combined features of both Chinese and foreign languages [5]. The previous studies of MAWs have mainly focused on language policy issues [6, 7], action techniques [4, 5, 8, 9], and the evolutionary traits of MAWs [10, 11]. More recently, scholars have begun to investigate the language behaviours of MAWs on morpho-phono-orthographical levels via experimental approaches [12, 13] and corpus-based approaches [4, 5, 14, 15]. Despite significant work, scholars still need to answer whether or to what extent MAWs adapt to the Mandarin language system.

Among the linguistic issues, it is well known that Chinese is a classifier language [16]. The selection of the Chinese classifier (hereafter CL) for a noun is driven by the semantic properties of the head noun [17, 18, 19]. The CL-noun agreement and the CL choice are the two dominant topics in recent research on Chinese CL. Unfortunately, this vital reference has not been linked to the linguistic manifestations of MAWs, although it is intriguing to consider how the two lines of lexicons interact.

Therefore, this paper addresses the relationship between CLs and MAWs by studying the CL distribution of the C-A type of MAWs with alphabetic letters as the morphological heads. The reason for investigating this C-A structure of MAWs was that we were curious about how well the alphabetic letters were integrated into the Chinese lexicon at the syntactical level, mainly when they played an important role as head morphemes (the MAWs mentioned in the rest of the paper refer only to C-A MAWs).

In summary, our work aimed to answer the following research questions:

(1) Does the CL distribution of MAWs follow the semantically motivated rule the same as Mandarin nominals do (e.g., 一台车载GPS/一台车载导航, 'a vehicle navigator' vs 一台电脑 'a computer')?

(2) Are the MAW's dominant CLs the semantically more specific ones rather than being the neutral classifier 个(GE)?

Since MAWs share a typical modifier-head compounding structure, we hypothesized that the CLs of MAWs should be motivated by the semantic meaning of the head regardless of the orthography of the head; thus, their CLs would not be replaced by the neutral CL 个(GE) as well.

The rest of the paper is structured as follows: The criteria for the data selection are described in Section 2, the method is introduced in Section 3, the results of the data analysis are presented in Section 4, and the paper ends with a discussion and a conclusion in sections 5 and 6.

[1] Like many languages, Chinese compound nouns or nominal phrases often have the "modifier-head" structure; for example, 火车站 'train station' has the modifier 火车 'train' modifying the head 站 'stop, station', and the syntactic status of the word is classified by the head 站 'stop, station' instead of by the modifier 火车 'train'.

## 2 Selection of the MAWs and CLs

### 2.1 Criteria for MAW selection

The selection of the MAW seed words for the present study was based on two corpus-based wordlists and one published dictionary. The first wordlist is built by [4] from both the Sinica Corpus [20] and the Chinese Gigaword Corpus [21]. The second one is the wordlist built by [5] from Sina Weibo, a predominant social media platform in China. These two wordlists are up-to-date and considerable in scale. However, the words were not manually checked, so we also consulted the *dictionary of Lettered-words* [22] to balance the bias of word source. The dictionary has over 2000 MAW entries covering all structures of A-C, C-A, C-A-C, and pure alphabetical words with different POS statuses.

However, to capture the essential properties of MAWs relevant to the research questions, the MAWs were shortlisted using the following criteria:

- the MAWs are nouns or compound nouns
- the form of the MAWs is C-A (character-alphabet) type
- the MAWs can co-occur with at least three CLs in the corpus
- the MAWs have over 200 hits in the corpus

During the selection process, it is noticed that there are subcategories of MAWs. Most MAWs have alphabetic letters as the heads (e.g., 蓝光 DVD ‘Blu-ray DVD’ has DVD as the head), but few present structures differently. For example, 卡拉 OK ‘karaoke’ is single morpheme word, where 卡拉 and OK phonetically transcribed the first two and the last syllables in *karaoke*, respectively; 维生素 C ‘vitamin C’ in which 维生素 ‘vitamin’ has properties more as the semantic head and C indicates the subtype of vitamin; 甲 A 'First Division Group A-League’, and 甲 B ‘First Division Group B League’, instead, having the alphabetic letters, A and B modifying the First Division Group to show its sub-levels.

Considering that the semantic head of a noun is a crucial factor that affects the selection of Chinese CLs, the MAWs are sorted into two groups. One group has alphabetic letters as the head; the other group contains other structures.

After careful selection, 30 focused MAWs are listed in Tables 1 and 2. The amount of seed words in this study is limited due to two facts. First, the MAWs are much less than the others by nature. For example, in *The Dictionary of Lettered-words,* there are only 38 MAWs out of the total over 2000 entries. Second, the candidates went through a strict filtering process to retain only those with high frequency and at least three co-occurring CLs in the corpus.

**Table 1.** MAWs with the lettered-head.

| MAWs | Hits in corpus[2] | English translations |
|---|---|---|
| 阿 SIR | 358 | policeman, the cop (Cantonese) |

[2] The Chinese Web 2017 Simplified corpus (zhTenTen17) [23]: https://www.sketchengine.eu/zhtenten-chinese-corpus/

|  |  |  |
|---|---|---|
| 短 T | 516 | short-sleeved T-shirt (short T) |
| 长 T | 354 | long-sleeved T-shirt (long T) |
| 深 V | 8173 | deep V-neck dress |
| 气垫 BB | 1523 | air cushion blemish balm |
| 蓝光 DVD | 1134 | blue ray digital video disk |
| 车载 GPS | 5291 | vehicle global position system |
| 电台 DJ | 1040 | radio disc jockey |
| 人均 GDP | 34488 | per capita GDP |
| 硬盘 DV | 237 | hard drive digital video |
| 终极 PK | 1993 | the ultimate player killing |
| 真人 CS | 5949 | cosplay of counter-strike |
| 量贩式 KTV | 451 | buffet-style KTV |
| 螺旋 CT | 8829 | spiral computed tomography |
| 亲子 DIY | 591 | parent-child Handmade (do it yourself) |
| 企业 HR | 3144 | Human Resource manager of an enterprise |
| 职场 OL | 538 | office lady in the workplace |
| 高清 MV | 737 | music TV of high-definition |
| 白光 LED | 5841 | white light LED (light-emitting diode) |

**Table 2.** MAWs without lettered-head.

| MAWs | Hits in corpus | English translations |
|---|---|---|
| 甲 A | 5822 | First Division Group A |
| 甲 B | 2084 | First Division Group B |
| 傻 B | 1839 | silly person |
| 卡拉 OK | 25526 | Karaoke |
| 维生素 A | 53996 | Vitamin A |
| 维生素 B | 29566 | Vitamin B |
| 维生素 C | 105054 | Vitamin C |
| 维生素 E | 44331 | Vitamin D |
| 维生素 D | 45499 | Vitamin E |
| 维 C | 8994 | short form of 'Vitamin C' |
| 维 E | 1757 | short form of 'Vitamin E' |

## 2.2 CL selection rules

In general, a CL tends to occur before a noun in Mandarin Chinese when it needs to be individualized for counting [17, 18], including the individual reading 一本书 'a book', the kind reading 一种希望 'a hope', and event reading 一次比赛 'a competition' [19]. It is generally agreed that CLs are distinguished from measure words [24, 25]. According to Tai [24], CLs "pick out the salient perceptual properties associated with the noun". On the other hand, measure words only "denote the quantity of nouns and are used to count or measure". Based on the CL & measure word dichotomy, Ahrens and Huang [18] proposed the Chinese CLs taxonomy shown in Fig. 1. Based on this categorization, only sortal CLs (individual CLs, kind CLs, and event CLs) are relevant in the current study, while measure words are not investigated. The detailed interpretations of each CL mentioned in the rest of the paper (including CLs in the tables) are listed in the Appendix.

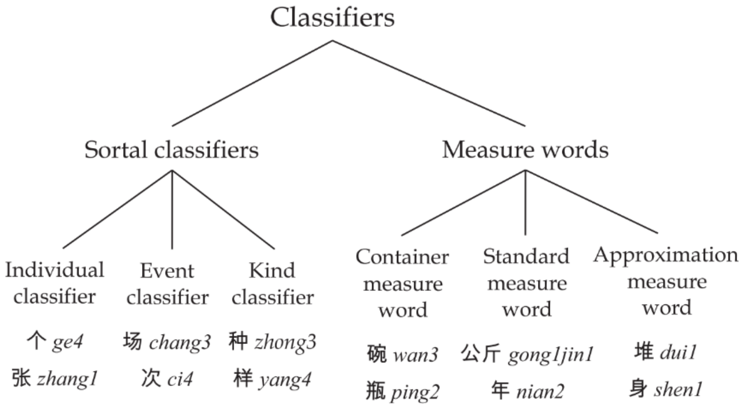

**Fig. 1.** Taxonomy of classifiers [18].

# 3 Data

The data of this study were generated from an up-to-date Chinese corpus: the Chinese Web 2017 Simplified corpus (zhTenTen17) [23], which can be accessed through the platform Sketch Engine [26]. This corpus has the advantages of considerable size (15 billion tokens), genre and style diversity, adequate functions, and linguistic annotations.

## 3.1 Data extracting process

Native Chinese speakers can name a few CLs of a specific noun easily by intuition but may not be able to retrieve all the possible CLs; however, a corpus-assisted extraction of CLs can help to provide a more comprehensive CL set. The Sketch Engine platform has a function to get a CL reference list for a specific noun to gain an ideal range of CLs. That is the “word sketch” function. For instance, a list of CLs (indicated as “measure”) co-occurring with the noun 维 C ‘vitamin C’ can be obtained together with the Mutual Information (MI) scores indicating the strength of the collocation. After getting the CL lists suggested by the corpus, the next step is to obtain qualified CL-MAW data.

The whole process of extracting data has three steps. The first step was to obtain the concordances of a target noun in the corpus. In the second step, we filter the concordance examples by each co-occurring CL. Then, each filtered data set was shortlisted by manual checking.

Noted that, in step three, if the data set was too large, such as containing over 1000 concordance lines, a random sample of 100 concordances was extracted automatically, and a rate of valid data was gained by counting the valid data out of the 100 examples. For example, suppose 3950 sentence examples contain both the noun 维生素 A 'vitamin A' and the CL 种 'kind'; 100 random samples will be checked to see how many valid examples there are. If there are four valid examples out of 100, the valid data rate would be 0.04 and the total number of 维生素 A 'vitamin A' and 种 'kind' combination will be 3950*0.02, that is 158. In our dataset, seven MAW-CL combinations underwent this random sample process; the others were checked based on the full concordance results.

### 3.2 Data examples

Based on the primary semantic functions of CLs, which are to individuate and to identify the units for enumeration or reference [18], the common types of CL-MAWs are shown below with authentic examples taken from the corpus of zhTenTen:

(1) CLs with a numeral preceding the noun:

| | | | | | |
|---|---|---|---|---|---|
| 一 | 件 | 短T | 就 | 够 | 了。 |
| yi | jian | duanT | jiu | gou | le。 |
| one | CL | short T | just | enough | LE |

'One short T is enough.'

(2) CLs be used with a referential demonstrative without a numeral:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 这 | 款 | 气垫 BB | 一共 | 有 | 两 | 个 | 色号。 |
| zhe | kuan | qidian BB | yigong | you | liang | ge | sehao。 |
| this | CL | air cushion BB | altogether | have | two | GE | colour type |

This air cushion BB has two colour types altogether.'

(3) CLs with ordinal numbers preceding the noun:

| | | | | | | |
|---|---|---|---|---|---|---|
| 1994 | 年 | 研制 | 出 | 第一 | 颗 | 白光 LED。 |
| yi jiu jiu si | nian | yanzhi | chu | diyi | ke | baiguang LED。 |
| 1994 | year | develop | out | first | CL | white light LED |

'The first white LED was developed in 1994.'

(4) The numeral-CL combinations are placed after the noun:

| | | | |
|---|---|---|---|
| 卡拉 OK | 300 元 | 一 | 场。 |
| kala OK | san bai yuan | yi | chang。 |
| Karaoke | 300 Yuan | one | CL |

'The Karaoke is 300 Yuan one time.'

(5) CLs and nouns are separated by other modifiers (e.g., a relative clause):

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 一 | 件 | 带 | 有 | 萌宠 | 图案 | 的 | 短 T。 |
| yi | jian | dai | you | mengchong | tu'an | de | duanT。 |
| one | CL | take | have | cute pet | motif | DE | short T |

‘A short T-shirt with a cute pet motif.’

## 4 Results

The main results of MAWs’ CL distributions are listed in Table 3 and 4. Table 3 focuses on the CLs of MAWs with alphabets as a head, while Table 4 is about the CL results of single morpheme MAWs or MAWs with characters as a head.

In Table 3, the results show us that first, MAWs can co-occur with multiple CLs covering all three types of sortal CLs. For example, the individual CLs 个 (GE), 张 ‘to classify objects that are flat, two-dimensional, and horizontal or objects’, 件 ‘to classify clothes’; the event CLs 场 ‘to classify scheduled events’, 次 ‘a general classifier for events’, 局 ‘to classify the occurrence of games’; and the kind CLs 种 ‘to classify the kinds of entities’, 类 ‘to classify kinds of entities’, 款 ‘to classify the categories of products’. Second, all three types of sortal CLs can be dominant CLs of a specific MAW. The most dominant CLs are individual CLs, followed by the kind and event CLs. Third, the dominant CL of each MAW is consistent with the semantic meaning of the head morpheme, which is also the alphabetical part of the word.

As examples to the third point, 阿 SIR ‘policeman’ and 职场 OL ‘office lady’ has the semantic meaning of certain kind of people. The semantic meaning is denoted by the head morphemes SIR and OL; according to the heads, the CLs selected are CLs for people, such as 个 (GE), 位 ‘to classify people with a polite sense’, and 名 ‘to classify people when their social role refers to them'. Interestingly, for the two words of 真人 CS ‘cosplay of counter strike’ and 亲子 DIY ‘parent-child handmade’, although the character parts indicate people, their collocating CLs ‘to classify scheduled events’, 次 ‘a general classifier for events’, and 期 ‘to classify events that involve stages of completion’ are consistent with the alphabetical head morphemes CS ‘counter strike’ and DIY ‘do it yourself’ which refer to specific activities.

**Table 3.** CL-distributions of MAWs with the lettered-head.

| MAW | English | Frequency of CLs |
|---|---|---|
| 阿 SIR | Policeman, cop (Cantonese) | 个 9，位 6，名 2，群 1 |
| 短 T | short-sleeved T-shirt (short T) | 件 55，款 15，种 3，个 3，套 1 |
| 长 T | long-sleeved T-shirt (long T) | 件 27，款 20，个 2，种 2，身 1 |
| 深 V | deep V-neck dress | 件 14，个 5，款 3，种 3，袭 1 |
| 气垫 BB | air cushion blemish balm | 款 132，个 13，种 6，套 1，波 1，批 1 |
| 蓝光 DVD | blue ray digital video disk | 张 13，台 3，个 2，代 2，批 2 |

| | | |
|---|---|---|
| 车载 GPS | vehicle global position system | 款 75，台 13，个 10，种 6，部 2 |
| 电台 DJ | radio disc jockey | 位 37，名 34，个 33，家 5 |
| 人均 GDP | per capita GDP | 项 5，种 2，条 1，类 1 |
| 硬盘 DV | hard drive digital video | 款 20，代 2，台 1 |
| 终极 PK | the ultimate player killing | 场 60，轮 25，次 22，个 5，番 2 |
| 真人 CS | cosplay of counter-strike | 场 29，次 12，个 5，局 3，种 1，款 1 |
| 量贩式 KTV | buffet-style KTV | 家 11，个 2，种 2 |
| 螺旋 CT | spiral computed tomography | 台 167，次 31，个 9，种 4，款 2 |
| 亲子 DIY | parent-child Handmade (do it yourself) | 场 5，种 3，个 1，项 1，份 1，期 1 |
| 企业 HR | Human Resource manager of an enterprise | 位 44，名 27，个 24，类 1 |
| 职场 OL | office lady in the workplace | 个 11，位 2，名 2 |
| 高清 MV | music TV of high-definition | 首 9，个 5，部 2，支 1 |
| 白光 LED | white light LED (light-emitting diode) | 个 110，颗 54，种 50，只 44，类 6，组 6，支 6，款 4，粒 2，串 1，排 1 |

The CLs of the second group (Table 4) are also diverse in type and plentiful in numbers. According to Table 4, the CLs of these MAWs are also semantically motivated by the head morphemes. For example, event CLs 轮 ‘to classify sequences of events’, 场 ‘to classify scheduled events’, and 届 ‘to classify scheduled events’ are classifying events of 甲 A ‘First Division Group A’, and 甲 B ‘First Division Group B’; the individual CLs 群 ‘to classify a swarm of people’, 帮 ‘to classify a group of people’ and 名 'to classify people when their social role refers to them' are classifying the people of 傻 B ‘silly people’. In addition, the single-morpheme word 卡拉 OK ‘Karaoke' is mainly classified by 家 ‘to classify institutions’, indicating that it is often related to the venue. Meanwhile, it is also referred to as songs, singing events, or singing equipment because there are other CLs such as 首 ‘to classify songs or music’, 场, and 代 ‘to classify people, technologies or products of an equal generation’.

The vitamin MAWs, the characterized 维生素 ‘vitamin’ or its short form 维 ‘V-’ denotes the core physical meaning, whereas the letters indicate the categories of vitamins. We can see that the coverage of CLs to these vitamin words is similar, as all of them have the shape CLs such as 颗 ‘to classify round objects’, 粒 ‘to classify small round objects’, and 片 ‘classify objects that are flat and thin’ classifying their common feature of small round physical shapes. Additionally, several kinds of vitamins select the kind CL 种 ‘to classify the kinds of entities as their dominant CL; this may be due to their common feature of being referred to as kinds of vitamins in most contexts.

**Table 4.** CL-distributions of MAWs without lettered-head.

| MAW | English translations | Frequency of CLs |
|---|---|---|
| 甲 A | First Division Group A | 轮 57，届 23，场 7，支 1，局 1 |

| | | |
|---|---|---|
| 甲 B | First Division Group B | 轮 14，场 2，届 1 |
| 傻 B | silly person | 个 177，群 27，种 11，帮 11，名 1 |
| 卡拉 OK | Karaoke | 家 74，个 36，次 30，首 30，场 19，曲 7，种 6，代 1 |
| 维生素 A | Vitamin A | 种 158，粒 11，颗 2，个 2 |
| 维生素 C | Vitamin B | 片 238，种 149，粒 46，颗 12，款 9，个 2 |
| 维生素 E | Vitamin C | 粒 85，种 67，片 45，颗 27，个 4 |
| 维生素 D | Vitamin D | 种 53，粒 3，款 3，类 2，片 1 |
| 维生素 B | Vitamin E | 种 154，片 104，粒 9，颗 6，类 2 |
| 维 C | Vitamin C | 片 25，粒 7，颗 4，个 4，种 3，枚 1 |
| 维 E | Vitamin E | 颗 7，粒 7，片 4，种 1，款 1 |

Generally, the data results above verify our first hypothesis that the CL selection rule of MAWs conforms to the Mandarin grammar rule, which is highly semantic motivated. Meanwhile, the CLs can demonstrate the lexical semantics of MAWs or coerce different semantic readings of MAWs.

## 5 Discussion

Based on the above results, we elicit further observations to compare the CL behaviours of MAWs with their traditional Chinese canonical forms.

First, for the group of MAWs with lettered-heads, we generated the CLs of some of the equivalent Chinse canonical forms in Table 5. Note that only some MAWs have canonical forms with good CL collocations. From Table 5, the MAWs and the Chinese canonical words behave similarly regarding CL selection in most cases. Both can select a group of CLs to emphasis different semantic facets in an authentic context. For example, 件 'to classify clothes' is preferred by clothes; 个 (GE), 位 'to classify people with a polite sense', and 名 'to classify people with social role' are preferred by people; 款 'to classify the categories of products' and 个 are selected by products, etc.

Nevertheless, within each pair, the preference difference of a dominant CL exists. For instance, 警察 'policemen' has a particular preference for the CL 位, while 阿 SIR 'policemen (Cantonese)' mostly choose the neutral CL 个. In the sense that 位 is more polite than 个 when classifying a person, it may indicate that 警察 is more respectful than 阿 SIR in most contexts. Most importantly, there is no evidence that the MAWs have a trend of using more 个 than the canonical forms.

**Table 5.** CL-distributions of letter-head MAWs and their canonical forms.

| Keywords | English translations | Frequency of CLs |
|---|---|---|
| 短 T | short T-shirt | 件 55, 款 15, 种 3, 个 3, 套 1 |
| 短袖衫 | short T-shirt | 件 159, 款 21, 种 10, 身 2 |
| 长 T | long T-shirt | 件 25, 款 12, 个 2, 种 2, 身 1 |

| | | |
|---|---|---|
| 长袖衫 | long T-shirt | 件 76, 款 11, 条 3, 种 2 |
| 气垫 BB | air cushion blemish balm | 款 132, 个 13, 种 6, 套 1, 波 1, 批 1 |
| 气垫粉底 | air cushion foundation | 款 72, 个 10, 种 6, 类 1, 套 1, 批 1,件 1 |
| 车载 GPS | vehicle global position system | 款 75, 台 13, 个 10, 种 6, 部 2 |
| 车载导航 | vehicle global position system | 款 43, 台 3, 个 10, 代 7, 种 3, 部 3 |
| 终极 PK | the ultimate player killing | 场 60, 轮 25, 次 22, 个 5, 番 2, 版 1 |
| 终极对决 | the ultimate player killing | 场 110，次 14，轮 7，个 2，局 1 |
| 企业 HR | Human Resource manager | 位 44, 名 27, 个 24, 类 1 |
| 企业人事 | Human Resource manager | 位 3, 名 3, 个 3 |
| 白光 LED | white light LED | 个 110, 颗 54, 种 50, 只 44, 类 6, 组 6, 支 5, 款 4, 粒 2, 串 1, 批 1, 排 1, 条 1 |
| 白光发光二极管 | white light LED | 个 5, 种 2, 只 1, 代 1, 层 1 |
| 阿 SIR | policeman (Cantonese) | 个 9, 位 6, 名 2, 群 1 |
| 警察 | policeman | 名 36219, 个 18382, 位 5697, 群 1165, 帮 452, 批 391, 种 207, 类 44, 波 16, 枚 4 |
| 亲子 DIY | parent-child Handmade DIY | 场 5, 种 3, 个 3, 项 1, 份 1, 期 1 |
| 亲子手工 | parent-child Handmade | 次 3, 个 3, 场 1, 份 1, 种 1, 款 1 |

Second, to have a more general view of the two forms of words, the CLs are sorted into three types: the general 个 (GE), other individual CLs (IN), and kind CLs (KD), and each CL-type percentage is calculated in Table 6. Here, the reason that the general CL 个 is separated from individual CLs is that it is different from other individual CLs, which are decided explicitly by the salient semantic meaning of the noun; 个 is vague in semantic meaning and sometimes only used to fulfil the CL function; also, it can replace any other individual CLs grammatically. To some extent, the rich range of individual CLs of a noun is a good symbol of their identity as a typical Chinese word. It is also noticed that the event CLs are not being investigated here since most words are not event nouns.[3]

**Table 6.** Percentages of three-type CLs.

| English | MAW | GE | IN | KD |
|---|---|---|---|---|
| short T-shirt | 短 T | 3.90 | 72.73 | 23.38 |
| long T-shirt | 长 T | 3.85 | 53.85 | 42.31 |
| air cushion BB | 气垫 BB | 8.50 | 1.31 | 90.20 |
| Vehicle GPS | 车载 GPS | 9.43 | 14.15 | 76.42 |
| The ultimate player killing | 终极 PK | 83.33 | 16.67 | 0 |
| Human Resource manager | 企业 HR | 25.00 | 73.96 | 1.04 |
| white light LED | 白光 LE | 38.73 | 40.14 | 21.13 |
| policeman | 阿 SIR | 50.00 | 50.00 | 0.00 |
| parent-child Handmade | 亲子 DIY | 28.57 | 28.57 | 42.86 |

[3] Event nouns are a subtype of nouns that lexically encode process readings [19]. For example, 真人 CS ‘cosplay of counter strike’.

| | **Average** | **27.92** | **39.04** | **33.04** |
|---|---|---|---|---|
| **English** | **Canonical** | **GE** | **IN** | **KD** |
| short T-shirt | 短袖衫 | 4.95 | 79.7 | 15.35 |
| long T-shirt | 长袖衫 | 0.00 | 85.87 | 14.13 |
| air cushion BB | 气垫粉底 | 8.55 | 1.32 | 90.13 |
| Vehicle GPS | 车载导 | 11.24 | 37.08 | 51.69 |
| The ultimate player killing | 终极对决 | 100.0 | 0.00 | 0.00 |
| Human Resource manager | 企业人事 | 33.33 | 66.67 | 0.00 |
| white light LED | 白光发光二极管 | 55.56 | 22.22 | 22.22 |
| policeman | 警察 | 29.35 | 70.24 | 0.4 |
| parent-child Handmade | 亲子手工 | 50.00 | 16.67 | 33.33 |
| | **Average** | **32.74** | **42.2** | **25.25** |

From Table 6 above, it is highlighted that firstly, the average difference between the three kinds of CLs is minor between the MAW group and the canonical group, less than 10% each. Secondly, the MAW and canonical groups share that the individual CLs (IN) have the highest mean among all three CL types. That is to say, the similarities testify to our second hypothesis that there are no major differences between MAWs and the canonical forms regarding the CL selecting methods and the regular range of 个 (GE).

The other MAWs group can also enhance the generalization stated above. We consulted the CL distributions of 甲 A 联赛 'First Division Group A-League', 甲 B 联赛 ‘First Division Group B League’, 傻子 ‘fool’, and 维生素 ‘vitamin’ as the character-headed comparative items, to look at whether their CLs behave differently. Table 7 shows that the dominant CLs and the range of CLs between these MAWs and their counterparts are highly consistent. For instance, event CLs such as 轮 ‘to classify sequences of events’, 场 ‘to classify scheduled events’, 届 ‘to classify scheduled events’ are shared by football games; individual CLs 个 (GE), 群 ‘to classify a swarm of people’, 种 ‘kind’ is used to classify foolish people, and the CLs of 维生素 ‘vitamin’ can mostly be shared with the words of sub vitamin categories. It is also noticeable that these character-headed words are taking more CLs than the MAWs, probably because, for these particular cases, the emergence of the MAWs is later than their counterparts, and they also have a broader and stronger application in the Chinese language. Still, it does not affect MAWs' tendency toward traditional Chinese lexicon, considering the same set of CLs is involved in syntax.

**Table 7.** CL-distributions of MAWs without lettered-head and their canonical forms.

| Keyword | English | Frequency of CLs |
|---|---|---|
| 甲 B | First Division Group B | 轮 14，场 2，届 1 |
| 甲 B 联赛 | First Division Group B | 轮 22，场 5，届 3，个 1 |
| 甲 A | First Division Group A | 轮 57，届 23，场 7，支 1，局 1 |
| 甲 A 联赛 | First Division Group A | 轮 143，届 62，场 24，个 3，次 1 |

| | | |
|---|---|---|
| 傻B | silly person | 个177，群27，种11，帮11，名1 |
| 傻子 | silly person | 个7135，群271，种132，帮79，位26，批15，名8，波2，枚1，类1 |
| 维生素A | Vitamin A | 种158，粒11，颗2，个2 |
| 维生素C | Vitamin B | 片238，种149，粒46，颗12，款9，个2 |
| 维生素E | Vitamin C | 粒85，种67，片45，颗27，个4 |
| 维生素D | Vitamin D | 种53，粒3，款3，类2，片1 |
| 维生素B | Vitamin E | 种154，片104，粒9，颗6，类2 |
| 维C | short form of ‘Vitamin C’ | 片25，粒7，颗4，个4，种3，枚1 |
| 维E | short form of ‘Vitamin E’ | 颗7，粒7，片4，种1，款1 |
| 维生素 | Vitamin | 种8499，类555，粒93，款47，片42，颗39，个28，份8 |

The data results above may have pedagogical implications for L2 Chinese learning and teaching. While Chinese-English code-mixing is the inevitable result of the contact between Chinese and English [27], teachers should hold an open mind towards the code-mixed MAWs. From the learners’ side, on the other hand, in the sense that the CLs are highly dependent on the semantic meaning of the head morpheme, the L2 learners should be aware that when coming across the situation of a MAW co-occurring with a CL, it is not always acceptable to attach the general GE to it; instead, the CL choice should obey the same rule of other typical Chinese words.

## 6 Conclusion

Mandarin Alphabetical Word (MAW) is a unique lexicon in the Chinese language complex system where the Chinese language code-mixes with another foreign language at the lexical level. It is certain that when two languages come into contact, words and even syntactic structures of the foreign language will appear in the dominant language. What is still being determined is the extent to which the dominant language formally accepts the foreign language units. The MAW entries documented by official dictionaries are one hint; the other evidence is their similar linguistic behaviours in Chinese grammar. Thus, this paper has utilized a corpus-based method to empirically explore the lexical semantic status and language usage of MAWs from the perspective of their classifier selection tendency, a characteristic feature of the Chinese grammar system.

The corpus evidence indicated that MAWs with character-alphabet structure follow the traditional Chinese modifier-modified (head) morphological rule, and the classifiers are semantically motivated by the semantic feature of the head without overusing the general classifier 个 (GE). This phenomenon confirms the essential role of alphabets that are loan into the Chinese lexicon; in other words, people may accept MAWs, especially the alphabets, as the Chinese lexicon in a unique way that the distribution of CL collocations is comprehensive, productive, and dynamic.

This work may present certain limitations as a preliminary investigation on the opaque issue of classifier selection in MAWs. For example, the MAW seed words are limited in number due to the strict selection criteria in semantics, which leads to occasional sparseness in the corpus. Second, only the C-A type of MAWs has been studied, while another essential direction would be to probe into the A-C type, for instance, characters as head of nouns, or to investigate the pure alphabetical words' CL selections. Third, the manual checking of data examples can ensure a higher quality of the data, yet it may introduce bias to the data selection. Hence, future studies on this topic shall consider employing automatic metrics or tools in finding more seed words and their CL collocation samples, and we are also interested in applying the current line of research to a broader scope of MAW structures based on more prosperous language facts in corpus and with more reliable data.

## References

1. Muysken, P.: Two linguistic systems in contact: Grammar, phonology, and lexicon. The Handbook of Bilingualism and Multilingualism 193-216 (2013)
2. Grosjean, F.: Life with two languages: An introduction to bilingualism. Harvard University Press (1982).
3. Ritchie, W. C., Bhatia, T. K.: Social and psychological factors in language mixing. The Handbook of Bilingualism and Multilingualism 375-390 (2012)
4. Huang, C. R., H. Liu.: Jiyu Yuliaoku De Hanyu Zimuci Zidong Chouqu Yu Fenlei [Corpus-based Automatic Extraction and Analysis of Mandarin Alphabetic Words]. Yunnan Shifan Daxue Xuebao (Zhexue Shehui Kexue Ban) [Journal of Yunnan Normal University (Humanities and Social Sciences Edition)] (2017)
5. Xiang, R., Wan, M., Su, Q., Huang, C. R., Lu, Q.: Sina Mandarin Alphabetical Words: A Web-driven Code-mixing Lexical Resource. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing 833-842 (2020)
6. Su X., and X. Wu.: Zimucci De Shengmingli Yu Juxianxing Jianlun Xiandai Hanyu Cidian De Chuli Zimuci De Shenzhong Zuofa [Vitality and Limitation of Lettered Words -- the Discretion of lettered Words in Modern Chinese Dictionary]. Beihua Daxue Xuebao (Shehui Xueke Ban) [Journal of Beihua University (Social Sciences)] (2013)
7. Zhang, T. W.: Zimuci Shiyong Shi Yuyan Jiechu De Zhengchang Xianxiang [The Use of Chinese Lettered-words is a Normal Phenomenon of Language Contact]. Beihua Daxue Xuebao (Shehui Xueke Ban) [Journal of Beihua University (Social Sciences)] (2013)
8. Jiang, S., Dang, Y.: Zidong Tiqu Han Zimuciyu De Xinyu Xinshuyu Yanjiu [Research on Automatic Extraction of New Terms in the Field of Alphabetic Words]. Computer Engineering 47-49 (2007)
9. Zheng, Z.Z., Zhang, P., Yang, J. G.: Research on the Automatic Extraction of Letter Words Based on Corpus. Journal of Chinese Information Processing 78-85 (2005)

10. Kozha, K.: Chinese via English: A case study of "lettered-words" as a way of integration into global communication. In Chinese Under Globalization: Emerging Trends in Language Use in China 105-125 (2012)
11. Miao, R.: Loanword adaptation in Mandarin Chinese: Perceptual, phonological and sociolinguistic factors (Doctoral dissertation) (2005)
12. Ding, H., Zhang, Y., Liu, H., Huang, C. R.: A Preliminary Phonetic Investigation of Alphabetic Words in Mandarin Chinese. In Interspeech 3028-3032 (2017)
13. Li, X. H.: Zai Tang Zimuci De Duyin Wenti [Talk about the pronunciation of letter words]. Yuyan Wenzi Yingyong [Applied Linguistics] (2002)
14. Riha, H.: Lettered Words in Chinese: Roman Letters as Morpheme-Syllables. Columbus: The Ohio State University 93-99 (2010)
15. Riha, H., Baker, K.: Lettered Words: Using Roman Letters to Create Words in Chinese (2010)
16. Allan K.: Classifiers. Language (1977)
17. Tai, J.H.Y.: Chinese classifier systems and human categorization. In honor of William S.-Y. Wang: Interdisciplinary studies on language and language change 479-494 (1994)
18. Ahrens, K., Huang, C.-R: Classifiers. In C.-R. Huang., D. X. Shi (Eds.): A reference grammar of Chinese. United Kingdom: Cambridge University Press 169-198 (2016)
19. Wang, S., Huang, C. R.: Towards an event-based classification system for non-natural kind nouns. Chinese Lexical Semantics: 13th Workshop, CLSW 2012, Wuhan, China, July 6-8, 2012, Revised Selected Papers 13. Springer Berlin Heidelberg 381-395 (2013)
20. Chen, K. J., Huang, C. R., Chang, L. P., Hsu, H. L.: Sinica corpus: Design methodology for balanced corpora. In Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation 167-176 (1996)
21. Huang, C.R.: Tagged Chinese gigaword version 2.0, ldc2009t14. Linguistic Data Consortium (2009)
22. Liu, Y. Q.: Zimuci Cidian [The Dictionary of Lettered-words]. Shanghai Cishu Chubanshe [Shanghai Lexicographical Publishing House] (2001)
23. Jakubíček M., Kilgarriff A., Kovář V.: The TenTen corpus family. 7th International Corpus Linguistics Conference CL. Lancaster University 125-127 (2013)
24. Tai, J.H.Y: Variation in classifier systems across Chinese dialects: towards a cognition-based semantic approach. Chinese Language and Linguistics 587-608 (1992)
25. Croft, W.: Semantic universals in classifier systems. Word 145-171 (1994)
26. Kunilovskaya, M., Koviazina, M.: Sketch engine: A toolbox for linguistic discovery. Journal of Linguistics 503-507 (2017)
27. Lu, D. H.: Hanyu Zhong De Zimuci Yinyici He Hunhe Yuma [Lettered and transliterated words and code-mixing in Chinese]. Waiguoyu [Journal of Foreign Languages] 59 – 65 (2010)

## Appendix

| CL | Category | Interpretation/ translation | CL | Category | Interpretation/ translation |
|---|---|---|---|---|---|
| 个 | General | A general individual classifier | 群 | individual | to classify a swarm of people or animals |
| 位 | individual | to classify people with politeness | 帮 | individual | to classify a group of people |
| 批 | individual | to classify a group of objects or people | 名 | individual | to classify different identities of people |
| 枚 | individual | to classify circular objects | 粒 | individual | to classify small granular objects |
| 颗 | individual | to classify granular objects | 片 | individual | to classify flake-shaped objects |
| 项 | individual | to classify distributed items in group | 份 | individual | to classify distributed items in group |
| 支 | individual | to classify long rod-shaped objects, also groups or teams | 串 | individual | to classify items in clusters |
| 排 | individual | to classify entities in the form of rows | 条 | individual | to classify long thin and soft objects |
| 代 | individual | to classify objects of a certain generation | 层 | individual | to classify layers of substance |
| 版 | individual | to classify the editions of a book or product | 件 | individual | to classify clothes, or general implements |
| 套 | individual | to classify a pack/set of artifacts | 身 | individual | to classify wearing clothes |
| 家 | individual | to classify families or institutions | 首 | individual | to classify poems, songs, or music |
| 曲 | individual | to classify songs or music | 部 | individual | to classify machines, automobiles, volumes of books or movies |
| 台 | individual | to classify machines or large appliances | 袭 | individual | to classify clothes especially long dresses |
| 波 | event | to classify series of actions | 轮 | event | to classify sequences of events |
| 场 | event | to classify scheduled events | 届 | event | to classify events held on a regular basis |
| 局 | event | to classify the occurrence of games | 期 | event | to classify events that involve stages of completion |
| 番 | event | to denote times of a repeated event | 种 | kind | to classify kinds of entities |
| 类 | kind | to classify entities in categories | 款 | kind | to classify categories of manufactured products |