# Homophone Density and Word Length in Chinese

San Duanmu and Yan Dong

Abstract

Chinese has many disyllabic words, such as 煤炭 *meitan* ‘coal’, 老虎 *laohu* ‘tiger’, and 学习 *xuexi* ‘study’, most of which can be monosyllabic, too, such as 煤 *mei* ‘coal’, 虎 *hu* ‘tiger’, and 学 *xue* ‘study’. A popular explanation for such length pairs is homophone avoidance, where disyllabic forms are created to avoid homophony among monosyllables. In this study we offer a critical review of arguments for the popular view, which often rely on cross-linguistic comparisons. Then we offer a quantitative analysis of word length pairs in the Chinese lexicon in order to find out whether there is internal evidence for a correlation between homophony and word length. Our study finds no evidence for the correlation. Instead, the percentage of disyllabic words is fairly constant across all degrees of homophony. Our study calls for a reconsideration of the popular view. Alternative explanations are briefly discussed.

## 1. The issue: elastic word length in Chinese

Chinese has many synonymous pairs of word forms, one short and one long. The short one is monosyllabic and the long one is made of the short one plus another morpheme. Since the long form has the same meaning as the short (evidenced by their mutual annotation in the dictionary), the meaning of its extra morpheme is either redundant or lost. Some examples are shown in (1), where Chinese data are transcribed in Pinyin spelling (tones are omitted unless relevant). In each case, the extra part of the long form is shown in parentheses. For ease of reading, a hyphen is added between morphemes.

(1) Words of elastic length in Chinese

| Pinyin | Character | Literal | Gloss |
|---|---|---|---|
| *mei-(tan)* | 煤(炭) | coal-(charcoal) | ‘coal’ |
| *(lao)-hu* | (老)虎 | (old)-tiger | ‘tiger’ |
| *xue-(xi)* | 学(习) | study-(practice) | ‘study’ |
| *ji-(shu)* | 技(术) | skill-(technique) | ‘skill’ |
| *(shang)-dian* | (商)店 | (business)-store | ‘store’ |

In ‘coal’, the original meaning of 炭 *tan* ‘charcoal’ is absent, because 煤炭 *mei-tan* simply means ‘coal’, not ‘coal and charcoal’. Similarly, in ‘tiger’, the original meaning of 老 *lao* ‘old’ is absent, because 老虎 *lao-hu* simply means ‘tiger’, not ‘old tiger’. To say ‘old tiger’, one must add another 老 *lao* ‘old’, i.e., 老老虎 *lao lao-hu*; similarly, ‘young tiger’ is 小老虎 *xiao lao-hu*, where 小 *xiao* means ‘young’. In ‘study’, the meaning of 习 *xi* ‘practice’ is redundant, because studying presumably involves practice. In ‘skill’, both parts of the long form have their own meanings, but the meanings are repetitive, which means that the second is redundant. Finally, in ‘store’, 商 *shang* ‘business’ has its own meaning, too, but it is again redundant, because stores are for business.

The long form looks like a compound and has been so called in some studies (e.g. Karlgren 1918; Jespersen 1930; Mullie 1932; Chao 1948; Sproat and Shih 1996). On the other hand, since the long and short forms have the same meaning, some linguists consider them to be variants of the same word and such words are said to have ‘elastic length’ (Guo 1938; Pan 1997; Huang and Duanmu 2013).

The existence of such length pairs is well known. For example, Karlgren (1918: 50-51) observes that, although most Chinese words are monosyllabic, many also have a corresponding long form, which he calls ‘synonym compound’ (synonymkomposita) or ‘elucidative compound’ (förtydligande sammansättningar). Similarly, Chao (1948: 33) observes that, while Chinese morphemes are dominantly monosyllabic, Chinese words as they are used are usually polysyllabic (Chao considers the former units to be words in the Chinese sense and the latter units to be words in the English sense). Moreover, Pan (1997: 140) suggests that nearly all Chinese words have elastic length.

However, the explanation for the property is not obvious. In the orthodox view, the property is the result of homophone avoidance: since most Chinese morphemes are monosyllabic, there are too many homophones, and disyllabic forms are created to avoid ambiguity. In support of the view, cross-linguistic evidence has been offered. We review arguments for the orthodox view and show that the evidence is limited and inconclusive. We then offer an analysis of language-internal data and show that the orthodox view makes wrong predictions. Finally, we offer a brief discussion of alternative motivations for the creation of word length pairs.

## 2. The orthodox view: homophone avoidance

The abundant use of disyllabic words in Chinese was observed by Karlgren (1918), who also proposed an explanation for it. His view has been echoed by many others (Jespersen 1930; Guo 1938; Lü 1963; T'sou 1976; Li and Thompson 1981; Ke 2006; Jin 2011; Nettle 1995; 1999) and has been called the orthodox view (Kennedy 1955).

In the orthodox view, the syllable inventory of modern Chinese is too small, especially for a language whose morphemes are mostly monosyllabic. In particular, Middle Chinese (about AD 600) used to have over 3,000 distinct syllables (including tonal contrasts), whereas modern Standard Chinese has just 1,300. In comparison, English has many times more syllables (estimated to be 158,000 by Jespersen 1930: 347). As a result, Chinese has many homophones. To avoid ambiguity in speech, 'elucidative compounds' are created, such as 看见 *kan-jian* 'look-see' for 看 *kan* 'see', 技术 *ji-shu* 'skill-technique' for 技 *ji* 'skill', and 学习 *xue-xi* 'study-practice' for 学 *xue* 'study'. A comparable example in English would be for those who have the same pronunciation for *pen* and *pin* to use *ink-pen* for the former and *thumb-pin* for the latter.

Some proponents of the orthodox view (e.g. Karlgren 1918; Guo 1938; Lü 1963) offer little quantitative evidence; instead, they often rely on the reader's positive response to what seems to be a plausible idea. Other proponents have offered quantitative evidence from cross-linguistic data. For example, Jespersen (1930) notes a striking difference between the number of possible syllables in English and that in Chinese and concludes that Chinese must have a much higher degree of homophony than English. Similarly, T'sou (1976) examines syllable inventories in Cantonese and Mandarin and the amount of disyllabic words they use. He reports that (i) Cantonese has more syllables than Mandarin and (ii) Cantonese speakers use fewer disyllabic words than Mandarin speakers. He concludes that the result supports of the orthodox view. Moreover, Ke (2006) examines 20 dialects of Chinese and reports a 'correlation between the degree of homophony and the degree of disyllabification'. Finally, Jin (2011) compares Mandarin, Cantonese, English, and Japanese and argues that there is a constant relation between the size of the syllable inventory and the percentage of monosyllabic words, regardless of the language (i.e. S/M = C, where S is the number of distinct syllables of a language, M the percentage of monosyllabic words in that language, and C a near constant).

Central to the orthodox view is the prediction that there is a correlation between homophone density (i.e. the degree of homophony) and word length, which we state in (2), where homophone density is measured in terms of how many homophones a monosyllabic word or morpheme has.

(2) Prediction of the orthodox view (homophone avoidance):
There is a positive correlation between homophone density and the percentage of disyllabic words.

In the next section, we examine previous arguments for the prediction and see whether it is adequately supported by quantitative data.

## 3. Problems in the orthodox view

In this section, we consider three quantitative studies in support of the orthodox view: T’sou (1976), Ke (2006), and Jin (2011).

T’sou (1976) compares Mandarin (Standard Chinese) and Cantonese. First, he observes that, excluding tones, Mandarin has about 400 different syllables whereas Cantonese has 700. Then he offers an experiment in which Mandarin and Cantonese speakers were asked to tell two short stories. Percentages of disyllabic words were then counted, and it was found that Cantonese speakers used fewer disyllabic words than Mandarin speakers. Consider the data from one of the stories, shown in Table 1.

| | | Type/Token | Poly type% | Poly token% |
|---|---|---|---|---|
| **Mandarin** | Extended | 1:3.2 | 28.9% | 14.7% |
| | Standard | 1:2.8 | 27 % | 16.8% |
| **Cantonese** | Extended | 1:3.2 | 20.3% | 11.5% |
| | Standard | 1:2.7 | 18.7% | 12.9% |

Table 1: Percentages of polysyllabic words in Mandarin and Cantonese, in type and token counts, in the story *The boy who cried ‘Wolf!’* (T’sou 1976: 82). ‘Extended’ refers to speakers who told the story with elaboration. ‘Standard’ refers to speakers who told the story at normal length.

More than one speaker was used in each dialect. The average percentages of polysyllabic words are clearly quite different between the two dialects. However, some questions remain. First, it was not reported how many speakers were used in each dialect. Second, there are no statistics on the variation among speakers of a dialect, or on whether the difference between the dialects is significant. Third, the data size is not reported, but it is likely to be rather small. For example, a typical Chinese version of the story in Table 1 is about 400 graphs (including punctuations), or about 200 words in English. Therefore, it is unclear whether such a data size can adequately reflect the overall difference between the two dialects. Fourth, word segmentation was probably done by the author himself, but the method was not reported. In particular, word segmentation is a notorious problem in Chinese and decisions on whether a disyllabic unit is a phrase (two words), a compound (possibly two words), or a single word are not always obvious. The author does give some examples, but they are not always clear. For example, Mandarin N+*er* (noun with the *–er* suffix) is counted as a disyllabic word (T'sou 1976: 77), but such forms are normally pronounced as a single syllable (Duanmu 2007). In summary, while T'sou (1976) is innovative in introducing quantitative argument for the orthodox view, its conclusion remains open, since its data size is small and its method is rather casual.

Ke (2006) is a more ambitious study. She deals with three issues: (i) the relation between phoneme inventory size and word length in general, (ii) the relation among syllable complexity, homophones, and disyllabic words in English, German, and Dutch; and (iii) the relation among syllable complexity, homophones, and disyllabic words in twenty Chinese dialects.

The discussion of (i) follows a proposal by Nettle (1995; 1999), which has two problems. First, Nettle uses UPSID (Maddieson and Precoda 1990) as the source for phoneme inventories, but phoneme inventories are notoriously inconsistent in UPSID (see, for example, Vaux 2009). Consider the African language !Xũ, cited in Nettle (1995), which is listed in UPSID as having 46 vowels. However, an inspection shows that over 30 of them are diphthongs or long vowels. If we follow a common practice in generative phonology, where diphthongs are combinations of two vowels each and a long vowel is a regular vowel linked to two timing slots, then !Xũ has only a fraction of the reported vowel number. Such uncertainties in the size of a phoneme inventory directly affect the accuracy of the predictions. For example, let us consider the analysis of Mandarin. According to Nettle (1999: 145), the relation between the average word length (L) and the phoneme inventory size (P) is L = 17.54 $P^{-0.30}$. In UPSID, Mandarin has 25 consonants and 7 vowels [i u y a ɤ a˞ ɚ]. According to Nettle (1995), the vowel inventory is the product of the number of vowels times the number of tones, 7 x 4 = 28 in Mandarin. Thus, Mandarin has 25 + 28 = 53 phonemes, and the predicted average word length is 5.33 segments (L = 17.54 $P^{-0.30}$ = 17.54 x $53^{-0.30}$ = 5.33). Now, it can be argued that Mandarin only has five vowels [i u y a ɤ], where [a˞ ɚ] are either marginal or derived (Duanmu 2007). In addition, Mandarin has just 19 consonants, where [j w ɥ] are positional variants of the high vowels [i u y], and [ȶɕ ȶɕʰ ɕ] are variants of the clusters [tsj tsʰj sj] (Duanmu 2007). If so, Mandarin has just 19 + 5 x 4 = 39 phonemes, and the predicted average word length should be 5.84 segments (L = 17.54 $P^{-0.30}$ = 17.54 x $39^{-0.30}$ = 5.84). Thus, there is almost a 10% error margin, which is quite large for a small list of 10 languages. The second problem in Nettle's analysis is how word length is determined. For Mandarin, Nettle samples 50 words from a dictionary. However, as we shall see below, if we exclude true compounds, such as 书房 *shu-fang* 'book-room (study)', 午饭 *wu-fan* 'noon-meal (lunch)', and 开会 *kai-hui* 'hold meeting', about 40% of Chinese words have elastic length, i.e. they can be either monosyllabic or disyllabic, such as 煤-煤炭 *mei-meitan* 'coal'. Clearly, there is a very large difference between counting the length of the short form and counting the length of the long form of such length pairs.

For (ii), Ke (2006: 140) assumes that all pronounced consonants are in a syllable. Therefore, the maximal possible syllable in English is thought to be CCCVCCCC. However, it has been shown that the VCCCC part is found only in word-final position; in medial positions, the English rime is limited to VV or VC (Borowsky 1986; Duanmu 2008). Therefore, some linguists have proposed that certain word-final consonants can be excluded from the final syllable, under the notion of 'extrametricality' (Hayes 1982). It has also been shown that the CCCV part is limited to word-initial position (Duanmu 2008). If we exclude extra consonants at word edges, then the basic English syllable size is CRVX (where R is an approximant), which is not very different from that of Chinese. Thus, her comparison between Chinese and Germanic languages needs to be reevaluated.

Ke's discussion of (iii) is more persuasive, because it is easier to compare syllables, homophones, and word length in Chinese dialects. Ke proposes two correlations, shown in (3).

(3) Two correlations found among 20 Chinese dialects (Ke 2006)
- a. There is a strong negative correlation between the number of distinct syllables and the degree of homophony.
- b. There is a strong positive correlation between the degree of homophony and the degree of disyllabification.

The correlation in (3a) is not controversial. Specifically, complete lists of syllables of Chinese dialects are available, and such lists indeed vary in size. For example, if we include tones, Shanghai has about 900 syllables (Xu and Tao 1997), Standard Chinese has 1,300 (Duanmu 2007), and Cantonese has 1,800 (Kao 1971). If we assume that Chinese dialects have a similar number of morphemes, most of which are monosyllabic, then the degree of homophony (or homophone density) will clearly differ from dialect to dialect.

The correlation in (3b) is less obvious though. Let us consider what is reported in Ke (2006), shown in Table 2.

| Dialect | Syllable | Homo% | Disyl%1 | Disyl%2 |
|---|---:|---:|---:|---:|
| Taiyuan | 828 | 0.70 | 0.60 | 0.40 |
| Wuhan | 870 | 0.72 | 0.62 | 0.40 |
| Chengdu | 938 | 0.70 | 0.62 | 0.42 |
| Yangzhou | 947 | 0.68 | 0.61 | 0.40 |
| Hefei | 976 | 0.68 | 0.61 | 0.40 |
| Changsha | 981 | 0.67 | 0.62 | 0.41 |
| Suzhou | 999 | 0.64 | 0.61 | 0.40 |
| Shuangfeng | 1,001 | 0.67 | 0.63 | 0.43 |
| Wenzhou | 1,048 | 0.65 | 0.53 | 0.31 |
| Ji’nan | 1,063 | 0.69 | 0.59 | 0.36 |
| Xi’an | 1,084 | 0.69 | 0.61 | 0.41 |
| Nanchang | 1,111 | 0.66 | 0.60 | 0.38 |
| Beijing | 1,125 | 0.67 | 0.62 | 0.41 |
| Jian’ou | 1,241 | 0.63 | 0.55 | 0.31 |
| Meixian | 1,304 | 0.60 | 0.60 | 0.39 |

| | | | | |
|---|---|---|---|---|
| Yangjiang | 1,319 | 0.61 | 0.51 | 0.24 |
| Guangzhou | 1,367 | 0.59 | 0.50 | 0.24 |
| Fuzhou | 1,413 | 0.61 | 0.51 | 0.25 |
| Chaozhou | 1,759 | 0.52 | 0.50 | 0.23 |
| Xiamen | 1,855 | 0.54 | 0.54 | 0.29 |

Table 2: Syllable counts, percentages of syllables with homophones, and percentages of disyllabic words (counted in two ways) in 20 Chinese dialects (Ke 2006: 150). Disyllabic monomorphemes are counted as disyllabic words in Disyl%1 and as monosyllabic words in Disyl%2.

The correlation between degree of homophony and the degree of disyllabification is found to be statistically significant, even though there is some variation among the dialects and the differences seem small. Still, there are some questions.

First, the number of syllables and the percentage of homophones are mostly based on Peking University (1989), which consists of some 3,000 common characters, whereas the percentage of disyllabic expressions is based on Peking University (1995), which consists of 905 common lexical entries. Therefore, the data sets are not only different but are rather small. For example, in *Modern Chinese Dictionary* (XDHYCD 2005), there are 10,000 characters and over 60,000 entries. Second, many of the 905 entries in Peking University (1995) are true compounds, such as 猪肉 *zhu rou* ‘pig meat (pork)’, 素菜 *su cai* ‘vegetarian dish’, 开水 *kai shui* ‘boiled water’, 午饭 *wu fan* ‘noon meal (lunch)’, 电筒 *dian tong* ‘electric tube (flashlight)’, 阴天 *yin tian* ‘cloudy day’, and 明年 *ming nian* ‘next year’. Such compounds have to be at least disyllabic in most dialects, regardless of the degree of homophony. Therefore, it would be better to exclude them from calculation. Third, of the 20 dialects in Table 2, only 16 are from Peking University (1995), and two dialects in Peking University (1995) are not found in Table 2. It is not explained where the disyllabic data for the 4 added dialects come from, nor why the data of two dialects are left out. It is worth noting, too, that of the four added dialects, two have the fewest number of syllables (Taiyuan and Wuhan), which could have changed the statistics. Fourth, the 905 lexical entries in Peking University (1995) are based on Standard Chinese. It is not entirely clear whether they fairly represent the basic vocabulary of other dialects. Fifth, as just mentioned, Peking University (1995) contains 905 lexical entries, yet in Ke's calculation, there are 1,236 entries (Ke 2006: 149). It is not explained where the 300 extra words come from.

Next we consider Jin (2011), who proposes a linear relation between the size of the syllable inventory and the percentage of monosyllabic words, regardless of the language. This is shown in (4).

(4) The relation between syllable inventory size (S) and the percentage of monosyllabic words (M), regardless of the language (Jin 2011):
$S/M \approx C$, where C a constant

Jin illustrates her proposal with data from Mandarin, Cantonese, English, and Japanese. For example, she argues that $S_{Can}/M_{Can} \approx S_{Eng}/M_{Eng}$ (Jin 2011: 49), where $S_{Can}$ is the number of syllables in Cantonese, $M_{Can}$ is the percentage of monosyllabic words in Cantonese, $S_{Eng}$ is the number of syllables in English, and $M_{Eng}$ is the percentage of monosyllabic words in English.

Once again, several questions can be raised. First, as in Ke (2006), the English analysis is problematic, because Jin (2011) also assumes that all consonants are syllabified. For example, the largest English rime is thought to be VCCCCC, as in *angsts* (Jin 2011: 40). However, as mentioned above, this view of the English syllable is controversial. Second, Japanese has very different morphology from Chinese; the former has regular suffixes whereas the latter does not. Therefore, their word lengths should not be directly compared, because required inflections (rather than homophony) would lengthen many words in Japanese. Third, the data for Mandarin are based on materials that cover a wide range of subjects. In contrast, the data for Cantonese are based on two textbooks on the Cantonese language. Therefore, the data for the two dialects are not parallel in content or style; this is problematic, because the average word length can vary a lot from style to style. Fourth, the difference in the percentage of monosyllabic words is not always very large between Mandarin and Cantonese. This can be seen in Table 3, where the Mandarin data are based on 3,000 most frequent words in a corpus of over 200 million characters, and the Cantonese data are based on two textbooks on Cantonese.

| Language | Source | Words | Mono | Mono% |
|---|---|---|---|---|
| Mandarin | High frequency words | 3,000 | 1,000 | 33.3 |
| Cantonese | Textbooks on Cantonese | 2,291 | 796 | 34.7 |

Table 3: Percentages of monosyllabic words in the basic lexicons of Mandarin and Cantonese (Jin 2011: 38).

Jin (2011: 39) argues that the difference between Mandarin and Cantonese can be made larger (25.5% in Mandarin and 31.4% in Cantonese) if we only calculate nouns, verbs, adjectives, and adverbs, without function words. It is not obvious whether such manipulations are justified.

In summary, quantitative evidence for the orthodox view has been inconclusive. Besides, there are two other problems. First, as Chao (1948: 34) points out, ambiguity rarely arises in context, despite the presence of homophones. For example, there is hardly a context where the English words *son* and *sun* would cause ambiguity. Therefore, it remains to be shown how much homophony would start forcing a language to introduce longer words. Second, according to Karlgren (1923), 'elucidative compounds' are used in speech only, not in writing, because Chinese has enough distinct written graphs and one rarely encounters ambiguity in reading. But if elucidative compounds are not recorded in writing, it is hardly possible to verify a fundamental claim of the orthodox view, i.e. the increase of disyllabic words (or 'elucidative compounds') in spoken Chinese is a modern phenomenon, as the result of the loss of syllable contrasts. We shall return to this point.

## 4. Method

If homophony can motivate the creation of disyllabic words, there ought to be language-internal evidence. Specifically, morphemes with few homophones should have a low percentage of disyllabic counterparts, and morphemes with many homophones should have a high percentage of disyllabic counterparts. In this study, we examine such evidence.

We shall focus on lexical data, which contain information on both homophony and word length. In addition, we focus on Standard Chinese, for which there is a large amount of high-quality data. For example, *Modern Chinese Dictionary* (XDHYCD 2005, hereafter MCD) has over

65,000 entries. In contrast, a typical dialectal dictionary normally has only 8,000 entries (Li 2002). In this study, we conduct a complete examination of all monomorphemic units in MCD. The information of interest is shown in (5).

(5) Information of interest for the present study
  a. The part-of-speech (POS) of a morpheme
  b. The number of homophones of a morpheme
  c. Whether a monosyllabic morpheme has a disyllabic form (i.e. elastic length)

Such information is not always readily available in MCD and requires much manual annotation. Let us consider some details.

Elastic length pairs, such as 煤 *mei* and 煤炭 *meitan* for 'coal', have often been cited in the literature, and some linguists believe that most Chinese words have the property (e.g. Pan 1997). However, there is no dictionary that lists such pairs. In our annotation, we rely on two sources. First, if a monosyllabic morpheme in MCD is defined by a disyllabic one (or vice versa), we consider them to be a length pair (see examples below). Second, if no length pair is offered in MCD, we consult at least three native speakers to see if a length pair is available. In most cases, the native speakers agree on the solution (see Huang and Duanmu 2013 for more discussion of the method).

To obtain POS information, we need to distinguish three kinds of items: orthographic shapes (graphs), entries, and senses. A graph can have one or more entries depending on meaning or pronunciation. Entries differ from each other in meaning (such as financial *bank* vs. river *bank* in English) or in etymological origin (such as *petrol* vs. *gas*). Each entry in turn is divided into senses, which differ in shades of meaning or in POS categories. A graph from MCD is shown in (6), where the digit after the spelling of an entry indicates tone, angle brackets indicate POS, and […] indicates sample usage, which we omit. We also ignore the literal translation of the extra syllable of the long form, since its meaning is irrelevant.

(6) Entries and senses for the graph 供 in MCD
  Entry 1: 供 *gong1* (basic meaning: 'supply')
  a. <verb> 供给 *gongji*; 供应 *gongying* 'supply': […]
  b. <verb> 提供 *tigong* 'provide (conditions for use)': […]
  c. <noun> surname

Entry 2: 供 *gong4* (basic meaning: ‘offer’)

a. <verb> 供奉 *gongfeng* ‘offer (to the sacred or deceased)’: […]

b. <noun> 供品 *gongpin* ‘offering on display (to the sacred or deceased)’: […]

Entry 3: 供 *gong4* (basic meaning: ‘confess’)

a. <verb> ‘confess (by the accused)’: […]

b. <noun> 口供 *kougong*; 供词 *gongci* ‘confession’: […]

The graph has three entries and two pronunciations, indicated by Pinyin spelling and tone. The pronunciation of the first entry is *gong1*, which has tone 1 and three senses. The first sense has two disyllabic forms, both of which satisfy the criteria for elastic length. The second sense has one disyllabic form. The third sense is a surname and has no disyllabic form. The pronunciation of the second entry is *gong4*, which has tone 4 and two senses, each having a disyllabic form. The pronunciation of the third entry is also *gong4*, which has two senses. The first sense is not listed with a disyllabic form, while the second sense is listed with two disyllabic forms. In our annotation, when a sense has two (or more) disyllabic forms, we normally choose the first one. When a disyllabic form is not listed for a sense, we check whether an unlisted one is available, to be confirmed by three native speakers. For example, for the first sense of entry 3, there is a disyllabic form 供认 *gongren* ‘confess’, which we shall add.

Let us consider another graph 学, which has one entry *xue2* (with tone 2) and six senses, shown in (7).

(7) Senses of the entry 学 *xue2*

a. <verb> 学习 *xuexi* ‘study’: […]

b. <verb> ‘imitate’: […]

c. 学问 *xuewen* ‘knowledge’: […]

d. 学科 *xueke* ‘discipline’: […]

e. 学校 *xuexiao* ‘school’: […]

f. <noun> surname

Four of the senses have a disyllabic form and two do not. In addition, three of the senses have POS annotation and three do not. Based on their meanings and the examples, it is easy to see that the three senses without POS annotation are all nouns, which we add accordingly.

MCD contains a total of 65,381 entries. Their basic information is summarized in Table 4 and Table 5.

| | |
|---|---:|
| Regular | 64,687 |
| Rare graph | 626 |
| ‘Archaic’ | 68 |
| Total | 65,381 |

Table 4: Entry types and counts in MCD.

| Length | Count | % | Senses |
|---|---:|---:|---:|
| 1 | 10,244 | 16% | 2.1 |
| 2 | 42,163 | 65% | 1.2 |
| 3 | 5,977 | 9% | 1.1 |
| 4 | 5,761 | 9% | 1.0 |
| 5+ | 542 | 1% | 1.0 |
| All | 64,687 | 100% | 1.3 |

Table 5: Distribution of regular entries by length (in syllables), along with the average number of senses for each length.

The type ‘rare graph’ refers to those not available in our word processor; most of them are found in written or dialectal vocabulary. The type ‘archaic’ refers to historical forms that have been replaced by modern ones. These two types do not include many members and are excluded in the discussion below.

Regular entries can be divided into several cases, depending on their length and content. This is shown in Table 6, where we use ‘simple word’ to refer to a monomorphemic unit.

| Length | Content | Example (tones omitted) |
|---|---|---|
| 1 | Simple word | 煤 *mei* 'coal' |
| 1 | Pointer | 咖 *ka* see 咖啡 *kafei* 'coffee' |
| 1 | Pointer | 啡 *fei* see 咖啡 *kafei* 'coffee' |
| 2+ | Compound | 草帽 *caomao* 'straw-hat' |
| 2+ | Pseudo-comp. | 煤炭 *meitan* 'coal-(charcoal)' |
| 2+ | Idiom; set phrase | 牛头马面 *niutou mamian* 'cattle-head horse-face (ugly looking thugs)' |
| 2+ | Simple word | 咖啡 *kafei* 'coffee' |
| 2+ | Simple word | 儒艮 *rugen* 'dugong' |

Table 6: Types of regular entries according to length and content, where a 'simple word' is a monomorphemic word.

Most monosyllabic entries are simple words. A monosyllabic entry can also be a pointer, which is not a word by itself but points to a polysyllabic simple word, of which the pointer is a part. For example, both parts of 咖啡 *kafei* 'coffee' point to the disyllabic entry, because neither part is a word by itself. Most polysyllabic entries are compounds, such as 草帽 *caomao* 'straw-hat' or 有名 *you-ming* 'have-name (famous)'. Some polysyllabic entries look like compounds but are in fact long forms of monosyllabic words, which we can call pseudo-compounds. For example, 煤炭 *mei-tan* 'coal-(charcoal)' is the long form of 煤 *mei* 'coal' and 学习 *xue-xi* 'study-(practice)' is the long form of 学 *xue* 'study'. Such long forms are already covered when we annotate the length of the short forms. Some polysyllabic entries are idioms or set phrases. A polysyllabic entry can also be a simple word, pointed to by a monosyllabic entry, such as 'coffee'. Finally, some polysyllabic entries are simple words and not pointed to by a monosyllabic one, because all of its parts are independent words. For example, the first part of 'dugong' 儒 *ru* is the word for 'Confucianism' and the second part 艮 *gen* is a surname, and neither part points to 'dugong'.

We shall focus on simple words, because compounds, pseudo-compounds, and idioms are made of simple words. Once we know the length properties of simple words, we can understand those of compounds. As seen above, simple words occur in three places: (a) monosyllabic entries, (b) polysyllabic entries pointed to by a monosyllabic entry, and (c) polysyllabic entries not pointed to by a monosyllabic entry. Therefore, we adopt the procedure in (8).

(8) Procedure
a. Annotate all monosyllabic entries.
b. Annotate all polysyllabic entries pointed to by a monosyllabic entry.
c. Examine polysyllabic entries for other simple words not yet covered.

For step (8c), we sample 1,000 randomly selected polysyllabic entries and obtain the result in Table 7.

| Type | Count |
|---|---|
| Same as | 9 |
| Complex | 870 |
| Repeat | 113 |
| Simple word | 8 |
| All | 1,000 |

Table 7: Types of polysyllabic entries in 1,000 randomly selected samples. See text for explanations of the types.

The type 'same as' refers to an orthographic alternative to a recommended standard. The type 'complex' refers to a compound, a set phrase, or an idiom. The type 'repeat' refers to a simple word already covered, i.e. it is either pointed to by a monosyllabic entry or it is the long form of a monosyllabic word. Finally, out of the 1,000 samples, there are eight simple words not yet covered. The eight simple words are shown in Table 8, none of which belongs to a set that has many members. For example, while MCD includes about 2,000 surnames, most of them are monosyllabic and less than 10% are disyllabic.

| | In sample | In all | Example (tones omitted) |
|---|---|---|---|
| Onomatopoeia | 3 | 163 | 烘烘 *honghong* (sound of flame) |
| Name (noun) | 2 | 109 | 赫连 *Helian* (surname) |
| Loan (noun) | 3 | 163 | 儒艮 *rugen* ‘dugong’ |
| All | 8 | 435 | |

Table 8: Polysyllabic simple words not covered or pointed to by a monosyllabic entry. Their numbers in all polysyllabic entries are estimated from 1,000 samples.

Let us now consider annotation. Because we are interested in POS, and because POS annotation is based on senses, we must annotate senses, not entries. The contents of annotation are shown in (9) and (10).

(9) Contents of annotation for each sense
- a. POS (part-of-speech)
- b. Source of POS (original in MCD or added)
- c. Word length property
- d. Source of word length property (original in MCD or added)
- e. Style

(10) Sample ‘style’ labels

| Label | Example |
|---|---|
| Written | 藟 *lei* ‘vine’ |
| Loan | 咖啡 *kafei* ‘coffee’ |
| Dialect | 伲 *ni* ‘I/we’ |
| Surname | 王 *Wang* ‘Wang’ |
| Same as | 化 *hua* same as 花 ‘spend’ |
| See | 榈 *lü* see 棕榈 ‘palm’ |

Some style labels are given in MCD. Most labels are self-explanatory. The label ‘same as’ means the item is an alternative to a recommended (orthographic) form. The label ‘see’ is a pointer to another entry; it is used under the second syllable of a disyllabic simple word, which is listed under the first syllable.

In Table 9 we show the annotation of two sample entries, where ‘L-S’ to the source of ‘Length’ and style is omitted.

| Entry | Sense | Length | L-S | POS | Meaning |
|---|---|---|---|---|---|
| 学 *xue* | 1 | 学习<br>*xuexi* | MCD | V, MCD | ‘study’ |
| | 2 | 1 | MCD | V, MCD | ‘imitate’ |
| | 3 | 学问<br>*xuewen* | MCD | N, added | ‘knowledge’ |
| | 4 | 学科<br>*xueke* | MCD | N, added | ‘discipline’ |
| | 5 | 学校<br>*xuexiao* | MCD | N, added | ‘school’ |
| | 6 | 1 | MCD | N, MCD | surname |
| 儒艮<br>*rugen* | 1 | 儒艮<br>*rugen* | MCD | N, MCD | ‘dugong’ |

Table 9: Sample annotation of two entries, where ‘L-S’ refers to the source of ‘Length’.

By comparing the ‘Entry’ and ‘Length’ columns, we can obtain three kinds of length categories, shown in Table 10.

| Category | Entry | Length | Example |
|---|---|---|---|
| Mono-only | Mono. | Mono. | Sense 2 of 学 *xue* |
| Poly-only | Poly. | Poly. | Sense 1 of 儒艮 *rugen* |
| Elastic | Mono. | Poly. | Sense 1 of 学 *xue* |

Table 10: Three length categories, illustrated by examples from Table 6.

## 5. Results and discussion

We first report an overview of the results. Then we report our findings on the relation between homophone density and word length.

## 5.1. Overall statistics

We begin with the total number of entries and senses. This is shown in Table 11, where ‘pointed’ refers to polysyllabic simple words pointed to by a monosyllabic entry.

| | Entries | Senses |
|---|---|---|
| Monosyllabic | 10,243 | 20,533 |
| Pointed | 255 | 323 |
| Polysyllabic | 435 | 435 |
| All | 10,933 | 21,291 |

Table 11: Entries and senses of simple words in MCD. ‘Monosyllabic’ and ‘pointed’ entries and senses are all individually annotated. ‘Polysyllabic’ entries and senses are based on sampling, discussed earlier.

Next we consider the sources of annotation, excluding 1,404 senses that are orthographic alternatives to other forms or pointers to other entries or senses. The result is shown in Table 12.

| | Length | POS |
|---|---|---|
| Original | 16,133 | 11,418 |
| Added | 3,754 | 8,469 |
| All senses | 19,887 | 19,887 |

Table 12: Senses and sources of annotation, excluding 1,404 senses that are orthographic alternative forms, or pointers to other entries or senses.

It can be seen that most length information is original. In addition, nearly half of POS information is added.

Next we consider the distribution of POS categories and their length properties. MCD uses twelve POS categories for words, which are noun, verb, adjective, adverb, measure, mood, pronoun, preposition, interjection, numeral, conjunction, and onomatopoeia. In addition, there is a category called ‘affix’. The result is shown in Table 13.

| POS | Count | POS % | Mono % | Poly % | Elastic % |
|---|---|---|---|---|---|
| Noun | 9,559 | 48.1% | 52.5% | 8.9% | 38.6% |
| Verb | 5,904 | 29.7% | 56.1% | 1.8% | 42.1% |
| Adj. | 2,709 | 13.6% | 53.3% | 9.4% | 37.2% |
| Adverb | 429 | 2.2% | 72.5% | 0.2% | 27.3% |
| Measure | 411 | 2.1% | 91.0% | 0.5% | 8.5% |
| Onom. | 291 | 1.5% | 18.6% | 74.6% | 6.9% |
| Mood | 121 | 0.6% | 96.7% | 1.7% | 1.7% |
| Pronoun | 116 | 0.6% | 90.5% | 6.9% | 2.6% |
| Prep. | 103 | 0.5% | 97.1% | 0.0% | 2.9% |
| Interj. | 79 | 0.4% | 97.5% | 2.5% | 0.0% |
| Conj. | 69 | 0.3% | 63.8% | 0.0% | 36.2% |
| Num. | 64 | 0.3% | 96.9% | 0.0% | 3.1% |
| Affix | 32 | 0.2% | 100.0% | 0.0% | 0.0% |
| All | 19,887 | 100% | 55.6% | 7.2% | 37.2% |

Table 13: POS counts, POS percentages (POS %), and percentages of words that are monosyllabic only (Mono %), polysyllabic only (Poly %), and elastic in length (Elastic %), in all senses in MCD.

It can be seen that, or 19,887 senses, only 32 are affixes. In fact, the number of affixes is likely to be much smaller, because some ‘affixes’ can be seen as locative nouns, such as 边 *bian* ‘side’ and 面 *mian* ‘face’, and some occur as the semantically empty part in the long form of a length pair, such as 老 *lao* in (老)虎 *(lao)-hu* ‘(old)-tiger’, and 头 *tou* in 木(头) *mu-(tou)* ‘wood-(head)’.

Some mono-only nouns in Table 13 rarely occur as monosyllables. Instead, they are used in disyllabic forms. For example, family names, such as 王 *Wang*, are not used alone but usually occur in 老王 *lao Wang* ‘old Wang’ or 小王 *xiao Wang* ‘little Wang’. Similarly, mountain names, such as 华 *Hua* and 太 *Tai*, nearly always occur as 华山 *Hua Shan* ‘Hua Mountain’ and 泰山 *Tai Shan* ‘Hua Mountain’. We can refer to such words as ‘elastic in use’. In what follows we shall treat such words as having elastic length.

## 5.2. Homophone density and word length in nouns

In this section, we examine the correlation between homophony and word length. We shall focus on nouns. There are two reasons. First, nouns form the largest POS category and constitute nearly half of all word senses. Second, homophones within the same POS category are more likely to cause ambiguity, whereas homophones across POS categories are less likely to (T’sou 1976; Ke 2006). To simplify the matter, we exclude nouns that are polysyllabic-only and focus on monosyllabic nouns in order to see which of them have elastic length, and whether the presence of elastic length is correlated with homophone density.

There are 8,706 monosyllabic nouns, 65% of which have elastic length. The information on their homophone density and word length is given in Table 14 and the relevant statistics are shown in Table 15.

| Homo | 1-only | Elastic | All | Elastic% |
|---:|---:|---:|---:|---:|
| 1 | 75 | 92 | 167 | 55% |
| 2 | 91 | 125 | 216 | 58% |
| 3 | 101 | 199 | 300 | 66% |
| 4 | 101 | 219 | 320 | 68% |
| 5 | 139 | 226 | 365 | 62% |
| 6 | 129 | 267 | 396 | 67% |
| 7 | 147 | 252 | 399 | 63% |
| 8 | 139 | 237 | 376 | 63% |
| 9 | 141 | 264 | 405 | 65% |
| 10 | 87 | 203 | 290 | 70% |
| 11 | 115 | 248 | 363 | 68% |
| 12 | 132 | 240 | 372 | 65% |
| 13 | 133 | 270 | 403 | 67% |
| 14 | 133 | 203 | 336 | 60% |
| 15 | 105 | 225 | 330 | 68% |

| | | | | |
|---|---|---|---|---|
| 16 | 65 | 159 | 224 | 71% |
| 17 | 98 | 174 | 272 | 64% |
| 18 | 72 | 126 | 198 | 64% |
| 19 | 117 | 225 | 342 | 66% |
| 20 | 81 | 139 | 220 | 63% |
| 21 | 54 | 93 | 147 | 63% |
| 22 | 96 | 168 | 264 | 64% |
| 23 | 49 | 89 | 138 | 64% |
| 24 | 59 | 85 | 144 | 59% |
| 25 | 37 | 63 | 100 | 63% |
| 26 | 52 | 104 | 156 | 67% |
| 27 | 77 | 166 | 243 | 68% |
| 28 | 42 | 98 | 140 | 70% |
| 29 | 8 | 21 | 29 | 72% |
| 30 | 17 | 43 | 60 | 72% |
| 31 | 22 | 40 | 62 | 65% |
| 32 | 34 | 62 | 96 | 65% |
| 33 | 11 | 22 | 33 | 67% |
| 34 | 10 | 24 | 34 | 71% |
| 35 | 20 | 50 | 70 | 71% |
| 36 | 26 | 46 | 72 | 64% |
| 37 | 24 | 50 | 74 | 68% |
| 40 | 17 | 23 | 40 | 58% |
| 42 | 15 | 27 | 42 | 64% |
| 44 | 14 | 30 | 44 | 68% |
| 47 | 35 | 59 | 94 | 63% |
| 48 | 19 | 29 | 48 | 60% |
| 49 | 16 | 33 | 49 | 67% |
| 51 | 8 | 43 | 51 | 84% |
| 56 | 23 | 33 | 56 | 59% |

| | | | | |
|---|---|---|---|---|
| 62 | 30 | 32 | 62 | 52% |
| 64 | 19 | 45 | 64 | 70% |
| All | 3,035 | 5,671 | 8,706 | 65% |

Table 14: Homophone density (Homo) and word length of nouns in MCD, excluding nouns that are polysyllabic-only. Homophone density refers to the number of homophones a noun has among all nouns in Table 14, where 1 means no other homophone. Word length information includes the number of nouns that are monosyllabic-only (1-only), the number of nouns that have elastic length (Elastic), the sum of 1-only and Elastic (All), and the percentage of elastic nouns (Elastic %), for each level of homophone density. Elastic nouns include those that are ‘elastic in use’, such as family names and mountain names.

| | |
|---|---|
| Correlation: | 0.089 |
| Confidence interval (95%): | -0.066  0.122 |
| Multiple R-squared: | 0.008 |
| F-statistic: | 0.3628 on 1 and 45 degrees of freedom |
| P-value: | 0.55 |

Table 15: Statistics on the data in Table 14, which show no correlation between homophone density and word length.

As can be seen in Table 15, the correlation value is small, the 95% confidence interval for the correlation crosses zero, the R-squared value is very small, and the probability of the null hypothesis (i.e. there is no correlation between homophony and word length) is high ($p$ = 0.55). Therefore, there is no correlation between homophone density and word length.

The result just seen includes nouns of all styles, many of which have restricted use. Let us now consider regular nouns only, where we exclude personal names, chemical elements or compounds, abbreviated administrative units, and dialectal, written, or archaic vocabulary. The result is shown in Table 15 and Table 16.

| Homo | 1-only | Elastic | All | Elastic% |
|---:|---:|---:|---:|---:|
| 1 | 49 | 127 | 176 | 72% |
| 2 | 79 | 197 | 276 | 71% |
| 3 | 94 | 230 | 324 | 71% |
| 4 | 90 | 242 | 332 | 73% |
| 5 | 108 | 282 | 390 | 72% |
| 6 | 105 | 261 | 366 | 71% |
| 7 | 75 | 198 | 273 | 73% |
| 8 | 101 | 267 | 368 | 73% |
| 9 | 118 | 242 | 360 | 67% |
| 10 | 87 | 233 | 320 | 73% |
| 11 | 50 | 148 | 198 | 75% |
| 12 | 78 | 102 | 180 | 57% |
| 13 | 86 | 174 | 260 | 67% |
| 14 | 53 | 129 | 182 | 71% |
| 15 | 44 | 91 | 135 | 67% |
| 16 | 36 | 60 | 96 | 63% |
| 17 | 24 | 78 | 102 | 76% |
| 18 | 23 | 67 | 90 | 74% |
| 19 | 21 | 36 | 57 | 63% |
| 20 | 24 | 56 | 80 | 70% |
| 21 | 13 | 29 | 42 | 69% |
| 22 | 35 | 75 | 110 | 68% |
| 23 | 13 | 56 | 69 | 81% |
| 24 | 9 | 39 | 48 | 81% |
| 28 | 7 | 21 | 28 | 75% |
| 30 | 2 | 28 | 30 | 93% |
| 33 | 12 | 21 | 33 | 64% |
| 43 | 22 | 21 | 43 | 49% |
| All | 1,458 | 3,530 | 4,968 | 49% |

Table 16: Homophone density (Homo) and word length of regular nouns in MCD. Homophone density refers to the number of homophones a noun has among the nouns in Table 16, where 1 means no other homophone. Word length information includes the number of nouns that are monosyllabic-only (1-only), the number of nouns that have elastic length (Elastic), the sum of 1-only and Elastic (All), and the percentage of elastic nouns (Elastic %), for each level of homophone density. Nouns of special styles are excluded, such as personal names, chemical elements or compounds, abbreviated administrative units, and dialectal, written, or archaic vocabulary.

| | |
|---|---|
| Correlation: | -0.126 |
| Confidence interval (95%): | -0.41 0.21 |
| Multiple R-squared: | 0.016 |
| F-statistic: | 0.4167 on 1 and 26 degrees of freedom |
| P-value: | 0.54 |

Table 17: Statistics on the data in Table 16, which show no correlation between homophone density and word length.

We see again that there is no correlation between homophone density and word length among regular nouns either. It would be interesting to ask, as a reviewer suggests, whether word length is correlated with token frequencies. In particular, would more frequent nouns have a higher the percentage disyllabic (or elastic) forms? This would require a separate study though. We venture to guess, based on random examples, that there is no correlation between token frequency and word length. For example, some very frequent words remain monosyllabic (not elastic), such as 水 *shui* ‘water’, 人 *ren* ‘person’, 狗 dou ‘dog’, 猪 zhu ‘pig’, and 饭 *fan* ‘meal’.

## 6. Non-homophone factors that create word length pairs

We have seen that homophone avoidance cannot explain the presence of word length pairs in Chinese. If so, what other factors could there be that have created word length pairs in Chinese? We would like to point out two mechanisms here, truncation and prosody.

Truncation is a well-known process in language, where a long word is truncated to a short one. Some examples from English are shown in (11).

(11) Truncation in English

| Original | Truncated |
|---|---|
| professional | pro |
| Patrick | Pat |
| demonstration | demo |
| situation (comedy) | sit(com) |

Truncation creates length pairs directly. In some pairs, both length forms can be used as free words, such as *Patrick-Pat*, *professional-pro*, and *demonstration-demo*. In other pairs, the short form is used with another form only (often in a set expression), such as *sit* in *sitcom*. Truncation occurs in Chinese, too. Some examples are shown in (12), where truncated syllables and their literal meanings are shown in parentheses.

(12) Truncation in Chinese

| Form | Literal meaning | Gloss |
|---|---|---|
| 加(拿大) Jia(nada) | | ‘Canada’ |
| 中(学) zhong-(xue) | ‘middle (school)’ | ‘middle school’ |
| 小(学) xiao-(xue) | ‘small (school)’ | ‘elementary school’ |
| 轮(船) lun-(chuan) | ‘wheel (boat)’ | ‘powered ship’ |
| 电(视) dian-(shi) | ‘electricity (view)’ | ‘television’ |
| (眼)镜 (yan)-jing | ‘(eye) mirror | ‘eye glasses’ |

In each case, the long form is the original. There is little doubt about foreign names, such as 加拿大 *Jianada* ‘Canada’. There is little doubt either that we can see, from the literal meaning, that 中 *zhong* for ‘middle school’ is a truncated form of 中学 *zhong-xue*. Similarly, it is clear from the literal meaning that 轮船 *lun-chuan* ‘powered ship’ is the original form and 轮 *lun* is the truncated form. It is worth noting, too, that some truncated forms only occur in set phrases, such as 镜 *jing*, which is not used alone but occurs in 墨镜 *mo-jing* ‘ink-glass (sunglasses)’

Let us now consider the role of prosody. It is well known that babies (or their parents) prefer disyllabic words (the reason for which we do not pursue here). To satisfy the preference, a suffix-like final *–y* is often added to monosyllabic nouns, as seen in (13).

(13) Prosody motivated length pairs in English

| Short | Long |
|---|---|
| mom | mommy |
| pot | potty |

In contrast, *-y* is not added to disyllabic nouns, such as **sistery* (from *sister*) or **rabbity* (from *rabbit*). The prosodic preference for disyllabic words creates length pairs, such as *mom-mommy*, *pot-potty*, and *dog-doggy*.

Prosody exerts an effect in Chinese, too. As Guo (1938) summarizes, many Chinese scholars, as far back as SHEN Kuo (1031-1095) and ZHENG Qiao (1104-1162), have observed that disyllabic words are needed in certain positions and monosyllabic ones in others. For example, Guo (1938: 7) suggests that monosyllabic words are needed in positions where one needs to speak quickly, whereas disyllabic words are needed in positions where one needs to speak slowly. In the perspective of present-day phonology, it can be shown that disyllabic words are needed in prosodically strong positions and monosyllabic words cannot be used in such positions (Lu and Duanmu 2002; Duanmu 2007). Consider [N N] compounds where both nouns have elastic length. An example is shown in (14), where 1 indicates a monosyllabic form and 2 a disyllabic form.

(14) Length patterns in **[**N N] compounds

| Length | Characters | | Pinyin | |
|---|---|---|---|---|
| 2+2 | 煤炭 | 商店 | *meitan* | *shangdian* |
| 2+1 | 煤炭 | 店 | *meitan* | *dian* |
| *1+2 | 煤 | 商店 | *mei* | *shangdian* |
| 1+1 | 煤 | 店 | *mei* | *dian* |
| | | | ‘coal | store’ |

In [N N], 2+2, 2+1, and 1+1 are generally good while 1+2 is generally bad (Lü 1963; Lu and Duanmu 2002; Feng 1998; Duanmu 2007; Duanmu 2012; Huang and Duanmu 2013). The reason is phonological. Specifically, compound stress falls on the first N, and Foot Binarity requires it to be disyllabic (as in 2+2 or 2+1), unless both Ns are monosyllabic (i.e. 1+1), which can form a binary foot, too. If the function of disyllabic words is simply to avoid ambiguity, the conditions on their use are not explained.

## 7. When did disyllabic words start to appear in Chinese?

The homophone-avoidance theory (the orthodox view) makes two fundamental assumptions in, shown in (15). They are proposed in Karlgren (1918; 1923) and adopted by many others.

(15) Two assumptions of the homophone-avoidance theory

a. Classic Chinese is a monosyllabic language (i.e. most words or morphemes are monosyllabic).

b. Disyllabic words appeared in Chinese only after massive losses of syllable contrasts.

Questions can be raised for both assumptions. With regard to (15a), it is true that classic Chinese is basically a monosyllabic language, but so is modern Chinese, in the sense that most morphemes in Chinese are monosyllabic and free. What is clear, as we have shown, is that many words in modern Chinese have ‘elastic’ length (to borrow a term from Guo 1938), i.e. they can be disyllabic or monosyllabic. What is unclear is whether words in classic Chinese also have elastic length. The orthodox view assumes that classic Chinese does not, but little evidence has been presented. In contrast, Guo (1938) argues that Chinese words have always had elastic length. Guo cites many previous scholars who made similar comments, along with many examples from classic Chinese. Why then is there a common belief that classic Chinese does not use, or uses fewer, disyllabic words? There are two possible reasons. First, as suggested by Kennedy (1955), traditional Chinese dictionaries are 字典 *zi-dian* ‘character-book’, whose entries are characters (monosyllabic morphemes). In contrast, 词典 ci-dian ‘word-book’, whose entries are morphemes and words, is a modern tradition. Second, classic Chinese may not be a faithful record of spoken language, but a telegraphic version of it. Therefore, disyllabic words may appear at a lower rate. It is worth noting though that in poetry, which is a better reflection of spoken language, elastic words appear quite often. Consider a poem from the Tang Dynasty by 贺知章 (He Zhizhang 744), shown in (16) and (17).

(16) Elastic words in a Tang poem (underline indicates long forms of elastic words)

| Poem | Wds | Gloss |
|---|---|---|
| 少小离家老大回 | 5 | I left home young and returned old |
| 乡音无改鬓毛衰 | 6 | Accent unchanged, beard fading |
| 儿童相见不相识 | 4 | Children took me as a stranger |
| 笑问客从何处来 | 6 | Smiling, they asked, ‘Guest, where are you from?’ |

(17) Analysis of the elastic words in (16)

| | | |
|---|---|---|
| 少(小) | shao-(xiao) | young-(small) |
| 老(大) | lao-(da) | old-(big) |
| 鬓(毛) | bin-(mao) | beard-(hair) |
| (儿)童 | (er)-tong | (baby)-child |
| (相)识 | (xiang)-shi | (each)-know |
| 何(处) | he-(chu) | where-(place) |

Of the twenty-one words in the poem, seven are long forms of elastic words, as seen in (17). Some of the monosyllabic words are elastic, too. For example, 离 *li* ‘leave’ has the long form 离别 *li-bie* ‘leave-(farewell)’, 客 *ke* ‘guest’ has a long form 客人 *ke-ren* ‘guest-(person)’ and 回 *hui* ‘return’ has a long form 回来 *hui-lai* ‘return-(come)’. Thus, it seems that as early as the Tang Dynasty, elastic words were already in extensive use.

With regard to (15b), there is no consensus on when disyllabic words started to increase in Chinese. What is clear is that in the 隋唐 Sui-Tang period (600-800), Chinese has about 3,000-4,000 distinct syllables, including tonal contrasts (Li 1952). In contrast, modern Standard Chinese has just 1,300. Lü (1963) suggests that the increase of disyllabic words started only since the second half of the 19th century, long after Chinese lost most of its syllable contrasts. However, Lü (1963) offers no evidence for his estimate.

A further complication is that, even in the 隋唐 Sui-Tang period (600-800), the Chinese syllable inventory of 3,000-4,000 is already diminutively small, compared with that of English, estimated to be 158,000 by Jespersen (1930: 347). If the homophone-avoidance theory is correct, should Chinese already need massive numbers of disyllabic words then, or a long time before that?

In summary, there are serious problems with the fundamental assumptions of the homophone-avoidance theory, and much research is needed in order to determine the amount of elastic words in classic Chinese and whether the increase of elastic words corresponds to the loss of syllable contrasts.

## 8. Concluding remarks

We have offered a quantitative analysis of word length pairs in Chinese (such as 煤-煤炭 *mei-meitan* 'coal' and 虎-老虎 *hu-laohu* 'tiger'), which confirms previous observations of their abundance. We have also reviewed a popular view (the orthodox view), according to which word length pairs are motivated by homophone avoidance: because most Chinese morphemes are monosyllabic yet the syllable inventory of Chinese is rather small, there are too many homophones and disyllabic words are created to avoid ambiguity in speech.

Arguments for the popular view are reviewed and shown to be inconclusive. In addition, an exhaustive analysis of *Modern Chinese Dictionary* (XDHYCD 2005) shows that there is no correlation between homophone density and word length. In particular, among nouns, the percentage of word length pairs remains more or less constant regardless of how many homophones a noun has. We conclude, therefore, that homophone avoidance cannot explain the presence of word length pairs in Chinese.

Two alternative mechanisms for creating word length pairs are discussed: truncation and prosody. In addition, prosody needs the long form for some positions and the short form for other positions. The two mechanisms are sufficient to account for not only the creation of length pairs but also why both forms are kept.

## Acknowledgments:

We would like to thank the editors Yuchau Hsiao and Lian-Hee Wee for inviting us to contribute to this volume. We are also grateful to Xinting Zhang, who annotated some of the data, to the Phonetics and Phonology Group at the University of Michigan for their feedback, and to an anonymous reviewer for helpful comments.

## References:

Borowsky, Toni. 1986. *Topics in the lexical phonology of English*. Doctoral dissertation, University of Massachusetts, Amherst.

Chao, Yuen Ren. 1948. *Mandarin primer, an intensive course in spoken Chinese*. Cambridge, MA: Harvard University Press.

Duanmu, San. 2007. *The phonology of Standard Chinese*. 2nd Edition. Oxford: Oxford University Press.

Duanmu, San. 2008. *Syllable structure: the limits of variation*. Oxford: Oxford University Press.

Duanmu, San. 2012. Word-length preferences in Chinese: a corpus study. *Journal of East Asian Linguistics* 21.1: 89-114.

Feng, Shengli. 1998. Lun Hanyu de "ziran yinbu" [On "natural feet" in Chinese]. *Zhongguo Yuwen* 1998.1 (262): 40-47.

Guo, Shaoyu. 1938. Zhongguo yuci zhi tanxing zuoyong [The function of elastic word length in Chinese]. *Yen Ching Hsueh Pao* 24: 1-34.

Hayes, Bruce. 1982. Extrametricality and English stress. *Linguistic Inquiry* 13.2: 227-276.

Huang, Lijun, and San Duanmu. 2013. Xiandai Hanyu ci chang tanxing de lianghua yanjiu [A quantitative study of elastic word length in Modern Chinese]. *Yuyan Kexue* [Linguistic Sciences] 12.1: 8-16.

Jespersen, Otto. 1930. Monosyllabism in English: Biennial Lecture on English philology. *Proceedings of the British academy* 14: 341-368. London: British academy.

Jin, Wen. 2011. A statistical argument for the homophony avoidance approach to the disyllabification of Chinese. *Proceedings of the 23rd North American Conference on Chinese Lingusitics (NACCL-23), Volume 1*, edited by Zhuo Jing-Schmidt, 35-50. University of Oregon, Eugene. http://naccl.osu.edu/sites/naccl.osu.edu/files/NACCL-23_1_03.pdf.

Karlgren, Bernhard. 1918. *Ordet och pennan i Mittens Rike*. Stockholm: Svenska Andelsförlaget.

Karlgren, Bernhard. 1923. *Sound & symbol in Chinese*. London: Oxford University Press.

Kao, Diana. 1971. *Structure of the syllable in Cantonese*. The Hague: Mouton.

Ke, Jinyun. 2006. A cross-linguistic quantitative study of homophony. *Journal of Quantitative Linguistics* 13.1: 129-159.

Kennedy, George A. 1955. The Butterfly Case. *Wennti papers* No. 8 (March, 1955), pp. 1-47. New Haven: Far Eastern Publications, Yale University.

Li, Charles N., and Sandra A. Thompson. 1981. *Mandarin Chinese: a functional reference grammar*. Berkeley and Los Angeles: University of California Press.

Li, Rong. 1952. *Qieyun yinxi*. Beijing: Kexue Chubanshe.

Li, Rong (editor-in-chief). 2002. *Xiandai Hanyu fangyan da cidian* [A large dictionary of modern Chinese dialects]. Nanjing: Jiangsu Jiaoyu Chubanshe.

Lu, Bingfu, and San Duanmu, 2002. Rhythm and syntax in Chinese: A case study. *Journal of the Chinese Language Teachers Association* 37.2: 123-136.

Lü, Shuxiang. 1963. Xiandai Hanyu dan shuang yinjie wenti chu tan [A preliminary study of the problem of monosyllabism and disyllabism in modern Chinese]. *Zhongguo Yuwen* 1963.1: 11-23.

Maddieson, Ian, and Kristin Precoda. 1990. Updating UPSID. *UCLA Working Papers in Phonetics* 74: 104-111.

Mullie, Joseph. 1932. *The structural principles of the Chinese language, an introduction to the spoken language (Northern Pekingese dialect), volume I*. Translated from the Flemish by A. Omer Versichel. Peiping, China: The Bureau of Engraving and Printing.

Nettle, Daniel. 1995. Segmental inventory size, word length, and communication efficiency. *Linguistics* 33: 359 – 367.

Nettle, Daniel. 1999. *Linguistic diversity*. Oxford: Oxford University Press.

Pan, Wenguo. 1997. *Han Ying yu duibi gangyao* [An outline of comparisons between Chinese and English]. Beijing: Beijing University of Languages and Cultures Press.

Peking University. 1989. *Hanyu fangyin zihui* [A monosyllabic word list of Chinese dialectal pronunciations], 2nd edition, compiled by the Linguistics Section, Department of Chinese Language and Literature, Beijing University. Beijing: Wenzi Gaige Chubanshe.

Peking University. 1995. *Hanyu fangyan cihui* [A vocabulary for Chinese dialects], 2nd edition, compiled by the Linguistics Section, Department of Chinese Language and Literature, Peking University. Beijing: Yuwen Chubanshe.

Sproat, Richard, and Chilin Shih. 1996. A corpus-based analysis of Mandarin nominal root compound. *Journal of East Asian Linguistics* 5.1: 49-71.

T'sou, Benjamin K. 1976. Homophony and internal change in Chinese. *Computational Analysis of Asian & African Languages* 3: 67 – 86.

Vaux, Bert. 2009. The role of features in a symbolic theory of phonology.

In *Contemporary views on architecture and representations in phonology*, ed. Eric Raimy and Charles E. Cairns, 75-97. Cambridge, MA: MIT Press.

XDHYCD. 2005. *Xiandai Hanyu Cidian* [Modern Chinese Dictionary]. 5th edition. Compiled by the Institute of Linguistics, Chinese Academy of Social Sciences. Beijing: Shangwu Yinshuguan.

Xu, Baohua, and Huan Tao. 1997. *Shanghai Fangyan Cidian* [Dictionary of Shanghai Dialect]. (Xiandai Hanyu Fangyan Da Cidian [Dictionaries of Modern Chinese Dialects], editor-in-chief Rong Li.) Nanjing: Jiangsu Jiaoyu Chubanshe.