In this article, I review issues in defining words, such as distinctions between roots and affixes, between morphemes and words, between single words and compounds, and between compounds and phrases. I also discuss possible solutions to some problems and compare Chinese with English.

Keywords: affix, morpheme, word, compound, elastic word length, morpheme inventory, foot

1. Introduction

In English, it seems obvious what a word is: it is a meaningful unit written between spaces. In Chinese, there is a similar unit, also written between (invisible) spaces: it is called 字, a monosyllabic graph that in most cases has a meaning. Naturally, many people equate “word” in English with 字 in Chinese. For example, 马 (1898), a pioneer of modern linguistics in China and the first native scholar to write a grammar of Chinese, calls a verb as 动字 ‘action 字’, a noun 名字 ‘name 字’, a conjunction 连字 ‘connection 字’, and so on.

However, there are problems if we equate word with 字. For example, Karlgren (1918, 1923), Kennedy (1951), and Lin (1952) point out that 字 is rarely used as a free word in modern Chinese; instead, most Chinese words are disyllabic. In addition, many scholars, such as Chao (1968), 姚 (1981), 王洪军 (1999), 正 (2003), Pan et al. (2004), and Xu (2005), have argued, quite compellingly, that there is nothing in Chinese that corresponds to the notion of word. Indeed, the term is not found in the vocabulary of traditional Chinese linguistics, or in the vocabulary of ordinary Chinese speakers.

A crucial issue in defining words in Chinese is whether Chinese has affixes, which I shall discuss first. Then I discuss two views on words: (a) words can be defined for both English and Chinese, and (b) languages can differ in the basic unit of grammar, where it is word in English but 字 in Chinese. Next, I discuss a well-known property of Chinese, which is elastic word length. I shall show that this property offers an explanation why in Chinese it is hard to distinguish morphemes from words on the one hand and words and compounds or phrases on the other (see also the articles on modern morphology, Old Chinese morphology, and prosodic morphology).

2. Affixes in Chinese

A word in English is often made of a root and one or more affixes. In the English lexicon CELEX, 80% of uninflected non-compound words (called ‘lemmas’) are made this way.

A typical affix has a grammatical function but no referential meaning, and it cannot be used alone. For example, the English suffix –ly (as in slowly) has the function of changing an adjective to an adverb and it cannot be used alone. Some English affixes have some referential meaning, such as –er in writer (which refers to a person), but these are in the minority and not used as roots.

Linguists who look for words in Chinese have proposed various affixes, too. Pan et al.
reviews 14 such studies, which have proposed a total of 355 affixes in Chinese. However, most of them, such as cài 菜 ‘vegetable’, chǎng 厂 ‘factory’, and chē 车 ‘vehicle’ have a referential meaning and no grammatical function; in addition, they can serve as roots or words. Naturally, linguists hardly agree on whether they all count as affixes. For example, of the 355 proposed affixes, only 11 are agreed upon by nine or more of the 14 studies. Let us take a look at them, shown in (1).

(1) Chinese affixes that have been proposed by nine or more of 14 studies

<table>
<thead>
<tr>
<th>Affix</th>
<th>Studies</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>mén 们‘plural’</td>
<td>9</td>
<td>wǒmén 我们‘i-plural (we)’</td>
</tr>
<tr>
<td>huà 化‘change’</td>
<td>11</td>
<td>gōngyèhuà 工业化‘industry-change (industrialize)’</td>
</tr>
<tr>
<td>xìng 性‘nature’</td>
<td>12</td>
<td>gòngxing 共性‘common-nature’</td>
</tr>
<tr>
<td>zhē 者‘person’</td>
<td>13</td>
<td>zuòjiā 作家‘write-expert (writer)’</td>
</tr>
<tr>
<td>yuán 员‘member’</td>
<td>11</td>
<td>jiàoyuán 教员‘teach-member (faculty)’</td>
</tr>
<tr>
<td>dù 度‘degree’</td>
<td>10</td>
<td>yìngdù 硬度‘hard-degree (hardness)’</td>
</tr>
<tr>
<td>tóu 头‘head’</td>
<td>11</td>
<td>shítóu 石头‘stone’</td>
</tr>
<tr>
<td>zǐ 子‘son’</td>
<td>12</td>
<td>zhuōzǐ 桌子‘table’</td>
</tr>
<tr>
<td>ěr 儿‘son’</td>
<td>11</td>
<td>hāuér 花儿‘flower’</td>
</tr>
<tr>
<td>lǎo 老‘old’</td>
<td>9</td>
<td>lǎoshī 老师‘teacher’</td>
</tr>
</tbody>
</table>

The list is rather short. In addition, from an English perspective, only the first looks like like an affix, and possibly the second. Besides, mén 们 is mostly limited to the first, second, and third person pronouns, and not required otherwise (e.g., ‘those teachers’ can be zhèxiē lǎoshī 这些老师们). And huà 化 can be used as a word. The next five items in (1) all have referential meanings, so that the disyllabic example looks like a compound (a word made of two words). The last four items in (1) add no meaning or function to the root they attach to; indeed, they are often redundant, because the root can be used without them, e.g., fāng zhuō 方桌‘square table’ (without 子 for ‘table’) and huā ping 花瓶‘flower vase’ (without 儿 for ‘flower’).

Unlike the paucity of affixes in Chinese, English has hundreds of them, based on the data in CELEX. For example, there are three suffixes spelled as -acy, as in supremacy (where the suffix converts A to N), papacy (where the suffix converts N to N), and conspiracy (where the suffix converts V to N). Therefore, while Chinese may have some affixes, the number is strikingly small compared with that in English.

3. Defining words in Chinese

Many linguists have offered analyses of words in Chinese in terms of roots and affixes (see Pān et al. 2004 for a comprehensive review). However, as just noted, a problem for this approach is the lack of true affixes in Chinese.

Some linguists, such as Jespersen (1922:369), are aware of the problem and have concluded that Chinese words are essentially monosyllabic. However, as Karlgren (1918; 1923), Kennedy
(1951), and Lin (1952) have argued, the question for this view is why most monosyllables are not free and why most Chinese words occur in disyllabic forms.

In yet another approach, both Chinese and English have words, but they differ in morphology (Sproat and Shih 1996; Packard 2000). In English, most words are made of a root plus one or more affixes, whereas in Chinese most words are made of two (or more) roots (called ‘root compounds’). This approach also faces a question, namely, why most roots are not free in Chinese and why root compounds are rare in English.

4. Giving up words in Chinese
Some linguists have come to the conclusion that “word” is not a universal entity for every languages. Instead, languages can differ in the basic units of grammar. For example, according to Chao and Yang (1947) and Chéng (2003), both English and Chinese have morphemes, but only English has words. Unlike English, which uses morphemes to build words, Chinese uses morphemes to build phrases. Some linguists go even further. For example, Pān et al. (2004) and Xú (2005) argue that English and Chinese do not need to share any common category at all: English has morphemes and words, and Chinese has zì 字, which is neither a morpheme nor a word.

5. Elastic word length
Many Chinese words (or morphemes) can be long (disyllabic) or short (monosyllabic), where the former contains the latter. The property has been discussed by Guō (1938) and many other scholars (see Dong 2015; Duanmu and Dong 2015a, 2915b). Let us call the property elastic word length, which is defined in (2) and exemplified in (3).

(2) Defining elastic word length in Chinese
A word has elastic length if it has two forms A and B where
a. A is monosyllabic and B disyllabic (or longer);
b. A and B share the same base (morpheme) (i.e., B contains A);
c. A and B have the same meaning (i.e., B is semantically redundant or empty);
d. A and B are interchangeable in some contexts

(3) Sample words with elastic length

<table>
<thead>
<tr>
<th>Long</th>
<th>Short</th>
</tr>
</thead>
<tbody>
<tr>
<td>xué-xí 学习</td>
<td>xué 学</td>
</tr>
<tr>
<td>ji-shù 技术</td>
<td>ji 技</td>
</tr>
<tr>
<td>měi-tàn 煤炭</td>
<td>měi 煤</td>
</tr>
<tr>
<td>lăo-hǔ 老虎</td>
<td>hǔ 虎</td>
</tr>
<tr>
<td>yā-zǐ 鸭子</td>
<td>yā 鸭</td>
</tr>
</tbody>
</table>

The meaning of the long form can be limited to one of three kinds, which can be called XX, X0, and 0X. In XX, the meanings of the two parts are repetitive, such as xué-xí 学习 ‘study-practice’ and xūyào 需要 ‘need-want’. In X0, the meaning of the second part is empty, such as yā-zǐ 鸭子 ‘duck’. In 0X, the meaning of the first part is empty, such as lăo-hǔ 老虎 ‘tiger’.
There are two views on such long-short pairs. The first view sees the two forms of a pair are essentially synonymous and belong to the same word (e.g., Karlgren 1918; 1923; Guójù 1938; Chao and Yang 1947; Lín 1952; Pān 1997; Dong 2015; Duanmu and Dong 2015a, 2015b). The second view holds that the two forms differ in meaning and are different words (e.g., Lí 1990; Liu 1992; Wáng Cànlóng 2002; Wú 2003; Kē 2007).

It is true, as Lí (1990) and Wáng (2002) argue, that in some pairs, the long form is more formal (e.g., gòumăi 购买 vs. măi 买 ‘buy’), or more abstract (e.g., sīwáng 死亡 vs. sī 死 ‘death’), or of a larger quantity (e.g., shūjí 书籍 vs. shū 书 ‘books’). However, such differences do not always hold for other pairs. For example, there is no evidence that lăohǔ 老虎 ‘tiger’ is more formal or abstract, or implies a larger quantity, than hǔ 虎. Rather, there is broad consensus that, in most pairs, the two forms have the same referential meaning. This is evidenced by the fact that in dictionaries, the two forms of a pair are either listed under the same entry (e.g., Chao and Yang 1947) or used to annotate each other (e.g., Xiàndài Hányǔ Cídiǎn 现代汉语词典 [Dictionary of Modern Chinese]).

Karlgren (1918, 1923) suggests that the creation of long forms is motivated by the desire to avoid homophones, a view shared by many people. This view faces three problems. First, while some long forms are indeed expanded from short forms (such as lăohǔ 老虎 ‘tiger’ from hǔ 虎), the reverse process is also present, i.e., reducing long forms to short ones, such as rì 日 ‘Japan’ from rìběn 日本 or jī 机 ‘airplane’ from fēijī 飞机 ‘fly machine (airplane)’. Second, there are restrictions on where long and short forms can occur. For example, in a corpus study, Duanmu (2012) has shown that in [N N], 1+2 (monosyllable + disyllable) is disfavored, whereas in [V N], 2+1 is disfavored. Such patterns suggest that long forms are preferred in phonologically strong positions. For example, in [N N], the first N is strong, as predicted by the Compound Stress Rule, and in [V N], N is strong, as predicted by the Nuclear Stress Rule (Chomsky and Halle 1968). An analysis in terms of stress assignment and foot binarity can be found in Lu and Duanmu (2002) and (Duanmu 2007). Third, according to the homophone avoidance theory, the more homophones a syllable has, the more likely a disyllabic form will be created for it. However, it has been shown that such a correlation does not exist; instead, syllables with no homophones are just as likely to have elastic length as syllables with many homophones (Dong 2015; Duanmu and Dong 2015a, 2015b).

How many monosyllables in Chinese have a disyllabic form? Pān (1997) suggests that nearly all of them do, but he offers no actual count. Huáng and Duānmù (2013) sampled a random list of 2,000 morphemes (one tenth of those in Xiàndài Hányǔ Cídiǎn) and found that 61% of them have elastic length. Dong (2015) annotated the entire Xiàndài Hányǔ Cídiǎn and found that 47% of Chinese morphemes have elastic length. The data are summarized in (4), divided by parts of speech (POS), where 1-only means a morpheme only has a monosyllabic form, Poly-only means a morpheme only has a polysyllabic form, and Elastic means a morpheme has both a monosyllabic form and a disyllabic (or longer) form.
Two comments of interest can be made. First, it can be seen that there is a fairly high percentage of elastic forms among content words (nouns, verbs, adjectives, and adverbs). In contrast, the percentage of elastic forms among function words (i.e. ‘others’) is quite low. Second, the average percentage of elastic words, at 46.8%, is lower than previous estimates and contradicts with a popular view that most Chinese expressions, especially nouns, occur in disyllabic forms (e.g., Chao 1948; Pān 1997). However, Duānmù (2015) has shown that most 1-only nouns cannot occur alone. For example, monosyllabic names commonly occur with another syllable, such as Lǎo Wáng 老王 ‘Old Wang’ and Xiǎo Wáng 小王 ‘Little Wang’, whereas disyllabic names usually occur without the extra syllable, such as Ōuyáng 欧阳 ‘Ouyang’. Other 1-only words are often part of a set phrase, such as gū 辜 ‘guilt’ in wú gū 无辜 ‘without guilt (innocent)’ and jiā 佳 ‘fine’ in jiā rén 佳人 ‘fine person (beautiful woman)’. Therefore, as far as Chinese expressions are concerned, most of them are disyllabic (or have a disyllabic form) in actual use.

6. zì 字, morphemes, and words
It has been suggested that zì 字 is basically a morpheme (Chao and Yang 1947; Chao 1968; Chéng 2003). Several provisions are needed though. First, sometimes a zì has two or more meanings and represents two (or more) morphemes, such as què 却, which can mean ‘retreat’ or ‘but’. Second, there are some disyllabic names, such as wúgōng 蜈蚣 ‘centipede’, where neither half has a meaning by itself. Third, a zì can occasionally represent a disyllabic compound, such as qiān wă 千瓦 ‘thousand watt (kilowatt)’, but such cases are new creations and normally have a disyllabic alternative (in this case 千瓦). For the first case, we can consider them to be homographs, i.e. different zì that happen to look (and sound) the same. The second and third cases are fairly rare. Therefore, in most cases, each zì represents a morpheme.

Pān et al. (2004) and Xú (2005) argue that zì is more than a morpheme, mainly because zì also includes the graph of a morpheme, which plays an important role in Chinese. For example, because Chinese has more homophones than English, the disambiguating role of zì is greater than that of the English orthography. For example, ‘morning’ 旦, ‘but’ 但, ‘egg’ 蛋, and ‘thin/weak’ 淡 are all pronounced dàn but are distinguished by different zì. In addition, a Chinese graph often offers clues to its meaning (e.g., many verbs that involve the use of hands have a graphic component that indicates a hand). Indeed, when a literate Chinese speaker explains a zì to someone, he or she would often instinctively use the index finger to draw the character in the air. However, while orthography may interact with grammar in some ways, it needs not be part of a language (for example, for children or people who have not learned writing). If we set aside the
orthographical aspect of \( zi \), then each \( zi \) corresponds to a morpheme quite well.

It is interesting to compare the morpheme inventory sizes of English and Chinese. Consider the data in (5), calculated by myself, where English is based on CELEX and Chinese on Xiàndài Hányǔ Cídiǎn.

(5) Sizes of morpheme inventories in English and Chinese

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data source</td>
<td>CELEX</td>
<td>Xiàndài Hányǔ Cídiǎn</td>
</tr>
<tr>
<td>Zero derivations excluded</td>
<td>10,000</td>
<td>10,000</td>
</tr>
<tr>
<td>Zero derivations included</td>
<td>17,000</td>
<td>17,000</td>
</tr>
</tbody>
</table>

If we exclude zero derivations, e.g., counting study (noun) and study (verb) as one morpheme, then both English and Chinese have about 10,000 morphemes, of which two thirds are commonly used. If we include zero derivations, e.g., counting study (noun) and study (verb) as two morphemes, then both languages have about 17,000 morphemes. Thus, the morpheme inventories are strikingly similar in size.

Recall that there are two problems in defining words in Chinese. First, most monosyllables are not free and hence not words. Second, most words are disyllabic, which look like compounds. Given the property of elastic word length, we can account for both problems. First, a monosyllable is not free because phonology requires a minimal word to be a disyllabic foot, in which case we should choose the disyllabic form of a word, which is always free; thus, most words do have a free form. Second, the disyllabic forms of a word may look like a compound, but it is not, since its meaning is XX, X0, and 0X, whereas a true compound should be XY (made of two different meanings); thus, based on semantic structure, we can distinguish disyllabic words from disyllabic compounds.

7. Words vs. phrases

There are three difficulties in distinguishing words from phrases in Chinese. First, as just discussed, the long form of a word looks like a compound. Second, some nominal units look like a phrase, such as 小车 ‘small car’. Third, there are disyllabic units, called 离合词 ‘separable words’, which behave like a compound in some cases but can be split in a phrase elsewhere. For example, 有钱 ‘have money (rich)’ behaves like a compound (e.g., it can be modified by ‘very’, as in 很有钱 ‘very rich’), but it can be split in 有很多钱 ‘have a lot of money’.

We can account for the first case by drawing a distinction in semantic structure, where words are XX, 0X, and X0 and phrases are XY. For the second case, a distinction can be made, as discussed in Duanmu (1998) and references therein. Specifically, \([M N]\) is a word or compound and \([M \text{的} N]\) is a phrase, where M is a modifier of the noun N. For the third case, we can consider有钱 to be ambiguous between a compound ‘rich’ and a phrase ‘have money’, just as ‘black sheep’ can be a compound (an unusual person) or a phrase (a sheep that is black).
8. Summary
I have reviewed problems in defining words in Chinese, including difficulties in distinguishing roots from affixes, morphemes from words, single words from compounds, and compounds from phrases. I have suggested that many problems may be solvable, especially in view of the elastic property of word length. Therefore, Chinese may not be as different from English as it seems after all.

Primary Sources:
Xiàndài Hànyǔ Chángyòng Cíbiǎo: Xiàndài Hányǔ Chángyòng Cíbiǎo Kètìzǔ 现代汉语常用词表课题组, Xiàndài Hányǔ chángyòng cíbiǎo (cǎoàn) 现代汉语常用词表(草案) [A commonly used word list in Modern Chinese], Bēijīng 北京: Shāngwǔ Yìnhūguǎn 商务印书馆, 2008.

References:


Huang Lijun 黄丽君 and Duanmu San 端木三, Xiaodai Hanyu cichang tanxing de lianghua yanjiu 现代汉语词长弹性的量化研究 [A quantitative study of elastic word length in Modern Chinese], Yuyankexue 语言科学 12.1, 2013, 8-16.


Karlgren, Bernhard, Ordet och pennan i Mittens Rike [The word and the pen in the Middle Kingdom], Stockholm: Svenska Andelsförlaget, 1918.


Li Liding 李临定, Dongci fenlei yanjiu shouluo 动词分类研究说略 [A brief discussion on verb classification], Zhongguo Yuyen 中国语文 4, 1990, 248-257.


Mai Jianzhong 马建忠, Mashi wen tong 马氏文通 [Mr. M.a’s grammar], Shanghai 上海: Shangwu Yinshuguan 商务印书馆, 1898.


Pān Wénguó 潘文国, Yè Bùqīng 叶步青, and Hán Yáng 韩洋, Hányǔ de gòucífǎ yánjiū 汉语的
g构词法研究 [A study of Chinese morphology], Shànghǎi 上海: Huádōng Shìfàn Dàxué

Sproat, Richard, and Chilin Shih, A corpus-based analysis of Mandarin nominal root compound.

Wáng Cànlóng 王灿龙, Jūfǎ zhǔ hé zhōng dān shuāng yīnjié xuǎn zé de rènzhī jiěshì 句法组合中
单双音节选择的认知解释 [A cognitive explanation of the choice between monosyllables
and disyllables in syntactic units], Zhōngguó Yǔwén Zázhìshè 中国语文杂志社, ed., Yǔfǎ
Yánjiū hé Tànsuǒ, dì 11 Jí 语法研究和探索, 第 11 辑, Běijīng 北京: Shāngwù Yìnshūguǎn

Wáng Hóngjūn 王洪君, Hányǔ fēixiàn xìng yīnxì xué 汉语非线性音系学 [Chinese non-linear

Wú Wéishàn 吴为善, Shuāngyīnhuà, yǔfǎ huà hé yùnlǜ cí de zài fēnxī 双音化、语法化和韵律词
的再分析 [A reanalysis of disyllabification, grammaticalization, and prosodic words],

Xú Tōngqiāng 徐通锵, Hányǔ jiégòu de jīběn yuánlǐ — zì běnwèi hé yǔyán yánjìu 汉语结构的
基本原理 — 字本位和语言研究 [Fundamentals of the structure of Chinese: the zì-base and
language study], Qīngdāo 青岛: Zhōngguó Hǎiyáng Dàxué Chūbǎnshè 中国海洋大学出版社, 2005.