Wordhood*

San Duanmu

November 5, 2012

Summary

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.

Short biography

San Duanmu is Professor of Linguistics, University of Michigan. He received his PhD in Linguistics from MIT in 1990 and has held teaching posts at Fudan University, Shanghai (1981-1986) and the University of Michigan, Ann Arbor (1991-present). He is the author of *The Phonology of Standard Chinese* (2nd edition, Oxford 2007) and *Syllable Structure: The Limits of Variation* (Oxford 2008).

Terms for indexation (5-10): 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 zì, a monosyllabic graph that in most cases has a meaning. Naturally, many people equate word in English with zì in Chinese. For example, Mā (1898), a pioneer of modern linguistics in China
and the first native scholar to write a grammar of Chinese, calls a verb as 動字 dòng zì ‘action zì’, a noun as 名字 míng zì ‘name zì’, a conjunction as 連字 lián zì ‘connection zì’, and so on.

However, there are problems if we equate word with zì. For example, Kennedy (1951) and Lín (1952) point out that zì is rarely used as a free word in modern Chinese; instead, most Chinese words are disyllabic. In addition, many scholars, such as Chao (1968), Lǚ (1981), Wáng (1999), Chéng (2003), Pân et al. (2004), and Xú (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 zì 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.

2. Affixes in Chinese

A word in English is often made of a root and one or more affixes. In the CELEX lexicon of English (Baayen et al. 1995), 80% of uninflected non-compound words 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. Pān et al. (2004: 464-487) reviews fourteen such studies, which have proposed a total of 355 affixes in Chinese. However, most of them, such as 菜 cài ‘vegetable’, 廠 chǎng ‘factory’, 車 chē ‘vehicle’, etc., 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 eleven are agreed upon by nine or more of the fourteen studies. Let us take a look at them, shown in (1).

(1) Eleven proposed affixes that have the broadest consensus

<table>
<thead>
<tr>
<th>Affix</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>倫 mén ‘plural’</td>
<td>我們 wǒmén ‘i-plural (we)’</td>
</tr>
<tr>
<td>化 huà ‘change’</td>
<td>工業化 gōngyèhuà ‘industry-change (industrialize)’</td>
</tr>
<tr>
<td>性 xìng ‘nature’</td>
<td>共性 gòngxing ‘common-nature’</td>
</tr>
<tr>
<td>家 jiā ‘expert’</td>
<td>作家 zuòjiā ‘write-expert (writer)’</td>
</tr>
<tr>
<td>者 zhě ‘person’</td>
<td>作者 zuòzhě ‘write-person (author)’</td>
</tr>
<tr>
<td>員 yuán ‘member’</td>
<td>教員 jiàoyuán ‘teach-member (faculty)’</td>
</tr>
<tr>
<td>度 dù ‘degree’</td>
<td>硬度 yìngdù ‘hard-degree (hardness)’</td>
</tr>
<tr>
<td>頭 tóu ‘head’</td>
<td>石頭 shítóu ‘stone’</td>
</tr>
<tr>
<td>子 zǐ ‘son’</td>
<td>桌子 zhúzǐ ‘table’</td>
</tr>
<tr>
<td>兒 ér ‘son’</td>
<td>花兒 hāuér ‘flower’</td>
</tr>
<tr>
<td>老 shī ‘old’</td>
<td>老師 lǎoshī ‘teacher’</td>
</tr>
</tbody>
</table>

The list is rather short. In addition, from an English perspective, only the first looks like an affix, and possibly the second, although the latter can be used as a word. The next five all
have referential meanings, so that the disyllabic example looks like a compound (a word made of two words). The last four 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. 花瓶 huā píng ‘flower vase’ and 方桌 fāng zhōu ‘square table’).

Unlike the paucity of affixes in Chinese, English has hundreds of them (Baayen et al. 1995). 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 Kennedy (1951) and Lín (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 *zi*, 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. Let us call the property elastic word length. Some examples are shown in (2).

(2) Elastic word length in Chinese

<table>
<thead>
<tr>
<th>Long</th>
<th>Short</th>
</tr>
</thead>
<tbody>
<tr>
<td>煤炭 méi-tàn</td>
<td>煤 méi</td>
</tr>
<tr>
<td>學習 xué-xí</td>
<td>學 xué</td>
</tr>
<tr>
<td>技術 jì-shù</td>
<td>技 jì</td>
</tr>
<tr>
<td>老虎 lāo-hǔ</td>
<td>虎 hǔ</td>
</tr>
<tr>
<td>鴨子 yā-zǐ</td>
<td>鴨 yā</td>
</tr>
</tbody>
</table>
The meaning of the long form can be limited to one of three kinds, which I call XX, X0, and 0X. In XX, the meanings of the two parts are repetitive, such as 煤炭 méi-tàn ‘coal-charcoal’, 學習 xué-xí ‘study-practice’, and 需要 xūyào ‘need-want’. In X0, the meaning of the second part is empty, such as 鴨子 yā-zi ‘duck’. In 0X, the meaning of the first part is empty, such as 老虎 lāo-hǔ ‘tiger’.

There are two views on such long-short pairs. On one view, the two forms of a pair are essentially synonymous and belong to the same word (e.g. Guō 1938; Chao and Yang 1947; Lín 1952; Pān 1997). On another view, the two forms differ in meaning and are different words (e.g. Lǐ 1990; Liu 1992; Wáng 2002; Wú 2003; Kē 2007).

It is true, as Lǐ (1990) and Wáng (2002) argue, that in some pairs, the long form is more formal (e.g. 購買 gòumǎi vs. 買 mǎi ‘buy’), or more abstract (e.g. 死亡 sǐwáng vs. 死 sǐ ‘death’), or of a larger quantity (e.g. 書籍 shūjí vs. 書 shū ‘books’). However, such differences do not always hold for other pairs. For example, there is no evidence that 老虎 lǎohǔ ‘tiger’ is more formal or abstractness, or implies a larger quantity, than 虎 hǔ. 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 Hànyǔ Cídiǎn).

It is often suggested that the creation of long forms is motivated by the desire to avoid homophones. This view faces two problems. First, while some long forms are indeed expanded from short forms (such as 老虎 lǎohǔ ‘tiger’ from 虎 hǔ), the reverse process is also present, i.e. reducing long forms to short ones, such as 日 rì ‘Japan’ from 日本 riběn or 機 jī ‘airplane’ from 飛機 fēijī ‘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).

How many monosyllables in Chinese have a disyllabic form? Pān (1997) suggests that nearly all of them do, but an actual count has been lacking. Based on a small sample of thirty characters (54 lexical entries), Duanmu (2011) estimates that 90% of Chinese monosyllables have a disyllabic form. A more extensive study is offered by Huang and Duanmu (2012), based on a random sample of 2,000 monosyllabic senses (one tenth of the entire Xiàndài Hányǔ Cìdiǎn). It is found that about 70% of Chinese monosyllables have a disyllabic form. The data are summarized in (3), divided by parts of speech (POS), where N, V, and A are noun, verb, and adjective, respectively, and 1-only means a word only has a monosyllabic form.

(3) Word length elasticity in Chinese (Huang and Duanmu 2012)

<table>
<thead>
<tr>
<th>POS</th>
<th>% of all</th>
<th>1-only</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>41%</td>
<td>19%</td>
</tr>
<tr>
<td>V</td>
<td>38%</td>
<td>26%</td>
</tr>
<tr>
<td>A</td>
<td>14%</td>
<td>39%</td>
</tr>
<tr>
<td>Other</td>
<td>8%</td>
<td>85%</td>
</tr>
<tr>
<td>All</td>
<td>100%</td>
<td>29%</td>
</tr>
</tbody>
</table>

It can be seen that about 80% of nouns have elastic length, higher than that of verbs. In contrast, closed class POS categories (grouped under ‘other’), remain mostly monosyllabic. For example, prepositions are 100% monosyllabic. It is interesting to note, too, that there are nearly
as many verbs as nouns, a topic not to be pursued here.

6.  zi, morphemes, and words

It has been suggested that zi is basically a morpheme (Chao and Yang 1947; Chao 1968; Chéng 2003). Several provisions are needed though. First, sometimes a zi has two or more meanings and represents two (or more) morphemes, such as 却 què, which can mean ‘retreat’ or ‘but’. Second, there are some disyllabic names, such as 蜈蚣 wúgōng ‘centipede’, where neither half has a meaning by itself. Third, a zi can occasionally represent a disyllabic compound, such as 瓊 qiānwā ‘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 zi that happen to look (and sound) the same. The second and third cases are fairly rare. Therefore, in most cases, each zi represents a morpheme.

Pān et al. (2004) and Xú (2005) argue that zi is more than a morpheme, mainly because zi 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 zi is greater than that of the English orthography. For example, 旦 ‘morning’, 但 ‘but’, 蛋 ‘egg’, and 淡 ‘thin/weak’ are all pronounced dàn but are distinguished by different zi. 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). However, while orthography may interact with grammar in some ways, it is not a necessary 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 (4), calculated by myself, where English is based on CELEX (Baayen et al. 1995) and Chinese on Xiàndài Hànyǔ Cídiàn.

(4) Sizes of morpheme inventories in English and Chinese

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero derivations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>excluded</td>
<td>10,000</td>
<td>10,000</td>
</tr>
<tr>
<td>Zero derivations</td>
<td>17,000</td>
<td>17,000</td>
</tr>
<tr>
<td>included</td>
<td></td>
<td></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 elastic length property, we can account for both problems: 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. 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).

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 小車 xiao chē ‘small car’. Third, there are disyllabic units, called 離合詞 lihécí
‘separable words’, that behave like a compound in some cases but can be split in a phrase elsewhere. For example, 有錢 yǒu qián ‘have money (rich)’ behaves like a compound (e.g. it can be modified by ‘very’, as in 很有錢 hěn yǒu qián ‘very rich’), but it can be split in 有很多錢 yǒu hěnduō qián ‘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. For the third case, we can consider 有錢 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.

*Acknowledgments: I would like to thank some reviewers for their useful comments and Xinting Zhang for her help with formatting.

References:


