

Phonology and the Chinese Lexicon
San Duanmu, University of Michigan
Huili Zhang, Peking University and University of Michigan

Despite a wide-spread view that Chinese words are mostly monosyllabic (Jespersen 1922), defining the word as a ‘free morpheme’ is highly problematic in Chinese. Various proposals have been made in the literature, but none of them is satisfactory. We show that the notion of the word is inevitably intertwined with phonology. We also offer some statistical results and show that Chinese is not a monosyllabic language, nor is it a disyllabic language.

Traditional Chinese dictionary entries are based on monosyllabic characters, nearly all of which are morphemes and most of which can be seen as words, even though some are not completely ‘free’. For example, *hu* ‘tiger’ is rarely used alone, but often in *lao-hu* ‘(old) tiger’, where ‘old’ is semantically empty. However, when ‘tiger’ is used with another word, *hu* is often preferred, such as *meng hu* ‘fierce tiger’ and *hu shan* ‘tiger mountain’. Therefore, many dictionaries still list *hu* ‘tiger’ as a word, along with *lao-hu* ‘(old)-tiger’.

However, since morphemes like *hu* ‘tiger’ cannot be used alone and are not free, some linguists hesitate to call them words. For example, Sproat & Shih (1996) suggest that morphemes like *hu* are roots, which can combine with an affix to make a word, or with another root to make a ‘root compound’. It is worth noting that root compounds (such as ‘fierce tiger’ and ‘tiger mountain’) are highly productive and it is not possible for a dictionary to list all of them.

If we consider root compounds to be words, then Chinese may end up with many long words. For example, [*song hua*] *jiang* ‘[pine flower] river’ has to be a word, because *jiang* is a bound root, and [*song hua jiang*] *hu* ‘[pine flower river] tiger’ has to be a word, because *hu* is a bound root. To avoid such problems in word segmentation, most modern Chinese lexicons have adopted a policy of syllable count that favors disyllabic unit and avoid longer units (e.g. Lexicon of Common Words in Contemporary Chinese 2008, the largest corpus-based lexicon to date). Not surprisingly, 70% of the words in such lexicons are disyllabic.

An explicit reference to syllable count is clearly phonological. We examine some modern Chinese lexicons quantitatively and offer additional evidence that phonology plays a central role in word length. For example, there is a much higher ratio of disyllables among nouns than among verbs. In addition, we coded the top 3000 common disyllabic words and found that 40% of them are ‘pseudo-compounds’, in the sense that, in an appropriate context, they can be replaced by either the first syllable (e.g. *yu-(yan)* ‘language-(speech)’), or the second (e.g. (*lao*)-*hu* ‘(old) tiger’), or either (e.g. *xu-yao* ‘need-want’). We offer an analysis in which word length in Chinese is heavily driven by phonology, where a disyllabic foot is needed in a metrically strong position (such as O in VO) and a monosyllabic word is used in a metrically weak position (such as V in VO). When a monosyllabic morpheme occurs as a foot, it is usually made into a pseudo-compound by an added syllable. The apparent ‘boundness’ of many morphemes, therefore, are not morphological but phonological, the avoidance of a monosyllabic foot.

References:

- Jespersen, Otto. 1922. *Language: its nature, development and origin*. New York: Macmillan.
Sproat, Richard, and Chilin Shih. 1996. A corpus-based analysis of Mandarin nominal root compound. *Journal of East Asian Linguistics* 5: 49-71.