HOW MANY CHINESE WORDS HAVE ELASTIC LENGTH?

San Duanmu (端木三)
University of Michigan, USA
September 22, 2011

ABSTRACT
A word has elastic length when it has both a short form and a long form, such as 煤-煤炭 'coal' and 种-种植 'to plant'. The goal of this study is to find out how many Chinese words have this property. We sample 1/50 of all monomorphemic Chinese words, selected from those represented by the top 3,000 most common characters. It is found that 80%-90% of all Chinese words have elastic length. In addition, the percentage for verbs is higher than the average, and the percentage for nouns is higher still.

SUBJECT KEYWORDS
Elastic word length, Chinese lexicon, Word segmentation, Compound

1. INTRODUCTION

An unusual property of Chinese is that many of its words have elastic length, such as 煤-煤炭 'coal', 麦-小麦 'wheat', 种-种植 'to plant', and 学-学习 'to study'. Some English words also do, such as *John-Johnny* and *lab-laboratory*, but the scope is more limited in English than in Chinese. The property has been observed for a long time (see 郭绍虞 1938 for early references), but it remains unclear how many Chinese words have elastic length. 潘文国 (1997, p. 141) suggests that 'nearly all Chinese words' have elastic word length, but he provides no evidence. This study takes a close look

at the issue through a quantitative analysis of the Chinese lexicon.

2. WHAT IS ELASTIC WORD LENGTH?

Many Chinese words can be long (two syllables) or short (one syllable), with more or less the same meaning. 郭绍虞 (1938) uses the term 'elastic word length' to refer to this property. Duanmu (2007) uses the term 'dual vocabulary' to refer to the same thing. Some examples are shown in (1).

(1) Elastic word length in Chinese

2 syllables	1 syllable	
学习	学	'to study'
种植	种	'to plant'
煤炭	煤	'coal'
工人	工	'worker'
商店	店	'store'
老虎	虎	'tiger'
印度	印	'India'

The two forms of each pair are interchangeable in at least some context. For example: 'coal store' can be 煤炭店 or 煤店 and 'skill worker' can be 技术工人 or 技术工 (relevant word underlined). The two forms of a pair need not be interchangeable in all contexts. For example, 'coal mine' is 煤矿 and not 煤炭矿, and 'worker union' is 工会 and not 工人会. This shows that the actual usage may be partly influenced by convention.

The long form may look like a compound, but it is not. For example, the long form of 虎 'tiger' is 老虎, which literally means 'old tiger'. However, 老虎 simply means 'tiger' and not 'old tiger', because even a baby tiger can be called 老虎. Similarly, the long form of 煤 'coal' is 煤炭, which literally means 'coal-charcoal' but actually means 'coal', not 'coal and charcoal'. Therefore, we can call the long forms 'pseudo-compounds' (Duanmu 2007).

Word-length alternation is not limited to 2-1 pairs. For example, 'Canada' has a 3-1 pair 加拿大 and 加, and 'California' has a 5-1 pair 加利福尼亚 and 加. However, 2-1 pairs are clearly the most common.

Word-length alternation in Chinese has been observed for a long time (see 郭绍虞 1938 for early references and Duanmu 2007, Chapter 7, for recent ones). However, it remains unclear how many Chinese words have elastic length. 潘文国 (1997) suggests that nearly all Chinese words do, but he offers no evidence. In this study I address this question in detail through a quantitative analysis of the Chinese lexicon.

3. WORDS AND THE LEXICON

To find out how many words have elastic length, it is necessary to ask how many words there are in the lexicon. The latter question in turn invites another question: What is a word?

In English, words are usually separated by space, although some complications exist. For example, should compounds be included in the lexicon? What kinds of compounds should be included? How do we define compounds? Hockett (1958, p. 167) argues that *out-* in *outside* is a word and so *outside* is a compound, but in the CELEX lexicon (Baayan *et al*. 1995), *out-* in *outside* is a prefix and *outside* is not a compound. Even if we can set such cases aside, there are still different ways to define the lexicon. For example, CELEX offers three lexicons for English, shown in (2).

(2) Different definitions of the English lexicon by CELEX

Definition	Size (words)
Word shape, including inflection	160,595
Lemmas, excluding inflection	52,447
Monomorphemic words	7,401

If words are defined by their orthographic shapes (e.g. counting *cat*, *cats*, *cat's*, and *cats'* as different words), English has 160,595 words. If inflection is excluded (e.g. counting *cat* but not *cats*, *cat's*, or *cats'*), English

has 52,447 words (or 'lemmas'). We can even limit words to single morphemes, in which case there are 7,401 words.

The issue in Chinese is also problematic. Bloomfield (1926) defines a (minimal) word as a free morpheme, but the distinction is not always clear (e.g. 吕叔湘 1959, 1962, 1979; Xia 2000; and 王洪君 2001). In particular, monosyllabic forms, such as 虎 'tiger', are not always free, but they are not affixes either. According to Sproat and Shih (1996), most Chinese monosyllables are bound roots and the disyllabic forms are root compounds. But a Chinese monosyllable differs from a bound root or affix in English in an important way: the latter requires a (certain kind of) morpheme on a particular side, but the former does not. For example, *theo-* requires a morpheme to its right and *–ology* requires a morpheme to its left. In contrast, a Chinese root simply needs another syllable, on either side. For example, 猛虎 'fierce tiger' and 虎山 'tiger mountain' are both fine. Thus, the lack of freedom for Chinese monosyllables is not morphological but phonological, i.e. the need for a minimal word to be disyllabic. Therefore, we can continue to treat them as words.

As in English, the size of the Chinese lexicon depends on the definition. One difficulty is the lack of distinction between a word and a compound or phrase. For example, 鸡 'chicken' is a free word but 鸭 'duck' is not quite so, because it is often used in a disyllabic form 鸭子. If we treat 鸡蛋 'chicken egg' as a compound and 鸭蛋 'duck egg' as a single word, we lose a structural parallel. Similarly, it is controversial whether 有钱 'have money (rich)' is a word, a phrase, or sometimes a word and sometimes a phrase.

Modern compilers of the Chinese lexicon often follow a practice in which an expression is treated as a word if it occurs frequently, even if its constituents are free and the meaning is transparent (e.g. Xia 2000 and 王洪君 2001). For example, in 《现代汉语常用词表》 (现代汉语常用词表课题组 2008) and the Lancaster Corpus of Mandarin Chinese (McEnery and Xiao 2004), 鸡蛋 'chicken egg' is a word, so is 有钱 'have money (rich)'. According to this method, the size of the Chinese lexicon is similar to that in English, excluding inflections. This is shown in (3).

(3) The size of the Chinese lexicon

Size (words)	Source
56,064	《现代汉语常用词表》
45,429	Lancaster Corpus of Mandarin Chinese

According to 《现代汉语常用词表》 and the Lancaster Corpus of Mandarin Chinese, Chinese has about 50,000 words, a size comparable to that of the English lexicon (e.g. 52,447 in the 'lemma' lexicon of CELEX). The Lancaster Corpus of Mandarin Chinese has a smaller lexicon because it is based on a smaller corpus.

4. WORDS, ENTRIES, AND SENSES

Traditional dictionaries distinguish 'entries' (词条) and 'senses' (词义). An entry can be made of one or more words. For example, in the dictionary 《现代汉语词典》 (中国社会科学院语言研究所词典编辑室 2002), some entries headed by 煤 'coal' are shown in (4), where coal balls are made from coal powder for easy use in stoves. It can be seen that the entries 煤 and 煤炭 have the same meaning. They are the short and long forms of the same word.

(4) Some entries (词条) headed by 煤 in 《现代汉语词典》

煤	'coal'
煤层	'coal layer'
煤末	'coal powder'
煤气	'coal gas (natural gas)'
煤球	'coal ball'
煤炭	'coal'
煤田	'coal field'
…	

Sometimes an entry has several related meanings or represents different word categories. These are called senses (词义) and are grouped under the same entry. For example, in 《现代汉语词典》, the entry 青 has six senses, shown in (5), and the entry 俘 has two, shown in (6).

(5) Senses (词义) of the entry 青 in 《现代汉语词典》
   a. blue or green
   b. black
   c. green grass
   d. young age
   e. young person
   f. a family name

(6) Senses (词义) of the entry 俘 in 《现代汉语词典》
   a. to take prisoner of war
   b. prisoner of war

An entry made of two (or more) words can also have more than one sense. For example, the compound 青年 has two senses, shown in (7).

(7) Senses (词义) of the entry 青年 in 《现代汉语词典》
   a. young age (between fifteen and thirty)
   b. people in this age range

In this study, we focus on word entries, rather than word senses. Sometimes, we shall divide the senses of a word into two (or more), if they seem different enough. For example, ignoring the family name, the five senses of 青 can be divided into ‘color’ and ‘age’, where the former has a disyllabic form 青色 ‘blue/green/black color’ and the latter a disyllabic form 青年 ‘young age’. Similarly, the senses of 俘 have the disyllabic forms 俘虏 (a noun) and 俘获 (a verb).

5. PROCEDURE

Since we are interested in the elastic length of single words, we can focus on monosyllabic words, or 字 'characters'. The reason is that most polysyllabic words in Chinese are compounds or phrases, such as 鸡蛋 'chicken egg', 有钱 'have money (rich)', and 煤田 'coal field'. Their components are already represented by other words. Some disyllabic words are not true compounds but monomorphemic in nature, such as 煤炭 'coal', 大蒜 'garlic', and 学习 'study', which are also represented by their monosyllabic counter parts, i.e. 煤, 蒜, and 学 respectively. Therefore, excluding polysyllabic words will not affect the accuracy of our study. Finally, true polysyllabic words, such as 玛瑙 'amber' and 尴尬 'embarrassed', which are quite rare in Chinese, are listed under both characters and will not be missed (as long as we do not double-count them).

According to Da (2004), in a corpus of over 190 million character tokens in modern Chinese, there are 9,933 different characters, but 99% of the texts are covered by the top 3,000 most frequent ones. This is shown in (8).

(8) Coverage of Chinese characters (字) in modern texts (Da 2004)

Number of characters	Text coverage
First 500	75.81%
First 1,000	89.14%
First 1,500	94.55%
First 2,000	97.13%
First 2,500	98.45%
First 3,000	99.18%
First 4,000	99.74%
First 5,000	99.92%
First 6,000	99.98%
All 9,933	100.00%

Let us focus on the top 3,000 characters, some of which represent more than one word each. To determine how many of the words have elastic length, we use the procedure in (9).

(9) Procedure

a. Select the top 3,000 most frequent characters from Da (2004)
b. Randomize the 3,000 characters
c. Select every 50th character, yielding 60 characters in all
d. List all the words each selected character represents
e. Determine how many of the words have elastic length

First, each character was assigned a random number by the Excel function 'randbetween (1, 3000)'. Then the characters were sorted by the random number, and every 50th was selected. For each selected character, the dictionary 《现代汉语词典》 was used to determine the word(s) it represents.

6. RESULT

The result of our analysis is shown in (10). The columns provide information on each selected character, its frequency rank among all characters (smaller numbers indicate higher ranks, based on Da 2004), the gloss of each word a character represents, the short and long forms of each word, and whether the word can be a noun, a verb, or both (part of speech).

(10) Result

Char	Rank	Gloss	Short	Long	POS
安	232	peace	安	安全	N
		[W 书] how can	安		
		amp	安	安培	N
俺	2,749	[D 方] I/we	俺	俺们	
毙	2,534	die violently	毙	毙命, 枪毙	V
测	861	measure	测	测量	N, V
朝	593	court, face	朝	朝代, 朝着	N, V
		morning, day	朝		N

吵	2,040	shout	吵	争吵	V
		[D 方] noisy	吵	吵吵	
呈	1,563	display	呈	呈现, 呈上	V
池	1,709	pond	池	池子, 水池	N
雌	2,382	female	雌	雌性	N
赐	2,072	bestow	赐	赐予	V
摧	2,166	destroy	摧	摧毁	V
俄	975	[W 书] a while	俄	俄顷	
		Russia	俄	俄罗斯	N
俘	2,057	prisoner of war	俘	俘虏, 俘获	N, V
尬	2,729	embarrassed		尴尬	
该	319	should	该	应该	
		owe	该		V
		that	该	该个	
		[W 书] include	该	赅括	
肝	1,760	liver	肝	肝脏	N
阁	1,682	chamber	阁	楼阁	N
诡	2,578	tricky	诡	诡诈	
涵	2,330	contain	涵	涵盖	V
话	170	speak, speech	话	说话, 话语	N, V
及	198	reach	及		V
		and	及	以及	
奖	1,233	award	奖	奖励, 奖品	N, V
棵	2,108	[M 量]	棵		
坑	2,242	pit	坑	土坑, 坑害	N, V
哭	1,210	weep, cry	哭	哭泣	V
烂	1,754	rotten, broken	烂		
力	106	force, power	力	力量, 能力	N
莲	1,837	lotus	莲	荷莲, 莲子	N
鹿	2,056	deer	鹿	鹿子	N
仑	2,139	[W 书] order	仑	伦次	
蜜	2,014	honey	蜜	蜂蜜	N

寞	2,601	lonesome	寞	寂寞	
纳	684	collect	纳	纳入	V
		sew (shoe)	纳		V
偶	1,361	image	偶	偶像	N
		even (number)	偶	偶数	N
		accidental	偶	偶然	
擒	2,850	capture	擒	擒获	V
青	497	green/blue, young	青	青色, 青年	N
却	287	retreat	却	退却	V
		but	却	却是	
晒	2,630	sun bathe	晒	日晒	V
师	333	master	师	老师, 师傅	N
		division (military)	师		N
授	968	give, award	授	授予	V
蜀	2,602	Sichuan	蜀	蜀汉	N
司	278	administer, office	司		N, V
套	1,091	cover, sheath, set	套	套子, 套住	N, V
条	214	strip, item	条	条子, 条目	N
帖	2,892	fit snugly	帖	服帖	V
		card	帖	帖子	N
		model example	帖	帖模	N
		[D 方] prescription	帖		
惟	1,856	only	惟	惟一	
		[W 书] 助词	惟		
		thought	惟	思维	N
嗡	2,783	humming	嗡	嗡嗡	
窝	1,962	nest	窝	鸟窝, 窝窝	N
昔	2,388	past	昔		
闲	1,529	leisure, idle	闲	闲空, 空闲	N
限	613	limit	限	限度, 限制	N, V
新	161	new	新		
徐	1,313	slow	徐	徐徐	

选	499	choose	选	挑选	V
阳	650	sun, yang	阳	太阳, 阳性	N
颐	2,999	[W 书] chin	颐		
		[W 书] nourish	颐	颐养	
鹰	1,927	eagle	鹰	老鹰	N
志	542	will	志	志向	N
		[D 方] weigh	志		
		gazette, mark	志	志记, 标志	N
注	492	inject	注	注入, 注射	V
		comment	注	批注	N, V
钻	1,724	make/enter (a whole)	钻		V
		drill, diamond	钻	钻子, 钻石	N

The characters are listed alphabetically by Pinyin. When a character represents two or more words (entries), each is shown on a separate line. For example, 安 represents three words, shown on three lines. In the Gloss column, one or more senses are listed for each word. In the Gloss column, the label [W 书] indicates whether the word is for written language only, [D 方] indicates whether it is a dialectal term, and [M 量] indicates whether it is a measure word only. For each word, a short form is given if available, and one or more long forms are given if available. Finally, the part of speech (POS) column indicates whether a word is usually a noun, a verb, or both, excluding words that are dialectal or written only.

The percentage of words that have elastic length is shown in (11). It can be seen that about 80% of all words have elastic length, i.e. having both a short and a long form.

(11) Statistical result on elastic word length

	Count	Percent
Long form only	1	1.2%
Short form only	17	20.2%
Both forms available	66	78.6%
Total	84	100%

If we focus on nouns and verbs, the result is shown in (12), excluding words that are dialectal or for written language only.

(12) Elastic word-length in nouns and verbs

	Total	Elastic	% elastic
N	39	36	92%
V	29	24	83%

The result shows that there is a higher than average percentage of word with elastic length in nouns and verbs. The nouns and verbs without elastic length are shown in (13).

(13) Nouns and verbs without elastic word-length

N	朝	morning, day
N	师	division (military)
N	司	office
V	司	administer
V	该	owe
V	及	reach
V	纳	sew (shoe)
V	钻	make/enter (a whole)

Most of the words seem to have restricted usage. For example, 朝 is a less common word for morning (早上, 上午) or day (天, 日), 师 is a specific rank of military units, and 司 is a specific level of government office. Among

verbs, 司 is rarely used in modern Chinese, 该 is a less common (and probably dialectal) word for 欠, 及 is mainly used in 不及 ‘not as much as’ and 及格 ‘pass (exam)’, and 纳 is used for sewing the sole of a shoe. This means that, among commonly uses nouns and verbs, the percentage of words with elastic length would likely be still higher.

7. CONCLUDING REMARKS

In our study, 1/50 of the 3,000 most frequent Chinese characters were examined and the words they represent were determined. Based on the sample, it is found that 80% to 90% of Chinese words (monomorphemic entries in a dictionary) have elastic length, i.e. a monosyllabic form and a disyllabic (or longer) form. In addition, the percentage for verbs is higher than the average, and the percentage for nouns is higher still.

Our study has some theoretical implications. First, a word should not be equated to a character (字), as is sometimes done (e.g. 马建忠 1898). The reason is that most words have both a short form and a long form, and a character (字) only represents the short form.

Second, our study can help explain some problems in defining words in Chinese. In particular, 吕叔湘 (1979, pp. 491-492) notes two problems in applying Bloomfield’s (1926) definition of words to Chinese. The first is a lack of distinction between free and bound morphemes. The second is a lack of distinction between a word and a compound. Our study provides an explanation for both problems. Because Chinese requires a minimal expression to be disyllabic, the monosyllabic form of a word, such as 虎 ‘tiger’, is often not free. However, the disyllabic form of a word, such as ‘tiger’ 老虎, is free. This means that most Chinese words have a free (disyllabic) form and a non-free (monosyllabic) form, hence the lack of distinction between free and bound morphemes. In addition, the long form of a word often looks like a compound, such as 老虎 (literally ‘old tiger’) and 商店 (literally ‘business store’). However, such compounds, or pseudo-compounds, are semantically equivalent to single words (their monosyllabic counterpart), hence the lack of distinction between words and compounds.

Elastic word length can also explain apparent meaning changes in translation. For example, Tamil Tiger (an anti-government organization in Sri Lanka) is translated as 泰米尔猛虎, literally ‘Tamil Fierce Tiger’. Although ‘fierce’ may better describe what the rebels were, tigers are already fierce, and therefore the adjective is largely redundant. A better explanation for the extra word in Chinese, it seems, is the need for a disyllabic form of ‘tiger’ in this prosodic context.

Finally, our study provides a basis for more accurate studies of length patterns, and for the analysis of linguistic constraints. For example, given the percentage of nouns that have elastic length, we can predict the probabilities of 2+2, 2+1, 1+2, and 1+1 in noun-noun compounds (where 1 is a monosyllabic form and 2 a disyllabic one). Then we can compare the predictions against the actual frequencies. If the actual frequency of a pattern is lower than the prediction, there is likely a constraint against the pattern.

Our study has focused on elastic length in Chinese words (词条). It would be interesting to examine elastic length in different senses (词义) of a word. For example, 力 has four senses (excluding the use as a family name): ‘force’, ‘bodily strength’, ‘ability’, and ‘effort’, all of which seem to have a long form: 力量, 力气, 能力, and 努力 respectively. Similarly, 条 has six senses, ‘a twig’, ‘a strip’, ‘a strip shape’, ‘an item or entry’, ‘orderliness’, and a measure word, all of which except the last seem to have a long form: 枝条, 条子, 条形, 条目, and 条理. This topic is left for a separate future study.

REFERENCES

BAAYEN, R. Harald, Richard PIEPENBROCK, and L. GULIKERS. 1995. *The CELEX lexical database: release 2* (CD-ROM). Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

BLOOMFIELD, Leonard. 1926. A set of postulates for the science of language. *Language* 2.3: 153-164.

DA, Jun. 2004. *Chinese text computing*. Murfreesboro: Department of Foreign Languages and Literatures, Middle Tennessee State University. http://lingua.mtsu.edu/chinese-computing/, Accessed September 14, 2011.

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

郭绍虞. 1938. 中国词语之弹性作用. 《燕京学报》 24: 1-34.

HOCKETT, Charles F. I958. *A course in modern linguistics*. New York: Macmillan.

吕叔湘. 1959. 汉语里‘词’的概述. 《语言学问题》 （俄罗斯杂志 *Вопросы Языкознания*）， 1959 年 5 期. Chinese version reprinted in 1990 in《吕叔湘文集 第二卷：汉语语法论文集》, 359-369. 北京: 商务印书馆.

_____. 1962. 说‘自由’和‘粘着’. 《中国语文》 1962.1: 1-6. Reprinted in 1990 in 《吕叔湘文集 第二卷：汉语语法论文集》, 370-384. 北京: 商务印书馆.

_____. 1979. 《汉语语法分析问题》 北京: 商务印书馆.

马建忠. 1898. 《马氏文通》 上海: 商务印书馆.

MCENERY, Tony, and Richard XIAO. 2004. The Lancaster Corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC) 2004*, ed. M. T. LINO, M. F. XAVIER, F. FERREIRE, R. COSTA, & R. SILVA, 1175-1178. Lisbon, May 24-30, 2004.

潘文国. 1997. 《汉英语对比纲要》北京: 北京语言文化大学出版社.

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

王洪君. 2001. 《信息处理用现代汉语分词词表》的内部构造和汉语的结构特点. 《语言文字应用》2001 年 11 月 第 4 期: 90-97.

XIA, Fei. 2000. The segmentation guidelines for the Penn Chinese Treebank (3.0). http://www.cis.upenn.edu/~chinese/ctb.html. Accessed April 3, 2011.

现代汉语常用词表课题组. 2008. 《现代汉语常用词表》 (草案) 北京: 商务印书馆.

中国社会科学院语言研究所词典编辑室. 2002. 《现代汉语词典》 (2002 年增补本) 北京: 商务印书馆.

汉语有多少词的长度有弹性?

端木三

美国密西根大学

提要

词长的弹性指一个词即有一个短的形式，也有一个长的形式，如“煤-煤炭”，“种-种植”，等。本文的目的是确定汉语有多少词有这一特性。我们从汉语最常用三千字所代表的单语素词里，抽出五十分之一进行分析。分析发现，汉语百分之八十到九十的词有弹性长度。分析还发现，有弹性长度的动词的比例高于平均值，而有弹性长度的名词的比例超过动词。

关键词

弹性词长, 汉语词汇, 词界划分, 复合词