Word-length Preferences in Chinese: A Corpus Study

(Running title: **Word-length Preferences in Chinese**)

San Duanmu

November 23, 2011

(*Journal of East Asian Linguistics* 21.1: 89-114, 2012)

Department of Linguistics

University of Michigan

440 Lorch Hall

611 Tappan Street

Ann Arbor, MI 48109-1220

Email: duanmu@umich.edu

Note: some errors occurred in the published version:

Figure 2:	token → type
Figure 3:	[N N] → [V O]
Figure 4:	[N N] → [V O]
	token → type

Word-length Preferences in Chinese: A Corpus Study

Abstract Are there preferred word-length combinations in Chinese? If there are, are they motivated by semantics, syntax, prosody, or a combination of these? While the issue has been discussed for some time, opinions remain divided. This study offers a quantitative analysis of word-length patterns in Chinese [N N] and [V O] sequences, using the Lancaster Corpus of Mandarin Chinese. It is found that 1+2 is overwhelmingly disfavored in [N N] and 2+1 is overwhelmingly disfavored in [V O]. In addition, it is found that apparent exceptions, ranging between 1% and 2%, are limited to certain specific structures, and when these are factored out, both 1+2 [N N] and 2+1 [V O] are well below 1% in either token count or type count. The result bears on several theoretical debates, such as the validity of word-length preferences in Chinese, the motivation of the preferences, the extent and the nature of exceptions, and the interaction among syntax, semantics, and phonology.

Keywords Word-length preferences . Elastic word length . Corpus analysis . Syntax-phonology interaction . Prosody . Metrical structure

1 Introduction: word-length preferences in Chinese

Many words in Chinese can be long (disyllabic) or short (monosyllabic), with more or less similar meanings. Some examples are shown in (1).

(1) Word-length variation in Chinese (1 = monosyllabic, 2 = disyllabic)

2	1	
meitan	mei	‘coal’
xuexi	xue	‘to study’
gongren	gong	‘worker’
shangdian	dian	‘store’
laohu	hu	‘tiger’
Yindu	Yin	‘India’

The two forms of each pair are interchangeable in at least some contexts. For example: ‘coal store’ can be *meitan dian* or *mei dian*, where ‘coal’ can be long or short, with little syntactic or semantic difference. The two forms need not be interchangeable in all contexts. For example, ‘coal mine’ is *mei kuang* and not *meitan kuang*, which shows that the actual usage may be partly dependent on convention.

A remark is in order on what a word is. Bloomfield (1926) defines a word as a minimal free form, where the notion ‘free’ (or being able to occur alone) is not always unambiguous. For example, *the* and *at* are rarely used alone but are treated as words. Similarly, Hockett (1958, p. 167) uses a strenuous example to argue that *out*- in *outside* is a free word, while in Baayan et al. (1995) it is treated as a bound prefix. The issue is problematic in Chinese, too (e.g., Lü 1962, 1990; Xia 2000b; Wang 2001b). For example, monosyllabic forms, such as *hu* ‘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, *meng hu* 'fierce tiger' and *hu shan* '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. We shall therefore continue to call them words. (See Pirani (2008) for a similar view.)

Guo (1938) uses the term 'elastic word length' to refer to the alternation between long and short forms. He also documents various ways this can be achieved, such as reduplication, truncation, and addition of a redundant syllable. Duanmu (2007) uses the term 'dual vocabulary' to refer to the fact that many words have two length forms. It is an interesting question how many Chinese words have elastic length. Pan (1997) suggests that nearly all Chinese words do, but he offers no statistical evidence. Duanmu (2011) sampled 1/100 of the basic Chinese vocabulary and found that that 90% of Chinese words have elastic length.

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

Word-length alternation is not limited to monosyllabic-disyllabic pairs. For example, 'Canada' has a trisyllabic-monosyllabic pair, *Jianada* and *Jia*. In this article I focus on monosyllabic-disyllabic pairs, which are the most common.

The availability of elastic word length creates options of different length patterns. In a two-word expression, there are four possible length patterns, shown in (2).

(2) Four word-length patterns in a two-word expression

2+2 2+1 1+2 1+1

The patterns are not equally preferred. Specifically, it has been reported that in [N N] (noun-noun compounds), 1+2 is disfavored, and in [V O] (verb-object phrases), 2+1 is disfavored (e.g., Lü 1963; Wu 1986; Lu 1990; Lu and Duanmu 2002; Feng 1998; Duanmu 2007). Examples can be seen in (3) and (4); native judgment on the preference patterns is generally robust.

(3) Length preferences in two-word structures: 1+2 is disfavored in **[**N N**]**

2+2	meitan	shangdian	
2+1	meitan	dian	
*1+2	mei	shangdian	
1+1	mei	dian	
	coal	store	‘coal store’

(4) Length preferences in two-word structures: 2+1 is disfavored in **[**V O**]**

2+2	zhongzhi	dasuan	
*2+1	zhongzhi	suan	
1+2	zhong	dasuan	
1+1	zhong	suan	
	plant	garlic	‘plant garlic’

Some scholars believe that word-length preferences are motivated by prosody (e.g., Guo 1938; Feng 1998; Lu and Duanmu 2002; Duanmu 2007). Other scholars believe that syntax or semantics plays a primary role (e.g., Liu 1992; Ke 2007; Zhou 2007). Before we compare previous proposals, let us address a more basic issue, namely, whether the word-length preferences are real. Although the preference patterns have been proposed in several studies, there is so far little quantitative evidence, which leaves a number of questions unresolved, to be discussed next.

2 The need for quantitative study

Several problems have been noted. First of all, there are expressions that do not seem to follow the preferred length patterns. Consider the examples in (5) and (6).

(5) Exceptions to preferred length patterns in [N N]

	1+2	pi	shoutao	‘leather glove’
	1+2	mian	dayi	‘cotton coat’
	1+2	mu	diban	‘wood floor’
Cf.	*1+2	mei	shangdian	‘coal store’

(6) Exceptions to preferred length patterns in [V O]

	2+1	xihuan	qian	‘love money’
	2+1	yanjiu	gui	‘study ghost’
Cf.	*2+1	zhongzhi	suan	‘plant garlic’

Lu and Duanmu (2002) suggest that in the exceptions in (5), the first N serves as an adjective; therefore, such expressions are [A N] rather than [N N]. Duanmu (2007) suggests that, when neither word has flexible length, exceptional length patterns can be used, such as those in (6), because we have no other choice. However, such proposals lack quantitative evidence and questions remain. For example, how often do 1+2 [N N] and 2+1 [V O] occur? How often are they made of words without flexible length? If such exceptions are rare, they might need no further analysis. But if there are many of them, an explanation is needed.

A second problem is that it is unclear whether there are finer differences among the preferred length patterns. In particular, are 2+2, 2+1, and 1+1 equally preferred for [N N]? Are 2+2, 1+2, and 1+1 equally preferred for [V O]? If 2+1 or 1+2 is found to be less common, one might argue for a constraint against trisyllabic feet or against a free syllable (i.e., one outside of a foot). Without quantitative data, it is hard to address such questions.

A third problem is that sometimes both preferred and non-preferred length patterns are found. For example, the word for ‘school’ has flexible length, *xuexiao* and *xiao*, as does the word for ‘store’, *shangdian* and *dian*. Since we have length choices, we might expect that 1+2 is avoided for ‘school store’. However, Ke (2007) cites the examples in (7).

(7) Semantic differences for word-length differences

2+1	xuexiao	dian	(a store serving students and teachers)
1+2	xiao	shangdian	(a store owned by or located at school)
	‘school	store’	

According to Ke, the 2+1 form refers to a store whose main customers are students and teachers. The 1+2 form refers to a store owned by or located at a school, regardless of the customers. It is unclear whether this example is an isolated case, or a general one.

A fourth problem is that some scholars deny a prosodic constraint against 1+2 [N N] altogether (e.g., Wang 2001a; Ke 2007; Zhou 2007). For example, Ke (2007) argues that word-length patterns depend on semantics, where 2+1 is favored when the relation between the two nouns is close and 1+2 is favored when the relation is loose. There is nothing wrong with 1+2 [N N] itself, as long as the semantic or syntactic relation between the words is appropriate. A similar view is held by Zhou (2007). Such proposals expect a sizeable number of 1+2 [N N] expressions to be possible, a prediction yet to be verified.

A fifth problem is that, without quantitative data, we do not have a clear picture of what exceptions there are. Are there other kinds of 1+2 [N N], besides those in (5) and (7)? Are there other kinds of 2+1 [V O], besides those in (6)? It would be useful to know the answers.

Finally, it is useful to consider how to make use of corpus data and how to compensate for the shortcomings of a corpus.

In summary, traditional phonological analyses often draw generalizations from a small number of examples. The persuasiveness of the argument depends on the reader’s subjective judgment and theoretical orientation. In addition, some questions may remain open for a long time. Moreover, certain phonological facts are hard to uncover without quantitative evidence. For example, it used to be thought that native judgment on potential words in English is clear cut (Halle 1962), but quantitative studies found that the judgment is variable and gradient (e.g., Frisch et al. 2000). Quantitative studies, therefore, can provide new impetus for research, resolve some long-standing questions, and even help discover new issues.

Empirical data can be solicited in lab experiments or collected from actual usage. In this study we take the latter approach. Some quantitative studies on word-length preferences have recently appeared. For example, Yang (2005) offers a corpus analysis of different types of trisyllabic words. Kuang (2006) offers a corpus study of 2+1 [V N] and argues that it favors modifier-noun over verb-object. Ke (2007) examines [V O] in a dictionary and finds more 1+2 than 2+1. However, there has been no quantitative study of all patterns of 2+2, 2+1, 1+2, and 1+1 in [N N] and [V O]. This study intends to fill this gap.

3 The corpus

To address our question, we need a natural corpus that is balanced in coverage and reasonably large. In addition, the corpus must be segmented into words and each word labeled for its part-of-speech (POS). The Lancaster Corpus of Mandarin Chinese (LCMC, McEnery and Xiao 2004) meets the requirements and is used for our study. Let us consider some features of LCMC, including its shortcomings.

LCMC contains 1.5 million graphs and over 1.3 million Chinese characters. It is twice the size of the Penn Chinese Treebank (Palmer et al. 2004) but not the largest corpus available. For example, Xiandai Hanyu Changyong Cibiao Ketizu (2008) is based on a corpus of 250 million characters. However, over 135 million of them come from *People's Daily* 2001-2005, the official government newspaper, which may have over-represented one style of usage. Similarly, the Penn Chinese Treebank is heavily based on news texts (e.g., 83% of the articles come from Xinhua News Agency). In contrast, LCMC consists of fifteen text styles, from formal to colloquial on a wide range of topics, and offers a more balanced representation of modern usage.

The basic statistics of LCMC are shown in (8).

(8) Overall statistics of LCMC (token counts)

Words	839,006	(84%)
Punctuations	162,820	(16%)
Chinese characters	1,341,628	(89%)
Punctuation graphs	166,861	(11%)
All graphs	1,508,489	

Word segmentation and POS labeling are controversial in Chinese (see Lü (1990) and Xia (2000a, 2000b) for some discussion). One difficulty is the lack of a clear distinction between a free word and a bound morpheme. Another difficulty is the lack of a clear distinction between a word and a compound or phrase. For example, *ji* ‘chicken’ is a free word in Chinese but *ya* ‘duck’ is not quite so, because it is often used in a disyllabic form *yazi*. Now if we treat *ji dan* ‘chicken egg’ as a compound but *ya dan* ‘duck egg’ as a single word, we lose a structural parallel. Similarly, it is controversial whether *you qian* ‘have money’ is a word, a phrase, or sometimes a word and sometimes a phrase. In response, compilers of Chinese lexicons have settled on a few rules. A common practice is to treat an expression as a word if it occurs frequently, even if its constituents are free and the meaning is transparent (e.g., Xia 2000b; Wang 2001b). Thus, *ji dan* ‘chicken egg’ is usually segmented as a word, as is *you qian* ‘have money’. This practice heavily favors disyllabic words. For example, in LCMC, over 40% of all word tokens, and about 70% of all word types, are disyllabic.

Word segmentation and POS labeling are time consuming. LCMC used a tool to do so automatically. The tool is called the Chinese Lexical Analysis System, developed by the Institute of Computing Technology, Chinese Academy of Sciences. The result was then hand checked by the authors of LCMC, which “improved the annotation precision to over 98%” (McEnery and Xiao 2004). Some decisions are debatable though. For example, *da dianhua* ‘hit telephone (make phone call)’ is segmented as one word (a trisyllabic verb) while *da dianbao* ‘hit telegram (send telegram)’ is segmented as two words ([V N]), apparently because the former expression is frequent and the latter is not. Similarly, *Beijing Daxue* ‘Peking University’ is segmented as one word but *Zhijiang Daxue* ‘Zhejiang University’ is segmented as two, apparently because the former occurs more frequently.

LCMC offers no information on syntactic bracketing. For example, given two words A and B, LCMC does not indicate whether they form a constituent, i.e., whether their syntactic relation is [A B], [A [B…]], [[…A] B], or […A][B…]. We shall return to this issue below.

4 Analysis of [N N]

LCMC contains fifteen text styles. The largest two are ‘biographies and essays’ and ‘science: academic prose’. An examination shows that the former has a more neutral style and yields similar results as the entire LCMC. In this section, therefore, we examine [N N] and [V O] in ‘biographies and essays’.

4.1 Procedure

The procedure for analyzing [N N] is outlined in (9) and explained below.

(9) Procedure in analyzing [N N]

a. Defining N

b. Defining [N N]

c. Counting tokens and types of 2+2, 2+1, 1+2, and 1+1

d. Estimating error rate and making adjustment

e. Supplemental counts

We begin by asking what counts as N (a noun). LCMC uses a list of fifty POS labels (in contrast, the Penn Chinese Treebank uses thirty-three POS labels; see Xia 2000a). Twenty-one of the labels have various degrees of noun-likeness, shown in (10).

(10) Noun-like POS labels in the corpus ‘biographies and essays’

POS	Count	Description
an	494	adjective with nominal function
f, fg	2,957	direction/location (word or morpheme)
j	807	abbreviation
m, mg	5,287	numeral (word or morpheme)
n, ng	24,772	common noun (word or morpheme)
nr	7,279	personal name
ns	2,452	place name
nt	92	organization name

nx	57	letter string (mostly English words)
nz	294	other proper noun
q, qg	3,292	classifier (word or morpheme)
r, rg	9,241	pronoun (word or morpheme)
s	894	space word
t, tg	2,282	time (word or morpheme)
vn	2,581	verb with nominal function
All	62,781	

We limit our definition of N to common nouns, the largest category, plus words for location, direction, and space. We exclude time words (e.g., ‘morning’, ‘afternoon’, ‘then’, ‘now’, ‘today’, ‘Monday’, ‘that time’, and various numbers of years, days, months, or hours), which often occur as adverbials. We also exclude proper nouns (e.g. names of people, places, and organizations), which often do not have elastic word length and which some dictionaries have not collected. We also exclude English words, nominalized verbs, nominalized adjectives, and pronouns. Our definition is summarized in (11).

(11) Defining N

POS labels	Tokens
n, ng, f, fg, s	28,623

The length of N can range from one to several syllables. For example, *dan* ‘egg’ is a monosyllabic N, *deng zhao* ‘lamp shade’ is a disyllabic N, and *ladin zimu* ‘Latin letter’ is a four-syllable N. Sometimes, it is possible to split an N into two (or more) words, such as *ji dan* ‘chicken egg’, *xiao zu* ‘small group’, and *ladin zimu* ‘Latin letter’. Other times an N is not decomposable, such as *jueding* ‘decision’, *zuzhi* ‘organization’, and *mishu* ‘secretary’. We shall return to this issue.

Next, we define [N N], which is made of two adjacent N’s, shown in (12), where each N is as defined above.

(12) Defining [N N]:

Two N’s that are adjacent (not separated by punctuation)

The counting of tokens and types of various [N N] was done in Excel, where some short programing codes were used.

Since LCMC offers no annotation of syntactic bracketing, our definition of [N N] will yield some errors. Some examples are shown in (13).

(13) Errors in extracting [N N]

	Structure	Example	Gloss
Correct	[N N]	[yinyue jiaoshi]	‘music teacher’
Error	[N [N…]]	[guojia [banquan ju]]	‘state copyright bureau’
Error	[[…N] N]	[[xuesheng shitang] fujin]	‘student cafeteria vicinity’
Error	[…N][N…]	[zhezhi budui][yaoqiu yan]	‘this army requirement strict’

Here, ‘state copyright’ is an error from ‘[state [copyright bureau]]’, ‘cafeteria vicinity’ is an error from ‘[[student cafeteria] vicinity]’, and ‘army requirement’ is an error from ‘[[this army][requirement strict]]’ (i.e. ‘for this army, the requirement is strict’).

To compensate for such errors, a manual examination was performed on at least 10% of all [N N] tokens. The error rate for each length pattern was then determined and the frequency of each pattern adjusted. Besides errors involving syntactic boundaries, there are some that involve POS. For example, *xia xueqi* ‘next semester’ should be [A N] rather than [N N]. Such errors are uncovered by the manual examination, too.

In the final procedure ‘supplemental counts’, we examine [N N] units that are segmented as single nouns. As mentioned above, LCMC segments a polysyllabic expression as a single word if it occurs frequently, even if the expression is syntactically decomposable. The overall information on polysyllabic nouns is given in (14).

(14) Polysyllabic N’s in ‘biographies and essays’

Length	Tokens	Types
5-8 syllables	29	17
4 syllables	268	121
3 syllables	1,904	846
2 syllables	17,141	4,451
All	19,342	5,830

Nouns longer than four syllables are not many and are ignored. A 2-syllable N could be 1+1 [N N], a 3-syllable N could be 1+2 or 2+1 [N N], and a 4-syllable N could be 2+2, 1+3, or 3+1 [N N], where [N N] is as defined above. To obtain the additional counts, we manually examined all 4-syllable nouns (268 tokens), 10% of all 3-syllable nouns (190 tokens), and 1% of all 2-syllable nouns (171 tokens). Some examples are shown in (15).

(15) Supplemental counts: polysyllabic N's that are [N N]

Case	Example	Gloss
2-syllable N as 1+1 [N N]	ji dan	'chicken egg'
3-syllable N as 1+2 [N N]	dang zhibu	'party branch'
3-syllable N as 2+1 [N N]	shangye qu	'business district'
4-syllable N as 2+2 [N N]	lingdao banzi	'leadership team'
4-syllable N as 1+3 [N N]	(Not found)	
4-syllable N as 3+1 [N N]	jingjixue jia	'economics scholar'

If a polysyllabic N is not [N N], it is not included. Some examples are shown in (16). The free translation is given in parentheses, if it is not obvious from the literal translation.

(16) Examples of polysyllabic N's that are not [N N]

Length	Example	Gloss	Note
4-syll.	zhiming renshi	'famous person'	[A N]
3-syll.	zong shuji	'general secretary'	[A N]
3-syll.	shiyong zhe	'use person (user)'	[V N]
2-syll.	chao ren	'super person (superman)'	[A N]
2-syll.	fei ji	'fly machine (airplane)'	[V N]
2-syll.	di di	'brother-brother (brother)'	
2-syll.	peng you	'company friend (friend)'	

Many of them are [A N] or [V N], where A is an adjective and V is a (nominalized) verb. Some are reduplications, such as ‘brother’. Some are pseudo-compounds, i.e., the disyllabic equivalent of a monosyllabic word, such as ‘friend’.

The supplemental counts deal only with words not already covered by 2+2, 2+1, 1+2, and 1+1 [N N]. This is true for three-syllable and four-syllable nouns, since no word in 2+2, 2+1, 1+2, and 1+1 is longer than two syllables. In the supplemental count of two-syllable nouns, we have excluded those that already occur in 2+2, 2+1, and 1+2 [N N].

4.2 Result

The result of 2+2, 2+1, 1+2, and 1+1 [N N], in token counts, is shown in (17) and Figure 1. Other length patterns, such as 3+3, 3+2, 4+3, 4+2, etc., total about 15% and are not shown. The error rate is based on a manual check of 10% of the tokens of 2+2, 2+1, and 1+1 and all tokens of 1+2. It can be seen that the error rate for 1+2 is much higher than those for other patterns. An inspection showed that many raw units of 1+2 [N N] were misanalyses of [… N] N. For example, *hu cunmin* ‘family (of) villagers’ was a misanalysis of *[yi hu] cunmin* ‘[one family] (of) villagers’; *jie yidai* ‘street area’ was a misanalysis of *[48 Jie] yidai* ‘[48th Street] area’; and *ren fengfan* ‘person style’ was a misanalysis of *[Tang ren] fengfan* ‘[Tang person] style’. This shows that a 1+2 pair of nouns rarely form a syntactic unit.

(17) Token counts of 2+2, 2+1, 1+2, and 1+2 [N N] compounds

Pattern	Raw	Errors	Corrected	Supplemental	Final	%
2+2	1,807	28%	1,307	162	1,469	21.8%
2+1	1,202	40%	725	711	1,436	21.3%
1+2	409	92%	31	41	72	1.1%
1+1	679	40%	410	3,346	3,756	55.8%
All	4,097		2,473	4,260	6,733	100%

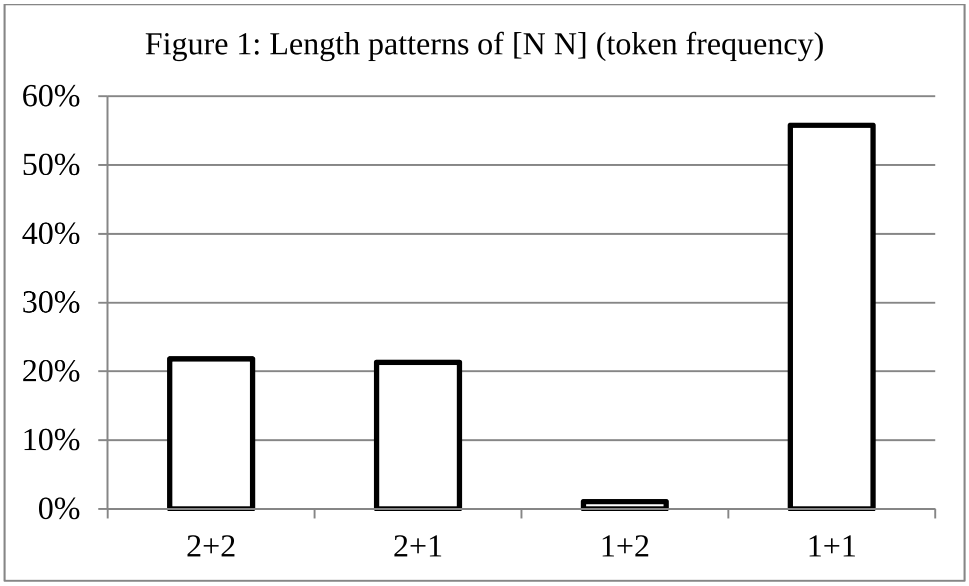


Figure 1: Length patterns of [N N] (token frequency)

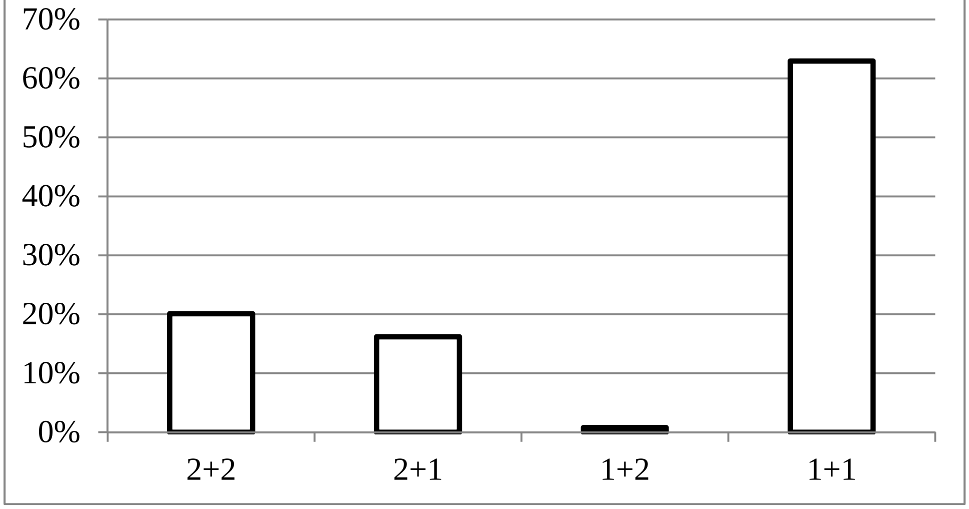
If word length is completely flexible, 2+1 and 1+2 should occur at similar frequencies. The result shows that 1+2 [N N] is clearly disfavored, which constitutes about 1% of all cases; we shall take a close look at examples of 1+2 [N N] below. We also see the prevalence of 1+1 [N N]. Finally, there is no obvious frequency difference between 2+2 and 2+1.

The above data show token counts, where an expression is counted as many times as it occurs. Let us now consider type counts, where each expression is counted just once, regardless of how many times it occurs. It has been argued that type frequency has a greater influence on grammatical patterns (Bybee 2006; Richtsmeier 2011). The result is shown in (18) and Figure 2. As noted above, the error rate for raw units of 1+2 [N N] is much higher.

(18) Type counts of 2+2, 2+1, 1+2, and 1+2 [N N] compounds

Pattern	Raw	Errors	Corrected	Supplemental	Final	%
2+2	1,516	28%	1,089	71	1,160	20.1%
2+1	860	53%	401	533	934	16.2%
1+2	367	93%	25	20	45	0.8%
1+1	504	33%	340	3,296	3,636	63.0%
All	3,247		1,855	3,920	5,775	100%

Figure 2: Length patterns of [N N] (type frequency)



In the present data, only about 15% of the tokens are repeated. Therefore, the result of type count is similar to that of token count. There is a slightly higher rate of repetition in 2+1 and hence a slightly lower type count of it, but not by a whole lot. We see still that 1+2 is clearly disfavored and 1+1 [N N] dominates.

5 Analysis of [V O]

As in the analysis of [N N], we examine [V O] in ‘biographies and essays’. Since we are looking at two-word units, we limit our analysis of [V O] to [V N], where the object is a single noun.

5.1 Procedure

The procedure in analyzing [V N] is similar to that in the analysis of [N N], outlined in (19) and explained below.

(19) Procedure in analyzing [V N]

a. Defining V and N

b. Defining [V N]

c. Counting tokens and types of 2+2, 2+1, 1+2, and 1+1

d. Estimating error rate and making adjustment

e. Supplemental counts

V and N are defined in (20), where N is the same as in the analysis of [N N] and V includes the LCMC POS labels v (verb, 27,519 tokens) and vg (verb morpheme, 490 tokens).

(20) Defining V and N

	POS labels	Tokens
V	v, vg	28,009
N	n, ng, f, fg, s	28,623

We define [V N] as a V followed by an N without a space or punctuation in between, given in (21).

(21) Defining [V N]:

[V N] is an adjacent VN pair (not separated by punctuation).

The counting of tokens and types of various [V N] was again done in Excel, where some short programing codes were used.

As in the case of [N N], our definition of [V N] will yield some errors, mostly involving syntactic bracketing. Some examples are shown in (22).

(22) Errors in extracting [V N]

	Structure	Example	Gloss
Correct	[V N]	[qing ke]	‘invite guest’
Error	[V [N…]]	[you [yishu tiancai]]	‘have art talent’
Error	[V [N…]]	[weihu [shijie heping]]	‘protect world peace’
Error	[[…V] N]	[[yi dong] shuhua]	‘mind control painting’

Here, ‘have art’ is an error from ‘[have [art talent]]’, ‘protect world’ is an error from ‘[protect [world peace]]’, and ‘control painting’ is an error from ‘[[mind-control] painting]’. There are also some errors with POS labels, owing to lack of inflection in Chinese. For example, *chuban gongsi* was labeled as [V N] ‘publish company’, while it should be ‘publication company’, where the first word is a nominalized verb.

To compensate for such errors, a manual examination was performed on at least 10% of the tokens of each length pattern. The error rate for each length pattern was then determined and the frequency of each pattern adjusted.

In the final procedure ‘supplemental counts’, we examine [V N] units that are segmented as single verbs. The information on polysyllabic verbs is given in (23). There are no verbs longer than four syllables.

(23) Polysyllabic V’s in ‘biographies and essays’

Length	Tokens	Types
4 syllables	8	8
3 syllables	175	89
2 syllables	13,153	3,769
All	28,009	4,904

A 2-syllable V could be 1+1 [V N], a 3-syllable V could be 1+2 or 2+1 [V N], and a 4-syllable V could be 2+2, 1+3, or 3+1 [V N]. To obtain the additional counts, we manually examined all 4-syllable and 3-syllable verbs (183 tokens) and 1% of all 2-syllable verbs (130 tokens). Some examples are shown in (24).

(24) Supplemental counts: polysyllabic V’s that are [V N]

Case	Example	Gloss
2-syllable V as 1+1 [V N]	shi ming	‘lose sight’
3-syllable V as 1+2 [V N]	diu mianzi	‘lose face’
3-syllable V as 2+1 [V N]	diandian tou	‘nod head’

If a polysyllabic V is not [V N], it is not included. Some examples are shown in (25).

(25) Examples of polysyllabic V’s that are not [V N]

Length	Example	Gloss	Note
4-syllables	lailai wangwang	‘come go’	reduplication
3-syllables	bu ya yu	‘no less than’	phrase
3-syllables	zhengchanghua	‘normalize’	derived verb
2-syllables	da po	‘hit break’	verb compound
2-syllables	gong sheng	‘co-grow’	derived verb
2-syllables	paipai	‘pat’	reduplication

The supplemental counts deal only with words not already covered by 2+2, 2+1, 1+2, and 1+1 [V N]. In particular, in the supplemental count of 2-syllable verbs, we have excluded those that already occur in 2+2, 2+1, and 1+2 [V N].

5.2 Result

The result of 2+2, 2+1, 1+2, and 1+1 [V N], in token counts, is shown in (26) and Figure 3.

Other length patterns, such as 3+3, 3+2, 4+3, 4+2, etc., total about 15% and are not shown. The error rate is based on a manual check of 10% of the tokens of 2+2, 1+2, and 1+1 and all tokens of 2+1. It can be seen that the error rate for 2+1 is much higher than those for other length patterns. An inspection showed that most raw units of 2+1 [V N] were misanalyses of the original syntax. For example, *yiwei ren* 'thought person' was a misanalysis of *yiwei [ren wanle]* 'thought [(the) person died]'; in *jinlai shi* 'enter time', 'time' is not the object of V but the head N of a relative clause, i.e. 'the time (someone) enters (the room)'; and in *konggu quan* 'controlling right', the syntax is not verb-object, but modifier-noun. This shows that a 2+1 pair of VN rarely form a syntactic unit.

(26) Token counts of 2+2, 2+1, 1+2, and 1+2 [V N] compounds

Pattern	Raw	Errors	Corrected	Supplemental	Final	%
2+2	1,325	46%	711	0	711	16.2%
2+1	508	88%	63	18	81	1.8%
1+2	1,479	46%	798	40	838	19.1%
1+1	1,020	43%	580	2,169	2,749	62.8%
All	4,332		2,152	2,227	4,379	100%

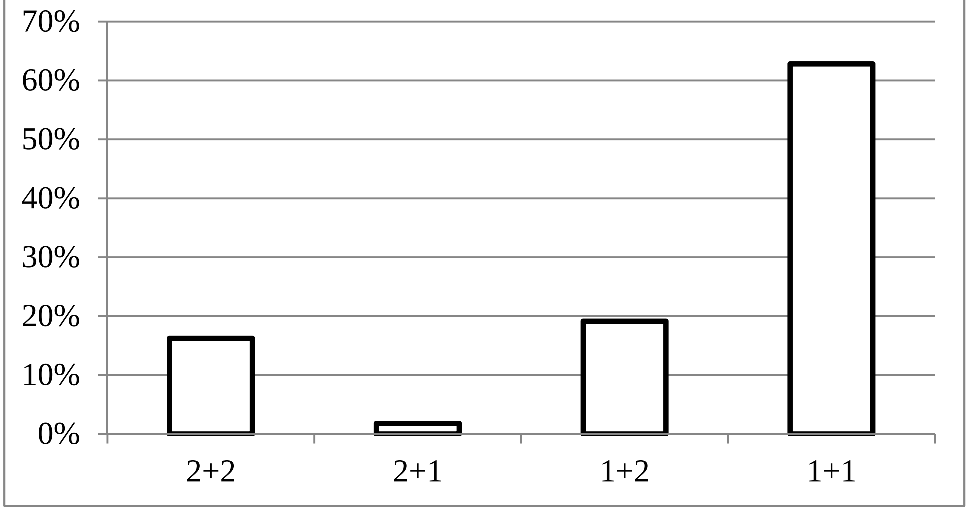
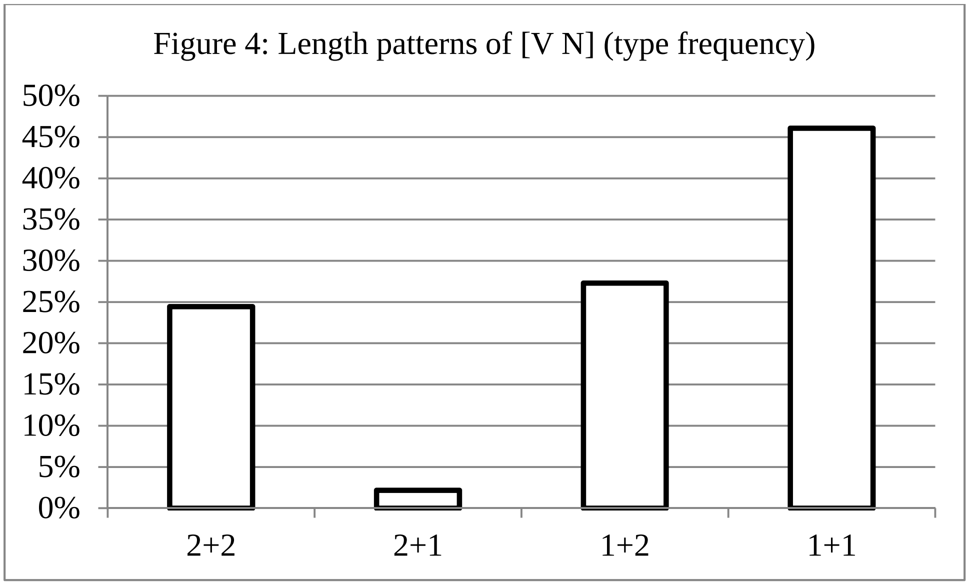


Figure 3: Length patterns of [V N] (token frequency)

Again, if word length is freely variable, we would expect to see as many 2+1 tokens as 1+2 tokens. The result shows that 2+1 [V N] is clearly disfavored, which constitutes about 2% of all cases; we shall take a close look at examples of 2+1 [V N] below. In addition, we see the prevalence of 1+1 [V N]. Finally, there are slightly fewer 2+2 tokens than 2+1 tokens. The type counts of various [V N] patterns are shown in (27) and Figure 4. We note again that the error rate for 1+2 is much higher than for other length patterns.

(27) Type counts of 2+2, 2+1, 1+2, and 1+2 [V N] compounds

Pattern	Raw	Errors	Corrected	Supplemental	Final	%
2+2	1,252	50%	630	0	630	24.5%
2+1	420	89%	47	9	56	2.2%
1+2	1,326	49%	676	27	703	27.3%
1+1	803	49%	406	781	1,187	46.1%
All	3,801		1,760	817	2,577	100%



Again, the repetition rate is fairly low. Therefore, the result of type count is similar to that of token count. In both cases, 1+2 slightly exceeds 2+2, 2+1 (at about 2%) is clearly disfavored, and 1+1 dominates others.

6 A look at the exceptions

In this section, we examine expressions whose length form is disfavored, i.e. 1+2 [N N] and 2+1 [V O]. (See Appendix for the full list.) Since the number of such expressions is small, ranging from 1% to 2%, it is possible that they are true exceptions and need no further explanation. On the other hand, it is also possible that there are further regularities that underlie the exceptions. Given the lack of previous studies in this area, the discussion of this section will be tentative.

6.1 Exceptional [N N]: 1+2 [N N]

There are altogether forty-five 1+2 [N N] compounds in the corpus ‘biographies and essays’, excluding repeated tokens. Most of them seem to fall into one of two structures. In the first, the first N is adjectival in nature, a point observed in Lu and Duanmu (2002). In the second, the first N is semantically the possessor of the second N (e.g., ‘party branch’ means ‘party’s branch’). This is summarized in (28).

(28) Structures of 1+2 [N N] (type count)

	Count (%)	Example	Gloss
N1 is adjectival	19 (42%)	er jiangjun	‘baby general’
		ming jizhe	‘famous reporter’
		he wuqi	‘nuclear weapon’
		gui difang	‘ghostly place’
N1 is possessor	23 (51%)	dang zhibu	‘party branch’
		xian yiyuan	‘county hospital’
		sheng zhengfu	‘province government’
		xi zhuren	‘department head’
Others	3 (7%)	zhu siliao	‘pig feed’
		guan huzi	‘officer beard’
		shi kuzi	‘feces pants’
All	45		

Since Chinese lacks overt inflection, some words can be interpreted as nouns or adjectives. For example, *gui difang* can be translated as ‘ghost place’ or ‘ghostly place’. Similarly, *ming jizhe* can be ‘fame reporter (reporter of some fame)’ or ‘famous reporter’. In 42% of the exceptions, the first N can be interpreted as an adjective of this kind. In another 51% of the exceptions, the first N is a possessor, which can be interpreted as an adjective, too. For example, *sheng zhengfu* can be translated as ‘province government’ or ‘provincial government’ and *xi zhuren* can be translated as ‘department head’ or ‘departmental head’. Some words, such as ‘party’ and ‘county’, do not have an adjective form in English, but they could be seen as adjectives with zero derivation (an empty suffix), in parallel to words like ‘province’ and ‘department’. If so, over 90% of the exceptions are [A N], rather than [N N], and true cases of 1+2 [N N] become diminishingly rare. It is an interesting question why [A N] tolerates 1+2 but [N N] does not. Some discussion is offered in Duanmu (2007). In addition, we shall review a proposal by Zhou (2007) below.

Feng (1997, pp. 19-20) proposes that the exceptional 1+2 [N N] pattern is limited to two structures: either the first N is a prefix, as in *nan yanyuan* ‘male actor’, or the final syllable is unstressed, as in *shu daizi* ‘book worm’. What Feng calls a prefix is similar to what we consider to be an adjective. Our analysis is preferable, as the first word can be material, such as ‘iron’ and ‘wood’, which are content words and not prefixes. (See Wang and Fu (2005) on difficulties in defining a prefix in Chinese.) There are two other shortcomings in Feng’s analysis. First, in our corpus, only a few cases of 1+2 [N N] have an unstressed final syllable. Second, Feng fails to mention the possessor-possessed structure, which accounts for half of all cases of 1+2 [N N].

6.2 Exceptional [V O]: 2+1 [V O]

There are fifty-six 2+1 [V N] units in the corpus ‘biographies and essays’, excluding repeated tokens. The most striking generalization is that forty-nine of them, or 88%, can be related to, or seen as derive from, 1+1 [V O]. Let us call them ‘derived 2+1 [V O]’. Some examples are shown in (29), where the extra syllable in the 2+1 form is shown in parentheses.

(29) Derived 2+1 [V O], covering 88% or all 2+1 [V O]

2+1	1+1	Gloss
ju(qi) shou	ju shou	‘raise (rise) hand’
shen(chu) shou	shen shou	‘stretch (exit) hand’
zhao(zhao) shou	zhao shou	‘beckon (beckon) hand’
bai(bai) shou	bai shou	‘wave (wave) hand’
xi(xi) shou	xi shou	‘wash (wash) hand’

The extra syllable in the 2+1 form could be a copy of the verb or a different morpheme for syntactic or morphological purposes. In fact, a wide range of modifications to the verb can be made. In (30) we show some derived 2+1 forms, based on 1+1 *hui jia* ‘return home’. Some of the derived forms are not found in our corpus but all of them are quite acceptable.

(30) Derived 2+1 [V O] from 1+1 [V O]: *hui jia* ‘return home’

mei-hui jia	‘didn’t-return home’
hui-guo jia	‘return-pass home’
hui-hui jia	‘return-return home’
bai-hui jia	‘useless-return home’
bu-hui jia	‘not-return home’
zou-hui jia	‘walk-return home’
pao-hui jia	‘run-return home’

It is not entirely clear why derived 2+1 [V O] forms cause no problem for native judgments. Some of them seem to be [1+[1+1]], instead of 2+1, such as *bu hui jia* ‘[not [return home]]’ and *bai hui jia* ‘[uselessly [return home]]’. Could other forms be analyzed in the same way? For example, could *zou hui jia* be analyzed as ‘[walk [return home]]’, i.e., ‘to return home by walking’? If so, many derived 2+1 forms are [1+[1+1]] instead of 2+1 and true 2+1 [V O] forms are far rarer.

There are seven 2+1 [V O] forms that do not seem to be derived from 1+1. They are shown in (31). They constitute 0.2% of all [V O] forms.

(31) 2+1 [V O] expressions not derived from 1+1

xihuan shi	‘like poetry’
xihuan qiang	‘like gun’
gesong dang	‘praise party’
chengli dang	‘establish party’
choujian kuang	‘plan-construct mine’
duo yu ren	‘more than people’
duo yu bing	‘more than soldier’

My own judgment is that the first five expressions are not nearly as good as those derived from 1+1. The last two expressions may not be [V O]; the verb seems to be an empty copula that precedes ‘more’, equivalent to ‘be’ in English.

Feng (2000, p. 119) suggests that 2+1 [V O] occurs only if the second syllable of the verb is unstressed (in Beijing Mandarin). In our data, about 20%-30% of 2+1 [V O] involve an unstressed second syllable of V (mostly when the verb is reduplicated). This means that Feng’s proposal can only account for a small portion of the data. In contrast, the notion of ‘derived 2+1 [V O]’, which covers nearly 90% of the cases, is a better generalization.

6.3 Summary

We have seen that exceptional length patterns are limited to a few specific structures. In [N N], 1+2 is mostly found when the first N is adjectival in nature, or when the relation between the words is possessor-possessed, in which the possessor can also be seen as an adjective. In [V O], 2+1 is mostly found when it is derived from 1+1.

7 Theoretical implications

The present study provides quantitative evidence that 1+2 [N N] and 2+1 [V O] are disfavored length patterns. Their frequencies of occurrence, in either token count or type count, are well below expected values, should word length be freely variable. In addition, exceptional cases are limited to a few specific structures. The rarity of 1+2 [N N] and 2+1 [V O] calls for an explanation. Let us consider some proposals.

7.1 Phonology-based analyses

A phonology-based analysis is one that mainly appeals to phonological requirements or constraints. Of course, some phonological requirements, such as the cyclic assignment of stress, make reference to syntax. Therefore, few analyses are purely phonological. In this section we discuss analyses that are mainly based on phonological requirements.

Several studies have proposed that word-length preferences are motivated by prosody. For example, Guo (1938) suggests that the long form is used where speech needs to be slow and the short form is used when speech needs to be fast, although he does not explain when speech needs to be slow and when it needs to be fast. Feng (1998) suggests that feet are built from left to right in [N N] and from right to left in [V O], which ultimately leads to different preference patterns in the two structures. Lu and Duanmu (2002) and Duanmu (2007) offer a more specific analysis in terms of stress assignment and foot structure. The analysis assumes four metrical requirements, shown in (32).

(32) Metrical requirements (S = syllable, boldface = phrasal stress)

a. Foot binary: A foot needs two syllables, i.e. (SS).

b. Stress assignment: [**N** N] and [V **O**]

c. Cyclicity: stress is assigned cyclically.

d. Every stress represents a foot.

Foot binarity (Prince 1980) is well known in metrical literature. The proposed stress assignment is similar to that in English, where main stress in [N N] goes to the first N and that in [V O] goes to O (Chomsky et al. 1956; Chomsky and Halle 1968). It is worth asking why compounds and phrases have opposite directions of stress. In Duanmu (1990), the two stress rules are unified under the notion Non-head Stress, where the syntactic non-head received stress. In other words, in any syntactic relation [X XP], XP is stressed. Similar ideas have been proposed by Gussenhoven (1983), Cinque (1993), Zubizarreta (1998), Truckenbrodt (2005), and Zubizarreta and Vergnaud (2006). Duanmu (2007) attributes Non-head Stress to a deeper reason, which is called the Information-Stress Principle, by which the syntactic constituent that has more information gets stress. Information in turn is defined by probability of occurrence (Shannon 1948). In [X XP], X is a word level unit, while XP is a phrase level unit. Since there are more phrases than words, the probability of each XP is lower (i.e. more unpredictable) than that of X, and hence each XP on average carries more information than an average X.

Cyclicity is another well-known requirement in metrical literature, according to which stress is assigned to smaller syntactic units first and then to larger ones (e.g., Chomsky et al. 1956; Chomsky and Halle 1968; Shih 1986; Chen 2000). Finally, stress is defined as the head of a foot, so that every stress implies a foot and vice versa (Halle and Vergnaud 1987; Hayes 1995). The analysis of [N N] is shown in (33).

(33) Metrical structures for [N N] (offending foot underlined)

2+2	(SS)(SS)
2+1	(SS)S
*1+2	*(**S**)(SS)
1+1	(SS)

In 2+2, each word forms a binary foot. In 1+1, the two words can form one binary foot, similar to a disyllabic word. In 2+1, the first word forms a binary foot; the second word need not form a foot, since it does not have compound stress. In 1+2, the second word forms a binary foot; the first word has compound stress and so must also form a foot, creating a monosyllabic foot, which is metrically ill formed. The analysis of [V O] is similar, except that main stress now falls on the second word instead of the first. This is shown in (34).

(34) Metrical structures for [V O] (offending foot underlined)

2+2	(SS)(SS)
*2+1	*(SS)(**S**)
1+2	S(SS)
1+1	(SS)

In 2+2, each word forms a binary foot. In 1+1, the two words can form one binary foot and serve as a compound (which often happens, as the word segmentation of LCMC confirms). In 1+2, the second word forms a binary foot; the first word need not form a foot, since it does not have phrasal stress. In 2+1, the first word forms a binary foot; the second word has phrasal stress and so must also form a foot, creating a monosyllabic foot, which is metrically ill formed.

The phonology-based analysis has two advantages. First, it can be seen as the null hypothesis, since it makes no special assumptions beyond familiar phonological requirements. Second, it covers a very wide range of cases. In our data, it covers at least 98% of all relevant patterns. Critics of the phonology-based analysis (e.g., Ke 2007; Zhou 2007) have focused on some exceptional patterns. The present study shows that the exceptional patterns are clearly disfavored and must be subject to some strong restrictions.

Among the good patterns, there is no obvious frequency difference between 2+2 [N N] and 2+1 [N N], or between 2+2 [V O] and 1+2 [V O]. This means that a trisyllabic unit is not by itself a problem in phonology. In other words, either trisyllabic feet can be formed (e.g., Shih 1986; Feng 1997, 1998; Chen 2000; Davis 2005), or some monosyllables can remain free, contrary to the prosodic licensing principle of Ito (1986). We note, however, that not any trisyllabic unit is acceptable. Problems occur when the syntactic non-head is a monosyllable, i.e., when the first N in [N N] or the object in [V O] is a monosyllable. In the metrical analysis, this occurs when the monosyllable is stressed and cannot form a disyllabic foot.

7.2 Syntax- or semantics-based analyses

A syntax- or semantics-based analysis is one that mainly appeals to syntactic or semantic requirements or constraints, even though some of the requirements may refer to phonology in some way.

Liu (1992) proposes that verbs should be monosyllabic and nouns should be disyllabic. The reason for the correlation is not fully explained, but it seems true statistically for English and Chinese. Given the requirement, Liu would predict 1+2 to be the most common length pattern of [V O] and 2+1 the least common. The prediction is only partly correct: while 2+1 [V O] is indeed uncommon, 2+2 and 1+1 are also common patterns for [V O] besides 1+2, which Liu does not expect. In addition, Liu expects [N N] to favor 2+2, which is not always the case. In particular, Liu cannot explain the frequency difference between 1+2 and 2+1.

Ke (2007) suggests that length patterns are determined by the closeness of semantic and prosodic relations. If the semantic relation between two words is close, we should choose 2+1, which is prosodically close. If the semantic relation is loose, we should choose 1+2, which is prosodically loose. In Ke’s analysis, there is nothing against 1+2 [N N] other than semantics: some meanings favor 1+2 and some 2+1. Ke’s analysis is not the null hypothesis, because it needs a separate theory for ranking semantic closeness, and one for ranking prosodic closeness. In addition, Ke fails to offer quantitative predictions, in particular why those semantic relations that call for 1+2 are so rare.

Zhou (2007) proposes that word length is determined by information load: the more information a word has, the longer the word should be. Information load in turn is based on word meaning, or the number of potential contrasts in a semantic category. For example, the semantic category ‘size’ involves only a few contrasts, and a word that refers to size has a low information load. In contrast, the semantic category ‘location’ involves more contrasts, and a word that refers to location has a higher information load. Zhou divides the information load of the modifier of N into three levels, shown in (35).

(35) Information, meaning, and word length for the modifier of N (Zhou 2007, p. 217)

Information	Meaning	Word length
Low	old/new, size, color, shape, smell	1 syllable
Medium	property, time, location, material	1 or 2 syllables
High	Function	2 syllables

However, Zhou offers no quantitative predictions of each length pattern, in particular why 1+2 [N N] is so rare. In addition, we have seen that many cases of 1+2 [N N] involve the possessor-possessed structure. In Zhou’s analysis, it is unclear to which semantic category possessors belong, and why possessors have low information load.

7.3 Interactions among syntax, semantics, and phonology

The word-length problem in Chinese involves phonology (i.e., syllable count and foot structure), syntax (e.g. [N N] vs. [V O]), and possibly semantics (e.g. whether the first N in [N N] indicates a property or the possessor of the second N). Naturally, different analyses have been proposed that appeal to syntactic requirements (e.g., Liu 1992), semantic requirements (e.g., Ke 2007; Zhou 2007), and phonological constraints (e.g., Duanmu 2007). I have argued that the phonology-based analysis seems the simplest and makes the best predictions.

Anttila et al. (2010) discusses a similar problem in English, traditionally known as heavy-NP shift, which also involves both phonology and syntax. Unlike some previous analyses, which appeal to semantic and pragmatic factors, Anttila et al. demonstrates that a phonology-based analysis works well and can make quantitative predictions correctly. The present study agrees with Anttila et al. in favoring a phonology-based analysis.

In the Chinese case, the distinction between good and bad structures is fairly clear cut. Therefore, the analysis is also fairly simple. In contrast, the data on heavy-NP shift are more variable and the analysis by Anttila et al. (2010) uses more intricate constraints. For example, *TERNARY prohibits a prosodic phrase from containing three units, where each unit can be a foot or a syntactic phrase; *P-PHRASE requires there to be as few prosodic phrases as possible; and *to prohibits the use of prepositions. None of these constraints is commonly found in phonological literature.

Feng (2003) proposes that phonology can constrain syntax. The word-length problem in Chinese and the heavy-NP shift problem in English seem to support Feng’s view, in the sense that phonology can prohibit the use of certain word orders (e.g., [[V NP1] NP2] when NP2 is unstressed) or certain syntactic structures (e.g. [N N] when the first N is shorter than the second). In contrast, Golston (1995) argues for the opposite, i.e., syntax outranks phonology and not vice versa. It is unlikely that both claims are correct. The choice between the claims may be a false one though, because it is not always clear whether syntax or phonology is really violated by the other. For example, one might argue that the double-object construction can be linearized as either [V NP1 NP2] or [V NP2 NP1], with little syntactic difference. Similarly, one might argue that the constraint against 1+2 [N N] has little to do with syntax, because [N N] is still available for 2+2, 2+1, or 1+1. On this view, the core requirements of phonology and syntax are both satisfied, and neither overrides the other. However, a critic might argue that different word orders represent different syntactic structures, which in turn yield different semantic interpretations, even though the difference may be subtle. If so, phonology can override syntax and semantics by making some word orders, and their semantic interpretations, unavailable. Clearly, more research is needed in this area.

8 Conclusions

Word-length preferences in Chinese (i.e., the choice between a monosyllabic vs. a disyllabic word) have been observed for a long time. (For historical literature, see Guo (1938) and references therein.) However, owing to a lack of quantitative studies, opinions remain divided over how strong the preferences are, whether they are motivated by phonological, syntactic, or semantic factors, what kinds of exceptions there are, and how to deal with the exceptions.

This study offers a quantitative examination of word-length patterns in Chinese [N N] (noun-noun compounds) and [V O] (verb-object phrase), using the Lancaster Corpus of Mandarin Chinese. It is found that 1+2 is overwhelmingly disfavored in [N N] and 2+1 is overwhelmingly disfavored in [V O]. Apparent exceptions, ranging between 1% and 2%, are limited to certain specific structures, and when these are factored out, both 1+2 [N N] and 2+1 [V O] are well below 1% in either token count or type count.

The result bears on several theoretical debates. First, word-length preferences are real in Chinese. Second, a phonology-based analysis (e.g., Lu and Duanmu 2002 and Duanmu 2007) is better than one based on syntax (Liu 1992) or semantics (Ke 2007; Zhou 2007). Third, it is found that exceptional cases of 1+2 [N N] are restricted to two structures: (a) when the first N is adjectival in nature; and (b) when the first N is the possessor of the second N. In addition, exceptional cases of 2+1 [V O] are restricted to one structure, which are derived from 1+1 [V O] (e.g., by adding a negation or a directional particle to V or by reduplicating V). These exceptional structures do not always fit the profiles proposed in previous studies, such as Feng (1997, 2000). Fourth, there is little preference difference between 2+2 and 2+1 in [N N] or between 2+2 and 1+2 in [V O], which means that a free syllable next to a binary foot is quite acceptable. This calls for a reconsideration of the prosodic licensing theory of Ito (1985), which disallows a free syllable outside a foot. It also questions the need for a trisyllabic super-foot (e.g., Shih 1986; Feng 1997; Chen 2000; Davis 2005).

Like Anttila et al. (2010), the present study supports the view that in problems that involve the interaction between phonology and syntax, phonological constraints play a major role. The present study also raises some new questions. For example, why are derived 2+1 [V O] so readily acceptable? Is it because prosodic requirements only hold at a deeper level, where a derived 2+1 is still 1+1? To what extent are word categories determined by meaning? For example, is a possessor (e.g., ‘county’ in ‘county hospital’) always adjectival in nature, even when an overt adjectival suffix is lacking? These questions await further research.

Acknowledgments For helpful comments, I am grateful to Shengli Feng, James Huang, Michael Kenstowicz, James Myers, Wenguo Pan, Shangzhi Zhou, three anonymous reviewers, and audiences at the Chinese University of Hong Kong, the City University of Hong Kong, Tongji University, University of Michigan, and IACL-19 at Nankai University.

References

Anttila, Arto, Matthew Adams, and Michael Speriosu. 2010. The role of prosody in the English dative alternation. *Language and Cognitive Processes* 25.7-9: 946-981.

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.

Bybee, Joan. 2006. *Frequency of use and the organization of language*. New York: Oxford University Press.

Chen, Matthew Y. 2000. *Tone sandhi: patterns across Chinese dialects*. Cambridge, UK: Cambridge University Press.

Chomsky, Noam, Morris Halle, and Fred Lukoff. 1956. On accent and juncture in English. In *For Roman Jakobson*, ed. Morris Halle, Horace Lunt, Hugh MacLean, and Cornelis van Schooneveld, 65-80. The Hague: Mouton.

Chomsky, Noam, and Morris Halle. 1968. *The sound pattern of English*. New York: Harper and Row.

Cinque, Guglielmo. 1993. A null theory of phrase and compound stress. *Linguistic Inquiry* 24.2: 239-297.

Davis, Stuart. 2005. Capitalistic v. militaristic: The paradigm uniformity effect reconsidered. In *Paradigms in phonological theory*, ed. Laura J. Downing, T.A. Hall, and Renate Raffelsiefen, 107-121. Oxford and New York: Oxford University Press.

Duanmu, San. 1990. *A formal study of syllable, tone, stress and domain in Chinese languages*. Doctoral dissertation, MIT, Cambridge, Mass.

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

Press.

Duanmu, San. 2011. Wordhood in Chinese. To appear in *Encyclopedia of Chinese Language and Linguistics*, ed. Wolfgang Behr, Gu Yueguo, Zev Handel, C.-T. James Huang, and Rint Sybesma. Leiden: Brill.

Feng, Shengli. 1997. *Hanyu de yunlu, cifa yu jufa* [Interactions between morphology, syntax and prosody in Chinese]. Beijing: Peking University Press.

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

Feng, Shengli. 2000. *Hanyu yunlu jufaxue* [Prosodic syntax in Chinese]. Shanghai: Shanghai Jiaoyu Chubanshe.

Feng, Shengli. 2003. Prosodically constrained postverbal PPs in Mandarin Chinese. *Linguistics* 41.6: 1085–1122.

Frisch, Stefan A., Nathan R, Large, and David B. Pisoni. 2000. Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language* 42.4: 481-496.

Golston, Chris. 1995. Syntax outranks phonology: Evidence from Ancient Greek. *Phonology* 12.3: 343-368.

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

Gussenhoven, Carlos. 1983. Focus, mode and the nucleus. *Journal of Linguistics* 19.2: 377-417.

Halle, Morris. 1962. Phonology in generative grammar. *Word* 18: 54-72.

Halle, Morris, and Jean-Roger Vergnaud. 1987. *An essay on stress*. Cambridge, MA: MIT Press.

Hayes, Bruce. 1995. *Metrical stress theory: Principles and case studies*. Chicago: University of

Chicago Press.

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

Ito, Junko. 1986. *Syllable theory in prosodic phonology*. Ph.D. dissertation, University of Massachusetts, Amherst.

Ke, Hang. 2007. *Xiandai hanyu dan shuang yinjie dapei yanjiu* [A study of monosyllabic and disyllabic usage in modern Chinese]. Ph.D. dissertation, Institute of Linguistics, Chinese Academy of Social Sciences, Beijing.

Kuang, Laying. 2006. ‘V shuang + N dan’ de xingzhi ji biaoshi pianzheng guanxi de youshi [The property of ‘disyllabic V + monosyllabic N’ and its preference for modifier-noun relation]. *Journal of South China Agricultural University (Social Science Edition)* 3.5: 95-98.

Liu, Feng-hsi. 1992. Verb and syllable in Chinese. Paper presented at the 25th International Conference on Sino-Tibetan Languages and Linguistics, Berkeley.

Lu, Bingfu. 1990. *The structure of Chinese nominal phrases*. M.A. thesis, University of Connecticut, Storrs.

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. 1962. Shuo ‘ziyou’ he ‘zhanzhao’ [On ‘free’ and ‘bound’]. *Zhongguo Yuwen* 1962.1: 1-6.

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.

Lü, Shuxiang. 1990. *Lü Shu-Xiang wen ji 2* [Collected papers by Lü Shu-Xiang, volume 2]. Beijing: Shangwu Yinshuguan.

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.

Palmer, Martha, Fu-Dong Chiou, Nianwen Xue, Tsan-Kuang Lee, and Jeremy LaCivita. 2004. *Chinese Treebank 4.0*. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

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

Pirani, Laura. 2008. Bound roots in Mandarin Chinese and comparison with European "semi-words". In *Proceedings of the 20th North American Conference on Chinese Linguistics (NACCL-20), 2008*, ed. Marjorie K.M. Chan and Hana Kang, Vol. 1, 261-277. Columbus, Ohio: The Ohio State University.

Prince, Alan. 1980. A metrical theory for Estonian quantity. *Linguistic Inquiry* 11: 511-562.

Richtsmeier, Peter T. 2011. Word-types, not word-tokens, facilitate extraction of phonotactic sequences by adults. *Laboratory Phonology* 2.1: 157–183.

Shannon, Claude E. 1948. The mathematical theory of communication. *Bell System Technical Journal* 27: 379-423 and 623-656, July and October. (Reprinted in Shannon and Weaver 1949 with minor revisions.)

Shih, Chi-lin. 1986. *The prosodic domain of tone sandhi in Chinese*. Ph.D. dissertation, University of California, San Diego.

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

Truckenbrodt, Hubert. 2005. Phrasal stress. In: Keith Brown, (Editor-in-Chief), Encyclopedia of

Language & Linguistics, Second Edition, volume 9, 572-579. Oxford: Elsevier.

Wang, Hongjun. 2001a. Yinjie danshuang, yinyu zhanlian (zhongyin) yu yufa jiegou leixing he chengfen cixu [The relations between the number of syllables, the tonal range of pitch (stress) and the grammatical structure in Chinese]. *Dangdai Yuyanxue* [Contemporary Linguistics] 3.4: 241-252.

Wang, Hongjun. 2001b. 《Xinxi chuli yong xiandai hanyu fenci cibiao》de neibu gouzhao he hanyu de jiegou tedian [The internal structure of *A Modern Chinese Lexicon for Information Processing* and structural properties of Chinese]. *Yuyanwenzi Yingyong* [Applied Linguistics] 2001.4: 90-97.

Wang, Hongjun, and Li Fu. 2005. Shilun xiandai hanyu de lei cizhui [On semi-affixes in modern Chinese]. *Yuyan Kexue* 4.5: 3-17.

Wu, Weishan. 1986. Xiandai Hanyu san yinjie zuhe guilü chutan [Preliminary discussion on trisyllabic structures in modern Chinese]. *Hanyu Xuexi* 1986.5: 3-4.

Xia, Fei. 2000a. The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0). <http://www.cis.upenn.edu/~chinese/ctb.html>. Accessed 19 April 2011.

Xia, Fei. 2000b. The segmentation guidelines for the Penn Chinese Treebank (3.0). <http://www.cis.upenn.edu/~chinese/ctb.html>. Accessed 19 April 2011.

Xiandai Hanyu Changyong Cibiao Ketizu [Common Lexicon in Modern Chinese Task Group]. 2008. *Xiandai Hanyu changyong cibiao (cao an)* [Common lexicon in Modern Chinese (Draft)]. Beijing: Shangwu Yinshuguan.

Yang, Shujun. 2005. *Xiandai hanyu sanyinjie ciyu yanjiu* [A study of trisyllabic words in modern Chinese]. Ph.D. dissertation, Each China Normal University, Shanghai.

Zhou, Ren. 2007. Xingxi liang yuanze yu hanyu jufa zuhe de yunlu moshi [The principle of information load and the prosodic model of syntactic combinations in Chinese]. *Zhongguo Yuwen* 2007.3 (318): 208-222.

Zubizarreta, Maria Luisa. 1998. *Prosody, focus, and word order*. [Linguistic Inquiry Monograph 33] Cambridge, MA: MIT Press.

Zubizarreta, Maria Luisa, and Jean-Roger Vergnaud. 2006. Phrasal stress, focus, and syntax. In *The Blackwell Companion to Syntax: Vol. III*, ed. Martin Everaert and Henk Van Riemsdijk, 522-568. Malden, MA: Blackwell.

Appendix: Expressions of exceptional length patterns in LCMC

1+2 [N N] (45 in all)

<中><草药>	Chinese grass-medicine	<军><分区>	army branch-district
<儿><官僚>	baby official	<区><党委>	district party-committee
<儿><将军>	baby general	<县><医院>	county hospital
<名><演员>	fame actor	<坑><两旁>	pit both-sides
<名><记者>	fame reporter	<山><大王>	mountain king
<把><兄弟>	sworn brother	<市><政府>	city government
<木><化石>	wood fossil	<市><领导>	city leadership
<机><帆船>	machine sail-boat	<师><政委>	division commissar
<核><战争>	nuclear war	<拳><宗师>	boxing founding-master
<核><武器>	nuclear weapon	<清><政府>	Qing government
<毛><大衣>	fur coat	<班><集体>	class collective
<电><真空>	electricity vacuum	<省><内外>	province inside-outside
<电><风扇>	electricity fan	<省><政府>	province government
<竹><书架>	bamboo bookshelf	<省><议员>	province legislator
<钱><票子>	money note	<系><主任>	department head
<鬼><地方>	ghost place	<脑><组织>	brain tissue
<鬼><天气>	ghost weather	<车><后架>	bike rear-rack
<唐><三彩>	Tang three-color	<门><背后>	door back-side
<亚><运会>	Asia sports-meet	<鸭><胸脯>	duck breast
<党><代表>	party representative	<猪><饲料>	pig feed
<党><支部>	party branch	<官><胡子>	official beard
<党><组织>	party organization	<屎><裤子>	excrement pants
<关><内外>	fort inside-outside		

2+1 [V O] (56 in all)

<举起><手>	raise-rise hand	<不容><情>	not-tolerate feeling
<了解><诗>	understand poem	<不见><底>	not-see bottom
<伸出><手>	stretch-exit hand	<不顾><家>	not-care family
<停下><车>	stop-descend car	<没有><事>	not-have matter
<变为><狗>	change-be dog	<没有><人>	not-have person
<变成><狼>	change-become wolf	<没有><山>	not-have mountain
<回到><家>	return-reach home	<没有><底>	not-have bottom
<回过><头>	return-pass home	<没有><水>	not-have water
<定下><心>	Decide-settle heart	<没有><脸>	not-have face

<忘记><党>	forget-remember party	<没有><路>	not-have road
<扬起><帆>	raise-rise sail	<没有><钱>	not-have money
<找上><门>	seek-approach door	<招招><手>	beckon-beckon hand
<抬起><头>	lift-rise head	<摆摆><手>	wave-wave hand
<拿到><药>	take-reach medicine	<摇摇><头>	shake-shake head
<换成><药>	change-become medicine	<摇晃><腿>	shake-sway leg
<接到><信>	catch-reach letter	<散散><步>	spread-spread walk
<接过><书>	catch-pass book	<洗洗><手>	wash-wash hand
<接过><信>	catch-pass letter	<消消><汗>	cool-cool sweat
<推开><门>	push-open door	<点点><头>	nod-nod head
<收到><信>	receive-reach letter	<耸耸><肩>	shrug-shrug shoulder
<站住><脚>	stand-stop foot	<教教><书>	teach-teach book
<翻看><书>	flip-read book	<喜欢><枪>	like gun
<诱惑><人>	lure-puzzle person	<喜欢><诗>	like poem
<转过><头>	turn-pass head	<多于><人>	more-than person
<遇到><事>	encounter-reach matter	<多于><兵>	more-than soldier
<遇到><人>	encounter-reach person	<成立><党>	establish party
<需要><人>	need-want person	<歌颂><党>	song-praise party
<需要><钱>	need-want money	<筹建><矿>	plan-build mine