Available online at www.sciencedirect.com

ScienceDirect

Lingua 198 (2017) 1–21

Lingua

www.elsevier.com/locate/lingua

A judgment study of word-length preferences in Chinese NN compounds

Zuxuan Qin [a,b,*], San Duanmu [b,1]

[a] Southwest Minzu University, China
[b] University of Michigan, USA

Received 11 January 2017; received in revised form 21 June 2017; accepted 23 June 2017
Available online 2 July 2017

Abstract

Some seemingly robust linguistic generalizations remain controversial for years, often for lack of experimental verification. As a case study, we examine word length preferences in Chinese NN compounds. Since the original observation of Lü (1963), there has been a broad consensus that 1+2 (monosyllabic + disyllabic) is ill formed, while 2+1, 2+2 and 1+1 are well-formed. In addition, it has been shown that the preferences can be derived from metrical principles. However, a judgment experiment is yet to be offered to verify the preferences, and without it some scholars remain skeptical. We offer such an experimental study, in which 15 native speakers are asked to rate the acceptability of 1,132 randomly generated NN compounds, with 283 each of the four length patterns. It is found that 2+2 is the most acceptable, followed by 2+1, while 1+2 and 1+1 are the least acceptable. The high rating of 2+2 and the low rating of 1+2 support previous predictions. The rating for 2+1 is slightly lower than expected, and that of 1+1 is most unexpected. It is also found that four other factors (the naturalness of the referent, homograph ambiguity, the boundness of a morpheme, and frequency) have significant effects on acceptability judgment, although their sizes are smaller than that of the length effect.

Keywords: Word-length preferences; Noun-noun compounds; Homograph ambiguity; Frequency; Acceptability experiment

1. Introduction

Many Chinese words can be monosyllabic (short) or disyllabic (long), with little difference in meaning (Karlgren, 1918; Guo, 1938; Chao, 1948; Kennedy, 1951; Lü, 1963; Sproat and Shih, 1996; Pan, 1997; Lu and Duanmu, 2002; Huang and Duanmu, 2013; Dong, 2015; etc.). For example, Sproat and Shih (1996:49) consider 蝇-苍蝇 *ying-cangying* to be two forms of the same word and have the same meaning. Guo (1938) considers such words to have 'elastic length'. For convenience, we shall call such synonymous short-long pairs 'elastic words'. For illustration, several elastic words are shown in (1), where parentheses indicate the part whose meaning is redundant.

* Corresponding author at: School of Foreign Languages, Southwest Minzu University, #16 South Section, 1st Ring Road, Chengdu 610041, China.
E-mail addresses: simonqzx@163.com (Z. Qin), duanmu@umich.edu (S. Duanmu).
[1] Address: Department of Linguistics, The University of Michigan, 440 Lorch Hall, 611 Tappan Street, Ann Arbor, MI 48109-1220, USA.

http://dx.doi.org/10.1016/j.lingua.2017.06.012
0024-3841/

(1) Chinese words with elastic length

Character	Monosyllabic	Disyllabic	Gloss
(看)见	jian	(kan)-jian	(look)-see
(苍)蝇	ying	(cang)-ying	(blue)-fly
(老)虎	hu	(lao)-hu	(old)-tiger
技(术)	ji	ji-(shu)	skill-(technique)
煤(炭)	mei	mei-(tan)	coal-(charcoal)
敌(人)	di	di-(ren)	enemy-(person)
权(力)	quan	quan-(li)	power-(force)
机(器)	ji	ji-(qi)	machine-(device)

The long forms resemble compounds and have been called so. For example, Packard (2000:82–83) cites the last three items in (1) as compounds. However, a true compound typically has two properties: (i) both parts contribute to the meaning of the whole and (ii) neither part is optional. For example, in *bath-tub*, both *bath* and *tub* contribute to the meaning of the whole (i.e. *bath-tub* is for taking a *bath* and is a kind of *tub*), and neither can be omitted (i.e. *bath-tub* is not equal to *bath* or *tub*).

In contrast, the two forms of an elastic word have the same meaning. For example, in '(blue)-fly', 'blue' is redundant, because a black fly is also called 'blue-fly', and because the part 'blue' can easily be omitted. Similarly, in the Chinese form '(old)-tiger', 'old' contributes no meaning to the whole (i.e. '(old)-tiger' need not be old, and a baby tiger is also an '(old)-tiger'), and 'old' can be omitted (i.e. '(old)-tiger' is the semantically equal to 'tiger'). Semantic identity is often reflected as mutual annotations in Chinese dictionaries. For example, in the dictionary XDHYCD (2005), the entry 煤 *mei* 'coal' is annotated as 煤炭 *meitan* 'coal-charcoal', and the entry 煤炭 *meitan* 'coal-charcoal' is annotated as 煤 *mei* 'coal'. Based on an exhaustive examination of XDHYCD (2005), Dong (2015) found that there are some 20,000 monosyllabic senses, of which 49% have elastic length. Indeed, it can be shown that, because the redundant part of an elastic word is often arbitrary, they often introduce exceptions to morphological generalizations, such as the Headedness Principle of Packard (2000). If elastic words are treated as monomorphemic, as suggested by Dong (2002:1), the Headedness Principle could be a lot more general.

The presence of elastic words creates word length choices. When two elastic words form a NN compound, there are four length pattern choices, namely, 1+1, 1+2, 2+1 and 2+2, where 1 is a monosyllable and 2 a disyllable. However, there is a general consensus that, of the four choices, 1+2 is bad while the other three are good (Lü, 1963; Feng, 1996, 2013; Duanmu, 1999, 2007, 2012; Wang, 2001; Lu and Duanmu, 2002; Wang, 2002; Wu, 2006; Huang and Duanmu, 2013; Dong, 2015; among many others). The generalization is shown in (2), along with an example. For simplicity, we omit annotating the semantically redundant part of the disyllabic form. Also, for visual clarity, we use a hyphen to separate the two nouns in a compound (but not the two parts within an elastic word).

(2) Generalization on word length preferences in NN compounds

	Generalization	Example
a.	1+1 is well formed	技-工 ji-gong
b.	*1+2 is ill formed	*技-工人 ji-gongren
c.	2+1 is well formed	技术-工 jishu-gong
d.	2+2 is well formed	技术-工人 jishu-gongren
		'skill worker'

A phonological explanation for the length preferences has also been offered. According to Lu and Duanmu (2002) and Duanmu (2007), the length preferences can be accounted for by the phonological requirements in (3).

(3) Phonological requirements
- a. In NN, the first N is stressed.
- b. A disyllabic syntactic constituent forms a foot
- c. A foot contains two beats (Foot Binarity).
- d. Stress is the head of a foot.

(3a) is similar to the Compound Rule in English (Chomsky and Halle, 1968). (3b) achieves the cyclic effect of rule application (Chomsky et al., 1956). (3c) is often known as Foot Binarity. (3d) is the metrical definition of what a foot is (Halle and Vergnaud, 1987; Hayes, 1995). Given (3), the length patterns in NN can be analyzed in (4), where S is a syllable and **S** is a stressed syllable in a trochaic foot.

(4) Analysis of length patterns in NN compounds

Length	Foot	Comment
2+2	(**S**S)-(**S**S)	By (3b)
2+1	(**S**S)-S	Second N need not be stressed
*1+2	(**S**)-(**S**S)	By (3a), First N has stress, violating (3c)
1+1	(**S**-S)	By (3b)

In 2+2, each disyllabic word forms a binary foot; (3a) is satisfied as well, since the first N is a foot and has stress. In 2+1, the first N forms a binary foot, satisfying (3a) as well. The monosyllabic second N can stay free, which does not violate any requirement in (3). In 1+2, the second N forms a binary foot by (3b). By (3a), the first N has stress, which means it is also a foot, but this foot has only one syllable, violating (3c). This explains why 1+2 is ill formed. In 1+1, a binary foot is formed by (3b), and all requirements are met.

Nevertheless, counterexamples to the generalization in (2) have been noted. For example, Sproat and Shih (1996:61) cite some well-formed 1+2 NN compounds, such as 脑-组织 *nao-zuzhi* 'brain tissue' and 肠-细胞 *chang-xibao* 'intestine cell'. In addition, Wang (2002) rejects the generalization in (2) and the phonological analysis in (4) entirely. He cites a number of well-formed 1+2 NN compounds, such as 纸-飞机 *zhi-feiji* 'paper airplane', 鬼-故事 *gui-gushi* 'ghost story', and 金-项链 *jin-xianglian* 'gold necklace', and concludes that 1+2 compounds are in principle well formed. Oddly enough, Wang (2002:163) also proposes that 2+2 NN compounds, such as 手表-工厂 *shoubiao-gongchang* 'watch factory' and 煤炭-商店 *meitan-shangdian* 'coal-store', are ill formed, contrary to the view of other native scholars. Like Wang (2002), Ke (2007) rejects the generalization in (2), too, and argues that both 1+2 and 2+1 NN compounds are in principle well formed.

Most of the counterexamples have been addressed by Duanmu (2012). Using the Lancaster Corpus of Mandarin Chinese (McEnery and Xiao, 2004), Duanmu (2012) shows that, among all length patterns of NN compounds, the occurrence of 1+2 is below 2%, either by token count or by type count. In addition, the occurring 1+2 compounds are mostly restricted to three specific cases: (i) the first noun is the material of the second (e.g. 'gold necklace', 'paper airplane', 'wood floor', etc.), (ii) the first noun is the possessor of the second (e.g. 'brain tissue (brain's tissue)', 'intestine cell (intestine's cell)', etc.), and (iii) neither noun is elastic and so there is no other length choice (e.g. 'ghost story').

Nevertheless, there has been no judgment study on the generalization in (2). Therefore, some important questions remain. First, the validity of (2), regardless of how intuitive it may seem, remains hypothetical and is open to challenges. Second, it remains open whether 2+2, 2+1, and 1+1 are equally good, as previously claimed. For example, according to (4), 2+1 contains a free syllable, outside of a foot. Would this make it slightly less acceptable than 2+2 or 1+1? Third, it is unclear whether the length pattern is the only factor that determines the acceptability of a compound.

To address these questions, we conducted a judgment experiment. Section 2 explains the data and the procedure. Section 3 presents a preliminary study of the length effect. Section 4 examines judgment variability and exceptional expressions in order to determine additional factors that may play a role. Section 5 offers an expanded study that includes five new factors identified in section 4. Section 6 offers concluding remarks.

2. Test materials, subjects, and procedure

We chose a written judgment experiment, in which native subjects were given a list of NN compounds and asked to offer an acceptability score to each.

2.1. Test materials

The criteria we used in choosing the compounds for the judgment experiment are summarized in (5).

(5) Criteria used in choosing NN compounds for the judgment experiment
- a. Each NN compound should be made of two nouns with elastic length, in order to offer four length patterns, 2+2, 2+1, 1+2, and 1+1.
- b. The compounds should contain commonly used words.
- c. The list of NN compounds should be representative and comprehensive.
- d. The length of the list should be feasible for the experiment.

Given the above criteria, we aimed at a list of about 1,000 items. Given four length patterns per compound, we aimed at choosing 300 NN compounds. To do so, we took the steps in (6).

(6) Steps in choosing NN compounds for the judgment experiment
   a. Gather all nouns with elastic length whose style is common and whose monosyllabic form is 'free' (as defined by XDHYCD, 2005).
   b. Select 200 most frequent ones from the above list.
   c. Exhaustively combine the 200 nouns to generate 200 × 200 = 40,000 NN units.
   d. Manually exclude NN units that are clearly uninterpretable.
   e. Randomly choose 300 from the set of interpretable NN units.
   f. Exclude those that contain a kinship word, which leaves 283 NN compounds.
   g. Generate a list of 1,132 NN compounds by making four length patterns for each of the 273 NN compounds, i.e. 2+2, 2+1, 1+2, and 1+1.

First, we gathered all elastic nouns from a comprehensive list of elastic words in Chinese provided by Dong (2015), totaling 3,686. Among the list, we further selected those whose style is common (e.g. not limited to 'dialectal', 'written', or 'archaic' styles) and whose monosyllabic form is 'free', as defined by XDHYCD (2005), where a 'free' form is annotated for part-of-speech but a bound form is not; this yields a total of 1,269. Second, we chose the top 200 most frequent elastic nouns, based on the word frequency list of Guojia Yuwei (2012). Third, we used the 200 nouns to freely combine with each other, generating 200 × 200 = 40,000 NN units. Fourth, we manually excluded expressions that are clearly uninterpretable, in order to focus on the effect of word length and to minimize the effect of interpretability. As exemplified in (7), uninterpretable expressions are in general easy to identify.

(7) Interpretable and uninterpretable NN expressions

Interpretable?	NN expressions	Gloss
No	国家-鼻子 guojia-bizi	'nation nose'
No	价格-小麦 jiage-xiaomai	'price wheat'
Yes	小麦-价格 xiaomai-jiage	'wheat price'
Yes	食盐-房子 shiyan-fangzi	'salt house'

We found it hard to interpret 'nation nose'. Similarly, we found it hard to interpret 'price wheat', in contrast to the fully interpretable 'wheat price'. Some NN compounds seem interpretable but the referent may be uncommon or unnatural, such as 'salt house', which could be 'a house to store salt' or 'a house made of salt', neither of which is common (to our judges). For completeness, we have kept such units in our data, although we shall consider the factor of 'naturalness' of the referent, to be discussed below. This yielded 3,407 semantically interpretable NN units.

Fifth, we randomly selected 300 NN units from the result of (6d). Sixth, we excluded 17 units that involved a kinship item, such as 妹妹-牙齿 *meimei-yachi* 'sister tooth' and 弟弟-帽子 *didi-maozi* 'brother hat'. Such expressions normally requires a possessive particle 的 *de* and they behave differently from NN compounds, as noted by Wang (2013) and two reviewers. Finally, from the remaining 283 NN compounds, we generated three additional length patterns each. For example, from the 2+2 form 小麦-价格 *xiaomai-jiage* 'wheat price', we generated 2+1 *xiaomai-jia*, 1+2 *mai-jiage*, and 1+1 *mai-jia*. This yielded a total of 1,132 NN compounds, to be used as test data.

2.2. Subjects

Fifteen native speakers of Chinese, who either already obtained or were pursuing a bachelor degree, participated in the judgment experiment.

2.3. Procedure

The 1,132 NN compound items, written in Chinese, were made into a randomized list, each on a separate line in a spreadsheet file. In order to facilitate judgment, when a noun is monosyllabic, the intended disyllabic meaning was provided. In addition, for each compound, four acceptability degrees were given and the subjects were asked to choose one. An example is shown in (8) in the original and glossed in (9).

(8) A sample item in the original test sheet

词组	补充说明	得分	评分标准
日麦	日=日本		1 完全不可以接受
	麦=小麦		2 基本不可以接受
			3 基本可以接受
			4 完全可以接受

(9) English gloss of the sample item (8)

Compound	Notes	Score	Criteria
ri-mai	ri = riben 'Japan'		1 Completely unacceptable
'Japan wheat'	mai = xiaomai 'wheat'		2 Somewhat unacceptable
			3 Somewhat acceptable
			4 Completely acceptable

In order to avoid fatigue and obtain more reliable results, the list was divided into six sessions, with 1/6 of the compound items per session. There was an interval from one to three days between two sessions. All the questionnaires were distributed and collected through email.

3. A preliminary study

3.1. Interpreting acceptability scores

Given that the judgment choices were verbal categories, yet they were ordered in a numeric scale 1–4, we used two ways to interpret the data, shown in (10).

(10) Two ways to interpret the data
- a. Categorical: Counting the percentage of the subjects who rated a compound form to be acceptable, including both 3 (somewhat acceptable) and 4 (completely acceptable).
- b. Numeric scale: Interpreting the scores 1, 2, 3, and 4 as 0%, 33%, 67%, and 100% respectively on the acceptability scale.

In the categorical interpretation, we consider the percentage of the subjects who regard a compound form to be acceptable; for example, we can calculate the percentage of subjects who gave 3 (somewhat acceptable) or 4 (completely acceptable) as a measure of acceptability. In the numeric scale interpretation, we regard the judgment scores 1–4 to be at equal distances on a scale, corresponding to 0%, 33%, 67% and 100% respectively. The conversion of 1 'completely unacceptable' and 4 'completely acceptable' into 0% and 100% is uncontroversial, because they represent the two ends of the scale. The treatment of 2 as 33% and 3 as 67% is based on the fact that their meanings fall between the two extremes and the verbal choices were intended to be more or less at equal distance from each other, even though the interpretation of verbal choices could be somewhat subjective and the exact distance between two choices may not to be exactly equal.

After trying out both approaches, we found the results to be similar. Therefore, in what follows, we only discuss results in the numeric scale interpretation.

3.2. Length and subject effects

ANOVA (analysis of variance) is used to determine main effects by length, subject, and their interaction. The statistics are shown in (11). The mean acceptability score of each length pattern for all subjects and their pairwise comparisons are shown in (12) and (13), respectively.

(11) Length and subject effects

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Length	3	628.5	209.51	2681.6	<2e−16***
Subjects	14	115.3	8.23	105.4	<2e−16***
Length:Subjects	42	81.1	1.93	24.72	<2e−16***
Residuals	17040	1331.3	0.08		

(12) Mean acceptability score of each length pattern for all subjects

Length	Acceptability	St. dev	N
1+1	47.8%	18.6%	283
1+2	49.6%	17.4%	283
2+1	70.2%	16.1%	283
2+2	95.1%	7.4%	283

(13) Tukey multiple comparisons of means (at 95% family-wise confidence level)

	difference	lower range	upper range	p value
(1+2)-(1+1)	0.0185	−0.0151	0.0521	0.4892
(2+1)-(1+1)	0.2242	0.1906	0.2579	0.0000
(2+2)-(1+1)	0.4734	0.4397	0.5070	0.0000
(2+1)-(1+2)	0.2057	0.1720	0.2393	0.0000
(2+2)-(1+2)	0.4548	0.4212	0.4885	0.0000
(2+2)-(2+1)	0.2491	0.2154	0.2828	0.0000

First, it can be seen that length has a large effect size on acceptability, while the effect sizes of subjects and the interaction between length and subjects are both small. Second, it can be seen that 2+2 has the highest degree of acceptability, followed by 2+1, whereas 1+1 and 1+2 have the lowest degrees of acceptability. In addition, there is no significant difference between 1+2 and 1+1, while there is a highly significant difference between any other pair of length types. Third, it can be seen that there is more consistency in the rating of 2+2, shown by the small value of standard deviation, while the rating of other length patterns is more variable, shown by the large values of standard deviation. Such within-length variation can be visualized in the boxplot in Fig. 1, where dots indicate outliers.

We see again that, if we exclude outliers, 2+2 has the highest median, with the least amount of variation. In addition, 2+1 has the second highest median, but many members in this category were ranked below the mean, over a fairly wide range. Finally, 1+1 and 1+2 show the lowest medians and the greatest range of variation, both above and below their means.

The subject effect can be seen in Fig. 2, which shows the mean acceptability scores of each length pattern by each subject.

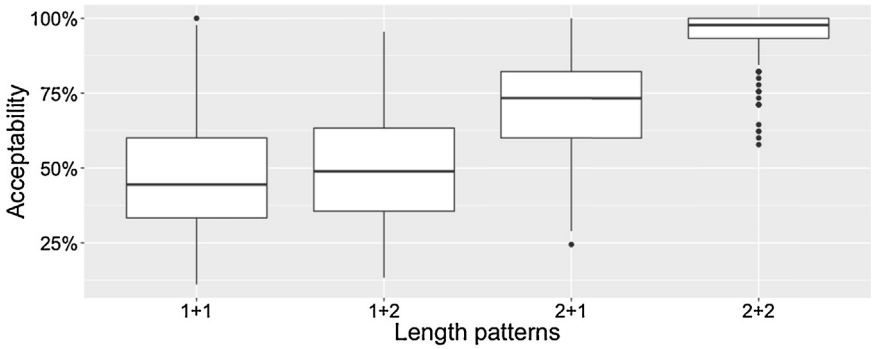

Fig. 1. Boxplot of acceptability of each length pattern of NN compounds.

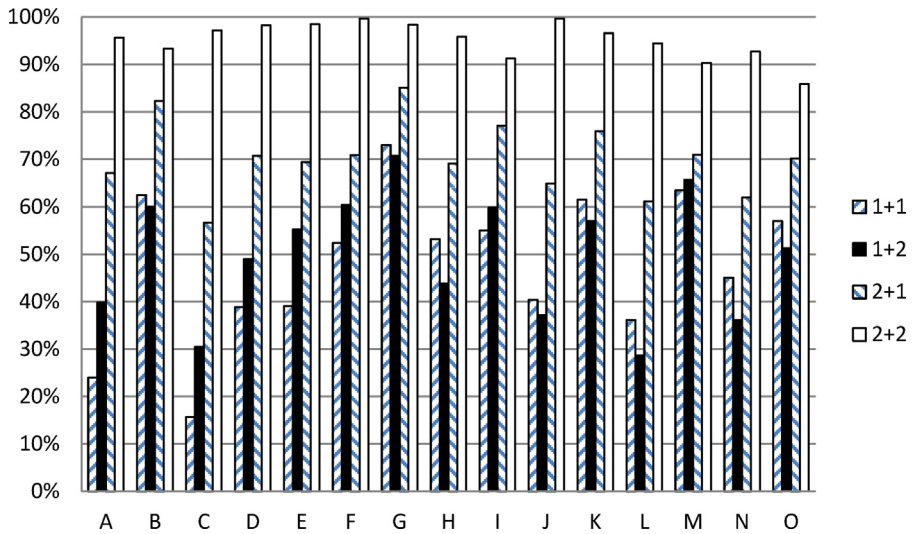

Fig. 2. Mean acceptability scores for each length pattern by each subject (A–O).

While all subjects gave high scores to 2+2, they differ considerably with regard to the other three length patterns. For example, the mean score for 1+1 is merely 16% by subject C but 73% by subject G, a difference of more than four folds. Pairwise comparisons between subjects are shown in (14) and (15), where ---, *, **, and *** indicates p-values >0.05, between 0.01 and 0.05, between 0.001 and 0.01, and <0.001 respectively. It can be seen that significant differences are found between most subject pairs.

(14) Pairwise comparisons between subjects

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
B	***													
C	***	***												
D	***	***	***											
E	***	***	***	---										
F	***	---	***	***	*									
G	***	***	***	***	***	***								
H	***	***	***	---	---	**	***							
I	***	---	***	**	*	---	***	*						
J	---	***	***	---	*	***	***	*	***					
K	***	---	***	***	***	---	***	***	---	***				
L	---	***	*	***	***	***	***	***	***	*	***			
M	***	---	***	***	***	---	***	***	---	***	---	***		
N	---	***	***	*	***	***	***	***	***	---	***	---	***	
O	***	***	***	---	---	*	***	---	---	**	***	***	***	***

(15) Summary of pairwise comparisons between subjects

Difference	---	*	**	***	All
Pairs	23	9	3	70	105

We shall discuss below that the between-subject difference may be influenced by how a subject perceives the meaning of a compound.

4. Discussion and additional factors

In this section we focus on two issues. First, we compare our preliminary results with the predictions of previous studies. Second, we examine exceptional cases and see whether additional factors play a role in acceptability judgment.

4.1. Predictions and acceptability results

A comparison between the present results and the predictions of the assumed general patterns in (2) is shown in (16).

(16) Comparison between predictions and acceptability results

Length	Predictions by (2)	Acceptability	Comment
2+2	Good	Highest	Expected
2+1	Good	High	Partly expected
1+2	Bad	Low	Expected
1+1	Good	Low	Unexpected
All	Clear cut judgment	Within-length variation	Unexpected

As can be seen, previous predictions are only partially confirmed by the acceptability results in the judgment experiments. Specifically, the results for 2+2 and 1+2 are expected, where 2+2 is the most acceptable and 1+2 the least. The result for 2+1 is partially expected, in that it has a higher acceptability score than 1+2 as expected, yet its lower score than 2+2 is unexpected. The result for 1+1 is completely unexpected, since it is previously thought to be a good pattern, and it is found to be the most frequent occurring NN pattern in corpus data, yet it has the lowest acceptability score. Finally, previous studies assume that the judgment of well-formedness is clear cut, yet the acceptability data show that there is a wide range of variation within each length type, especially for 2+1, 1+2, and 1+1.

4.2. Variability and exceptional cases

Let us now take a look at variability exceptional cases, in order to see whether some explanation can be offered. Specifically, we are interested the cases in (17).

(17) Variability cases of interest
- a. Comparison between the best and the worst 2+2 compounds
- b. Comparison between the best and the worst 2+1 compounds
- c. Comparison between the best and the worst 1+2 compounds
- d. Comparison between the best and the worst 1+1 compounds

(17a) examines why some 2+2 compounds, previously expected to be good, are ranked low on the acceptability scale, (17b) examines why some 2+1 compounds, previously expected to be good as well, are ranked low on the acceptability scale. (17c) examines why some 1+2 compounds, previously expected to be bad, are ranked high on the acceptability scale. (17d) examines why some 1+1 compounds are ranked near the top on the acceptability scale while some near the bottom.

4.2.1. 2+2 compounds

We begin with 2+2 compounds with the highest and the lowest mean acceptability scores, shown in (18) and (19). The subject scores show the raw scores on a four-point scale by each of the fifteen subjects. For visual clarity, we use + to separate the two nouns in a compound. For lack of space, we omit the Pinyin transcription of the examples.

(18) 2+2 compounds with top 10 mean acceptability scores

Compound	Subject scores	Score
棉花-类型 'cotton type'	4,4,4,4,4,4,4,4,4,4,4,4,4,4,4	100%
坟墓-数量 'tomb quantity'	4,4,4,4,4,4,4,4,4,4,4,4,4,4,4	100%
医院-等级 'hospital grade'	4,4,4,4,4,4,4,4,4,4,4,4,4,4,4	100%
小麦-价格 'wheat price'	4,4,4,4,4,4,4,4,4,4,4,4,4,4,4	100%
歌曲-质量 'song quality'	4,4,4,4,4,4,4,4,4,4,4,4,4,4,4	100%
蜜蜂-名字 'bee name'	4,4,4,4,4,4,4,4,4,4,4,4,4,4,4	100%
亚洲-节日 'Asia festival'	4,4,4,4,4,4,4,4,4,4,4,4,4,4,4	100%
食盐-质量 'salt quality'	4,4,4,4,4,4,4,4,4,4,4,4,4,4,4	100%
非洲-医学 'Africa medicine'	4,4,4,4,4,4,4,4,4,4,4,4,4,4,4	100%
家庭-集会 'family party'	4,4,4,4,4,4,4,4,4,4,4,4,4,4,4	100%

(19) 2+2 compounds with bottom 10 mean acceptability scores

Compound	Subject scores	Score
鲤鱼-口袋 'carp bag'	4,4,2,4,4,4,2,4,4,4,1,4,1,4,2	73%
蜜蜂-牙齿 'bee tooth'	1,4,4,4,3,4,4,3,4,4,4,2,2,1,3	71%
脖子-名字 'neck name'	1,3,4,4,4,4,3,2,4,4,4,2,3,3,2	71%
日本-界限 'Japan border'	3,1,4,3,3,4,4,4,4,4,3,2,3,2,3	71%
老虎-食盐 'tiger salt'	4,4,2,4,4,4,3,3,3,4,3,3,3,1,2	71%
菊花-口袋 'daisy bag'	1,2,3,2,4,4,4,2,4,4,4,3,2,1,4	64%
用处-数量 'usefulness quantity'	4,2,2,4,4,4,3,4,2,4,2,1,3,3,1	62%
血液-池塘 'blood pond'	3,1,2,3,2,4,4,4,4,4,3,2,3,1,3	62%
食盐-房子 'salt house'	1,2,2,4,4,4,4,2,4,4,1,2,2,3,3	60%
棉花-盒子 'cotton box'	1,2,3,3,3,4,1,2,4,4,3,1,3,3,4	58%

It can be seen that, for the high-score compounds, the referents are common objects or concepts and the subject scores are consistent. In contrast, for most low-score compounds, the referents seem less natural and the subject scores are less consistent. For example, while 'bee tooth', 'salt house', and 'blood pond' are conceivable to some, they are probably less conceivable to others. Similarly, 'cotton box' could be interpreted as (i) 'a box to hold cotton' or (ii) 'a box made of unwoven cotton', and feedback from subjects show that those who ranked it low assumed (ii), which is an unlikely object.

It is worth noting that the low-score examples in (19) are all statistical outliers. Therefore, let us consider examples of lowest acceptability scores that are not outliers. 10 examples are shown in (20).

(20) 2+2 non-outlier compounds with bottom 10 mean acceptability scores

Compound	Subject scores	Score
鸽子-图画 'pigeon picture'	4,4,4,4,4,4,4,2,4,4,4,4,3,3,3	89%
口袋-仓库 'bag depot'	4,2,4,4,4,4,4,4,4,4,4,4,4,4,1	89%
城市-石灰 'city lime'	4,4,2,4,4,4,4,4,4,4,4,4,3,4,2	89%
光线-数量 'ray quantity'	4,4,4,4,4,4,4,4,2,4,4,4,3,4,2	89%
秋季-舞蹈 'autumn dance'	4,3,4,4,4,4,4,3,4,4,4,3,3,4,2	87%
瓶子-盒子 'bottle box'	4,4,3,4,4,4,3,2,4,4,4,4,3,4,3	87%
价格-牌子 'price board'	1,4,4,4,4,4,3,4,4,4,4,3,4,3,4	87%
日本-棍子 'Japan stick'	1,4,4,4,4,4,4,4,3,4,4,4,4,2,4	87%
肿瘤-瓶子 'tumor bottle'	4,4,4,4,4,4,4,4,4,4,1,3,4,4,1	84%
图画-影子 'picture shadow'	4,4,4,4,4,4,4,4,2,4,4,1,4,4,2	84%

Most of the examples received high scores from most subjects, although one or two subjects gave very low scores. It is worth noting that the last subject (subject O), who offered the lowest mean score for 2+2 compounds (see Fig. 2), gave low scores to eight out of ten of examples in (20). We conclude, therefore, that there is no obvious factor that affects the acceptability of regular (non-outlier) 2+2 compounds, except that of subject variability.

In summary, the plausibility or 'naturalness' of the referent of a compound seems to have influenced some subjects in their acceptability judgment.

4.2.2. 2+1 compounds

Next we consider 2+1 compounds. Unlike disyllabic words, which are generally unambiguous, a monosyllabic word form often has more than one meaning. For illustration, consider 质 *zhi*, which has three common meanings as a noun, distinguished by three disyllabic forms, shown in (21). It also has four other meanings, not included here, one being an adjective, two being a verb, and one being a literary noun that is limited to written language only.

(21) Ambiguity of monosyllabic forms: Three noun meanings of 质 *zhi*

Disyllabic	Monosyllabic	Gloss	Order
(本)质 (ben)zhi	质 zhi	'(original) nature'	1
质(量) zhi(liang)	质 zhi	'quality (measure)'	2
(物)质 (wu)zhi	质 zhi	'(object) matter'	3

The last column in (21) shows the order of commonness among the three meanings, where 1 means the most common/frequent/salient meaning. The commonness of a word is based on the dictionary XDHYCD (2005), the word frequency list of Guojia Yuwei (2012), and the BCC corpus (Xun et al., 2016). In what follows, we shall add annotation for whether the most common meaning of a monosyllabic word is the same as the intended meaning in the given compound.

2+1 compounds with the highest and the lowest acceptability scores are shown in (22) and (23). In the third column, we indicate whether the intended meaning of the monosyllabic second noun (N2) is its most common meaning, where 'yes' means that the most common meaning is the same as the intended meaning in the compound, and 'no' means that it is not.

(22) 2+1 compounds with top 10 mean acceptability scores

Compound	Subject scores	Score	N2
中国-舞 'China dance'	4,4,4,4,4,4,4,4,4,4,4,4,4,4,4	100%	yes
棉花-帽 'cotton hat'	4,4,4,4,3,4,4,4,4,4,4,4,4,4,4	98%	yes
政党-名 'party name'	4,4,4,4,4,3,4,4,4,4,4,4,3,4,4	96%	yes
小麦-价 'wheat price'	4,4,4,4,4,3,4,4,4,4,4,3,4,4,4	96%	yes
股票-名 'stock name'	4,4,4,4,4,4,4,4,4,3,4,4,4,3,4	96%	yes
亚洲-货 'Asia goods'	4,4,3,4,4,4,4,4,3,4,4,4,4,4,4	93%	yes
亚洲-棉 'Asia cotton'	4,4,4,4,4,4,4,4,4,4,4,4,3,2,4	93%	yes
被子-价 'quilt price'	4,4,3,4,4,3,4,4,4,4,4,4,3,4,4	93%	yes
凳子-名 'stool name'	4,3,4,4,4,3,4,3,4,4,4,4,4,4,4	93%	yes
兔子-厂 'rabbit factory'	4,4,3,4,4,3,4,4,4,4,4,4,4,4,3	93%	yes

(23) 2+1 compounds with bottom 10 acceptability scores

Compound	Subject scores	Score	N2
城市-灰 'city lime'	1,3,1,1,1,3,3,2,3,1,3,2,3,2,2	36%	no
辫子-影 'braid shadow'	1,1,1,3,3,3,2,3,2,1,2,2,4,1,1	33%	yes
盒子-利 'box profit'	1,2,1,3,1,3,3,2,2,1,1,3,3,1,3	33%	no
非洲-计 'Africa plan'	1,2,1,3,2,3,3,2,2,1,1,2,3,1,3	33%	no
歌曲-质 'song quality'	1,2,1,2,2,3,3,2,2,2,1,1,3,2,3	33%	no
工厂-气 'factory air'	1,3,1,4,1,3,3,1,2,1,2,2,3,1,1	31%	no
兔子-利 'rabbit profit'	1,2,1,2,1,2,3,2,2,2,3,2,3,2,1	31%	no
路线-计 'route plan'	1,2,1,4,1,2,4,2,1,1,1,1,3,1,3	29%	no
秋季-气 'Autumn air'	1,2,1,2,2,2,3,2,1,1,3,1,3,2,2	29%	no
蜜蜂-力 'bee strength'	1,1,1,2,2,3,2,2,2,1,2,1,3,2,1	24%	no

The high-score compounds received fairly consistent subject scores. In addition, the most common meaning of N2 is the same as the intended meaning in the compound. The low-score compounds received more variable subject scores. There seem to be two reasons. First, in all of them except one, the most common meaning of N2 differs from the intended one in the given compound. For example, the most common meaning of 计 *ji* is not 'plan' but 'idea', and the meaning for 'plan' is given in the specific context of 百年大计 *bai nian da ji* 'hundred-year big plan'. Similarly, the most common meaning of 质 *zhi* is not 'quality' but 'nature', the most common meaning of 气 *qi* is not 'air' but 'gas', the most common meaning of 利 *li* is not 'profit' but 'benefit', the most common meaning of 力 *li* is not 'strength' but 'force', and so on. Subjects who offered lower scores may have focused on the more common (but unintended) meaning, while subjects who offered higher scores may have thought about the intended (albeit less common) meaning.

Second, as seen in 2+2 compounds, the commonness (or naturalness) of the referent may have played a role. For example, the competing meaning for 'bee strength' is 'bee force', which could seem unnatural: bees are of small sizes, whereas force is usually associated with objects much larger than them. Similarly, a girl's braid(s) can surely have a shadow, but it is not something one commonly notices.

4.2.3. 1+2 compounds

Next we look at 1+2 compounds with the highest and the lowest acceptability scores, shown in (24) and (25).

(24) 1+2 compounds with top 10 acceptability scores

Compounds	Subject scores	Sore	N1
药-瓶子 'drug bottle'	4,4,4,3,4,4,4,4,4,4,4,3,4,4,4	96%	yes
盐-利润 'salt profit'	4,4,4,4,4,4,3,4,4,3,4,4,4,4,2	91%	yes
台-军舰 'Taiwan warship'	4,4,4,4,4,3,4,3,4,2,4,4,4,4,4	91%	no
煤-重量 'coal weight'	4,4,3,3,3,3,4,4,4,4,4,4,4,3,4	89%	yes
党-名称 'party name'	4,4,4,4,4,3,3,3,4,4,4,2,3,4,4	87%	yes
货-质量 'goods quality'	4,4,3,4,4,3,3,3,4,4,4,2,3,4,4	84%	yes
煤-等级 'coal grade'	4,3,4,3,4,3,4,4,4,3,3,4,3,3,4	84%	yes
桥-等级 'bridge grade'	4,4,3,4,4,4,3,4,4,3,4,2,3,3,3	82%	yes
药-数量 'drug quantity'	4,3,3,4,4,3,3,4,4,3,4,4,4,3,3,3	82%	yes
国-界限 'nation border'	4,4,3,4,4,3,4,3,4,3,3,4,3,3,3	82%	yes

(25) 1+2 compounds with bottom 10 acceptability scores

Compounds	Subject scores	Score	N1
中-鸽子 'China pigeon'	1,3,1,1,1,2,3,2,1,1,1,1,3,2,2	22%	no
子-名称 'seed name'	1,1,1,1,1,2,3,1,2,2,2,1,3,2,1	20%	no
部-家庭 'army family'	1,2,1,1,2,2,3,1,1,1,2,1,3,1,2	20%	no
非-城市 'Africa city'	1,2,1,1,1,2,3,1,2,1,2,1,3,1,2	20%	no
中-同伴 'China friend'	1,2,1,1,1,3,3,2,2,1,1,1,3,1,1	20%	no
礼-屋子 'gift room'	1,1,1,2,1,2,2,2,2,1,2,1,3,1,1	18%	yes
黑-舞蹈 'night dance'	1,1,1,1,1,2,2,1,2,1,3,1,2,1,2	16%	no
非-池塘 'Africa pond'	1,1,1,3,1,2,2,2,1,1,2,1,2,1,1	16%	no
中-城市 'China city'	1,1,1,1,1,3,3,2,2,1,1,1,2,1,1	16%	no
用-数量 'usefulness quantity'	1,1,1,1,3,1,2,1,1,1,1,1,3,1,2	13%	no

We see again a clear difference in meaning between the high-score set and the low-score set. In the former, in nine of the ten cases, the most common meaning of the monosyllabic N1 is the same as the intended meaning in the compound. In contrast, in the low-score set, in nine of the ten cases, the most common meaning of the monosyllabic N1 differs from the intended meaning in the compound.

For the two exceptional cases, a reasonable explanation could also be offered. The low-score exception 'gift room' has no problem in the meaning of N1, but its referent (a room for or full of gifts) may seem uncommon for most of the judges. This explains why it has a low score. The high-score exception 'Taiwan warship' could have been reinterpreted from 1+2, literally '[Taiwan [military warship]]', to 2+1 '[[Taiwan military] warship]', in which case the monosyllabic term is *jian*, which only has one meaning and it is the same as that in the compound. Indeed, 'drug bottle' and 'nation border' can be reinterpreted as 2+1 as well, which are literally '[[drug bottle] diminutive]' and '[[nation border] line]' respectively.

A further difference between the two sets of 1+2 compounds can be observed. In the high-score set, the monosyllabic first N is mostly free (in that it can serve as the subject of a sentence, for example), whereas in the low-score set, the monosyllabic first N is usually not free. We shall consider this factor again in section 5.

4.2.4. 1+1 compounds

Finally, let us consider 1+1 compounds with the highest and the lowest acceptability scores, shown in (26) and (27).

(26) 1+1 compounds with top 10 acceptability scores

Compounds	Subject scores	Sore	N1; N2
军-歌 'army song'	4,4,4,4,4,4,4,4,4,4,4,4,4,4,4	100%	yes; yes
歌-名 'song name'	4,4,4,4,4,3,4,4,4,4,4,4,4,4,4	98%	yes; yes
国-事 'state affair'	4,4,4,4,4,3,4,4,4,4,4,4,4,4,4	98%	yes; yes
棉-帽 'cotton hat'	4,4,4,4,3,4,3,4,4,4,4,4,4,4,4	96%	yes; yes
国-界 'nation border'	4,4,4,4,4,3,4,4,4,4,4,4,3,4,4	96%	yes; yes
掌-声 'palm sound'	4,4,4,3,2,3,4,4,4,4,4,4,4,4,4	91%	yes; yes
药-瓶 'drug bottle'	4,3,4,4,3,3,4,4,4,4,4,4,3,4,4	91%	yes; yes
蜂-名 'bee name'	3,4,3,2,4,3,4,4,4,4,4,4,4,4,4	89%	yes; yes
煤-重 'coal weight'	4,4,3,4,4,3,4,4,4,3,4,4,3,3,4	89%	yes; yes
虎-纹 'tiger stripe'	4,4,3,4,3,3,4,4,4,3,4,4,3,3,4	87%	yes; yes

(27) 1+1 compounds with bottom 10 acceptability scores

Compounds	Subject scores	Sore	N1; N2
城-灰 'city lime'	1,2,1,1,1,2,4,2,2,1,2,1,3,1,1	22%	no; no
子-名 'seed name'	1,3,1,1,1,2,3,1,2,2,2,1,2,2,1	22%	no; yes
美-桌 'America table'	1,1,1,3,1,2,2,1,2,2,3,1,2,2,1	22%	no; yes
亚-地 'Asia region'	1,2,1,1,1,2,3,2,2,1,1,1,2,2,3	22%	no; no
日-绳 'Japan rope'	1,2,1,1,2,2,2,2,3,2,1,1,2,1,2	22%	no; yes
城-帽 'city hat'	1,1,1,2,2,1,3,2,2,1,2,1,2,1,2	20%	no; yes
亚-节 'Asia festival'	1,3,1,1,1,2,3,1,2,2,1,1,2,2,1	20%	no; no
非-城 'Africa city'	1,1,1,1,1,2,3,1,2,1,1,1,3,1,3	18%	no; no
非-计 'Africa plan'	1,2,1,1,2,2,3,2,2,1,1,1,2,1,1	18%	no; no
非-道 'Africa road'	1,1,1,2,1,2,2,1,1,1,1,1,2,1,2	11%	no; yes

A striking difference between the two sets of compounds again lies in meaning. In the high-score set, the most common meaning of every monosyllabic noun matches the intended meaning in the compound. In contrast, in every low-score compound, the most common meaning of at least one monosyllable does not match the intended one in the compound.

A reviewer suggested that two additional factors might play a role. First, besides semantic ambiguity based on homographs (such as 日 *ri* 'Japan' and 日 *ri* for 'sun', which are homophones as well), we should also consider the role of homophones that are not homographs (such as 舞 *wu* 'dance' and 午 *wu* 'noon'). In other words, for each N in a NN compound, we should consider whether there is another (more common) word that is written differently but sounds the same, in order to determine whether the meaning of that word interferes with the intended interpretation of the compound. Second, we should consider whether a compound we generated already exists in the language, because it is possible that an existing compound may have a higher score than a newly created one. Both suggestions are adopted, to be discussed below.

4.3. Summary

By comparing high-score and low-score items in each length type, we have found five possible new factors that influence acceptability judgment, shown in (28).

(28) Possible new factors that influence acceptability scores
- a. The 'naturalness' factor: whether the referent is common or natural.
- b. The 'homograph ambiguity' factor: whether the most common meaning of a written graph is the same as the intended one in the given compound.
- c. The 'non-homograph ambiguity' factor: whether there is a more common word that has the same pronunciation as but a different graph and a different meaning from the intended one in a given compound.
- d. The 'boundness' factor: whether a monosyllabic word is bound or free.
- e. The 'frequency' factor: whether a generated NN compound already exists in the language.

In the next section, we offer an expanded analysis that takes the new factors into consideration, along with the original factor of length.

5. An expanded study with new factors

We first discuss the annotation of each new factor and its effect individually. Then we offer a linear mixed effects model analysis that considers all factors together. Finally, we offer a further examination of 1+1 compounds in order to explore a possible explanation of the discrepancy between the common intuition that they are productive and the fact that they are ranked rather low in our study.

5.1. The 'naturalness' factor

This factor refers to whether the referent of a compound is common or natural. The annotation is based on the majority judgment of three native speakers (not including the authors). Consider the examples in (29).

(29) Examples of 'naturalness'

Compound	Gloss	'Natural' votes	Annotation
小麦-价格	'wheat price'	3/3	natural
楼房-表面	'building surface'	2/3	natural
肿瘤-瓶子	'tumor bottle'	2/3	natural
凳子-名字	'stool name'	1/3	unnatural
心脏-质量	'heart quality'	0/3	unnatural

If a compound is deemed to be natural by all three judges, such as 'wheat price', or by two of them, such as 'building surface' and 'tumor bottle', it is annotated as 'natural'. Otherwise, a compound is annotated as 'unnatural', such as 'stool name' and 'heart quality'.

The statistical effect of the 'naturalness' factor on acceptability judgment is shown in (30), (31), and Fig. 3.

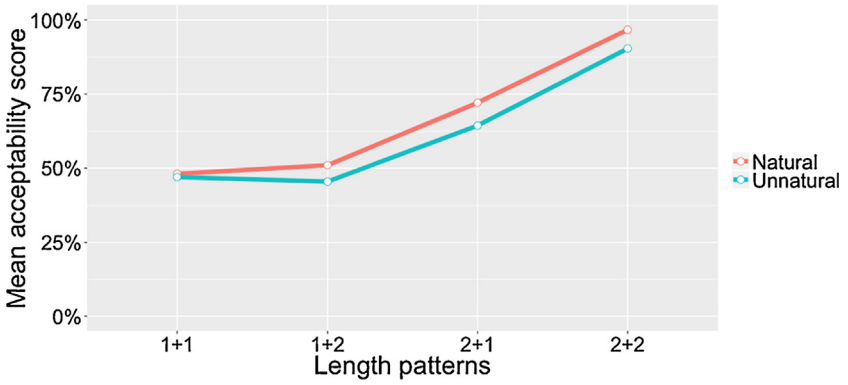

Fig. 3. Mean acceptability scores of natural and unnatural compounds of each length pattern.

(30) Overall 'naturalness' effect on acceptability scores

Natural	Acceptability scores	St. dev	N	p-value
Yes	66.9%	24.8%	864	0.002122
No	61.7%	23.6%	268	**

(31) The effect of 'naturalness' on acceptability scores by length pattern, where 'N' in the first column means natural and 'U' means unnatural

Compounds	Acceptability scores	St. dev	N	p-value
2+2 N	96.6%	5.5%	216	4.854e−06
2+2 U	90.3%	10.0%	67	***
2+1 N	72.0%	15.7%	216	0.000744
2+1 U	64.3%	15.9%	67	***
1+2 N	50.9%	17.7%	216	0.01759
1+2 U	45.4%	15.9%	67	*
1+1 N	48.0%	18.9%	216	0.6647
1+1 U	46.9%	17.7%	67	

It can be seen that 'naturalness' has a significant effect overall, where 'natural' compounds have a higher mean acceptability score. In addition, for three of the four length patterns (1+2, 2+1, and 2+2), the same effect is found as well.

5.2. The '*homograph ambiguity*' factor

This factor refers to whether a word has a more common competitor, written the same and usually pronounced the same as well, whose meaning differs from the intended one in a compound. The most common meaning of a graph is determined by consulting the word frequency list of Guojia Yuwei (2012) and the BCC Corpus (Xun et al., 2016). In general, disyllabic words have no homographic competitor, but monosyllabic words often do. Consider the examples in (32).

(32) Examples of 'homograph ambiguity' in 'salt profit'

Compound	Target N	Intended	Competitor	Annotation
食盐-利润	食盐	'salt'	(none)	unambiguous
食盐-利润	利润	'profit'	(none)	unambiguous
盐-利润	盐	'salt'	(none)	unambiguous
食盐-利	利	'profit'	'advantage'	ambiguous

The disyllabic 食盐 'salt' and 利润 'profit' have no competing meanings, nor does the monosyllabic 盐 'salt'. However, the monosyllabic 利 'profit' has a more common competitor, which is 'advantage'. Therefore, the intended meaning 'profit' in the 2+1 form of 'salt profit' is ambiguous.

The statistical effect of 'homograph ambiguity' on acceptability judgment is shown in (33), (34), and Fig. 4. Since disyllabic words are unambiguous, the 'homograph ambiguity' factor is not applicable to 2+2 compounds.

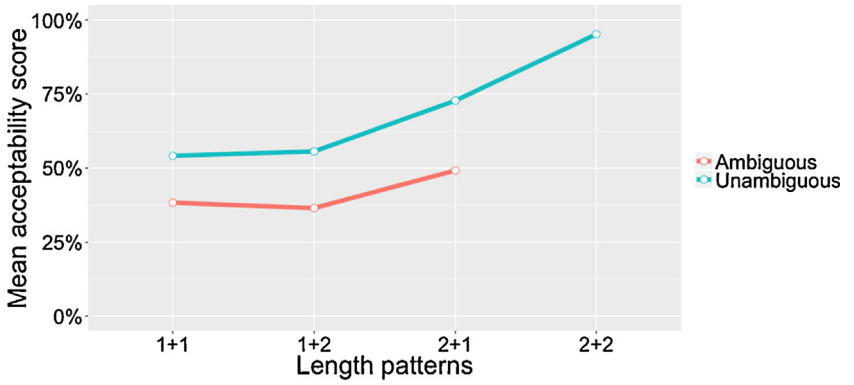

Fig. 4. Mean acceptability scores of homographically ambiguous and unambiguous compounds of each length pattern.

(33) Overall 'homograph ambiguity' effect on acceptability scores

Ambiguity	Acceptability scores	St. dev	N	p-value
Yes	39.0%	13.9%	231	<2.2e−16
No	62.2%	18.4%	618	***

(34) The effect of 'homograph ambiguity' on acceptability scores by length pattern, where 'A' means ambiguous and 'N' means unambiguous

Compounds	Acceptability scores	St. dev	N	p-value
2+1 A	49.1%	14.7%	30	6.702e−10
2+1 N	72.7%	14.3%	253	***
1+2 A	36.5%	12.2%	88	<2.2e−16
1+2 N	55.6%	16.2%	195	***
1+1 A	38.3%	13.8%	113	1.44e−14
1+1 N	54.1%	18.7%	170	***
2+2 U	95.1%	7.4%	283	

It can be seen that 'ambiguity' has a significant effect overall, where ambiguous compounds have a lower mean acceptability score. In addition, 'ambiguity' has a significant effect for the length patterns 2+1, 1+2, and 1+1, where ambiguous compounds have a lower mean acceptability scores.

5.3. The '*non-homograph ambiguity*' factor

This factor refers to whether a word has a more common competitor, pronounced the same but written differently, whose meaning differs from the intended one in a compound. The most common meaning of a pronunciation is determined by consulting a word frequency list (Guojia Yuwei, 2012). In our data, disyllabic words have no such competitors, but monosyllabic words often do. Consider the examples in (35).

(35) Examples of 'non-homograph ambiguity' in 'bee tooth'

Compound	Target N	Homophone	Competitor?	Annotation
蜜蜂-牙齿	蜜蜂 'bee'	密封 'sealed'	No	unambiguous
蜜蜂-牙齿	牙齿 'tooth'	(none)	No	unambiguous
蜂-牙齿	蜂 'bee'	风 'wind'	Yes	ambiguous
蜜蜂-牙	牙 'tooth'	芽 'bud'	No	unambiguous

The disyllabic 蜜蜂 'bee' has a non-homographic homophone 密封 'sealed', but the latter is less frequent and so not a competitor. The disyllabic 牙齿 'tooth' has no homophone in our frequent list and so no competitor either. The monosyllabic 蜂 'bee' has a non-homographic homophone 风 'wind', which occurs more frequently and is a competitor. Finally, the monosyllabic 牙 'tooth' has a non-homographic homophone 芽 'bud', but the latter is less frequent and not a competitor.

The statistical effect of 'non-homograph ambiguity' on acceptability judgment is shown in (36), (37), and Fig. 5. Since no disyllabic word is found to be ambiguous, the 'non-homograph ambiguity' factor is not applicable to 2+2 compounds.

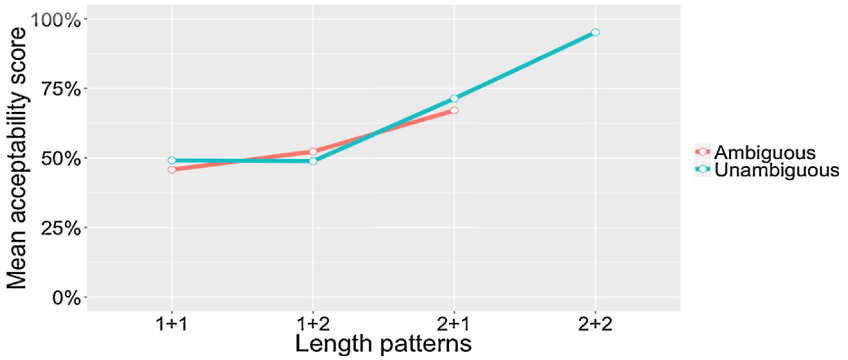

Fig. 5. Mean acceptability scores of compounds with or without a more common competitor that is homophonous but non-homographic, for three of the four length patterns.

(36) Overall 'non-homograph ambiguity' effect on acceptability scores

Ambiguity	Acceptability scores	St. dev	N	p-value
Yes	53.5%	18.3%	251	0.01977
No	56.9%	20.8%	598	*

(37) The effect of 'non-homograph ambiguity' on acceptability scores by length pattern, where 'A' means ambiguous and 'U' means unambiguous

Compounds	Acceptability scores	St. dev	N	p-value
2+1 A	67.0%	16.9%	72	0.06134
2+1 U	71.3%	15.7%	211	
1+2 A	52.3%	14.1%	62	0.1184
1+2 U	48.9%	18.2%	221	
1+1 A	45.8%	16.4%	117	0.1273
1+1 U	49.2%	19.9%	166	
2+2 U	95.1%	7.4%	283	

It can be seen that the overall effect of 'non-homographic ambiguity' is marginal. In addition, no significant effect of this factor is found for any specific length pattern. This is perhaps expected for an experiment with written data, rather than auditory data.

5.4. The '*boundness*' effect of monosyllabic words

The 'boundness' factor refers to whether a morpheme is 'free' or 'bound'. According to Bloomfield (1933), a morpheme is free if (i) it can occur alone, or (ii) if it can serve as an independent syntactic constituent; otherwise it is bound. It can be seen that a morpheme that satisfies (i) also satisfies (ii), while some morphemes only satisfy (ii) but not (i). This is true in both English and Chinese, as illustrated in (38).

(38) Two definitions of boundness (Bloomfield, 1933)

	Morpheme	(i) Alone?	(ii) Syntactic unit?	Free?
English	*that*	Yes	Yes (*know* <u>*that*</u>)	Yes
English	*the*	No	Yes (<u>*the*</u> *question*)	Yes
English	*un-*	No	No	No
Chinese	水 'water'	Yes	Yes (喝<u>水</u> 'drink water')	Yes
Chinese	华 'China'	No	Yes (访<u>华</u> 'visit China')	Yes
Chinese	们 'plural'	No	No	No

This view is adopted in many studies of Chinese, such as Packard (2000:12) and XDHYCD (2005). However, the definitions do not reveal differences in other syntactic environments, such as a possessive structure, shown in (39).

(39) Boundness in a possessive structure [A de B]

Before *de*	Gloss
水的质量 shui de zhiliang	'water's quality'
*华的人口 Hua de renkou	'China's population'

Although both 'water' and 'China' can serve as an object, 'water' can serve as A in [A de B], while 'China' cannot. In order to examine the role of such additional difference in boundness, we adopt an idea from Sproat and Shih (1996:67) and annotated all monosyllabic items in 2+1, 1+2, and 1+1 according to (40).

(40) New boundness annotation:
A monosyllabic item is free if it can serve as A in [A *de* B]; otherwise it is bound.

Our annotation was based on the majority judgment of three native speakers. Two independent native speakers annotated every item first. When they agree, their decision is used. When they disagree, one of the present authors broke the tie. The statistic effect of boundness is shown in (41), (42), and Fig. 6. Since disyllabic words are free, the 'boundness' factor is not applicable to 2+2 compounds.

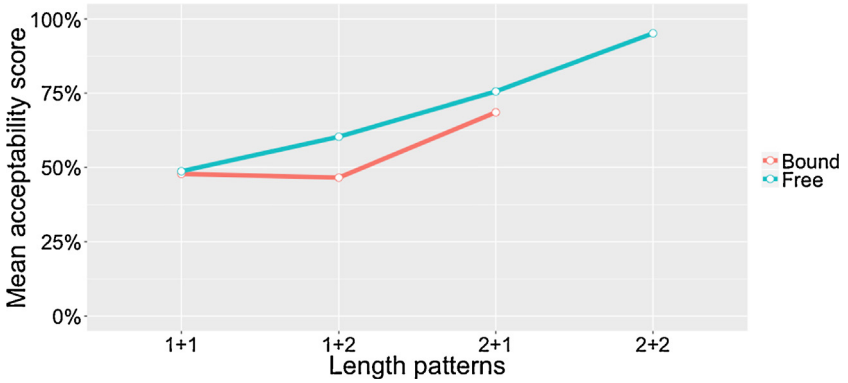

Fig. 6. Mean acceptability scores of each length pattern of compounds by boundness.

(41) Overall 'boundness' effect on acceptability scores

Compounds	Acceptability	St. dev	N	p-value
Free	66.6%	17.5%	142	2.713e−13
Bound	53.7%	20.0%	707	***

(42) Boundness effect on acceptability scores of each length pattern

Compounds	Acceptability	St. dev	N	p-value
2+1 free	75.5%	12.0%	67	0.0002544
2+1 bound	68.6%	16.8%	216	***
1+2 free	60.3%	17.8%	64	2.768e−07
1+2 bound	46.5%	16.1%	219	***
1+1 free	48.6%	14.7%	11	0.8507
1+1 bound	47.8%	18.8%	272	
2+2 free	95.1%	7.4%	283	

The analysis shows that the 'boundness' factor has a significant effect overall. In addition, the factor has a significant effect for the length patterns 2+1 and 1+2, but not for 1+1, probably because there are too few 1+1 compounds where both parts are free.

5.5. The '*frequency*' factor

A reviewer suggested that we consider whether a generated compound already exists, in order to find out whether is any familiarity effect on acceptability judgment. Since dictionaries in general do not collect all compounds, and Chinese dictionaries in particular rarely collect 2+2 compounds (unless they have a special meaning), we decided to look up every compound we used in the BCC Corpus (Xun et al., 2016), a text corpus consisting of about 10 billion words. In addition, since the BCC Corpus offers a frequency count for each search item, we annotate every compound by its frequency (error returns were manually examined and excluded). Some examples are shown in (43).

(43) Frequencies of some compounds, obtained from the BCC Corpus

Length	Compound	Gloss	Annotation
2+2	小麦-价格	'wheat price'	118
2+1	小麦-价	'wheat price'	0
1+2	麦-价格	'wheat price'	0
1+1	麦-价	'wheat price'	10
2+2	歌曲-质量	'song quality'	3
2+1	歌曲-质	'song quality'	0
1+2	歌-质量	'song quality'	0
1+1	歌-质	'song quality'	1

We used Pearson's analysis for the correlation between acceptability judgment and compound frequency. The result is shown in (44).

(44) Pearson correlation analysis between frequency and acceptability judgment

Length	Correlation	p-value
All	0.06639914	0.02548*
2+2	0.04175702	0.48410
2+1	0.1363444	0.02178*
1+2	0.1523414	0.01028*
1+1	0.2582944	1.078e−05***

It can be seen that there is a weak overall correlation between frequency and acceptability judgment. Among individual length patterns, the correlation is significant and strongest for 1+1, followed by 1+2 and 2+1, but not significant for 2+2.

If we treat frequency as a binary factor, converting the value 0 to 'non-existing' and 1 or more to 'existing', the result is similar, as shown in (45), (46), and Fig. 7.

(45) Overall effect of a binary 'frequency' factor on acceptability scores

Compounds	Acceptability	St. dev	N	p-value
Existing	78.2%	21.1%	256	<2.2e−16
Non-existing	62.0%	24.4%	876	***

(46) Binary 'frequency' effect on acceptability scores of each length pattern

Compounds	Acceptability	St. dev	N	p-value
2+1 Existing	76.6%	14.6%	68	0.0001023
2+1 Non-existing	68.2%	16.0%	215	***
1+2 Existing	62.8%	22.3%	17	0.02096
1+2 Non-existing	48.8%	16.8%	266	*
1+1 Existing	61.0%	18.4%	81	1.005e−12
1+1 Non-existing	42.5%	15.9%	202	***
2+2 Existing	97.9%	3.2%	90	1.056e−08
2+2 Non-existing	93.8%	8.3%	193	***

The binary frequency factor has a significant effect overall. In addition, it has a significant effect on all length patterns; it is strongest for 1+1, followed by 1+2 and 2+1, and weakest for 2+2.

5.6. A linear mixed effects model analysis

We considered the individual effects of five new factors: naturalness, homographic ambiguity, non-homographic ambiguity, boundness, and frequency. In this section we consider their effects together. Following Norman (2010), Gibson et al. (2011) and Kizach (2014), we assume that a linear mixed-effects model is applicable to categorical scores. We fitted a

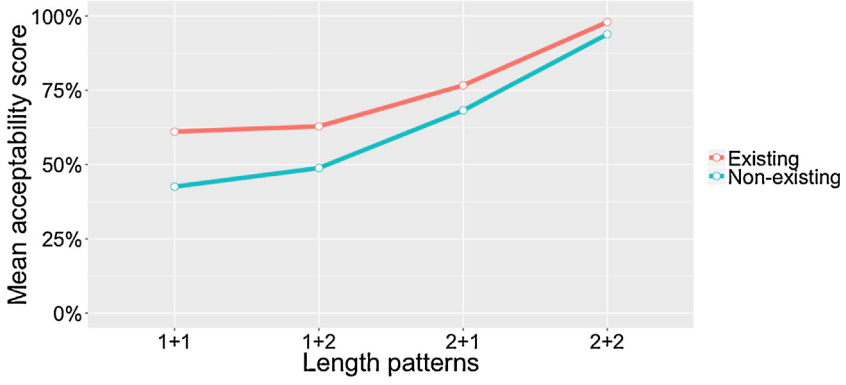

Fig. 7. Average acceptability scores of each length pattern of compounds by 'frequency', when it is treated as a binary factor (existing vs. non-existing).

linear mixed-effects model of the relationship between acceptability scores and all the relevant factors, using RStudio (2012) and lme4 (Bates et al., 2015). Specifically, we included length patterns, naturalness, homograph ambiguity, boundness, and frequency as fixed effects and subjects and items as random effects. We used 'frequency' as a binary factor ('existing' vs. 'non-existing'). Because 'non-homograph ambiguity' was found to have little effect (section 5.4), we did not include it here. We obtained intercepts for both subjects and items and by-subject and by-item random slopes for the effect of length patterns. The main results of fixed effects are shown in (47). For lack of space, other statistics are not shown.

(47) Fixed effects in a linear mixed-effects model where length, homograph ambiguity, boundness, naturalness and frequency are fixed effects and subjects and items as random effects. 1+1 is used as the intercept.

	Estimate	Std.error	Df	t.value	Pr(>|t|)
(Intercept)	0.48989	0.04089	17.800	11.980	5.85e−10***
Length1+2	0.00771	0.02608	24.400	0.296	0.76997
Length2+1	0.16703	0.02875	22.300	5.809	7.21e−06***
Length2+2	0.36230	0.04577	19.400	7.916	1.68e−07***
H ambiguityY	−0.17897	0.01242	792.200	−14.415	<2e−16***
BoundnessY	0.03942	0.01393	659.200	2.829	0.00481**
NaturalnessY	0.05641	0.00752	675.800	7.503	1.97e−13***
FrequencyY	0.05434	0.00737	538.200	7.369	6.50e−13***

As can be seen, all the fixed effects in (47) are significant predictors of acceptability scores. With regard to length, the results are similar to what we saw in the preliminary study (section 3), namely, when all else being held constant, 2+2 and 2+1 have significantly higher acceptability scores than 1+1, while there is no significant difference between the acceptability scores of 1+1 and 1+2.

The estimated co-efficient values also tell us the effect sizes of the factors. Specifically, length has the greatest effect size, a 16.7% jump in acceptability score from 1+1 to 2+1, and a 36.2% jump from 1+1 to 2+2. The 'homograph ambiguity' factor has the next greatest effect, a drop of 17.9% when a compound is ambiguous. The effects of boundness, naturalness, and frequency are much smaller, all of which are below 6%.

5.7. Further examination of 1+1 compounds

A reviewer points out that the low ranking of 1+1 compounds in our study does not agree with a common observation that 1+1 compounds are highly productive, and urges us to look for a further explanation. Specifically, the reviewer suggests that the culprit could be the presence of elastic words: when the first noun of a 1+1 compound has no elastic length, the compound seems perfect, such as 猪肉 *zhurou* 'pig-meat (pork)' and 牛肉 *niurou* 'cattle-meat (beef)'. The reviewer further suggests that many 1+1 compounds are similar to 'pork' and 'beef', which may have given rise to the intuition that 1+1 compounds are productive and good overall, despite the presence of some 1+1 compounds made of elastic words.

We agree with the reviewer that, with regard to 1+1 compounds, our result applies to those made of elastic words only, and not to all 1+1 compounds. To consider the effect of elastic length, and its interaction with all the factors that we have considered above, clearly requires another full-scale study, which is likely to be larger than the present one. Nevertheless, we find the reviewer's proposal to be of interest and conducted a further study to verify it.

The first step is to find out how many 1+1 compounds involve elastic words and how many do not. To find out, we used the Lancaster Corpus of Mandarin Chinese (LCMC, McEnery and Xiao, 2004) and extracted all 1+1 NN compounds from Corpus G 'biographies and essays', whose style is the most neutral and whose size is the largest among all the styles. The preliminary result yielded some 20,000 tokens (i.e. including repetitions of the same compound), which represented some 5,000 types (i.e. excluding repetitions of the same compound). Then we randomly selected 10% of them for manual inspection of errors. This yields a total of 161 valid 1+1 NN compounds (representing 1,610 1+1 NN compounds in all).

Next we manually annotated the 161 compounds for its elastic pattern, where 'EE' means both nouns have elastic length, 'NX' means the first noun is not elastic (regardless of the second noun), and EN means the first noun is elastic but the second is not. The result is shown in (48).

(48) Type counts of 1+1 noun-noun compounds in LCMC, corpus G

Length	Occurrence	Example
EE	51.6%	身(体)-前(面) 'body-front (front of body)'
NX	36.6%	猪肉 'pig-meat (pork)'
		脚-力(气) 'foot-strength'
EN	11.8%	竹(子)-纸 'bamboo-paper'

Next we added estimated acceptability values for each length pattern, which is shown in (49).

(49) Estimated acceptability values for each length pattern

Length	Acceptability	Comments
EE	48%	Present study
NX	100%	Reviewer's estimate
EN	74%	Average of EE and NX
All	70%	=51.6%*48%+36.6%*100%+11.8%*74%

The acceptability value of EE is based on the experimental result of the present study, as shown in (12). The acceptability value of NX is based on the estimate of the reviewer, which seems reasonable to us. The acceptability value of EN is based on the average of the acceptability values of EE and NX, based on the assumption that elasticity lowers acceptability. The average acceptability of length patterns is based on the percentage of each length pattern and its acceptability.

The result shows that, as suggested by the reviewer, the estimated average acceptability of all 1+1 NN compounds, with or without elastic words, is 70%, which is similar to that of 2+1. This result is consistent with the fact that many previous scholars regard 2+1 and 1+1 to be good NN patterns, better than 1+2, which is widely regarded to be a bad pattern.

6. Concluding remarks

Word-length preferences in Chinese NN compounds have been observed at least since Lü (1963) and many native scholars assume there to be a generalization, given in (2) and repeated in (50).

(50) A popular generalization on length preferences in NN compounds, from (2)
- a. 1+1 is well formed.
- b. *1+2 is ill formed.
- c. 2+1 is well formed.
- d. 2+2 is well formed.

However, no judgment experiment has been conducted to confirm the generalization. In addition, some scholars remain skeptical. Moreover, some specific questions remain unexplored. For example, are there subtle differences in acceptability judgment among the 'well-formed' forms (2+2, 2+1, and 1+1)? Are there other factors, besides prosody (word length), that influence acceptability judgment?

In this article we report two judgment studies. In a preliminary study, we examined the word length effect only. Then in an expanded study, we examined the effects of five additional factors: 'naturalness' of the referent, 'homograph ambiguity' of a compound, 'non-homograph ambiguity' of a compound, 'boundness' of the component nouns, and the 'frequency' of a compound. Major findings are summarized in (51) and (52).

(51) Major findings of the present studies
- a. Prosody (word length patterns) has a significant effect on acceptability judgment, so do four other factors, namely, 'homograph ambiguity', 'naturalness', 'boundness', and 'frequency'.
- b. Prosody has the largest effect size, followed by 'homograph ambiguity'. The effect sizes of 'naturalness', 'boundness', and 'frequency' are all fairly small.

(52) Comparing present findings with common views
- a. 1+1 also has the lowest mean acceptability score (statistically the same as 1+2), contrary to (50a).
- b. 1+2 has the lowest acceptability score, confirming (50b).
- c. 2+1 has the second highest acceptability score, partially confirming (50c).
- d. 2+2 has the highest acceptability score, confirming (50d).
- e. The effects of 'homograph ambiguity', 'naturalness', 'boundness', and 'frequency' on acceptability judgment have not been proposed or demonstrated in an experimental study.

Our results offer several points of interest. First, it provides experimental confirmation of a popular observation that prosody (length patterns) has a strong effect on the acceptability of NN compounds in Chinese. Second, unlike what is often assumed, such as the generalization in (50), acceptability judgment is not clear cut but gradient, in agreement with

what is found in English (Coetzee, 2004, 2006, 2008). Third, we have identified new factors that influence acceptability judgment, in particular the naturalness of the referent, homograph ambiguity, the boundness of each component noun, and the frequency of a compound. Fourth, the interaction among the factors offers an explanation why acceptability judgment is gradient. For example, with regard to its length pattern, 2+1 does not seem to have a metrical problem, as seen in (3) and (4), but its monosyllabic term may suffer from homograph ambiguity and boundness. Therefore, 2+1 compounds are expected to have a lower mean acceptability score than 2+2. Finally, as seen in section 5.7, the present results are based on compounds that are made of elastic words and should not be taken to be true of all NN compounds. Acceptability judgment of compounds without elastic words requires a separate study, which we intend to conduct in the future.

As a final remark, let us consider the question of whether compounding is a productive process. Dai (1992) proposes that true compounding is a syntactic process and is productive, whereas compounding involving bound roots is a lexical process and need not be productive. In contrast, Sproat and Shih (1996) propose that 'root compounds', which consist of at least one bound root, is productive in Chinese, and that there is no difference between 'true compounds and root compounds', both being productive in Chinese. Our study seems to refute the proposal of Sproat and Shih (1996), since the generated 1+1 compounds in our study, most of which are root compounds, have a low mean acceptability rating and cannot reasonably be called productive. However, we believe there is a more nuanced answer. Given the notion of elastic words, every compound in our study has four length patterns, and so every compound is productive, in the sense that every compound has at least one good length pattern. Thus, the notion of elastic words may have provided an answer to a standing controversy.

Acknowledgements

This study is supported by the China Scholarship Council [Grant No. 201500850002] and the Fundamental Research Funds for the Central Universities [Grant No. 2016SZYQN41]. Part of this paper was presented at the phonetics and phonology group meeting at the University of Michigan. We would like to thank the audience for their comments, in particular George Allen, Pam Beddor, and Andries Coetzee.

References

Bates, D., Maechler, M., Bolker, B., Walker, S., 2015. lme4: Linear Mixed-Effects Models Using Eigen and S4. R Package Version 1.1-12.
Bloomfield, L., 1933. Language. Henry Holt, New York.
Chao, Y.R., 1948. Mandarin Primer, an Intensive Course in Spoken Chinese. Harvard University Press, Cambridge, MA.
Chomsky, N., Halle, M., 1968. The Sound Pattern of English. Harper and Row, New York.
Chomsky, N., Halle, M., Lukoff, F., 1956. On accent and juncture in English. In: Halle, M., Lunt, H., MacLean, H., van Schooneveld, C. (Eds.), For Roman Jakobson. Mouton, The Hague, pp. 65–80.
Coetzee, A.W., 2004. What It Means to be a Loser: Non-optimal Candidates in Optimality Theory (doctoral dissertation). University of Massachusetts, Amherst.
Coetzee, A.W., 2006. Variation as accessing 'non-optimal' candidates. Phonology 23.3, 337–385.
Coetzee, A.W., 2008. Grammaticality and ungrammaticality in phonology. Language 84.2, 218–257.
Dai, J.X., 1992. Chinese Morphology and Its Interface with the Syntax (doctoral dissertation). The Ohio State University.
Dong, X., 2002. Cihuihua: Hanyu chuangyinjie ci de yansheng he fazhan (Lexicalization: The Origin and Evolution of Chinese Disyllabic Words). Sichuan Minzu Chubanshe, Chengdu.
Dong, Y., 2015. The Prosody and Morphology of Elastic Words in Chinese: Annotations and Analyses (doctoral dissertation). University of Michigan, Ann Arbor.
Duanmu, S., 1999. Zhongyin lilun he Hanyu de cichang xuanze (Metrical theory and word length choices in Chinese). Zhongguo Yuwen 4, 246–254.
Duanmu, S., 2007. The Phonology of Standard Chinese, 2nd edition. Oxford University Press, Oxford.
Duanmu, S., 2012. Word-length preferences in Chinese: a corpus study. J. East Asian Linguist. 21 (1), 89–114.
Feng, S., 1996. Lun Hanyu de yunlüci (On the prosodic word in Chinese). Zhongguo Shehui Kexue 1, 161–176.
Feng, S., 2013. Hanyu yunlü jufaxue (zengding ben), (Chinese Prosodic Syntax). expanded edition. Shangwu Yinshuguan, Beijing.
Gibson, E., Piantadosi, S., Fedorenko, K., 2011. Using mechanical Turk to obtain and analyze English acceptability judgments. Lang. Linguist. Compass 5 (8), 509–524.
Guo, S., 1938. Zhongguo yuci zhi tanxing zuoyong (The function of elastic word length in Chinese). Yen Ching Hsueh Pao 24, 1–34.
Guojia Yuwei[State Language Commission], 2012. Xiandai Hanyu Yuliaoku Ciyu Fencilei Pinlü Biao. (A Corpus-Based Frequency List of POS-Distinguished Words in Modern Chinese). www.cncorpus.org (accessed on 10.05.16)
Halle, M., Vergnaud, J.-R., 1987. An Essay on Stress. MIT Press, Cambridge, MA.
Hayes, B., 1995. Metrical Stress Theory: Principles and Case Studies. The University of Chicago Press, Chicago.
Huang, L., Duanmu, S., 2013. Xiandai Hanyu ci chang tanxing de lianghua yanjiu (A quantitative study of elastic word length in Modern Chinese). Yuyan Kexue 12 (1), 8–16.
Karlgren, B., 1918. Ordet och pennan i Mittens Rike. Svenska Andelsförlaget, Stockholm.

Ke, H., 2007. A Study of Monosyllabic and Disyllabic Usage in Modern Chinese (doctoral dissertation). Institute of Linguistics, Chinese Academy of Social Sciences, Beijing.
Kennedy, G.A., 1951. The monosyllabic myth. J. Am. Orient. Soc. 71.3, 161–166.
Kizach, J., 2014. Analyzing Likert-scale data with mixed-effects linear models – a simulation study. In: Poster presented at Linguistic Evidence 2014. . University of Tübingen, Germany.
Lü, S., 1963. Xiandai Hanyu dan shuang yinjie wenti chu tan (A preliminary study of the problem of monosyllabism and disyllabism in modern Chinese). Zhongguo Yuwen 1, 10–22.
Lu, B., Duanmu, S., 2002. Rhythm and syntax in Chinese: a case study. J. Chin. Lang. Teach. Assoc. 37 (2), 123–136.
McEnery, T., Xiao, R., 2004. The Lancaster Corpus of Mandarin Chinese: a corpus for monolingual and contrastive language study. In: Lino, M.T., Xavier, M.F., Ferreira, F., Costa, R., Silva, R. (Eds.), Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC) 2004, pp. 1175–1178.
Norman, G., 2010. Likert scales, levels of measurement and the ‘‘laws’’ of statistics. Adv. Health Sci. Educ. 15 (5), 625–632.
Packard, J.L., 2000. The Morphology of Chinese: A Linguistic and Cognitive Approach. Cambridge University Press, Cambridge, UK.
Pan, W., 1997. Han Ying duibi gangyao (An Outline Comparison of Chinese and English). Beijing University of Languages and Cultures Press, Beijing.
RStudio, 2012. RStudio: Integrated Development Environment for R (Version 0.99.902) [Computer Software]. Boston, MA. Retrieved May 13, 2016. Available from http://www.rstudio.org/
Sproat, R., Shih, C., 1996. A corpus-based analysis of Mandarin nominal root compound. J. East Asian Linguist. 5 (1), 49–71.
Wang, H., 2001. Yinjie danshuang yinyu zhanlian (zhongyin) yu yufa leixing jiegou he chengfen cixu. Dangdai Yuyanxue 4, 241–252.
Wang, C., 2002. In Jufa zuhe zhong de danshuang yinjie xuanzhe de rengzhi jieshi (A cognitive account of the choice between monosyllabic and disyllabic words in syntactic combination). In: Zhongguo Yuwen Zazhi She (Eds.), Yufa yanjiu he tansuo, (Studies on Grammar).vol. 11. Shangwu Yinshuguan, Beijing, pp. 151–168.
Wang, Y., 2013. Jufa zuhe songjin he ‘de’ de yinxian (Constituency closeness and the occurrence of ‘de’). Hanyu Xuexi 4, 29–34.
Wu, W., 2006. Hanyu yunlü jufa tansuo (A Study of Prosodic Grammar). Xuelin Chubanshe, Shanghai.
XDHYCD, 2005. Xiandai Hanyu Cidian, (Modern Chinese Dictionary. Compiled by the Institute of Linguistics, Chinese Academy of Social Sciences). 5th edition. Shangwu Yinshuguan, Beijing.
Xun, E., Rao, G., Xiao, X., Zang, J., 2016. Da shuju beijing xia BCC yuliaoku de yanzhi (The construction of the BCC Corpus in the age of Big Data). Yuliaoku Yuyanxue 3 (1), 93–118.