Proceedings of the 21st North American Conference on Chinese Linguistics (NACCL-21). 2009.
Volume 1. Edited by Yun Xiao. Smithfield, Rhode Island: Bryant University. Pages 13-27.

Core Vocabulary in Spoken Mandarin and the Integration of Corpus-Based Findings into Language Pedagogy*

Hongyin Tao
University of California, Los Angeles

A key issue in language acquisition is to improve native-like proficiency in vocabulary use. One solution to this is to identify word frequencies (especially in conjunction with core vocabulary) and collocation patterns based on native speakers discourse. In this paper, I first discuss some of the puzzles presented in some long-standing and recent quantitative observations of the Mandarin lexicon. I then discuss high frequency clusters in terms of their unique forms and functions as a way of solving some of the puzzles. Finally I discuss the implications of these findings for language teaching, especially vocabulary teaching.

0. Introduction

One of the most commonly encountered issues in language acquisition is to improve native-like proficiency in vocabulary use, whereby not only individual words are used appropriately, but word combinations are used in ways that are close to how native speakers deploy them in actual communicative contexts. This kind of research has been conducted along the lines of collocation, phraseology, idiom, fixedness, formulaic language, the Idiom Principle, and Lexical Priming, to name just a few (Pawley and Syder 1983, Sinclair 1991, Nattinger and DeCarrico 1992, Lewis 1993, Howarth 1998, McCarthy 1998, Erman and Warren 2000,Wray 2002, Hoey 2004). A key solution to this issue is to identify word frequencies (in conjunction with core vocabulary) and collocation patterns based on native speaker discourse. Fortunately, with the availability of electronic corpora and corpus analysis tools, such tasks have become increasingly manageable (Sinclair 1991, O'Keeffe, McCarthy, and Carter 2007).

Previous research on statistical properties of Chinese has tended to focus on the frequency of use, as well as the standards, of Chinese characters, due understandably to the prominence of characters in the Chinese writing system (GJYW 1988, Chen 1989, 1993, GJHB 1992). More recent work has begun to examine distributional properties of the language itself. Thus the well-known Frequency Dictionary of Modern Chinese compiled by the Beijing Language University (YYXY 1986) provides useful frequency information about various types of lexical items in different genres, as do the recently published Xiao et al. (2009): A Frequency Dictionary of Mandarin Chinese: Core Vocabulary for Learners as well as the frequency dictionary in Liu et al. (1990). However,

* I wish to thank Yun Xiao for the opportunity to present the paper at the NACCL conference and for encouraging me to think along the lines of language pedagogy. All errors are of course mine.

a major drawback with such studies is the lack of natural conversation data, limiting the scope mainly to written texts and spoken prose (Abercrombie 1963). Furthermore, very few quantitative studies have attempted to provide in-depth analyses of patterns of language use beyond simple character/word lists.

Looking beyond the Chinese linguistics realm, we can find that, in the area of word frequency distribution, as early as in the 1930s George Zipf (1935) had made influential proposals about statistical distributional properties of the lexicon, widely known as Zipf's Law. Interestingly, his work also involved data from Beijing Chinese. Among the phenomena discussed by Zipf, the relation of Beijing syllables to the shape of its words is described as high frequency words tending to have fewer syllables ("shorter") while low frequency words tending to have more syllables ("longer"). He pointed out that overall the variety of high frequency words is smaller than that of the low frequency words. These patterns of course fit well with what Zipf observed of vocabulary in general: 1) a small number of lexical items have very high frequencies in natural texts; and 2) in general the magnitude of words tends to stand in an inverse relationship to the number of occurrences. A recent study in Wang (2009) also shows that Zipf's Law applies to the variety of word senses: the more senses a word has, the shorter (and more frequent) it tends to be. While mathematicians have found Zipf's Law to apply to a wide range of physical and social phenomena (e.g. populations of cities), few linguists have attempted to understand the underlying reasons for the observed tendencies other than reiterating Zipf's (1935, 1949) "least effort" principle (Wang 2009). This paper is an attempt at elucidating some of the properties of lexical use, with a goal to demonstrate their relevance to Chinese language pedagogy.

In what follows I will first describe the database of this study. Then general findings from the data will be presented and explanations will be offered. At the end of the paper implications of the findings for Chinese language education will be discussed.

1. Data

My data come from 54 face-to-face conversations, recorded between the 1980s and 2005. The conversations are between native speakers of Mandarin who are generally familiar with each other in various locations in mainland China, Hong Kong, and overseas.

The data were word-segmented and tagged for parts-of-speech (POS) information by the software program ICTCLAS (Zhang, Liu, Zhang, and Cheng 2002, Xiao, Rayson, and McEnery 2009: 3-4), which uses algorithms based on statistical models. A total of 344,141 words were identified by the program.

2. General Patterns

A search of the data shows that there is a general dominance of a small number of lexical types in the corpus. Here, a type is taken to be a unique word as identified by the ICTCLAS program, while a token is any occurrence of the type in the corpus. From this point of view, the data show that the top 100 types account for near 80% of the running words.

	Type	Token	Proportion of tokens in corpus
High frequency	top 100	268,979	78%
Low Frequency	below top 100= 16,940	75,162	22%
Total	17,040	344,141	100%

Table 1: Type-token distribution: top 100 vs. the rest

This finding is clearly in line with Zipf's observation of Beijing Mandarin and other languages. Figure 1 provides another perspective. It gives a breakdown of the top 300 words and their proportions in the corpus: there are 3 words with a frequency of 10,000, 6 with a frequency of 5000, 51 with a frequency of 1000, and so forth. Together they make up a large majority of the corpus. On the other hand, there are over 14,000 words that occur just once in the corpus.

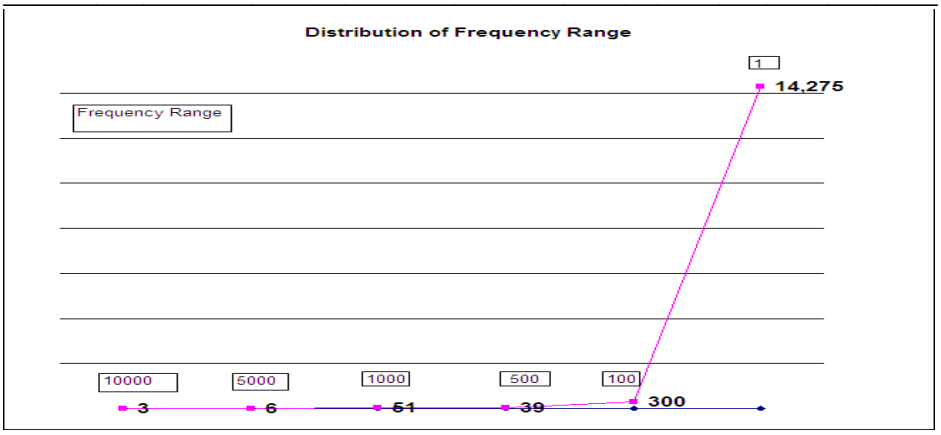

Figure 1: Major bands of words and their proportions in the corpus

In other words, a small number of high frequency words dominate over a large variety of low frequency words.

Given the high concentration of a few high frequency words in spoken discourse, it is natural for us to ponder: What are these words? What categories they may belong to? McCarthy (1999) and McCarthy and Carter (2003) show that in spoken (British) English, the following major categories are common in their data: 1) modal items, e.g. *can, could, should, will, look, seem, sound*, etc.; 2) delexical verbs, i.e. verbs that have low semantic content, e.g. *do, make, take, get*, etc.; 3) interactive markers which are central to spoken communication: *just, whatever, really, things*; 4) discourse markers which organize and monitor the talk, e.g. *I mean, right, so, good, you know*; 5) deictic words which refer to spatial and temporal points, e.g. *this, that, now, ago, away*; 6) basic nouns, e.g. *person, problem, situation, door, water, house, car*, etc. 7) basic adjective, e.g. *good, bad, different, lovely, terrible*; 8) basic adverbs, e.g., *today, yesterday, eventually, finally, usually, normally, quickly, slowly*, etc.; 9) basic verbs, e.g. *sit, give, say, leave, stop, help, feel, put*, etc.

For Mandarin, Tseng (2001:168, 2006:104) identifies 36 high frequency words as the core vocabulary on the basis of a small sample (less than ten thousand words) of spoken Chinese. Her classification is as follows:

7 verbs: 在 zai ‘be in/at’, 是 shi ‘copula’, 就是 jiushi ‘that is’, 说 shuo ‘say’, 去 qu ‘go’, 要 yao ‘want’, 有 you ‘have’;

6 discourse particles: 哦 o, 嗯 en, 哎 ai, 啦 la, 啊 a, 嘛 ma;

5 adverbs：也 ye ‘also’, 就 jiu ‘then’, 都 dou ‘all’, 很 hen ‘very’, 对 dui ‘right’;

4 grammatical particles: 呢 ne; 吗 ma; 了 le; 的 de;

4 nouns: 话 hua (words), 时候 shihou ‘time point’, 人 ren ‘person’, 小孩子 xiaohaizi ‘kids’;

3 na and zhe words: 这样 zheyang ‘this way’, 那个 nage ‘that one’, 那 na ‘that’;

3 pronouns 他 ta ‘he’, 我 wo ‘I’, 你 ni ‘you’;

2 negation: 不 bu ‘not’, 没有 meiyou ‘have not’;

1 adjective 好 hao ‘good’;

1 connective 所以 suoyi ‘so’.

For my data, the top 50 plus items are listed under Table 2. As can be seen from the raw frequencies, a few major groups emerge, with some overlapping with those on Tseng’s list while others not. An initial taxonomy of the core vocabulary can be established as follows.

1) Pronouns: 我 wo ‘I’，你 ni ‘you’，他 ta ‘he’
2) Low content verbs: 是 shi ‘be’，有 you ‘have’
3) Speech act verbs: 说 shuo ‘say’
4) Cognitive verbs: 觉得 juede ‘feel’，知道 zhidao ‘know’，看 kan ‘see, think’
5) Motion verbs: 去 qu ‘go’，到 dao ‘go to’，上 shang ‘get’
6) Adverbs: 就 jiu ‘then’，就是 jiushi ‘then’，都 dou ‘all’，也 ye ‘also’, 很 hen ‘very’，还 hai ‘also’
7) Numeral/Classifiers: 一 yi ‘one’，一个 yige ‘one’
8) Modal expressions: 要 yao ‘would, will, should’
9) Negation: 不 bu ‘not’，没有 meiyou ‘not have’
10) Deixes: 这 zhe ‘this’，这个 zhege ‘this one’，那 na ‘that’，那个 nage ‘that one’
11) Temporal deictic: 然后 ranhou ‘then’，现在 xianzai ‘now’
12) Reactive tokens: 哦 o，嗯 en，啊 a, 对 dui
13) Particles: 吧 ba，呢 ne，嘛 ma，啊 a
14) Interrogatives: 什么 shenme ‘what’
15) Conjunctions: 所以 suoyi ‘so’，而且 erqie ‘and’，但是 danshi ‘but’
16) General nouns: 人 ren ‘person’
17) Basic adjectives: 好 hao ‘good’

1) 的.....13245	19) 那个.....3154	37) 到.....1666
2) 是.....12047	20) 然后.....3076	38) 她.....1606
3) 我.....10052	21) 在.....3067	39) 没.....1590
4) 就.....7782	22) 什么.....3064	40) 吧.....1539
5) 不.....7743	23) 这.....3027	41) 多.....1490
6) 你.....7658	24) 这个.....2772	42) 它.....1474
7) 了.....7484	25) 很.....2373	43) 没有.....1438
8) 那.....6846	26) 哦.....2245	44) 得.....1412
9) 啊.....5792	27) 看.....2197	45) 呢.....1384
10) 个.....4696	28) 人.....2100	46) 跟.....1336
11) 他.....4385	29) 还.....2093	47) 他们.....1335
12) 对.....4285	30) 嗯.....1953	48) 儿.....1326
13) 就是.....3920	31) 好.....1939	49) 上.....1235
14) 有.....3816	32) 要.....1871	50) 吗.....1200
15) 都.....3760	33) 我们.....1847	51) 现在.....1176
16) 说.....3677	34) 去.....1824	52) 知道.....1135
17) 一.....3497	35) 一个.....1814	53) 嘛.....1112
18) 也.....3186	36) 觉得.....1694	54) 但是.....1082

Table 2: Top 50 plus high frequency words in the corpus

3 Understanding core vocabulary in spoken Chinese

3.1. General questions

If, as the results shown above indicate, a limited number of words are doing most of the work in spoken communication, how is this possible? Especially intriguing are the following properties that can be detected from the data:

-That many of the core vocabulary items are not real lexical or high content words. This is illustrated by words such as copula verbs, negation markers, and general nouns.

-That most of them can not stand alone. This is illustrated by words such as conjunctions, particles, and adverbs. One cannot typically make up an utterance with these words alone, as they rely heavily on the context provided by other words and expressions.

Given the above, why, then, would these lexical items be so frequent and be able to make up much of the talk/text?

Clearly, some of the usage patterns are transparent given the nature of conversation. For example, utterance-final particles are probably not too surprising given that one can practically not produce a spontaneous utterance in Chinese without attaching a final particle to indicate its pragmatic nuance. We can also safely anticipate the use of

person pronouns, which typically indicate speaker roles, and the use of reactive tokens, which regulate speaker interaction (Clancy et al. 1996). Yet, many on the top list demand an explanation. For example,

-Why would there be so many copulas?

-Why cognitive verbs?

-Why so many conjunctions if spoken language is supposed to be fragmented, short, and simple?

-Why so many negatives?

-Why do distal demonstratives outnumber proximal ones if conversation is supposed to be about "here and now"?

While there are no quick answers to any of the above questions, and a full-fledged study is certainly beyond the scope of the present paper, we can at least explore some possibilities with a few selected items here.

3.2. A proposal

In contrast with the dominant approach to meaning and vocabulary that emphasizes the single lexical words as a unit of meaning (Chao 1968), I propose that the key to a proper understanding of the puzzles presented in the quantitative data is to look beyond the single words and take multi-word units as a valid unit of meaning (Sinclair 1991, 1996, McCarthy 2002). That is, in addition to the meanings and grammatical patterns typically found in dictionary definitions and grammatical descriptions of individual words, most of these lexical items have special collocation patterns, constituting fixed or semi-fixed expressions; often they combine with one another and function as expanded phrasal units. These units tend to have specialized pragmatic meanings and functions and often play multiple roles in spoken discourse, resulting in mismatches between lexical forms and functions.

In other words, the individual frequency when used separately, the frequency of combinations involving these lexical items, as well as the extended meanings and functions beyond the lexical meanings, give rise to the statistical and functional prominence of these lexical items in spoken discourse.

3.3. A case study of cognitive verbs: 知道 zhidao 'to know'

In this section, I take on the case of one cognitive verb and demonstrate how individual items and the associated combinations work to create high frequency expressions.

Cognitive verbs such as *zhidao* 'to know' are typically taken to indicate mental states, cognitive abilities, and so forth. They are considered syntactically interesting as they can take a variety of objects, including complements (Meng et al. 1999). E.g.,

(1) 我也不是学西医，知道一点而已。
'I'm not a specialist in Western medicine, so I know just this much.'

(2) 其中有一个问题就是问他们打 - 有没有打流感预防针，然后说知不知道要多久

打一次第 - 流感预防针,

'One of the questions they asked them was whether or not they had had the flu shot. Then they asked whether they knew how often flu shots were given.'

In the first example, the object is a simple nominal, while in the second a complement clause. In both cases, the verb *zhidao* denotes a cognitive meaning, i.e. the possession of knowledge or lack thereof.

However, discourse data show that the attested patterns are quite different from the expected syntactic behaviors. In a previous study, Tao (2003) shows that half of the *zhidao* cases in the conversation corpus do not take any objects.

With Objects	55	47%
Without Objects	58	50%
Other	4	3%
Total	117	100%

Table 3. The syntax of zhidao in conversation

Furthermore, there are numerous combinations which function as special constructions with special meanings beyond the typical lexical semantics of the verb. One common collocation is 不知道 bu zhidao 'don't know'. Many of these combinations indicate an epistemic meaning, where the speaker is taking a stance to show a lack of commitment as to the source or truthfulness of the statement. E.g.,

(3)男: 那个梅, 梅市长我不知道为什么那个..升的真快, 他..

'The mayor, Mayor Mei, I don't know why he was promoted to fast, he must be..'.

In this segment, taken from a reporter's conversation with a colleague after they both interviewed a mayor, shows an apparent lack of knowledge. However, upon further examination of the conversation, one can see that the same speaker continued the conversation with an explanation of the mayor's rise to prominence. This shows that the lack of commitment is not due to cognitive deficiencies such as memory lapses, as the speaker did provide a full account of the mayor's professional history, but rather is a lack of epistemic commitment. The likely motivation here is that the speaker was trying to avoid creating an impression that he was in possession of knowledge that was lacked by his fellow reporter. As the concordance lines show, a combination of 我也不知道 wo ye bu zhidao 'I just don't know', though not all being an epistemic phrase, contributes to the high frequency of four of the top items on the frequency list: *wo*, a person pronoun; *ye*, an adverb; *bu*, a negator; and *zhidao*, a cognitive verb.

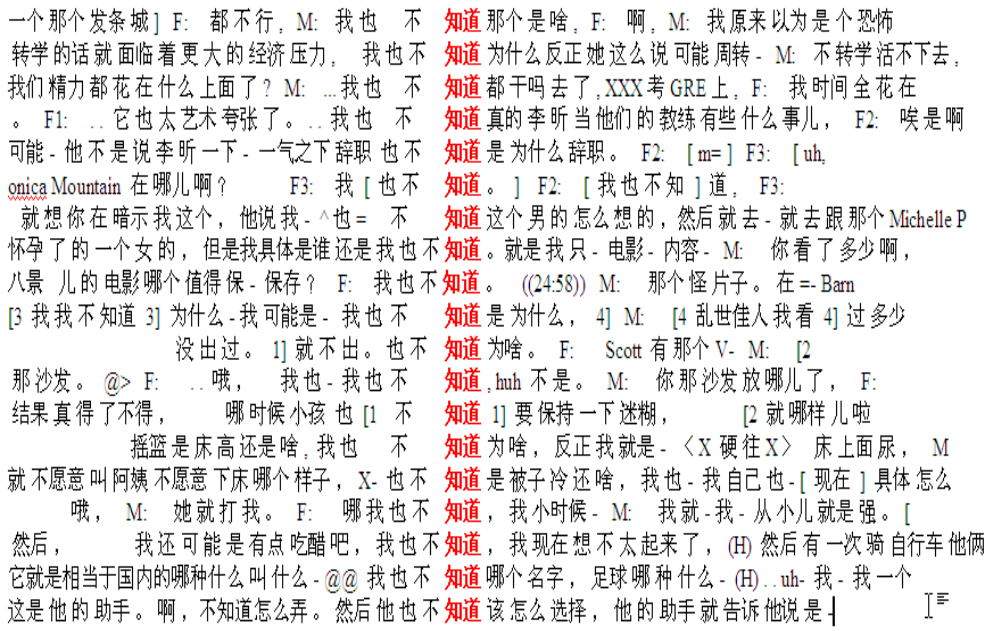

Figure 2: Concordance lines of *(wo) bu zhidao*.

Another common collocation involving *zhidao* is the phrasal unit 你知道 ni zhidao 'you know'. This expression functions in similar ways as the English discourse marker 'you know' (Schiffrin 1988) in that they both function as an involvement device to draw the address's attention. However, what is interesting in Mandarin Chinese is that there is usually an interrogative particle 吗 ma or 吧 ba attached to the subject-verb structure, making it apparently an interrogative form. However, in actual use it is not always a genuine question – and in fact it is usually not. Here is an example of *ni zhidao*.

(4) M: 那那是夹竹桃吧,
F: 不是, 是桃花啊, 你知道吗?
M: 夹竹桃吧,
'M: That looks like oleander. F: No, it's peach blossom, you know? M: It seems more like oleander.'

In this example, since the firs speaker begins by asking for confirmation, the second speaker's use of the apparent question with *zhidao* can only be interpreted as a confirmation token rather than a genuine question.

If we analyze the composition of examples such as (4), we can see that three common items on the high frequency list can be accounted for: *ni*, a second person pronoun; *zhidao*, a cognitive verb; and *ma*, a final particle. Again a phrasal unit with a special construction status and with special pragmatic meanings account for the high frequency of multiple lexical items. Of course this is not to suggest that such environments are the only ones in which the three items are used, but this does point to at

least one common place that contributes to the high frequency of the component elements in Mandarin conversation.

One way to show the fixedness of these phrasal units, *wo bu zhidao* and *ni zhidao (ba/ma)*, etc., is to look at the flexible positions they take in the stream of speech. That is, rather than taking a complement or any objects at all, they often appear at the end of a completed clause, rendering them a parenthetical status. Here is an example of *wo bu zhidao*:

(5)他这最多可以写多少字我也不知道，但是我反正曾经写过三十个字。

'How many characters he can write this way, I am not really sure about, but I used to write about 30.'

In this case the whole *wo bu zhidao* construction appears right after a complex clause. In the following example, *ni zhidao* is placed in the middle of a longer utterance:

(6) B: 而且我们这儿你知道不知道人家线路怎么走，看车辆牌子全一样。
A:对（笑）。

'B: In here we, as you know, we don't know how the locals get around; all those bus stop signs look the same. A: Exactly.'

那 那是 ya- 竹 桃 吧， F: 不 是， 是 桃花 啊， [你 知道 吗 ？] M: [夹竹桃 吧，] F: 桃花 也 有 啊，
[2@@@2] F2: [1 问问， 对 啊 = ， 1] [2 你 知道 吗 ? 真的 很 2] 残忍， 在 我们 那边 发生 了 ^ 好 多 起 = ，
从 国内 到 美国 来 , 这 感觉 这 不一样 啊 , 你 知道 我 说什么 , 这 肯定 是 不一样 的 , 我 要是 就是
- 非常 - 难受 的 地方 。 M1: 可是 他 有 人 ^ 喜欢 你 知道 吗 , M2: 真的 吗 ? M1: 对 啊 , M2:
M: @@@ F: 那时候 根本 不能 喝水 你 知道 为啥 ? 不能 喝水 ? 因为 根本 没办法 上厕所 , 蹲 不 下去 ,
[干 的 ?] F: 非常 好吃 , F: 嗯 , 你 知道 , M: 干 的 辣椒面 儿 还 调制 啊 ? F: 你 知道 我
有 -a^ccounting 有 个 <Xgoodwill X> 嘛 , 你 知道 , 还有 -um,.. 还有 就 说 =.. 还有 股票 , 一些 股票 啊 =,
怎么 一个 原因 , M1: 你 说 - F: 因为 你 知道 他们 高中生 你 平时 就 要 做 好 多 好 多 的 实验 , 什么 =
F: 真的 是 很 孤单 很 孤单 , 而且 那个 , 因为 你 知道 那 不是 你 最后 落脚 的 地方 , 那些 人 也 不 把 你 当 成 -
M2: 你 - 你 不 知道 ^ 打 哪儿 ^ 块 = 你 知道 吗 , @@@@ 你 不 知道 怎么 , 然后 我 就 - ^ 把 他 ^ 全
- 因为 我 可 不 习惯 女孩 儿 之间 这 一套 你 知道 吗 , M: uh= F: 因为 我们 X 家的 女孩儿
M2: 没有 , 外国 女人 没有 女人味 你 知道 吗 , 个个 像 男人 一样 。 M1: .. 没有 , F:
[它 这个 收音 ,] F: 小时候 你 知道 那 报纸 上 什么 什么 什么 什么 少年 什么 那 一类 的 报纸
F3: 小 卫生间 , 后来 那个 味道 - 就是 那 - 就 你 知道 那个 气 , F2: 对 , F3: 觉得 不对头 , F
买 2] 辆 新车 它 不 会 开 到 这 车 -... 就 ^ 一直 开 你 知道 吗 , M2: 对 , M1: 他 - 开 几 年 他 就
M: 习惯 了 。 F: 你 不 用 ^ 想 = 你 知道 吗 , 看 那 种 东西 你 就是 更 更 更 更 就 看 电视 一样 就
就 要 蹦 出 一个 英语 单词 来 那 <@ 一 种 感觉 你 知道 吗 ?@> huh 特别 - 特别 难受 看 了 那 一 台 那 个
我 接 - 接受 吃 肉 , 但是 我 - 你 知道 我 吃得 太 多 我 不行 = , uh uh,(H) F3: 我 -
- 在 医学 上 叫做 ^ 异体 排斥 嘛 , 异体 排斥 你 知道 吗 ? 就是 说 , eh,tsk((orlighter'ssoun
哪 天 比赛 你 没 见 着 , . . . 明星 。 明星 教练 你 知道 吧 , ")) F2: qi- (Hx), 他 把 别人 当 白痴

Figure 3: Concordance lines involving *ni zhidao*.

For a full account of the syntactic, semantic, discourse, and phonological properties

associated with *zhidao* constructions, the reader is referred to Tao (2003). Suffice it to say here that this cognitive verb is by no means a rarity, and that there are multiple combinations involving a large number of common words found in the high frequency list, all having constructional meanings different from their individual parts. For example, a quick review of the literature in Chinese discourse studies suggests that similar behaviors have been observed of many other cognitive verbs (e.g. *juede*, Lim (this volume), Chiang 2004), copula expressions involving *shi* and *jiushi* (Biq 2001), low content verbs *you/meiyou* (Dong 2004), as well as the speech act verb *shuo* (Liu 1986, Meng 1982, Dong 2004). When we take into account both the lexical use and the multi-word constructional use it is possible to understand why all of the items in question have such high frequencies, yet individually they have little grounds to stand alone or be independent in constructing utterances.

4. Summary

I have shown with a case study of a cognitive verb that although the variety of the core lexicon may be small, their capacity to generate new lexical forms is high. The mechanisms are collocation and colligation: words combine with one another. Through combinations, new semiotic resource are created and serve to indicate subtle meanings in the conduct of social interaction. As a result, the frequencies of individual items in question also increase. This can be viewed as complementing the "least effort" principle as argued by Zipf (1935, 1949).

That words cluster is hardly a surprising finding. As research from corpus linguistics has repeatedly shown, a proper understanding of language must evoke some degree of fixedness or idiomaticity, as it is not possible for all language use to be computed on the fly and formulas and prefabs facilitate both speech production and comprehension. Researchers have reported that about 60-80% of spoken texts fall into some sort of formulaic sequences (Altenberg 1998, Erman and Warren 2000, Schmitt and Carter 2004). Research in this area has touched upon the issue of unit of meaning beyond single words (Sinclair 1991, 1996, McCarthy 2002), chunking (Bybee 2006, 2007), and formulaicity/idiomaticity (Wray 2002, Wulff 2008, Corrigan et al. 2009). Concerning formulaicity, Wray (2002:280) points out that "formulaicity bridges the gap between novelty and routine, and makes it possible for us to protect our own interests by producing language that is fluent and easily understood". Bybee (2006, 2007, 2009) points out that "'chunking' results when sequences of units that are used together cohere to form more complex units" and create frequency effects that facilitate production and comprehension. All this calls into question long-standing views of the nature of lexical and grammatical units, where individual words are seen as independent meaningful units, and provides an advantageous perspective for understanding the highly skewed distribution patterns that are widely observed in natural discourse.

5. Implications for Chinese language education

Turning now to the issue of integrating corpus-based findings into language pedagogy, an obvious application would be identify and focus on multiword sequences in pedagogy, as frequency effects of prefabs have also been shown to facilitate production

and comprehension in the L2 context (Wood 2002). However, even a cursory survey of the most commonly used Chinese teaching materials will show that Chinese language pedagogy has an overwhelming tendency to focus on individual characters and isolated words. Although sometimes correlated expressions such as paired conjunctions (e.g. 因为 yinwei 'because'…所以 suoyi 'therefore'，不但 budan 'not only'…而且 erqie 'but also', etc.) may be singled out, the discussion rarely goes beyond this. Thus in a lesson on eating out at restaurants found in a textbook series recently published in mainland China, which is also widely distributed internationally, the following text is found:

上餐馆

爸爸开车带我到了中餐馆，妈妈已经坐在里面等我们了。星期天，中餐馆里人很多， 空位子很少。中餐馆里的菜可多了，有鱼，有肉，还有各种海鲜和青菜。中国菜颜色美，味道香，又好看又好吃。我们坐下来，要了茶，接着点了三菜一汤，还点了鸡蛋炒饭。饭菜的味道好极了，我们都吃得很饱。这顿饭才花了二十多美元。爸爸付了钱，我们高高兴兴地离开了餐馆。

What follows, as are typical of Chinese textbooks, are lists of single characters, single words, along with a couple of key sentences:

生字
馆 空 菜 肉 鲜 青 味 茶 汤 鸡 炒 极 饱 顿

词语
餐馆 好吃 接着 鸡蛋 味道

句子
我们吃得很饱。
这顿饭才花了二十多美元。

Even though this lesson consists of a made-up text rather than authentic material, we can still identify a number of common multi-word expressions:

上餐馆	中餐馆	颜色美	味道香	(又)好看(又)好吃
要了茶	点了菜	三菜一汤	鸡蛋炒饭	味道好　这顿饭

All of these are attested phrasal expressions from written language corpora (e.g.

http://corpus.leeds.ac.uk/internet.html#). As with common multi-word expressions (Wray 2002), many of them contain core elements plus variable components. For example, 中餐馆 zhong canguan 'Chinese restaurant' could be substituted and become 西餐馆 xi canguan 'Western restaurant', 三菜一汤 san cai yi tang 'a set of three dishes and one soup' could be 四菜一汤 'a set of four dishes and one soup', and 鸡蛋炒饭 jidan chao fan 'fried rice with eggs' could be 虾仁炒饭 xiaren chao fan 'fried rice with shrimps' or 鸡蛋炒青椒 jidan chao qingjiao 'fried eggs with green peppers' etc. Yet the commonality of these expressions are undeniable. If these chucks are made aware of to the leaner, there is no doubt that it would be much easier for learners to grasp similar expressions when they next encounter them. Of course this is by no means to suggest that all of these items must be prioritized in instruction, and researchers are still debating the pros and cons of formulaic languguage instruction (se Wray 2002, Part IV) . However, the benefits of focusing on not just individual words/characters but also fixed chunks are beyond question (Nattinger and DeCarrico 1992, Howarth 1998, McCarthy 2002, Wood 2002). Perhaps what is ironic is that expressions such as 颜色美 yanse mei 'pretty colors' and 味道香 weidao xiang 'delicious tastes' are probably designed to be learned as fixed expressions given their adjacent and parallel features, yet they are nowhere to be seen in the vocabulary list, and nor are they ever integrated in pattern drills or any other types of pedagogical practices.

By way of conclusion, the findings reported in this paper, many of which have been discussed extensively in the literature, point to the following:

1) Rather than learning ever lengthening lists of new rare words, students may become more effective communicators by being exposed to combinations of words already internalized in new and useful ways;
2) Teachers should use every opportunity to raise the learner's awareness about existing and novel combinations and the mechanisms of such combinations;
3) When analyzing fixed formulas, emphasis should be placed on both key components and flexible substitutes. It is also important to contrast individual meanings with meanings of the whole chunk.

REFERENCES

Abercrombie, David. 1963. Conversation and Spoken Prose. The ELT Journal. XVIII: 10-16.

Altenberg, Bengt. 1998. On the phraseology of spoken English: The evidence of recurrent word-combinations. In A.P. Cowie (ed.), Phraseology: Theory, analysis, and applications. Oxford: Clarendon, pp. 101–122.

Biq, Yung-O. 2001. The Grammaticalization of *Jiushi* and *Jiushishuo* in Mandarin Chinese. Concentric: Studies in English Literature and Linguistics, 27.2: 53-74.

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. Language 82.4: 529-551 .

Bybee, Joan. 2007. Frequency of use and the organization of language. Oxford: Oxford University Press.

Bybee, Joan. 2009. Language, usage and cognition. Cambridge: Cambridge University Press.

Chao, Ruan. Y. 1968. A Grammar of spoken Chinese. Berkeley: University of. California Press.

Chiang, Ting-Yi. 2004. Affective chunk of Mandarin *Wo Juede* (我覺得) and its discourse-pragmatic functions. Paper presented at ROCLING XVI: Student Workshop II.

Clancy, Patricia, Thompson, Sandra, Suzuki, Ryoko and Tao, Hongyin. 1996. The conversational use of reactive tokens in English, Japanese, and Mandarin. Journal of Pragmatics, 26: 355-387.

Corrigan, Roberta, Edith Moravcsik, Hamid Ouali, and Kathleen Wheatley, eds. 2009. Formulaic language. Amsterdam: John Benjamins.

Erman, Britt and Beatrice Warren. 2000. The idiom principle and the open choice principle. Text 20.1: 29-62.

Hoey, Michael. 2004. Lexical priming and the properties of text. In Louan Harmann, John Morley and Alan Partington, eds., Corpora and Discourse, Bern, Peter Lang, 2004. 385-412.

Howarth, Peter. 1998. Phraseology and second language proficiency. Applied Linguistics 19.11: 24-44.

Lewis, Michael. 1993. The Lexical Approach: The state of ELT and a way forward. Hove UK: LTP.

Lim, Ni-Eng. Stance-taking with *Wo Jue De* in conversational Chinese. This volume.

McCarthy, Michael. 1998. Spoken language and applied linguistics. Cambridge: Cambridge University Press.

McCarthy, M. J. 1999. What constitutes a basic vocabulary for spoken communication?. Studies in English Language and Linguistics 1: 233-249.

McCarthy, Michael. 2002. This that and the other: Multi-word clusters in spoken English as visible patterns of interaction. Teanga: The Irish Yearbook of Applied Linguistics, vol. 21, 2002 [2004], 30-52.

McCarthy, M. J. and Ronald Carter. 2003. What constitutes a basic spoken vocabulary? Research Notes, 13.2: 5-7. Cambridge; Cambridge University Press.

Nattinger, James and Jeaneete DeCarrico. 1992. Lexical phrases and language teaching. Oxford: Oxford University Press.

O'Keeffe, Anne, Michael McCarthy, and Ronald Carter. 2007. From corpus to classroom: language use and language teaching. Cambridge: Cambridge University Press.

Pawley, Andrew and Frances H. Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Jack C. Richards and Richard W. Schmidt, eds., Language and communication, 191-268. London: Longman.

Schiffrin, Deborah. 1988. Discourse Markers. Cambridge: Cambridge University Press.

Schmitt, Norbert and Ronald Carter. 2004. Formulaic sequences in action: An introduction. In N. Schmitt, ed., Formulaic sequences: Acquisition, processing and use. Amsterdam: John Benjamins, 1–22.
Sinclair, John. 1991. Corpus, concordance, collocation. Oxford: Oxford University Press.
Sinclair, John. 1996. The search for the units of meaning. Textus IX: 75-106.
Tseng, S.-C. 2001. Highlighting utterances in Chinese spoken discourse. In Language, Information and Computation. PACLIC 15, 163—174.
Tseng, S.-C. 2006. Repairs in Mandarin conversation. Journal of Chinese Linguistics 34.1: 80-120.
Wood, David. 2002. Formulaic Language in Acquisition and Production: Implications for Teaching. TESL Canada Journal/Revue TESL du Canada, 20.1: 1-15.
Wray, Alison. 2002. Formulaic language and the lexicon. Cambridge: Cambridge University Press.
Wulff, Stefanie. 2008. Rethinking Idiomaticity: A Usage-based Approach. London/New York: Continuum.
Xiao, Richard, Paul Rayson, and Tony McEnery. 2009. A frequency dictionary of Mandarin Chinese: Core vocabulary for learners. Routledge Frequency Dictionaries. London and New York: Taylor and Francis Group.
Zhang, Huaping, Q. Liu, H. Zhang, and X. Cheng. 2002. Automatic recognition of Chinese unknown words based on role tagging. In Proceedings of the 1st SIGHAN Workshop, COLING 2002, 71-77, Taipei.
Zipf, George K. 1935. The Psycho-Biology of Language. Boston: Houghton Mifflin Company.
Zipf, George K. 1949. Human behavior and the principle of least effort: An introduction to human ecology. Cambridge, MA: Addison-Wesley Press.

陈原 (Chen) 主编，1993.《现代汉语用字信息分析》，上海：上海教育出版社.
陈原 (Chen) 主编，1989.《现代汉语定量分析》，上海：上海教育出版社.
董秀芳 (Dong)， 2004.《汉语的词库与词法》，北京：北京大学出版社。
国家语言文字工作委员会汉字处 (GJYW)，1988. 《现代汉语常用字表》，语文出版社.
国家对外汉语教学领导小组办公室、汉语水平考试部(GJHB)，1992.《汉语水平词汇与汉字等级大纲》，北京语言学院出版社.
刘源、梁南元等编（Liu et al.）, 1990. 《现代汉语常用词词频词典》，北京: 宇航出版社.
刘月华 (Liu) 1986. 对话中说想看的一种特殊用法，《中国语文》3: 168-172.
孟琮 (Meng). 1982.口语“说”字小集.《中国语文》, 1982.5.
孟琮、郑怀德、孟庆海、蔡文兰编写（Meng et al.），1999.《汉语动词用法词典》，北京：商务印书馆.
陶红印 (Tao). 2003. 从语音、语法和话语特征看“知道”格式在谈话中的演化. 《中国语文》4：291-302.

王惠 (Wang). 2009.〈词义·词长·词频——《现代汉语词典》(第5版)多义词计量分析〉，《中国语文》2：120-130.

北京语言学院语言教学研究所编（YYXY），1986.《现代汉语频率词典》. 北京: 语言学院出版社.