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'Keefe, 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.
TAO: CORE VOCABULARY IN CONVERSATION

<table>
<thead>
<tr>
<th>Type</th>
<th>Token</th>
<th>Proportion of tokens in corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>High frequency top 100</td>
<td>268,979</td>
<td>78%</td>
</tr>
<tr>
<td>Low Frequency below top 100</td>
<td>75,162</td>
<td>22%</td>
</tr>
<tr>
<td>Total</td>
<td>344,141</td>
<td>100%</td>
</tr>
</tbody>
</table>

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.

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 ‘have not’
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’
### 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 cannot 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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>的</td>
<td>13245</td>
</tr>
<tr>
<td>2</td>
<td>是</td>
<td>12047</td>
</tr>
<tr>
<td>3</td>
<td>我</td>
<td>10052</td>
</tr>
<tr>
<td>4</td>
<td>就</td>
<td>7782</td>
</tr>
<tr>
<td>5</td>
<td>不</td>
<td>7743</td>
</tr>
<tr>
<td>6</td>
<td>你</td>
<td>7658</td>
</tr>
<tr>
<td>7</td>
<td>了</td>
<td>7484</td>
</tr>
<tr>
<td>8</td>
<td>那</td>
<td>6846</td>
</tr>
<tr>
<td>9</td>
<td>啊</td>
<td>5792</td>
</tr>
<tr>
<td>10</td>
<td>个</td>
<td>4696</td>
</tr>
<tr>
<td>11</td>
<td>他</td>
<td>4385</td>
</tr>
<tr>
<td>12</td>
<td>对</td>
<td>4285</td>
</tr>
<tr>
<td>13</td>
<td>就是</td>
<td>3920</td>
</tr>
<tr>
<td>14</td>
<td>有</td>
<td>3816</td>
</tr>
<tr>
<td>15</td>
<td>都</td>
<td>3760</td>
</tr>
<tr>
<td>16</td>
<td>说</td>
<td>3677</td>
</tr>
<tr>
<td>17</td>
<td>一</td>
<td>3407</td>
</tr>
<tr>
<td>18</td>
<td>也</td>
<td>3186</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>With Objects</th>
<th>55</th>
<th>47%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without Objects</td>
<td>58</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>4</td>
<td>3%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>117</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

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.
Another common collocation involving zhidaoren ‘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 ma or 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 zhidaoren.


In this example, since the first speaker begins by asking for confirmation, the second speaker’s use of the apparent question with zhidaoren 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; zhidaoren, 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.’
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:

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

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:

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 si cai yi tang ‘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 chunks 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 language instruction (see 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


Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. Language 82.4: 529-551.
Lim, Ni-Eng. Stance-taking with Wo Jue De in conversational Chinese. This volume.

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、梁南元主编（Liu et al.），1990.《现代汉语常用词语词典》，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。
TAO: CORE VOCABULARY IN CONVERSATION

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.