汉语搭配定量分析初探*

孙茂松　黄昌宁　方　捷

提要　搭配在语言教学和语言信息处理中具有一定的应用价值。汉语搭配的研究仍停留在主要以人的主观判断为标准的定性分析阶段,缺乏定量数据的支待。本文借鉴了国外在语言学和语料库语言学两个方面关于搭配的研究成果,提出了包括强度、离散度及尖峰三项统计指标在内的搭配定量评估体系,构造了相应的搭配判断算法。作为对算法的初步测试,我们以一个约 710 万词次的新华社新闻语料库为工作平台,利用计算机对"能力"一词可能构成的搭配进行了全面分析。实验结果显示,就该词而言,算法自动发现搭配的准确率约为 33.94%。本项研究可望为语言学家客观、系统、一致地分析搭配提供定量辅助手段。

1. 意义

为什么我们说"穿衣"、"戴帽"而不说"穿帽"、"戴衣"? 为什么同是"看",当后接"电影"、"球赛"、"小说"、"朋友"时,英语译文必须分别以 see, watch, read 及 visit 与之对应? 显然,这是搭配(collocation) 的影响所致。搭配在语言教学(特别是第二语言教学)进而在语言交际中的作用,早已为人们所认识,并且日益得到重视。近年出版或再版的一些比较著名的英语通用词典 (general-purpose dictionary), 如 *Webster's New World Dictionary*, *Collins English Dictionary*, *Concise Oxford Dictionary*, *The Random House Dictionary of the English Language*, 均收录了一定数量的搭配,正反映了这种趋势。

搭配研究的一个新兴应用领域是语言信息处理。一切自然语言处理系统归根结底都是基于知识系统,如果希望计算机实现对一个句子的理解或翻译,那么人完成同样任务所需要的全部知识,严格说来,计算机一点儿也不能少。搭配知识则是所谓"全部知识"中有机组成部分之一。仍以"看电影"、"看球赛"、"看小说"、"看朋友"为例。对它们进行句法、语义分析,就会发现它们的句法结构和语义表示完全相同(均为动宾结构且宾语均为动作"看"的受事),必须嵌入相关的搭配知识才能体现出"差异",从而生成合适的译文。再如,汉语中双音节动词加双音节名词既可构成谓词性成分(如"生产化肥"、"生产汽车"),也可构成体词性成分(如"生产能力"、"生产资料"),前者具有语法上的普遍性,后者则有特异性,根据搭配知识很容易排除分析过程中遇到的此类歧义。

2. 对搭配的认识及其相关研究

什么是搭配? 似乎不同的理论角度与应用背景,人们对此问题存在着不同的理解,本文不打算展开讨论。在搭配领域最具影响的研究当推美国宾州大学 Benson 教授的工作及其负责编纂的 *BBI Combinatory Dictionary of English* (1985, 1986, 1989, 1990)。我们比较倾向 Benson 编纂的 BBI 时给出的关于搭配的定义:

* 清华大学青年科学基金、国家自然科学基金资助项目。

定义 1　搭配是一种具有任意性的、重复出现的词的组合。

(A collocation is an arbitrary and recurrent word combination.)

从 Benson 的定义可知搭配的两条重要性质:

性质 1　搭配是重复出现的。

这一性质决定了搭配应有一定的流通度,而非偶然的"个例"。

性质 2　搭配是任意的。

这里有必要引入两个与性质 2 密切相联的重要概念,即词的"自由组合"(free combination)与"约束组合"(bound combination)。按照 Benson 的观点,自由组合是指构成该组合的词并非以一种相对特异的方式相互约束,它们各自还可以与其它词自由地进行组合。例如,动词 condemn 可带相当多的名词 (the abduction, abortion, abuse of power, the acquittal 等) 作宾语,而名词 murder 也可出现在数以百计的动词 (abhor, accept, acclaim, advocate 等) 之后,故组合 condemn murder 是自由组合。自由组合是可预期的,一个学习第二语言的人,只要了解有关词的含义、语法属性及相应的语法组合规则,就可以在语言交际中根据需要很容易地"拼"出这种组合;约束组合的情形正好相反,具有一定的特异性,辖内的词至少有一个与其它词的组合受到较大限制。如组合 commit murder 中的动词 commit 只可能同屈指可数的几个名词 crime, wrongdoing 发生关系,故应是约束组合。约束组合 (搭配) 是不可预期的,在同样的语法、语义制约条件下,为什么非得这么讲,那么讲就不行,没有太多的道理,恐怕一般只能解释为习惯使然(如英语只说 make an estimate, warmest greetings 而不说 make an estimation, hot greetings)。这是语言教学,尤其是第二语言教学过程中最感困难的环节之一,基本上无规律可循。正是在这个意义上,我们称搭配具有任意性(或也可叫不可预期性)。

搭配的其它性质还有:

性质 3　搭配通常是具有一定结构的。

Benson (1989) 将英语的搭配分成语法搭配及词汇搭配两大类,语法搭配再分 26 个细类 (如 v + prep, n + prep, adj + prep 等),词汇搭配分 7 个细类(如 v + n, n + v, n + n, adv + adj, v + adv 等)。对某些搭配,所辖的两个词之间允许有间隔,甚至调序(如搭配 make - decision, to make a decision, decisions to be made, made an important decision 几种表述都可以),但仍保持一定的结构关系(Smadja, 1993)。

性质 4　搭配是与领域相关的。

除流行于日常交际中的常用搭配外,对应各专门领域,还有大量的、作用范围仅限于该领域的特定搭配,如某些专业技术术语及领域习惯用语(Smadja, 1993)。

搭配研究中遇到的主要问题是,搭配(约束组合)与非搭配(自由组合)之间的界限许多情况下并不清晰,容易引起混淆。即便是定义 1, 亦仅限于定性描述,缺乏定量标准:重复出现多少次才算足够,约束到什么程度才称得上具有任意性,具体操作起来,见仁见智,因人而异。Benson 曾指出,业已出版的若干英语搭配词典(如 *Rodale's Word Finder*)即充斥了相当多的不该收录的自由组合,同时却又有不少遗漏。

能否找到某些适当的定量数据作为判断搭配的参考或补充? 国外较早开展搭配的计算机定量分析研究的是 Choueka 等(1983)。他们把搭配定义为重复出现的相邻词串(大致对应于搭配的性质 1),从《纽约时代周刊》(*New York Times*) 约 1100 万词的文本中提取出了数以千计的英语常用搭配,如 fried chicken, home run, Magic Johnson 等。其主要缺陷是没有考虑搭

配中两个词可能被隔开的情形(如 make－decision),也基本上没有考虑搭配性质 2 的要求。Church 等(1991)定义搭配为一组相互联系的词对,借鉴信息论中"互信息"(mutual information)的概念来评价两个词的结合能力,进而针对一个 4400 万词左右的新闻语料库(AP Corpus)进行了实验。互信息这一统计量相当程度地体现了搭配的性质 1 和性质 2,并且对搭配中的词是否邻接理论上没有什么限制。该方法的弱点是未顾及搭配的性质 3,导致许多提取出来的词对,如 doctor－nurse,doctor－bill,doctor－hospital 等均非搭配(这些词由于意义上关联密切,经常在相同的上下文中共现,而它们之间却并不存在语法制约关系)。Smadja(1993)的 Xtract 系统是迄今为止关于搭配定量分析的最新和最完整的工作。Xtract 提出了自己的词对间强度计算公式(对应搭配的性质 1),引入了位置信息以及相关统计数据分布的离散度计算公式(试图反映搭配的性质 3),更集成了语料库语言学词性自动标注技术,取得了可喜进展。在一个规模为一千万词的股票市场新闻报告语料库上运行 Xtract 所得到的结果显示,搭配提取的准确率达到了 80%(如果不诉诸词性自动标注技术,即在与 Choueka 及 Church 大致对等的条件下工作,准确率约 40%)。不过,我们以为,Xtract 对搭配的性质 2 注意不够。

汉语的搭配研究方面,国内业已出版了若干部搭配词典(王砚农等, 1984, 1987;杨天戈等, 1990;张寿康、林杏光, 1990, 1992; 张卫国、冀小军等, 1994)。研究汉语搭配,不可避免地也要遇到搭配与非搭配之间界限不清的问题,而现阶段国内的研究基本上是在纯语言学的范畴中进行的,多依赖于人的主观判断,再加之汉语搭配现象本身的复杂性,产生混淆的程度较英语可能会更严重些。郭茜(1995)认为《现代汉语实词搭配词典》是现有搭配词典中比较好的一部。但这部词典仍收录了为数相当可观的自由组合(诸如"小伙子能干"、"经理能干"、"工人能干"、"领导的能力"、"学生的能力"、"知识分子的能力"、"称赞徒弟"、"称赞战士"、"称赞服务员"之类),也遗漏了不少真正的搭配(如收了"能力强"、"能力差"而未收"能力弱")。郭明确提出应采用 Benson 的定义作为判断汉语搭配的标准。

本文是在计算机技术和大型汉语语料库的支持下对汉语搭配进行定量分析的首次尝试。我们采用 1990 和 1991 年的新华社新闻语料库 XH－CORPUS(1000 余万字,计算机自动分词后得约 710 万词)作为工作平台。期望为判断搭配提供量化的依据,把一般被认为"只能意会,不能言传"的语感量化成为可以捉摸的东西,最终实现一个完整的搭配辅助分析工具。

3. 搭配定量分析设计

搭配的重复性、任意性和结构性对搭配的判断有直接的意义。如何使这些特征量化,将是达到本文所提出目标的关键。一旦实现了量化,判断搭配的算法也就水到渠成了。

3.1. 搭配的强度

Church 等(1991)指出,任意两个词 w, w_i 相互联系的疏密程度可用它们之间的互信息来衡量:

$$mi(w,w_i)=\log_2\frac{p(w,w_i)}{p(w)p(w_i)} \qquad (式\ 1)$$

其中 $p(w,w_i)$ 是 w, w_i 在给定上下文范围内的共现概率,$p(w)$,$p(w_i)$ 分别是 w, w_i 的独立概率。

根据本文的研究兴趣,不妨进一步设 w, w_i 为一个候选搭配。则由式 1,我们有结论:搭配的性质 1 和性质 2 可以在相当程度上通过互信息反映出来。

(1)关于性质 1:w, w_i 共现次数愈多,则 $p(w,w_i)$ 的值愈大,$mi(w,w_i)$ 的值亦随之变大,表

明 w,w_i 的重复性趋势增强。反之，情形正好相反；

(2)关于性质2：w,w_i 受约束程度愈深，意味着 w 或 w_i 与其它词共现的机会愈少，则 $p(w)$ 或 $p(w_i)$ 的值减小，在 $p(w,w_i)$ 值不变的条件下，会使 $mi(w,w_i)$ 的值变大，表明 w,w_i 的任意性趋势增强。反之，情形也正好相反。

对搭配研究，应该把认为 w,w_i 可能发生联系的上下文限定为句子。在句子范围内，允许 w,w_i 被任意多的其它词隔开(譬如，"穿衣服"、"穿红衣服"，"穿了一件新衣服"，"穿了一件崭新的红衣服"中的"穿"和"衣服"都应视作共现)。当然，相距越远，两个词关连的可能性一般来说就越小。Xtract 系统把英语中一个词的影响所及近似为该词的前后五个词，超出此范围的不予考虑。本文对汉语也作了同样的简化。令 $p_j(w,w_i)$ 表示 w_i 在与 w 相距 j 个词的位置处与 w 共现的概率($j=-5,-4,-3,-2,-1,1,2,3,4,5$。$w_i$ 在 w 的左侧，j 取负，在右侧，j 取正)，则在式1的基础上，可引入搭配评估的强度公式：

$$s(w,w_i)=\log_2\frac{\sum_{j=-5}^{+5}p_j(w,w_i)}{p(w)p(w_i)} \qquad (式2)$$

在一个大小为 N 的语料库中，若 w_i 与 w 相距 j 个词处共现的次数为 $r_j(w,w_i)$，w,w_i 独立出现的次数分别为 $r(w),r(w_i)$，则利用最大似然估计，式2可继续推导成：

$$s(w,w_i)=\log_2\frac{N\sum_{j=-5}^{+5}r_j(w,w_i)}{r(w)r(w_i)} \qquad (式3)$$

考察候选搭配："能力，弱"和"能力，大"。由 XH-CORPUS($N\approx7.1\times10^6$)统计得到：

r_{-3}(能力，弱)=1　　r_1(能力，弱)=3，　　r_2(能力，弱)=5，

r_j(能力，弱)=0($j=-5,-4,-2,-1,3,4,5$)，　　r(能力)=2241，　　r(弱)=177

于是有：$s(能力，弱)=\log_2\frac{7.1\times10^6(1+3+5+7\times0)}{2241\times177}=7.33$

r_{-5}(能力，大)=6，　r_{-4}(能力，大)=4，　r_{-3}(能力，大)=8，

r_{-2}(能力，大)=4，　r_{-1}(能力，大)=2，　r_1(能力，大)=9，

r_2(能力，大)=6，　r_3(能力，大)=4，　r_4(能力，大)=6，

r_5(能力，大)=5，　r(能力)=2241，　r(大)=19913

于是有：$s(能力，大)=\log_2\frac{7.1\times10^6(6+4+8+4+2+9+6+4+6+5)}{2241+19913}=3.10$

s(能力，弱)远远大于 s(能力，大)，故"能力""弱"比"能力""大"更"像"一个搭配。值得注意的是，虽然"能力"和"弱"仅共现了9次，而"能力"和"大"共现了54次，但由于 r(弱)为177，r(大)却为19913，计算下来，s(能力，弱)相对 s(能力，大)反而超出。这是搭配的任意性所乐见的。

仿上可算出 s(能力，强)=7.45，s(能力，差)=6.63，s(能力，小)=0.74。按降序排序，有：s(能力，强)>s(能力，弱)>s(能力，差)>s(能力，大)>s(能力，小)，显示它们成为搭配的可能性依次降低。其中 s(能力，强)，s(能力，弱)，s(能力，差)十分接近，值也比较高，"能力，强"、"能力，弱"、"能力，差"基本上可以认定是搭配；s(能力，大)则陡然下跌，值居中，"能力，大"能否成为搭配尚有待观察；s(能力，小)已几乎为零，"能力，小"显然不应视作搭配。

3.2 搭配的离散度

搭配往往具有一定的结构关系(性质3)。这隐寓着一层意思,即如果 w, w_i 构成搭配,则受结构关系的制约,w_i 与 w 在某个或某几个位置 j($-5 \leq j \leq 5$)上共现的机会,较其它位置可能会大得多,从而导致 $r_j(w, w_i)$ 的分布呈较大幅度的抖动。对非搭配,$r_j(w, w_i)$ 的分布则相对平坦(Smadja,1993)。图1显示了两组词"能力,丧失"和"能力,方面"的共现分布:前一组是搭配,其分布变化剧烈;后一组不是搭配,其分布变化和缓(r_{-4}(能力,丧失)=r_{-3}(能力,丧失)=1;r_{-2}(能力,丧失)=8;r_{-4}(能力,方面)=r_1(能力,方面)=2,r_{-3}(能力,方面)=3;r_{-2}(能力,方面)=r_{-1}(能力,方面)=r_2(能力,方面)=1;其余均为0。"能力,丧失"和"能力,方面"的共现次数相等,均为10)。

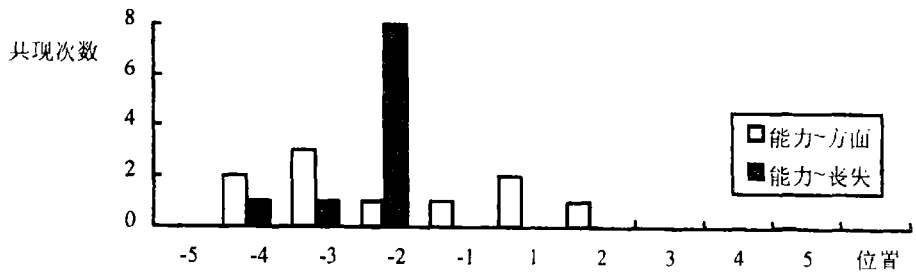

图1. 两个词共现的分布

显然,$r_j(w, w_i)$ 分布的抖动倾向可以用其方差予以描述。这也就是搭配评估的离散度公式(Smadja,1993):

$$u(w, w_i) = \frac{\sum_{j=-5}^{5} (r_j(w, w_i) - \bar{r}(w, w_i))^2}{10} \qquad (式4)$$

其中

$$\bar{r}(w, w_i) = \frac{\sum_{j=-5}^{5} r_j(w, w_i)}{10} \qquad (式5)$$

对图1中的两组词"能力,丧失"及"能力,方面",根据式4、式5,有:

$$\bar{r}(能力,丧失) = \frac{1+1+8}{10} = 1 \quad \bar{r}(能力,方面) = \frac{2+3+1+1+2+1+4\times 0}{10} = 1$$

$$u(能力,丧失) = \frac{(1-1)^2 + (1-1)^2 + (8-1)^2 + 7\times(0-1)^2}{10} = 5.60$$

$$u(能力,方面) = \frac{(2-1)^2 + (3-1)^2 + 3\times(1-1)^2 + (2-1)^2 + 4\times(0-1)^2}{10} = 1.00$$

值得注意的是,当分布的变化幅度很大时,在某些位置上可能还会出现明显的尖峰(如图1中"能力,丧失"的位置-2)。令 $z_j(w, w_i)$ 表示 $r_j(w, w_i)$ 的 z-测试:

$$z_j(w, w_i) = \frac{r_j(w, w_i) - \bar{r}(w, w_1)}{\sqrt{u(w, w_1)}} \qquad (式6)$$

则尖峰位置可由以下条件求出(Smadja,1993):

位置 j 上出现尖峰,如果 $z_j(w, w_i)$ 足够大的话,参考图1中位置-2:

$$z_{-2}(能力,丧失) = \frac{8-1}{\sqrt{5.60}} = 2.96$$

指示 r_{-2}(能力,丧失)高过 $\bar{r}$(能力,丧失)1.96个标准差,形成了一个尖峰。

针对汉语，我们将确定尖峰的算法进一步细化。

确定尖峰的算法 is_peak(w,w_i)

输入：任一词对 w,w_i 在各位置上的共现次数 $r_j(w,w_i)(j=-5,\cdots 5)$

输入：是否存在尖峰及其位置

计算 $\bar{r}(w,w_i)$及各位置的 $z_j(w,w_i)$ $(j=-5,\cdots 5)$；

对所有的 j，做：

如果$(0.30\leq\bar{r}(w,w_i)<1.00$ 且 $z_j(w,w_i)\geq 2.50)$ 或

$(1.00\leq\bar{r}(w,w_i)<5.00$ 且 $z_j(w,w_i)\geq 2.00)$ 或

$(5.00\leq\bar{r}(w,w_i)<10.00$ 且 $z_j(w,w_i)\geq 1.50)$ 或

$(\bar{r}(w,w_i)\geq 10.00$ 且 $z_j(w,w_i)\geq 1.00)$

则　位置 j 为尖峰位置　　否则　位置 j 非尖峰位置

该算法把 $\bar{r}(w,w_i)$划成几个区间，在不同的区间，位置 j 欲成为尖峰位置所要求的 $z_j(w,w_i)$的阈值也不同（区间的划分及阈值均根据实验测定）。$\bar{r}(w,w_i)$愈小，因统计数据不足而引起的干扰的影响就愈大，所以阈值应调高；反过来，统计数据比较充分、翔实，阈值可以适当调低。例如，"能力，丧失"在 XH－CORPUS 中仅共现了 10 次，$\bar{r}$(能力，丧失)＝1.00，数据噪音可能比较大，故阈值要高些(2.00)，结果只有 z_{-2}(能力，丧失)＝2.96 可以通过；而"能力，大"共现了 54 次，$\bar{r}$(能力，大)＝5.40，数据相对可靠，故阈值可以低些(1.50)，而 z_1(能力，大)＝1.84，于是也能顺利过关，位置 1 成为尖峰位置（综合节 3.1 已算出的强度值，现在有比较充分的理由判定"能力，大"是一组搭配。参阅下一节）。

离散度及尖峰这两个参数为定量研究搭配提供了结构方面的重要信息。不过，应该指出，说它们重要，并不意味着它们是判断搭配必不可少的条件。有些搭配，只要强度足够就行了，离散度并不一定很高，更不要求出现尖峰（显然尖峰是较离散度更为"苛刻"的要求）。当强度信息不足以据之作出裁决时，离散度及尖峰的作用就突出出来了。

3.3 判断搭配的算法

本文提出的搭配定量评估体系包括了强度、离散度及尖峰三项指标，其中强度公式（式 3）参照(Church et al 1991)给出，离散度（式 4）及尖峰（式 6）照搬(Smadja,1993)。评估指标的这种有机组合体现了我们对搭配的理解，既汲取了前人的经验，又使我们的工作仍具新的视域。判断搭配的算法根据该体系设计（区间的划分及阈值亦由实验测定）。

判断搭配的算法 is_collocation)(w,w_i)

输入：任一词对 w,w_i 的强度 $s(w,w_i)$、离散度 $u(w,w_1)$、均值 $\bar{r}(w,w_i)$及各位置的 $z_j(w,w_i)(j=-5,\cdots,5)$

输出：w,w_i 是否为搭配

如果 $\bar{r}(w,w_i)<0.30$ 则认为 w,w_i 不是搭配（否定条件 1）；

如果 $s(w,w_i)\geq 4.5$。

　　则认为 w,w_i 是搭配（肯定条件 1）；否则

如果$(3.50\leq s(w,w_i)<4.50$ 且 $u(w,w_i)\geq 10.00)$

　　则认为 w,w_i 是搭配（肯定条件 2）；否则

如果$(2.50\leq s(w,w_i)<3.50$ 且 $u(w,w_i)\geq 20.00)$

　　则认为 w,w_i 是搭配（肯定条件 3）；否则

如果 $s(w,w_i)\geq 2.00$

调用 is_peak(w,w_i);

如果发现尖峰,则认为 w,w_i 是搭配(肯定条件 4);

否则认为 w,w_i 不是搭配(否定条件 2)

否则认为 w,w_i 不是搭配(否定条件 3)

搭配的肯定条件有四:强度足够大,根本无须考虑离散度即可作出判断(肯定条件 1);随着强度的递减,对离散度的要求渐高(肯定条件 2,3);强度降到一定程度,更要求出现尖峰(肯定条件 4)。搭配的否定条件有三:两个词的共现次数太低,数据无统计意义(否定条件 1);强度较低又未发现尖峰(否定条件 2);强度太低,即使考虑离散度或尖峰也于事无补(否定条件 3)。肯定条件 1 与否定条件 3 是两个对称的极端。

4. 实验结果及其讨论

我们已用 Visual C++ 在中文 Windows 环境下实现了上述算法。工作平台是新华社语料库 XH-CORPUS。考虑到新闻语料的特点,我们选取"能力"一词作为对算法进行初步测试的研究对象,利用计算机详尽分析了"能力"形成搭配的种种可能变化。实验结果如下:在 XH-CORPUS 中,"能力"共出现了 2241 次(即 w=能力,r(w)=2241)。与"能力"在[-5,5]上下文范围内共现的不同的词计 1932 个,它们均应视作"能力"的候选搭配。三条否定条件共排除了 1317 个,其中否定条件 1 排除 962 个,否定条件 2 排除 201 个,否定条件 3 排除 154 个。经四条肯定规则确认的有 615 个,其中肯定条件 1 确认 411 个,肯定条件 2 确认 16 个,肯定条件 3 确认 11 个,肯定条件 4 确认 177 个。这些被确认的词中,包含 88 个根本不可能与"能力"构成搭配的数词、助词、连词、副词等,利用一个简单的过滤程序,可将它们筛掉。此外,包含因自动分词错误引起的误判 29 个。扣除这两项因素,机器最终认定的搭配共 498 个。通过人工审核,真正的搭配有 169 个。也就是说,就"能力"一词而言,本算法发现搭配的准确率(准确率指机器发现出来的搭配中真正为搭配者所占的比例)为 169/498=33.94%。

表 1 给出了部分实验数据。表 2 是机器发现并且经人核定的搭配的全部清单。

表 1 部分实验数据(w=能力 r(w)=2241)

序号	w_i	$r(w_i)$	$r(w,w_i)$	$s(w,w_i)$	$u(w,w_i)$	尖峰位置	机器判断是否为搭配	机器之判断是否正确	搭配之结构
1	吞吐	78	78	11.30	547.56	-1	是(肯 1)	对	定中
2	强	1651	91	7.45	138.29	1	是(肯 1)	对	主谓
3	弱	177	9	7.33	2.69		是(肯 1)	对	主谓
4	提高	6058	205	6.75	187.25	-4,-3	是(肯 1)	对	动宾、主谓
5	差	638	20	6.63	18.20	1	是(肯 1)	对	主谓
6	具有	2749	51	5.88	45.49	-3,-2	是(肯 1)	对	动宾
7	石油	1641	18	5.12	18.20	-2	是(肯 1)	错	
8	有	30354	212	4.47	578.36	-1	是(肯 2)	对	动宾
9	使	8701	62	4.49	57.30	-5,-4	是(肯 2)	错	

10	组织	7760	24	3.29	22.64	-1	是(肯3)	对	定中
11	问题	12476	28	2.83	24.16	-2	是(肯3)	错	
12	拥有	1141	5	3.80	0.85	-3	是(肯4)	对	动宾
13	大	19913	54	3.10	3.84	1	是(肯4)	对	主谓
14	不	17409	37	2.75	19.21	1	是(肯4)	错	
15	而	6908	17	2.96	3.01		否(否2)	对	
16	接受	1729	6	3.46	0.44		否(否2)	错	定中
17	小	9481	5	0.74	0.85	3	否(否3)	对	
18	活动	7428	8	1.77	2.36	-1	否(否3)	错	定中
19	民族	5473	2	0.21	0.16		否(否1)	对	
20	涵蓄	1	1	11.63	0.09		否(否1)	错	定中

表2 机器发现并经人核定的搭配(共169个)

(a) 培养 判断 鉴赏 生产 竞争 制造 运输 加工 支付 偿还 平衡 消化 吸收
繁殖 实际 具备 缺乏 提高 强 弱 大 差 有限 增强 具有 业务
劳动 适应 领导 组织 分析 保护 发挥 工作 技术 专业 管理 创造 运输
发电 丧失 防御 装卸 指挥

(b) 反应 扩大 综合 形成 达到 设计 抗灾 开采 影响 排水 客运 保障 承受
一定 执政 反应 安置 配套 不足 超过 出口 自力 创汇 动手 吞吐 增加
运行 足够 防务 操作 处理 作战 通信 同等 自给 自理 防守 减弱 现有
约束 作业 防卫 鉴别 通航 负重 不够 生存 隐蔽 科研 失去 抗病 炼油
腐蚀 后续 识别 抗旱 削弱 限制 识字 存储 自主 对抗 核算 机动 消费
分流 超出 防洪 自卫 干扰 免疫 再生 信任 过剩 供给 应急 饲养 运算
扑救 防疫 驾驭 筛选 参政 相应 采油 整体 通行 核定 载荷 维修 运载
接待 保存 分辨 保鲜 装备 耐寒 通车 转换 防范 自救 联运 决策 独到
起重 输送 新 有 开发 服务 群众 发展 测量 显示 突破 依靠 强化
控制 经营 供应 下降 监督 低 核 拥有

我们将机器发现(经人核定后)的搭配同《现代汉语辞海》(张卫国,冀小军等,1994)中词条"能力"下所列搭配作了比较(该词典对现代汉语常用词的搭配面貌进行了尽可能充分的描写)。表2中(a)是两者重叠的部分,(b)则是由机器发现出来而《现代汉语辞海》中没有的。

通过对实验结果的观察,可以初步得出几点结论:

1. 强度、离散度及尖峰是关于搭配定量研究比较合适的统计量。但它们只是一种相对的判断标准(表1中,每一个肯定条件下均列有误判的例子);

2. 统计数据的分布相当程度上反映了搭配的结构。

图2显示搭配"能力~具有"呈动宾结构(尖峰位置-3,-2);图3显示搭配"能力~差"呈主谓结构(尖峰位置1);图4搭配"能力~提高"既有动宾结构,又有主谓结构,且前者的用法

要多过后者(尖峰位置－4,－3);图 5 一峰突兀(尖峰位置－1),反映了汉语中一种很典型的搭配格式,即"动词＋名词"构成名词短语。

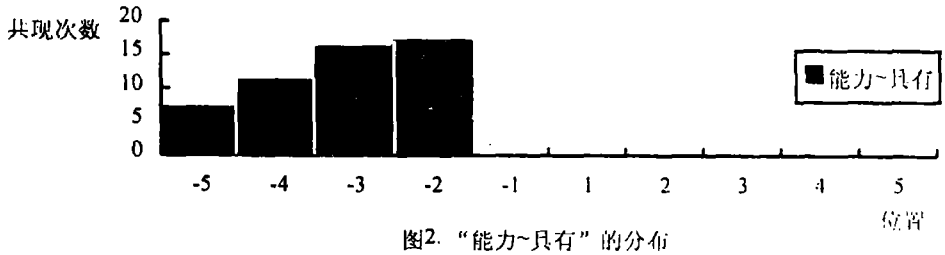

图2. "能力~具有"的分布

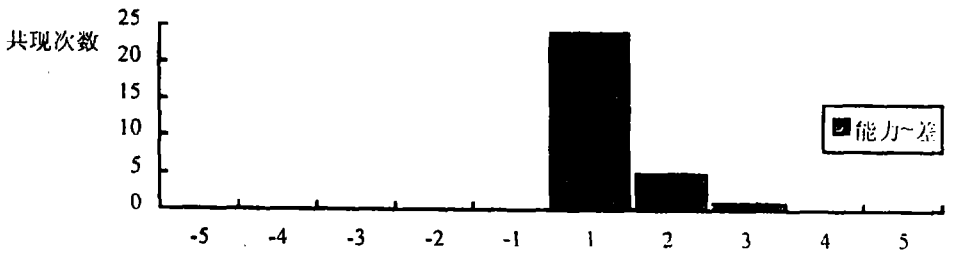

图3. "能力~差"的分布

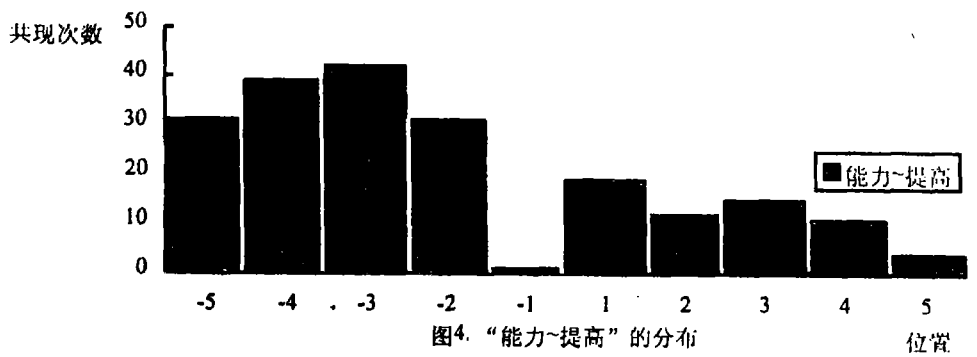

图4. "能力~提高"的分布

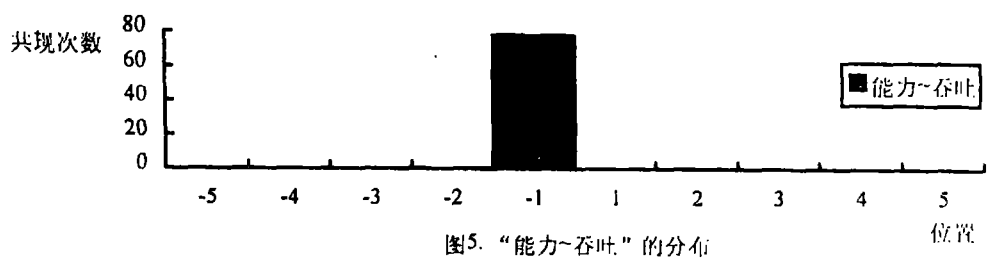

图5. "能力~吞吐"的分布

3. 搭配是受领域制约的(搭配性质 4)。如"阅读能力"、"写作能力"之类人们基本认可的搭配,在 XH-CORPUS 从未出现过。又如表 1 中的"接受能力"(强度 3.46)"活动能力"(尖峰位置－1)"涵蓄能力"(强度 11.63,但仅共现了一次),人们一般会认可它们是搭配,从统计数

据看，它们也确实"差不多"取得了成为搭配的资格，可惜共现于 XH-CORPUS 中的次数实在太低，只能舍弃。

4. 对阈值的设置，应在搭配发现的准确率与召回率（召回率指语料库所蕴含的搭配被机器正确发现出来的比例）之间寻找一个折衷。一般来说，阈值越高，准确率越高，召回率越低；阈值越低，准确率越低，召回率越高。本实验期望发现尽可能多的搭配，"宁滥勿缺"，故阈值定得比较保守。

有人会问：30%左右的准确率（当然，这仅是针对"能力"一词得到的局部数据），是否偏低了？但 Smadja(1993)指出，传统的办法手工编辑 *Oxford English Dictionary* (OED)的准确率大约只有 4%。相比之下，效率还是有较大幅度的提高。此外，传统的办法易受人为因素的干扰：研究者根据观察所得（包括语感）作出的判断很可能囿于各自的语言及知识背景而仅仅反映搭配现象的一个侧面，相互之间的协调也会因缺乏共同标准显得比较困难；随着社会、科技的突飞猛进，信息量急剧膨胀，语言材料也呈日新月异之势，即便是专家，及时掌握语言发展的全貌也绝非易事。利用计算机辅助技术，可望实现以语料库为支撑的，关于搭配的系统、全面、一致的分析，既减轻了语言学家的工作强度，又提高了搭配的质量和覆盖面。当然，也不能夸大统计分析的作用，其角色只是人的辅助或补充手段。人机协作可以采取两种形式：一是机器根据发现算法先自动给出一个搭配候选集，然后人对之进行评判；再一是对人已编纂好的搭配词典由机器给出定量信息，予以过滤。追求搭配的准确及少而精，应成为搭配研究的重心。

将来的工作将循两个方向展开：(1)引入词性标注及短语边界自动辨识等自然语言处理技术，以增加发现搭配的准确率；(2)扩大实验规模，进一步检验及改进算法。

参考文献

Benson, M. 1985 A combinatory dictionary of English, *Dictionaries: Journal of the Dictionary Society of North America*, 7.

Benson, M.; Benson, E.; And Ilson, R. 1986 *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*, John Benjamins.

Benson, M. 1989 The structure of the collocational dictionary, *International Journal of Lexicography*, 2(1).

Benson, M. 1990 Collocations and general-purpose dictionaries, *International Journal of Lexicography*, 3(1).

Choueka, Y.; Klein, T.; and Neuwitz, E. 1983 Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus, *Journal of Literary and Linguistic Computing*, 4.

Church, K.; Gale, W.; Hanks, P.; and Hindle, D. 1991 Using statistics in lexical analysis, *In Lexical Acquisition: Using On-Line Resources to Build a Lexicon, edited by Uri Zernik*. Lawrence Erlbaum.

Smadja, F. 1993 Retrieving collocations from text: Xtract, *Computational Linguistics*, 19(1).

Guo, Q. 1995 On the selection of target-language equivalent, *Master's Thesis*, Tsinghua University.

王砚农等 1984 《汉语常用动词搭配词典》，外语教学与研究出版社。

王砚农等 1987 《汉语动词—结果补语搭配词典》，北京语言学院出版社。

杨天戈等 1990 《常用词搭配词典》，外语教学与研究出版社。

张寿康，林杏光 1990 《简明汉语搭配词典》，商务印书馆。

张寿康，林杏光 1992 《现代汉语实词搭配词典》，商务印书馆。

张卫国，冀小军等 1994 《现代汉语辞海》，人民中国出版社。

（孙茂松　黄昌宁　方捷　北京　清华大学计算机科学技术系　100084）