《当代语言学》第4卷2002年第2期101－114页，北京

基于语料库和语料库驱动的词语搭配研究*

卫乃兴　　上海交通大学

提要　本文介绍和讨论语料库证据支持的词语搭配研究的基本方法和主要原则。基本研究方法可分为"基于语料库数据"和"语料库数据驱动"两类。基于数据的方法以语料库索引为基本依据，在传统的句法框架内对词项的搭配进行检查与概括；数据驱动的方法基本上不将句法结构作为主要参照，而设计和采用一套概念体系、步骤和程序提取和计算搭配词，凭借统计测量手段研究词语搭配的模式，或者采用技术手段提取和计算词丛。主要原则包括：以"自然发生数据"为基本依据，定量分析与定性分析相结合，采用词语中心的研究方法；以发现词组为目的等四项。文章首先讨论词语搭配研究的基本方法，然后概括研究应遵循的主要原则。

关键词　索引　类联接　搭配词　显著搭配　词丛　定量研究　数据驱动

1. 引言

词语搭配研究之父 Firth(1957：12)曾说："You shall know a word by the company it keeps."即词的意义从与它结伴同现的词中体现。根据 Firth 的观点，词项的结伴规律、结伴词项间的相互期待与相互吸引(mutual expectancy，mutual attraction)、搭配成份的类联接(colligation)关系等都是词语搭配的形式属性，都是词语搭配研究的重要内容。研究这些内容、描述词项的搭配情况一直是英国语言学的一个传统。但是，语料库问世前的研究，由于缺乏足够的自然数据，一般都基于直觉，很难深入下去，研究结果也有很大的局限性。语料库语言学的兴起为词语搭配研究开辟了崭新途径与广阔前景。在语料库研究中，词语搭配被赋予新的理念。研究者建立了一套概念体系，采用一系列方法与技术来提取信息、处理数据和描述搭配行为，研究的效度因此大大提高，搭配得到深入的探讨和详尽的描述。总的来说，语料库证据支持的词语搭配研究有两种基本方法：基于数据的方法(data-based approach)和数据驱动的方法(data-driven approach)。具体的作法有三种：(1)利用索引证据、参照类联接，检查和概括词项的搭配情况；(2)计算搭配词，采用统计测量手段，靠数据驱动研究词语搭配模式(pattening)；(3)采用技术手段，从语料库提取并计算词丛。本文根据笔者的语料库研究实践，并参照国内外同行的有关研究，对这三种方法逐一介绍和讨论，并概括语料库证据支持的词语搭配研究的重要原则。

2. 基于索引证据，参照类联接，检查和概括词项的搭配行为

类联接是词语搭配研究中的一个重要概念。它指的是文本中语法范畴间的结合。类联接不是与词语搭配平行的抽象，而是高一级的抽象(Firth 1957)。Mitchell(1975：120－122)认

*作者由衷感谢杨惠中对本文撰写的指导，他对文章的初稿提出了具体修改意见，并对成稿后全篇论文的观点、思路等提出了宝贵建议。

为，类联接是关于词语组合类别的抽象表述，搭配则是类联接的具体实现。简言之，类联接是词语搭配发生于其中的语法结构和框架。一个类联接代表了一个类别的词语搭配，可称为搭配类(collocational class)。人们常说的 N+V，V+N，N+N 等都是类联接，代表一类搭配，而 evidence suggests，perform analyses 和 corpus evidence 等则分别是这几个类联接的具体实例。在研究中，研究者可根据具体的研究内容和目的，界定繁简程度不同的类联接，如 DET+N+PREP 即为一个类联接，代表一类搭配，a sort of，a pair of，a couple of，a series of 等则是其具体实例(Renouf and Sinclair 1991：131－132)。词语搭配研究的通常作法之一就是以词项为中心，参照类联接框架，观察、概括和描述词项的搭配。

索引(concordance)，即语料库中含有所研究的关键词的句子片段。词语搭配研究所用的索引一般是 KWIC(key word in context)索引，它已成为语料库研究人员的最基本工具之一。研究人员每一次在语料库中查询，都要将自己即将研究的关键词输入计算机，计算机则按照编好的程序，显示出索引。在每一行索引中，关键词总是居中出现，而左右则是构成其语境的词语，研究者可据此分析其行为。提取索引时，随机的方法很重要；大型语料库中含有关键词的索引可能极多，全部提取出来不必要也不便于观察和描述。因此，可应用随机的方法提取索引，以使其具有代表性。比如，限定在每类文本中提取若干行索引，或者在所有的索引显示后，采用隔行提取、隔两行提取的办法获得索引。由于语言使用在很大程度上是个概率问题(Halliday 1991)，随机提取可以较好地反映某个使用特点的概率属性。下面是笔者随机从 COBUILD 语料库中提取的 commit 一词的部分索引：

1. merely by staying on, did not commit a criminal offence. Both the
2. she felt she would never be able to commit a serious sin again. Thirty-one
3. a knock-out or somebody's going to commit a major faux pas, and outside of
4. prison walls; or for governments to commit abuses that won't cost their
5. Exodus 20 14), "Thou shalt not commit adultery" has become "Be faithful
6. for a crime they knew he did not commit and you know, Monroe County is like
7. of an Account Holder or if you commit any breach of the Conditions,
8. for their own uses. They will commit any crime, but never in passion.
9. months inside for a crime he didn't commit. But in the meantime he had found
10. countries (and therefore free to commit crimes there), that country's right
11. Open daily, volunteers are asked to commit for at least one year; hours are
12. Mold, Clwyd, denies incitement to commit grevious bodily harm. The case
13. However, I feel that by refusing to commit itself to a date for the special
14. an action towards you cause you to commit murder?" [p] Sommers turned from
15. a determined and deliberate act to commit murder. It was well planned and
16. the activist students, he saw them commit no vandalism; on the contrary, they
17. had ended my life. [p] Jane I would commit suicide. I wouldn't want someone to
18. in and decide whether he should commit suicide or not, but his real
19. together, but his wife tried to commit suicide and his children started
20. voice suicidal thoughts or even commit suicide. They are prone to extreme
21. of my injuries, I badly wanted to commit suicide. only there was no way I
22. Graham and told him "I'm going to commit suicide" only 24 hours after
23. watch on him. Last week he tried to commit suicide, one waiter said. "Why?" He

24. you can move away or you can commit suicide. And that's one of the
25. the size of the new German army and commit the Allies to synchronized
26. claimed not only that Green did not commit the crime but that the body was not
27. is dead, that the detective didn't commit the crime, that the person who
28. congregation of young people to commit their lives to Christ. He called
29. of a computer keyboard button they commit their companies to high-risk
30. the Iraqi ambassador refused to commit to the resolutions. He called them

索引证据是研究赖以进行的基本依据。索引提取后，研究者便可对关键词的搭配情况进行观察。上述30个索引显示，commit可用于V+N+PREP，V+PREP 和V+N 等三个类联接。前两类的搭配例有索引11，13，25，28，29，30等6例。V+N搭配是多数，有24例，占全部索引的80%。其中，名词搭配词的情况如下：offence (1)，sin (2)，faux pas (3)，abuse (4)，adultery (5)，crime (6，8，9，10，26，27)，breach (7)，harm (12)，murder (14，15)，vandalism (16)，suicide (17－24)。研究者可以发现，这24个名词搭配词全是一些具有消极语义特点的词项，指生活中那些"消极"之事或"坏事"。无论"犯罪、谋杀、伤害、自杀、欺骗、通奸"，还是"故意破坏"都不是为人称道的事情。根据这些证据，研究者可对关键词commit的搭配特点进行概括和描述。

KWIC索引也可用于汉语的词语搭配研究。下面是笔者从上海交通大学一个汉语语料库中提取的"行为"一词的有关索引：

打击骗取出口退税的不法 行为 ，原告的诉讼请求不能成立，本院不
市工商系统已查处猪肉注水 行为 78起，并严厉惩处了67个不法分
冒产品、欺行霸市和骗买骗卖 行为 国家工商局局长王众孚近日在
保种子的质量，而且是一种违法 行为 对这种扰乱市场的行为要予以坚决
我为这种无所顾忌的以权谋私 行为 感到震惊。他们为什么能够
凸现出来，那种传统的文化 行为 方式和文化价值观念从总体和基本
告董事会的所为，属其企业内部 行为 与政府计划无关。
地产开发的市场规则，规范市场 行为 显得十分必要而且应当
很有必要对房地产开发交易 行为 尽快做出有关详尽的规定。
打击侵犯知识产权的违法 行为 此外，中国政府还分别与美国、欧盟、
顿秩序 规范交易 稳定物价 行为 工商局执法人员一连几天对
市场不但不得人心，而且是违法 行为 必须坚决反对
成了损失，而且，日本的侵略 行为 和殖民统治给亚洲近邻等各国的
服用兴奋剂只是少数人的个人 行为 这当然对整个中国游泳运动
抑肉菜价格，为了规范市场价格 行为 辽宁省从去年初开始普遍
需要规范市场主体的竞争 行为 维护市场秩序。三是改善和加强
用职权徇私、受贿、贪污等犯罪 行为 对国家、对社会危害极大。舆
范各级石油公司及加油站的经营 行为 减少流通环节；对重复建设的加
文化活动中的微观和宏观经济 行为 ，即整个社会的文化生产力和
不良影响，更属严重的侵权 行为 为此，刊登广告者必须立即
取得进展，双方就停止敌对 行为 等重大问题达成了协议。
经济生活中存在着的一些垄断 行为 排斥竞争，降低效率，抬高物价
事件，他们让文明道德 行为 在全市重新走红。让人们都来回答
条例的贯彻实施，在规范预算 行为 、加强预算管理、严肃财经纪律等

在规范土地使用权出让的政府　行为　上，实行土地供应总量控制，严
竞争的能力。一些有短期　行为　的企业，使国有资产流失。
重要的经营性中介服务收费　行为　，评估机构要按照“自愿委托、
职务时以殴打或者其他暴力　行为　造成公民身体伤害或者死亡的，违

由上述索引可知，在N+N搭配里，可作“行为”名词修饰语的词语十分丰富，主要有(1)表示“行为”的主体：“个人、政府、企业、市场”;(2)表示“行为”的性质：“垄断、违法、不法、犯罪、侵略、侵权、暴力、敌对”；(3)表示“行为”的内涵或具体内容：“文化、宏观经济、价格、交易、竞争、预算、服务收费、骗买骗卖、猪肉注水”;(4)其它：“短期、文明道德”等。索引证据显示，“行为”一词的名词搭配词大多与市场经济内容有关。

这种方法的主要环节可概括为：以关键词为中心，以语料库索引证据为依据，参照类联接框架，检查证据和概括关键词的搭配情况。语料库丰富的资源，可给研究者提供详实的证据，使他能够对关键词的搭配情况和特点进行较为扎实的概括，而不必求助于个人直觉。如果所使用的语料库有较大的容量，且有一定的代表性，那么研究结果就有较高的效度，真实语言使用中的词语搭配现象就会被挖掘出来并得以描述。而靠语言学家个人的直觉进行的词语搭配描述，其效度、深度和广度都往往有限。但是，研究限于在久已确立的语法结构之内进行。研究者所做的并不触动语法框架，而是依据证据对结构内词项的组合行为进行检查和概括。所以，这种方法可称之为“基于数据的方法”(data-based approach)。它较适用于实用性的研究活动，比如服务于语言教学的研究，也较易于让一般研究人员接受和掌握。但在大型语料库研究活动中，这种方法显得不尽科学和经济。

3. 计算搭配词方法

第二种常用的研究方法是计算搭配词方法(computing collocates)，即从语料库中将关键词的所有搭配词提取出来，然后用统计手段测量各搭配词与关键词共现的显著程度，以确定词项间在多大程度上存在着相互期待和相互吸引，从而概括、描述结伴的词项在多大程度上反映了词项的典型搭配情况。这种方法不太考虑词项所用于的句法结构，主要的研究环节和概括、描述都由数据驱动。

3.1 提取节点词在语料库中的所有搭配词

搭配词方法首先要从语料库中提取与关键词共现的所有词项。为此，研究人员设计和采用了一整套概念、思想和技术方法，包括节点词(node)、跨距(span)和搭配词(collocate)。节点词即研究人员要在语料库中检查其搭配的词项，也就是上面所说的关键词。语料库中的每个词都可以是节点词。选取哪些词为节点词完全由研究者根据其研究内容和研究目的而定。跨距指的是节点词左右的语境，以词形为单位计算，不包括标点符号。假如将跨距界定为-4/+4，意思是说在节点词左右各取4个词为其语境。以2中显示的commit一词的索引为例，在第一行索引中，commit左边的4个词 staying，on，did，not 和右边的4个词a，criminal，offence，Both 共同组成了节点词的跨距。跨距长度的界定直接关系到搭配词提取的结果。有些常见的词语搭配往往是习惯性地被别的词语分隔，是所谓的非连续搭配(discontinuous collocation)。比如，在LOB语料库中，different from这个搭配的观察频数(observed fre-

quency)为35，其中14次是以非连续搭配的形式出现的。而在两个非连续搭配实例中，different和from被6个词隔开(Kennedy 1990：225)，如：

Non-cooperators were not different in age or other conventional factors from the rest.

因此，跨距长度的界定必须考虑这些问题，一定要视所研究文本的题材领域(topic field)、文类(genre)以及文体风格(style)等诸多影响词语使用特点的因素而定。诸多研究表明，就普通英语文本和专业英语文本而言，将跨距界定为－4/＋4或－5/＋5是适宜的(Jones and Sinclair 1974；Martin *el al*. 1983)。另一个与跨距相关的概念是距位(span position)，指跨距内各个词项所居的位置，常用N－1，N－2，N－3，N－4；N＋1，N＋2，N＋3，N＋4，或者L1，L2，L3，L4；R1，R2，R3，R4等表示。其中，N－1和L1表示紧靠节点词左边的第一个位置，而N＋1和R1则表示紧靠节点词右边的第一个位置。在上面提到的commit一词的索引1中，staying，on，did，not分别位于N－4，N－3，N－2和N－1距位上；a，criminal，offence，both分别位于N＋1，N＋2，N＋3和N＋4距位上。所有落入跨距内的词都被视作节点词的搭配词(collocates)。按照所处位置，搭配词又可分为左搭配词和右搭配词。根据这一套思想设计好的程序，使得机器可对语料自动检索。

这种方法的基本思想是提取节点词的2SN个搭配词，用于观察和研究。其中S代表跨距，2S表示节点词左右两边的跨距，N代表节点词在语料库中出现的总频数，或者叫观察频数。节点词在语料库中出现了N次，就意味着查询时要有N行索引出现，每一行中有2S个搭配词被提取；那么，节点词在语料库中的所有搭配词就是2SN个(Halliday 1976：80)。下面是作者从上海交通大学JDEST的一个子库中提取的词形performed的有关搭配词。该子库词容为1812785，节点词performed的观察频数为272，界定跨距为－5/＋5。提取的搭配词包括2720(2＊5＊272)个形符(token)，这2720个形符含有60多个类符(type)，限于篇幅，现将其中频数最高的20个类符(也就是具体的搭配词)及其数据显示于表1。

N	WORD	TOT.	L.	R.	L5	L4	L3	L2	L1	*	R1	R2	R3	R4	R5
1	performed	272	0	0	0	0	0	0	0	272	0	0	0	0	0
2	the	199	76	123	25	28	15	8	0	0	8	57	20	18	20
3	and	70	34	36	10	12	8	3	1	0	7	2	8	9	10
4	are	34	26	8	0	1	2	3	20	0	1	1	2	0	4
5	were	26	24	2	1	0	0	2	21	0	0	0	2	0	0
6	that	25	14	11	6	3	2	2	1	0	1	3	3	3	1
7	can	24	21	3	1	0	3	17	0	0	0	1	0	1	1
8	for	22	3	19	0	3	0	0	0	0	9	0	2	2	6
9	operations	22	22	0	0	2	10	4	6	0	0	0	0	0	0
10	been	20	18	2	1	0	0	1	16	0	0	0	0	1	1
11	tests	19	16	3	1	3	4	6	2	0	0	1	2	0	0
12	have	19	19	0	0	0	3	6	10	0	0	0	0	0	0
13	with	18	2	16	2	0	0	0	0	0	8	2	1	1	4
14	was	16	14	2	0	0	0	0	14	0	1	1	0	0	0
15	experiments	16	14	2	2	0	2	7	3	0	0	2	0	0	0

16	work	14	12	2	0	1	3	2	6	0	0	0	0	2	0
17	analysis	12	11	1	0	0	3	6	2	0	0	0	1	0	0
18	when	10	7	3	1	3	2	0	1	0	0	0	1	2	0
19	measurements	9	7	2	1	1	3	0	2	0	0	1	1	0	0
20	operation	9	6	3	0	1	1	4	0	0	0	1	0	0	2

表 1　词形 performed 频数最高的 20 个搭配词及其数据

表中数据包含了许多有价值的信息：频数最高的几个搭配词(2－8)都是语法词，揭示了节点词所用于的语法结构；有些搭配词属于偶然搭配词(casual collocates)，它们对节点词没有预见作用，几乎平均分布于左右距位内，但由于语言使用的某种偶然因素而落入了界定跨距内，如 the，and，that 等；从多数搭配词的左右距位分布来看，节点词多用于被动句中：如 were 一词，与节点词共现 26 次，其中 24 次属于左搭配词，且 21 次出现在 L1 位置，紧靠节点词左面(were performed)；been 与节点词共现 20 次，其中 18 次属于左搭配词；was 与节点词共现 16 次，其中 14 次属于左搭配词。词汇词也占了搭配词的相当比例，但它们也多属于左搭配词：operations 一词的距位分布全部属于左搭配词(22/22)；tests 的 16/19 属于左搭配词；experiments 的 14/16 属于左搭配词；analysis 的 11/12 属于左搭配词。这些数据给我们提供了节点词在该库中搭配行为的大致规律和模式。

3.2 统计测量

词语搭配研究的是词项的典型共现行为。典型性(typicality)不同于可能性；在一定程度上，词项的任何组合都是可能的，甚至象 colorless green ideas sleep furiously(无色的绿思想愤怒地入睡)和 This lemon is sweet(这柠檬是甜的)这样的组合，在一定的语境中也不是不可能(McIntosh 1967：188)。Sinclair 说：There are virtually no impossible collocations，but some are more likely than others，即搭配无所谓"不可能"，只是出现的频率不同。所以，搭配词提取后就要进行统计测量，检验各搭配词与节点词之间的相互预见和相互吸引程度，判断它们的共现在多大程度上体现了词语组合的典型性。统计测量一般有两种手段，Z 值(或 T 值)测量和 MI 值测量。

3.2.1 Z 值测量

统计测量的基本理念是语言使用的概率观，按照 Halliday(1991：31)的观点，概率是语言的基本内在属性。在搭配词提取后，研究者要测量的是搭配序列在语料库中出现概率的显著程度。如果语料库的总词容为 W，某个搭配词在库中的观察频数为 C_1，那么，该搭配词在语料库中各个词位平均出现的概率则为 C_1/W。如果限定跨距为 S，该搭配词与每个节点词共现的概率则为 $C_1*(2S+1)/W$(2S 指的是节点词左右两边的跨距位置，1 为节点词所占的距位)。但是，当考虑该搭配词与观察频数为 N 的某个节点词共现的概率时，其理论上的概率应当是

$$P=\frac{C_1*(2S+1)}{W}*\frac{N}{W}$$

用这个理论上的共现概率乘以库容 W，便可求得该搭配词与节点词共现的期望频数(expected frequency)E。那么，搭配词与节点词共现的期望频数为：

$$E=\frac{C_1*(2S+1)*N}{W}$$

也就是说，期望频数的计算涉及4项数据：语料库包含的总词数W，某个搭配词在语料库中的观察频数C_1，限定跨距2S，节点词在语料库中出现的频数N。期望频数被用于Z值或T值的计算。Z值或T值表示的是节点词与搭配词相互预见或相互吸引的程度。在大样本的情况下，两种分值差别不大。计算Z值或T值需要先计算出搭配词在文本中分布的标准差SD：

$$SD=\sqrt{(2S+1)N*(1-C_1/W)*C_1/W}$$

然后用搭配词和节点词共现的实际频数C_2与期望频数E之差除以标准差，即可求得Z值，即：

$$Z=\frac{C_2-E}{SD}$$

Z值达到一定程度，搭配词即可被视为显著搭配词，它与节点词组成的序列则是显著搭配。在COBUILD项目和上海交通大学JDEST语料库研究中，研究者一般将Z=2.0取为显著值；因为用该值可将绝大部分偶然搭配词过滤掉，获得所有有意义的搭配。

所谓偶然搭配词，即那些对节点词没有预见作用，但由于语言使用的某种偶然因素而落入界定跨距的搭配词。从前面的例证可知，被称为搭配词的词语是所有那些落入界定跨距内的词项。其中有些词项可能和节点词没有相互吸引和相互预见关系。比如表1中的the，and，that等词和节点词performed的关系。严格说来，偶然搭配词对真正意义上的词语搭配研究没有太大意义，应当排除。排除的办法之一就是通过统计测量。一般来说，偶然搭配重现的概率较低。当一个词在语料库中出现的频数较高时，统计检验会突出它的典型搭配；它的偶然搭配不太可能具有统计意义上的显著性而被过滤掉。另外一个办法是对统计检验设置起点频数(threshold frequency)：搭配词只有和节点词共现达到一定频数时，才可进行统计检验。COBUILD项目和JDEST研究项目一般将起点频数定为3，因为研究表明，只出现一次的搭配序列可能是语言使用中的偶然行为，而与节点词共现两次的搭配词大多也都是偶然搭配词。偶然搭配词排除或过滤掉后，剩下的大都是和节点词具有相互期待和预见关系的显著搭配词，它们和节点词构成的搭配就是显著搭配。显著搭配基本上反映了典型的词语行为，可以据此勾画出节点词的搭配范围。

由上述计算原理可知，显著搭配通过比较搭配序列在语料库中的观察频数与其期望频数来进行。显著搭配实质上是节点词和搭配词实际共现的次数比人们根据它们在文本中各自出现的频数所作的预测还要多的词语序列。一个搭配序列在语料库中实际出现的频数与其期望频数的差额越高，它就越显著。Z分值(或T值)给研究者一种判断词项间预见和吸引程度的尺度，使他可以判断共现的词语间在多大程度上存在着搭配关系。分值越高，研究者对存在着搭配关系这一点就越有把握，因为计算的情况表明共现的频数已很高，足可以排除是偶然共现。表2显示的是3.1中讨论的节点词performed的有关搭配词的Z分值。

搭配词	与节点词共现的次数	总出现次数	Z值
experiment	17	174	31.19
operations	20	494	21.25

tests	15	465	16.25
calculation	6	97	14.44
engines	3	54	9.75
machines	9	476	9.27
task	6	231	9.10
computers	14	1190	8.59
analyses	4	143	7.81
tools	6	321	7.51
elements	8	692	6.42
systems	13	1614	6.34
functions	5	368	5.64
procedures	4	262	5.43
softwares	4	266	5.37
analysis	12	964	4.71
work	11	1778	4.71
activities	4	354	4.47
operation	6	810	4.03
unit	4	602	3.02
function	4	712	2.61
research	4	737	2.52
test	3	872	1.91

表2　词形 performed 有关搭配词的 Z 值统计测量数据

表2 中搭配词的 Z 值大都达到了显著标准，可视为显著搭配词。它们构成了节点词的搭配范围，反映了它的典型搭配情况。

T 值测量一般用于小样本数据。但 COBUILD 项目一直使用 T 值测量，尽管 COBUILD 基本上已是国际上最大的英语语料库，词容达 3.2 亿之多。表 3 显示的是 COBUILD 的一个子语料库中节点词 knowledge 的 15 个 T 值最高的搭配词的有关信息，子语料库的词容为四千五百万，节点词的观察频数为 3962。

搭配词	总出现次数	与节点词共现的次数	T 值
lack	4170	57	7.244933
common	6744	57	7.056728
prior	1445	42	6.357656
gained	1701	39	6.094638
basic	4347	36	5.600056
secure	2272	29	5.152265
public	19172	41	4.750265
detailed	1648	22	4.496458
general	15785	35	4.443188
gain	2507	22	4.395361
intimate	707	19	4.269362

academic	1885	19	4.120176
medical	6433	23	4.055357
acquired	1029	17	3.985337
practical	2765	18	3.882875

表 3　词形 knowledge 有关搭配词的 T 值统计测量数据

由上面两表数据可以看出，搭配词与节点词的共现频数是 Z 值或 T 值高低的关键因素。共现的频数越高，分值也就越高，也就越能说明搭配词与节点词间存在着搭配关系。

3.2.2 MI 值测量

相互信息值(MI value)原是信息科学领域常用的一个测量手段，现被用于词语搭配研究中，测量词语间的搭配强度。MI 值计算的是，一个词在语料库中出现的频数所能提供的关于另一个词出现的概率信息。Clear (1993)曾用下述例子说明 MI 值的原理：在一个 1000 万词的语料库中，词形 kin 出现了 10 次。这意味着 kin 在该语料库中出现的概率是 0.000001。但是，还是在这个语料库中，如果词形 kith 出现了 5 次，而且在 5 个实例中，kin 总是出现在 kith 之后(Kith and Kin)。那么，当我们看到 kith 时，我们就有 0.5 的概率看到 kin。这样，kith 的出现给我们提供了大量的信息，来揭示 kin 的出现(Clear：278)。MI 值的差异表明词语搭配强度的不同。MI 值的计算方法如下：

$$I(a,b)=\log_2\frac{P(a,b)}{P(a)\cdot P(b)}$$

其中，a 和 b 为语料库中的任意两个词形，P(a,b)是两者共现的概率，P(a)是词形 a 在库中出现的概率，P(b)是词形 b 在库中出现的概率。如果 a 和 b 的结合力很强，则 P(a,b)要比 P(a)．·P(b)大得多，两个词形的搭配强度 I(a,b)也就趋于正值。由此，

$$I(a,b)\geq 0$$

否则，

$$I(a,b)\leq 0$$

如果语料库的总词容为 W，F(a)为词形 a 的观察频数，F(b)为词形 b 的观察频数，F(a,b)为两个词形在库中共现的频数，则

$$P(a)=\frac{F(a)}{W}$$

$$P(b)=\frac{F(b)}{W}$$

而两个词形共现的概率则为

$$P(a,b)=\frac{F(a,b)}{W}$$

那么，

$$I(a,b)=\log_2\frac{W\cdot F(a,b)}{F(a)\cdot F(b)}$$

表 4 显示的是同一个 COBUILD 语料库中，与节点词 knowledge 共现的 MI 值最高的 15 个词项及其有关信息。

搭配词	总出现次数	与节点词共现的次数	MI 值
encyclopaedic	22	9	9.534417
tacit	80	8	7.501793
impart	70	6	7.279379
firsthand	65	5	7.123244
accumulating	80	4	6.501693
encyclopedic	60	3	6.501693
broaden	190	8	6.253741
acquiring	266	11	6.227743
requisite	104	4	6.123144
grammatical	79	3	6.104764
thorough	417	15	6.026561
intuitive	149	5	5.926324
prior	1445	42	5.719007
intimate	707	19	5.605893
gaps	338	9	5.592577

表 4　词形 knowledge 有关搭配词的 MI 值统计测量数据

将表 3 和表 4 中所列数据比较一下就会发现，MI 测量与 T 值测量的结果很不相同。prior 一词虽出现在了两个表中，但所处的次序位置极不相同：在 T 值表中位于第三，在 MI 表中却排至第十三位。MI 值高的搭配词不一定和节点词共现的频数就高。起决定作用的是搭配词与节点词共现的频数与各自单独出现频数之积的比值。MI 测量方法的缺点是，当一个搭配词和节点词共现的次数不多而仍有较高的 MI 值时，它可能表明了词语的搭配强度，也可能是由于语言使用者独特的个人用语特点或者某个语料库的特点所致。对此，研究者很难作出判断。MI 测量的优点在于它能较好地识别复合词、固定词组、科技术语等等。

计算搭配词方法本质上是数据驱动的方法，研究者没有太多的先入为主的观念，由统计数据驱动研究。这种方法适用于大型语料库研究，有利于发现语言使用中的新事实，词语行为的新特点等。

4. 提取词丛方法

传统的语言学描述试图使人们相信“句法第一，词汇第二”；词汇仅是用来填充一定语法结构的材料。但这不反映语言运作的真实情况。研究表明，语言使用中的形式选择在更大程度上是靠成语选择原则(idiom principle)而不是由句法驱动的逐词填缺(slot-filler)(Sinclair 1991：110)。也就是说，语言使用者主要是一次选择一个连续的词语搭配序列来表达意义。根据这一思想，词语搭配研究者设计了提取词丛(word cluster)的方法。词丛，即两个或两个以上的词形构筑的连续词语序列。序列可长可短，根据研究者的目的而定，有二词词丛、三词词丛等等，最长的可达七词词丛。一般的查询软件都有词丛提取功能。研究者将关键词或节点词输入后，再选择和设置所要词丛的长度，即可提取由节点词与其它词形组成的连续词丛。下面是从 LOB 语料库提取的 point 一词频数最高的 20 个四词词丛：

序号	词丛	频数	序号	词丛	频数
1.	from the point	12	11.	at a point on	3
2.	The point of	6	12.	point of view and	3
3.	on the point	5	13.	point of view the	3
4.	point of view	5	14.	the point at which	3
5.	To the point	5	15.	the point in question	3
6.	a point on	4	16.	the point of commencement	3
7.	commencing at a point	4	17.	the point where the	3
8.	in point of fact	4	18.	thereabouts from the point	3
9.	point on the foreshore	4	19.	this point of view	3
10.	a case in point	3	20.	to the point where	3

表 5　词形 point 频数最高的 20 个四词词丛

词丛提取后，可计算每个词丛的期望频数和期望频数与观察频数之比，并判断词丛在语料库中使用的是否显著。由于词丛是个连续序列，其期望频数的计算方法与上一节中讨论的搭配词的期望频数计算略有区别。如果词形 a 和词形 b 在词容为 W 的语料库中的观察频数分别为 F(a)和 F(b)，那么，两个词形在该语料库中组成序列的概率应是 F(a)/W 乘以 F(b)/W。用这个理论上的概率再乘以词容 W，即可求得期望频数之值，即

$$E=\frac{F(a)}{W}*\frac{F(b)}{W}*W=\frac{F(a)*F(b)}{W}$$

如果搭配序列有 n 个词形组成，分别为 a_1，a_2，a... a_n，则

$$E=\frac{F(a_1)*F(a_2)*\cdots*F(a_n)}{W^{n-1}}$$

据此，我们可计算出 point 一词频数最高的 20 个四词词丛在 LOB 语料库中的期望频数及其有关信息：

序号	词丛	名词词频	期望频数	观察频数	观察/期望
1.	commencing at a point	10/5810/22222/435	0.00000006	4	66666666
2.	thereabouts from the point	11/4520/65752/435	0.000001	3	3000000
3.	point on the foreshore	435/6798/65752/7	0.00001	4	400000
4.	the point of commencement	65752/435/34710/8	0.00001	3	300000
5.	this point of view	5043/435/34710/294	0.00002	3	150000
6.	a case in point	22222/492/20446/435	0.0001	3	30000
7.	point of view and	435/34710/294/26852	0.0001	3	30000
8.	point of view of	435/34710/294/34710	0.0002	5	25000
9.	the point of view	65752/435/34710/294	0.0003	6	20000
10.	the point in question	65752/435/20446/266	0.0002	3	15000
11.	point of view the	435/34710/294/65752	0.0003	3	10000
12.	at a point on	5810/22222/435/6798	0.0004	3	7500
13.	in point of fact	20446/435/34710/294	0.0009	4	4444
14.	the point at which	65752/435/5810/4274	0.0007	3	4285
15.	to the point where	26147/65752/435/994	0.0007	3	4285

16.	from the point of	4520/65752/435/34710	0.0045	12	2666
17.	the point where the	65752/435/994/65752	0.002	3	1500
18.	a point on the	22222/435/6798/65752	0.0043	4	930
19.	on the point of	6798/65752/435/34710	0.0067	5	746
20.	to the point of	26147/65752/435/34710	0.023	5	217

表 6　词形 point 频数最高的 20 个四词词丛的统计数据

词丛并不是真正意义上的搭配。在上述序列中，commencing at a point，thereabouts from the point，point on the foreshore 和 the point of commencement 四个序列显著值最高。但是，它们都集中出现在某一类特定的文本中，分布的语域很窄，因此不应视为常用搭配。另外，有些词丛，如 point of view and，at a point on，a point on the 等，明显地是不完整序列，如果增加丛长设置，便可提取出完整序列。但是，词丛方法毕竟能提取一些有用的搭配序列。研究者可以利用词丛提取手段及其有关数据，加上定性分析，发现语料库中常用的连续词语序列。

5. 基本原则与结论

上面几节讨论了基于语料库证据和语料库证据驱动的词语搭配研究的基本方法。我们主要介绍了三种研究方法的理念、环节、步骤和统计测量手段。这些方法体现和揭示了一些重要的原则。

(1)以“自然数据”(naturally occurring data)(Sinclair 1991)为基本依据。这是语料库证据支持的词语搭配研究的重要原则之一。语料库的自然数据来自真实语言交际活动，体现了语言使用的真实规律。当然，语料库语言学不排斥直觉的作用。在研究过程中，人们要使用语言直觉。但使用直觉不同于使用直觉数据。在 Birmingham(伯明翰)大学进行的多次实验表明，本族语者和语言学家个人的直觉具有很大的局限性，为说明和论述某种理论框架而生造的直觉数据不一定反映真实语言使用的客观情况。词语搭配研究的是词项使用的典型行为，不是可能行为(Hanks 1988：121)；直觉数据与可能性有关，但与典型性无关。因此，研究应当始终坚持基于语料库数据或者由语料库数据驱动，不能夹杂生造例子。唯此，才可保证研究的效度。

(2)定量分析与定性分析相结合的原则。语料库研究，就其主要本质特征来说，是基于定量分析的研究。在语料库中只出现一次的词语序列没有太大的意义，研究者注意的是反复共现的词项。然而，定量分析应与定性分析相结合。定性分析涉及词语搭配的界定体系。目前，词语搭配的界定体系很多，极不相同。但总的来说，多数研究人员都趋于强调搭配的因循性、句法限制性、统计度量性、语域限制性等界定特点(卫乃兴 2002)。在定量研究的基础上，进行定性研究，可对搭配序列进行语言学描述，提高一定的理论抽象度。而且，在很多环节上，定性分析可使研究省去一些不必要的步骤。比如，处理偶然搭配词时，一方面借助统计手段，可以将它们过滤掉，另一方面可根据词语搭配的句法限制性这一界定特点将它们排除。定性研究的必要性还体现在对少数非显著搭配的处理上。并非所有的非显著搭配都没有反映词语的典型行为。有些序列在进行统计测量时不够显著，是由于所依据的语料库代

表性不够，或者是语域、语体等具体因素所致。在这一点上，研究者应清楚地认识到，任何语料库，即使其容量再大，包括的文本再全面，从根本上说来，都只是对真实语言使用的有限抽样。所以总难免漏掉一些常见的语言现象，或者导致这些常见的语言现象未能在库中充分反映。对此，科学的作法是用概率的方法对搭配现象作出描述，并进行定性分析。没有定量分析及一整套研究手段和环节，研究就失去了坚实的数据基础，研究者也不可能发现大量的在真实语言交际活动中使用的搭配序列；没有定性分析，词语搭配的某些界定特点可能会得不到重视，研究结果可能缺乏系统性和理论抽象度。

(3)采用词项中心的研究方法(lexis-centred approach)。语言交际活动主要是在一定的交际环境中选择合适的词语来实现意义的一个过程。词语和语法的关系并不是首要(primary)和从属(secondary)的关系，而是共选(co-selection)的关系。为表达一定的意义，一旦词语被选定，相应的语法形式也就随之选定；词语是承载意义的主体，语法是组织词语，使交际有序进行的调节手段。因此，词语搭配研究应以词项为中心，通过观察、分析和概括词在一定语境中的典型行为，来发现它的搭配伙伴，它常用的语法形式及其常常实现的意义与功能等。

(4)以发现词组为目的(phrase-oriented)。词项不是单独或孤立使用的，它总是典型地和其它词结合一起使用，共同实现一定的意义。语言使用者主要是一次选择两个以上的词，作为一个单位来实现意义。词组包括固定词组，半固定词组，词块(lexical chunks)，成语等内容。词语搭配研究的重要目的之一就是发现和描述迄今尚未认真研究的半固定词组。某些领域，如学术英语领域里各种各样的半固定词组基本上仍是块尚未开垦的处女地，需要系统、深入地开掘。从这个意义上说，词语搭配研究又是一种语言学研究方法。用这种由词项到语境，再由语境到词项的方法(Francis 1993)，可以系统地研究语言使用中的词组和词块，对语言描述、自然语言处理、信息科学等相关领域都有重要意义。

本文介绍和讨论了基于语料库证据和语料库证据驱动的词语搭配研究的主要理念、方法、统计测量手段和技术措施，讨论了研究应遵循的基本原则。总的来说，在目前的语料库语言学研究中，词语搭配这个学术界久已谈论的话题被赋予了全新的内容、含义、理念和方法。笔者根据自己和同行进行的研究项目，对这些内容进行了概述、对含义进行了探讨，对理念进行了界定，对思想和方法进行了介绍。文中所谈仅是一家之言，难免挂一漏万，或失之偏颇。不过，作者旨在与同仁讨论，使我国的语料库语言学研究活动更快地开展起来。

参考文献

Aijmer, Karin and Altenberg, Bengt, eds. *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman.

Baker, M., G. Francis and E. Tognini-Bonelli, eds. *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins.

Clear, J. 1993. From Firth principles: computational tools for the study of collocation. In M. Baker, G.

Francis, and E. Tognini-Bonelli, eds., 271－292.
Firth, J. R. 1957. *Papers in Linguistics 1934－1951*. London: Oxford University Press.
Francis, G. 1993. A Corpus-driven approach to grammar: principles, methods and examples. In M. Baker, G. Francis, and E. Tognini-Bonelli, eds., 137－156.
Halliday. M. A. K. 1976. Lexical relations. In C. Kress, ed., *System and Function in Language*. Oxford: Oxford University Press.
----. 1991. Corpus studies and probabilistic grammar. In Karin Aijmer and Bengt Altenberg, eds., 30－43.
Hanks, P. 1988. Definitions and explanations. In J. Sinclair, ed., *Looking Up*. 116－136. London: Collins COBUILD.
Jones, S. and Sinclair, J. McH. 1974. English lexical collocations: a study in computational linguistics. *Cahiers de Lexicologie* 23: 2. 15－61.
Kennedy, G. D. 1990. Collocations: where grammar and vocabulary teaching meet. In S. Anivan, ed., *Language Teaching Methodology for the Nineties*. 215－229. Singapore: RELC.
Martin, W. J. R., B. F. P. Al, and P. J. G. van Sterkenburg. 1983. On the processing of a text corpus: from textual data to lexicographic information. In R. R. K. Hartmann, ed., *Lexicography: Principles and Practice*. 77－87. London: Academic Press.
McIntosh, A. 1967. Patterns and ranges. In A. McIntosh and M. A. K. Halliday, eds., *Patterns of Language: Papers in General, Descriptive and Applied Linguistics*. 181－199. Bloomington and London: Indiana University Press.
Mitchell, T. F. 1975. *Principles of Firthian Linguistics*. London: Longman.
Renouf, A. and Sinclair, J. 1991. Collocational frameworks in English. In Karin Aijmer and Bengt Altenberg, eds., 128－143.
Sinclair, J. 1966. Beginning the study of lexis. In C. E. Bazell, J. C. Catford, M. A. K. Halliday, and R. H. Robins, eds., *In Memory of J. R. Firth*. 410－430. London: Longman.
----. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
卫乃兴，2002，语料库证据支持的词语搭配研究。见杨惠中主编，《语料库语言学导论》上海：上海外语教育出版社。

作者通讯地址： 200030 上海 上海市华山路 1954 号 上海交通大学语言文字工程研究所

《当代语言学》第4卷2002年第2期157－158页，北京

Abstracts of Articles

Li, Aijun; Chen, Xiaoxia; Sun, Guohua; Hua, Wu and Yin, Zhigang, CASS: a phonetically transcribed corpus of spontaneous speech

A collection of Chinese spoken language has been collected and phonetically annotated to capture spontaneous speech and language effects. The Chinese Annotated Spontaneous Speech (CASS) corpus contains phonetically transcribed spontaneous speech. The purpose of the corpus was to collect samples of most of the phonetic variations in Mandarin spontaneous speech due to pronunciation effects, including allophonic changes, phoneme reduction, phoneme deletion and insertion, as well as duration changes. The speech was transcribed in a *three-tiered* annotation: the syllable tier, the semi-syllable tier, and the miscellaneous tier. In the syllable tier, *pinyin* and *tone* of each syllable is transcribed orthographically. In the semi-syllable tier, the *initial* and *final* of each syllable is labeled using SAMPA-C. Segmentation boundaries are also provided in the semi-syllable tier. Sound variability such as phoneme change, insertion and deletion are also transcribed on the semi-syllable tier. Tones after tone sandhi, or tonal variation, are attached to the finals. In the miscellaneous tier, paralinguistic and nonlinguistic phenomena of spoken discourse, such as coughing, laughing, and mouth noises, are transcribed. The percentage of sound variability is 42.2% for initials and 11.8% for finals and 27.2% for syllables.

Qin, Hongwu, Chinese 'V + Time-phrase Structure' from the perspective of situation type and boundedness

It is claimed that the Chinese 'V + Time-phrase Structure' can be interpreted by situation types. However, situation types alone cannot provide a full-scale explanation for the structure. To make a unitary account for this structure, this paper employs the theory of boundedness which is based on temporal perspective at a discourse level. The relativity of boundedness can help explain the reason for the ambiguity of time phrase reference as well as for the boundedness selection of NP serving for the argument of this structure. Moreover, it is argued that Chinese 'V + Time-phrase Structure' has many formats and the time phrase may have more possible referents accordingly.

Wei, Naixing, Corpus-based and corpus-driven approaches to the study of collocation

This paper addresses the features and issues in two major approaches to collocational study in corpus linguistics, a corpus-based approach and a corpus-driven approach. In the corpus-based approach, concordance data is extracted from a corpus on a random basis and examined within a colligational framework to generalize the collocational behaviour of a key word. In the corpus-driven approach, a set of notions and procedures, including the node, the span, the collocates and statistical measures, are designed and adopted to determine collocational patterning of words. Softwares can also be used to extract continuous lexical clusters from corpus in the latter approach. Major principles in conducting the two types of study are then discussed.