英语中词项搭配关系的定量研究

华中理工大学　冯跃进　汪腊萍

摘　要：建立在大量真实语料基础上的语料库为语言搭配的研究提供了强有力的工具。本文从信息理论中互信息量的测量法及统计学理论中实际观察值与期望观察值比值的意义两方面入手，介绍了词项之间相互关联的程度或搭配性强弱程度的定量计算方法，以期为对词项共现意义的判定提供数据化的依据，为广大语言教育者和学习者学习和掌握流畅而又地道的外语提供便捷而又有效的方法。

关键词：语料库、关联程度、频数、概率、互信息量

1. 绪论

词项搭配(collocation)并非新概念，早在五十年代 Firth(1951)在介绍自己有关意义的理论时就已提出：词义不仅通过音素、词素、语法形式、词境等表现出来，还通过词项搭配来体现，亦即通过一个词与另一个或另几个词共同出现的一种较固定的结构来体现。一个词项的搭配因而可以定义为这个词的约定俗成的或惯常的位置，它既不同于语境位置，也不同于语法位置(Firth 1968)。Firth 认为，词项搭配是一种"结伴关系"，且相结伴的词项彼此之间存在一种"共同期待"(mutual expectancy)性，这种靠"共同期待"而聚集在一起的"伙伴"决定了该词的特殊意义和用法。Firth 是从语义上而不是纯结构上去理解词项搭配的概念，这为尔后的语言学家从较高的层次上研究词项搭配奠定了理论基础。但当时的这种理解基本上是局限于语义层次上，分析和研究词项搭配只是为了更好地掌握词义(钱瑗，1997)。

随着语篇语言学的发展，人们开始从语篇衔接的角度去理解词项搭配。Halliday 等把词项搭配看作为语篇衔接的词汇手段之一；词项搭配被理解为词项共现(co-occurrence)，即通过"经常共同出现的词项间的联系"来实现语篇的衔接(Halliday & Hasan 1976)。语篇中的任何一个词项的出现都为另一个(些)词项的出现提供了条件。

综上所述，词项搭配是指常常共同出现于一种可预测的模式中的任何一群词项间的关系。它往往有两种形式：一种是语义层次上结构比较固定的词项搭配，如：

adjective+noun：	heavy/busy traffic
verb+object：	to answer the phone
verb+particle/preposition：	take off/in
noun+noun：	safety belt
verb+adverb：	to rain heavily

另一种是语篇层次上结构比较松散、意义比较连贯的词项共现(Halliday & Hasan 1976 及 W. McRoy 1992)，如 courtroom 与 defendant 常常共现于同一语义场。从另一角度来说，不同的语义场往往决定共现的词项具有不同的意义(Sinclair 1991)。Hanks 通过语料库调查发现，与 bank 共现的词大致可分为两组：

● money, notes, loan, account, investment, clerk, official, manager, robbery, vaults, working in a, its actions, First National, of England 等

● river, swim, boat, east, west, south, on top of 等

这两组不同的共现词与 bank 构成了不同的语义场(亦即"货币"和"堤岸"),同时这两个不同的语义场也决定 bank 具有不同的意义(Hanks 1987)。

对于外语学习者来说,要将真正有意义的词项共现和词项的偶然共现区别开来并非易事,即使是以该语言为母语的人有时也难作出判断。虽然在遣词造句为难之时可以参照各种词典以及名目繁多的工具书,但是这些参考书大多都是将大量的信息简单地按字母顺序把编撰者们认为常用的搭配结构罗列给使用者,使用起来甚是不便。至于搭配结构的常用性及重要性等等,目前还鲜有资料在这方面提供"引路导航"的帮助。

本文使用词料库语言学原则、信息论和统计学原理,旨在对真实语言资料中的词项共现进行数据化处理,通过计算和分析每一个词项共现的份额值大小来确定该词项与其搭配词项之间的搭配性(collocability)和搭配规律,以便为外语教学提供直观、有效的搭配依据。

2. 信息理论计量法在英语搭配研究中的应用

根据语料库语言学原则,我们可以选定任何一个词作为中心词(node word),并用索引软件在语料库中统计出该中心词出现的次数,检索出包含该中心词的所有索引行。假定 x 为中心词,它在语料库中的出现频数为 f(x),每一索引行中以 x 为中心的前后四个词均可视为中心词 x 的搭配词(collocate)。将"中心词±4"作为进行搭配常规研究的定长(span)是许多语料库语言学家的共识(Collier 1992),这是因为前后四个词的距离既大到可以足以看出各类搭配关系的特点,又小到可以依据邻近原则排除偶然共现的搭配词。这样,词项的出现频数×定长,即:f(x)×(4+4),可定义为对 x 进行搭配研究的搭配样库(sample corpus),且该样库中共计有 $N=8\times f(x)$ 个搭配词(次)。

互信息量(mutual information)(Fano 1961)是建立在信息理论计量基础之上的一个概念。它最初只是用来测量两个概念之间相互关联的程度(association ratio),而现在也用来测量动词与宾语、形容词与中心词、动词与小品词、合成名词等等之间的相互关联程度。根据信息理论原理,如果两个词 x 和 y 各自发生的概率是 P(x)和 P(y),其共同出现的概率为 P(x,y),那么 x 和 y 之间相互关联的程序可用互信息量 I(mutual information)表示为:

$$I(x,y)=\log_2\frac{P(x,y)}{P(x)P(y)}$$ (Church and Hanks 1990)

假定 x 和 y 之间存在真正有意义的搭配关系,则它们之间的共现概率 P(x,y)一定会远远大于单独出现概率 P(x)P(y),这时 I(x,y)>0;

假定 x 和 y 之间并不存在有意义的搭配关系,则 $P(x,y)\approx P(x)P(y)$,导致 $I(x,y)\approx 0$;

假定 x 和 y 之间的搭配性较小,彼此几乎成互补的分散关系,则 P(x,y)远远小于 P(x)P(y),导致 I(x,y)<0。

根据统计原理,I(x,y)>3 的词项共现常常都是有意义的词项搭配,而小于 3 的 I(x,y)所反映的词项搭配性不强。鉴于目前语料库大小的关系,一般很难查到 I(x,y)<0 的情形。这是因为:

设 x 和 y 在某一语料库中每百万词次分别都出现了 10 次,即:

$$P(x)=P(y)=10^{-5}$$

这样,从理论上 x 和 y 偶然共现的概率为:$P(x)P(y)=10^{-10}$。因此,若要 I(x,y)<0,则需:

$$\frac{P(x,y)}{P(x)P(y)}<1$$

即：

$$P(x,y)<10^{-10}$$

又因为：

$$P(x,y)=\frac{f(x,y)}{N}$$

故得：

$$N>10^{-10}$$

这就是说：语料库里至少是10^{10}即100亿词次才有可能查到x和y处于理想的互补分布，亦即$I(x,y)<0$的情形。而迄今为止我们所建造的最大语料库Bank of English也只有3.2亿词次(1996年底提供的数据)。

先看一看如何用互信息量来计算两个概念之间相互关联的程序。teachers是一个常见的概念名词，从一个总量为7301773(即：N=7301773)的语料库中检索得知出现频数为：f(teachers)=412。由此计算出teachers在语料库中的出现概率P为：

$$P(teachers)=\frac{f(teachers)}{N}=\frac{412}{7301773}$$

然后再统计出搭配样库中于teachers±4的定长内与之共现的所有概念实词的频数。这些搭配词构成了一个长长的搭配词表(其常用搭配词如表1所示)，其中，概念teachers与概念science共同出现的频数为：

$$f(teachers,science)=22$$

据此可得出teachers与science共现的概率为：

$$P(teachers,science)=\frac{f(teachers,science)}{N}=\frac{22}{7301773}$$

同样，检索出science在语料库中独立出现的频数为：

$$f(science)=1469$$

据此可计算出其在语料库中独立出现的概率为：

$$P(science)=1469/7301773$$

将所有这些数据代入互信息量公式中即得：

$$I(teachers,science)=9.9146$$

同理，可以计算出与teachers共现超过5次以上的所有概念实词的I(x,y)，I值最大的前10位列表如下：

x	f(x)	y	f(y)	f(x,y)	I(x,y)
teachers	412	Steiner	33	6	11.6539
teachers	412	shortage	107	7	10.1792
teachers	412	parents	406	22	9.9074
teachers	412	pupils	159	6	9.3854
teachers	412	students	595	14	8.7039
teachers	412	schools	582	9	8.0984
teachers	412	science	1469	20	7.9146
teachers	412	school	1064	12	7.6430
teachers	412	children	1920	6	6.7914
teachers	412	people	5080	6	4.3877

表1：与teachers相关联概念词的互信息量表(N=7301773)

表1列出的只是与teachers共现大于5次的概念词的I值。一般而言，I值越大，该词项与teachers的相关性就越强，共现于同一语义场的可能性也就越大。但是，如果f(x,y)较小或者f(y)较小时，计算出的I值往往不太可靠，这就是为什么I值最大的竟然是出现频数只有33次的专有名词Steiner。如果不考虑Steiner这一特例，只分析一下表1中有意义的其它9个概念词，就会看到一些十分有趣的现象：排在第一的是shortage，其中一些典型的索引行为：

1 …much a shortage of teachers as a shortage of…

2 …a shortage of Steiner-trained teachers,

causing problems in some…

3 Then the shortage of teachers is likely to bite…

4 There are not enough teachers. The teacher shortage problem…

5 …a shortage of science teachers: the Royal Society estimates…

这些索引行清晰地向人们展示了教师的缺乏是一个全社会关注的问题。余下的八个概念词中占比例最多的是 teachers 授业解惑的对象：pupils，students，children，people 以及其对象的父母亲 parents。其次是教学的场所 school 和教学的内容 science。正如认知语言学家所认为的那样，语言是社会现象的真实反映。基于语料库基础上的语言研究不仅可以看到真正的语言现象，而且可以了解到社会的发展及变迁。

动词 give（包括其各种变体）后可接 of，off，on，out，over，up 等小品词，以构成不同的短语动词。但是，give 到底同哪一个搭配词的关联程度最强呢？按同样的步骤和方法，先统计出 give 在语料库中出现的频数，然后统计出其后紧接着这些搭配词的共现频数，最后统计出这些搭配词在语料库中独自出现的频数，并计算出 I(x,y)，如表 2 所示：

x	f(x)	y	f(y)	f(x,y)	I(x,y)
give	4538	up	22818	202	3.8323
give	4538	off	2961	23	3.6432
give	4538	over	3773	27	3.5254
give	4538	out	4618	25	3.1228
give	4538	of	197345	327	1.4148
give	4538	on	58414	65	0.8403

表 2：与 give 相关联小品词的互信息量表（N＝7301773）

从表 2 可以看出，与 give 关联程度最强的搭配词依次是 up，off，over 和 out。正如 Sinclair（1991）所指出的那样，这种方法在某种程度上是可用来分辩短语动词的搭配性，也可为我们选择教学的重点和顺序提供了一定的参考资料。

3. COBUILD 采用的互信息量值计算方法

以上介绍的是建立在信息理论基础上的互信息量计算方法。虽然在统计搭配词时考虑到了搭配定长的限制，可以较科学、准确地计算出两词项间的搭配程度，但在进行概率计算时并未考虑到搭配样库的大小，而且在计算过程中由于对数运算的使用使得计算本身过于复杂。因此，英国伯明翰大学 COBUILD 语言研究中心采用了另一种既考虑样库大小、又易为大众所理解和接受的互信息量值计算方法。

根据这种简易而又科学的计算方法，假定两个词，即中心词 x 和搭配词 y，在语料库索引行中共同出现的观察频数 j_{obs}. 为 j_{obs}.(x,y)，则 x 和 y 之间的互信息量值为：

$$MI(x,y)=\frac{j_{obs}.(x,y)}{j_{exp}.(x,y)}$$

这里，j_{exp}.(x,y) 为两个词项在语料库中共同出现的期望频数。

以 save 为中心词作例。在 N＝7301773 词次的语料库中统计出 save 共出现 580 次，即 f(save)＝580。然后统计出在 save 左右 4 个词的定长内与 save 共同出现超过 5 次的词项的频数，即统计出 j_{obs}.(x,y)。与 save 共现 5 次以上的词有很多，其中：

j(save,the)＝216

j(save,time)＝31

j(save,money)＝24

然后再统计出这些搭配词在整个语料库中独立出现的频数 f(y)：

f(the)=429516

f(time)=4003

f(money)=1458

在这里，最原始的共现频数 j_{obs}.(x,y)虽然是按从小到大的顺序排列，但这种顺序并不能揭示出两词项之间搭配性的强弱。我们知道，英语中 the，and 和 of 是最常用的三个词，仅这三个词在语料库中的比例就高达 8%(Carter 1988)。定冠词 the 作为英语中第一常用词毫无疑问与 save 共现的机会是最大的，而且可以说，如果仅按原始共现频数 j_{obs}.(x,y)来对搭配性进行排序，那么，the、a、of、to 这样的词恐怕大多要排在任何这样一种搭配排序的最前列。所以现在要通过统计计算出，究竟哪些词才可称为是与 save 真正有意义的搭配词。

在进行推理和计算之前，取**零假设**如下：

假定词项 save 对其周围其它词项的出现无任何影响，亦即：无论 save 存在与否，其临近词项出现的频数保持不变。

从上面的统计资料得知：the 在 N=7301773 词次的语料库中出现的频数为 f(the)=429516。那么在进行 save 搭配研究的样库里，the 可期望出现的频数 j_{exp}.(save,the)在零假设的前提下，可根据下列比例等式计算得出：

$$\frac{f(the)}{\text{语料库词次总量 } N}=\frac{j_{exp}.(save,the)}{\text{样库词次总量 } n}$$

这里 n=(4+4)×save 的索引行数。将所有数据代入上述等式，即可得出在零假设的条件下 the 可期望出现的频数：

j_{exp}.(save,the)=272.9411

而事实上，实际观测到的 j_{obs}.(save,the)=216。两者的比值，即 MI 值为：

MI(save,the)=0.7914

由此可以看出，尽管 the 与 save 的共现频数是最高的，但它与 save 之间的 MI 值并不高。也就是说，save 并不一定要求有 the 修辞的名词短语作主语或宾语等。一般而言，这种 MI 值大致能够说明所观察到的共现频数与所期望的共现频数的不同程度。MI 值越大，说明两者的差异越大，就越能推翻零假设，也就越能证明中心词对自己搭配词的影响就越大。如法炮制，可以得出：

MI(save,time)=12.1865

MI(save,money)=25.9039

比较一下这三个词的 MI 值，可以说：save 与 money 和 time 共现的可能性要远远大于其与 the 共现的可能性。

其实，如果考虑到搭配的前后位置，以上的计算还可以进一步精确化。如果只需考察 save 后续词的词义和词性特点，那么在统计时只统计 save 后搭配词的频数，在计算预期频数时将定长由 8 改为 4，可得：

$$j_{exp}.(save,y)=\frac{f(y)\times 4\times 580}{N}$$

这样，在研究诸如 verb + particle，verb + preposition，adjective + noun 及 noun + noun 这样结构较固定的搭配时统计计算所得的结果将会更精确、更具说服力。

4. MI 值的解释

以上介绍了两种 MI 值的理解和计算方法。Church 和 Hanks 在信息理论的基础上建立了自己的互信息量计算公式，而 COBUILD 语言研究中心是在考虑了样库的大小后按数学统计的原理建立了自己的计算公式的。两者虽各有千秋，但其本质都是一样，都可以用来计算两词项之间的搭配强弱，特别是可以找出那些结构非常固定的习语及技术术语。但是这两种计量法均存在一个共同的缺陷：那就是当 f(x,y)较小时，MI 值可信度较低。有时候虽然 f(x,y)超过 5 次，但 f(y)却相对较小，这时的 MI 值也是不太准确的。从统计学的角度来讲，出现频率很小的事

件，小样本观察时是不会发生的；即使观察到了，也不能轻易否定该事件发生机率很小这一事实。前面讨论过的 Steiner 和 teachers 的 MI 值就是这种情况。虽然 MI(teachers, Steiner)的值是最高的，但由于 f(Steiner)特别小(表中所列出的共现词中频数是最小的一个)，这一最高的 MI 值并不能给我们任何根据来宣称 Steiner 与 teachers 的搭配性是很强的。因此，对于 f(x,y)或 f(x)较小的搭配词来讲，即使 MI 值较大，也不能冒然断言它与中心词的搭配性较强。这时我们需要采用 T 值来改进 MI 值计算方面的缺陷。

有一点需要特别指出的是，虽然语料库的不同可能会导致 MI 值的计算结果有一定程度的差异，但只要语料库相对较大且结构合理，其计算结果通常差异不大，是具有很强的说服力的。

5 结论

以大量真实的语言资料为基础的语料库语言学是人们研究和理解各种语言现象的强有力工具。以上介绍的用 MI 值识别搭配强弱的方法，虽然存在着一定的缺陷，但它确实能在一定程度上将词项间的搭配强弱程度数据化，为语言学习者提供直观且又科学的搭配资料，使得语言教育者和学习者在词汇学习时能够便捷而又准确地把握某一词项的语义特征和搭配特征，能够分清主次，同时提高教与学的效率，改善教学效果。因此对搭配现象的定量研究无论是对语言研究还是语言教学都极具意义。

主要参考书目

Carter, R. 1988. *Vocabulary and Language Teaching*. Longman Group UK Limited.

Church, K. and P. Hanks, 1990. "Word Association Norms, Mutual Information, and Lexicography." *Computational Linguistics* 16(1), PP22－29.

Collier, Alex. 1992. *Issues of Large-scale Collocational Analysis*. University of Birmingham.

Fano, R. 1961. *Transmission of Information: A statistical Theory of Communications*. MIT Press, Cambridge, MA.

Firth, J. R. (1951) "Modes of meaning". *Papers in linguistics 1934－51*, pp190－215, Oxford University Press, Oxford, UK.

Firth, J. R. (1968) "A synopsis of linguistic theory" In F. R. Palmer (Ed.), *Selected Papers of J. R. Firth 1952－59*. Bloomington, IN: Indiana University Press.

Haliday, M. A. K. & Hasan, R. 1976. *Cohesion in English*. Longman: London

Hanks, P. 1987. "Definitions and Explanations." in J. Sinclair (Ed.). *Looking up: An Account of the COBUILD Project in Lexical Computing*. Collins, London and Glasgow.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. OUP

McRoy, W. S. 1992. "Using Multiple knowledge Sources for Word Sense Discrimination. "*Computational Linguistics* 18(1), PP11－12

钱瑗，1997，"对 COLLOCATION 的再认识"，《外语教学与研究》，1997 年第 3 期，PP43－57。

(**通讯地址**：430074 武汉华中理工大学外语系)

(上接第 36 页)

在大学英语四、六级教学取得阶段性成果的今天，专业英语阅读课教学受到越来越广泛的重视。开展专业阅读课教学法的研究，对改进和提高该课程的教学效果，会有积极的作用。

参考书目

大学英语教学大纲·大学文理科英语教学大纲修订组·上海外语教育出版社·1986

国际金融实用英语教程·顾雪帆·沈泽群·上海外语教育出版社·1993·

(**通讯地址**：250014　济南市山东经济学院国际贸易系)