DOI:10.14100/j.cnki.65-1039/g4.2007.03.012

基于语料库的词语搭配个案研究
——以knowledge为例

秦平新

(平顶山工学院 科研外事处，河南 平顶山 467001)

摘　要:将词语搭配作为一个语言学术语正式提出来加以探讨的是Firth。他认为搭配是指词与词的结伴使用;是一种意义方式; 习惯性词语搭配的各伙伴相互期待和相互预见以及类联接是高于词语搭配的抽象。文章以knowledge的搭配为例,利用语料库进行研究,揭示其搭配规律,对英语词汇教学、翻译教学提供一个全新的研究途径。

关键词:语料库; 词语搭配; 类联接; 个案研究

中图分类号: H313　文献标识码: A　文章编号: 1005-9245(2007)03-0132-05

一、引言

自从Firth于半个世纪前提出collocation至今,语言学者关于词语搭配的研究一直在不断地发展、深入并取得了显著的成就。研究者使用的理论框架、研究方法各不相同,试图从不同的角度去探讨和界定词语搭配。学者们使用的术语也不统一。过去的词语搭配研究仅限于学者的直觉,难免带有个人的偏好,具有一定的局限性。而语料库语言学的兴起为词语搭配的研究带来了全新的理念和研究方法。搭配不再是语言学家头脑里的直觉存在;它是成百上千万词容的语料库连续文本(running text) 中的语言使用实体,是数据处理的结果显示。[1]1961年,Sinclair开始了第一个基于语料库证据的词语搭配研究项目。[2]此后,基于语料库的搭配研究在语言学界展开。目前,语料库语言学已形成一套研究词语搭配的方法和手段。

二、词语搭配的概念

根据Firth对collocation的界定: "You shall know a word by the company it keeps."[3]("由词之结伴可知其词"), 词语搭配是指词与词的结伴使用这样一种语言现象。Firth关于collocation的观点概括起来大致有四条: 第一, collocation 是指词与词的结伴使用; 第二, collocation 是一种意义方式; 第三, 习惯性词语搭配的各伙伴相互期待和相互预见; 第四, 类联接(colligation) 是高于词语搭配的抽象。Susan Hunston "…here it is sufficient to note that collocation is the statistical tendency of words to co-occur"[4](这足以说明词语搭配是具有统计意义的词语共现)。

三、国内关于词语搭配研究的文献综述

中国期刊全文数据库的检索(CNKI)结果显示,1999-2006共发表以词语搭配为题目的论文53篇。

卫乃兴的 《基于语料库和语料库驱动的词语搭配研究》[5]是近年来该研究领域的代表作,在学界引起了广泛的关注,对搭配研究产生了积极的影响。该文介绍和讨论了语料库证据支持的词语搭配研究的基本方法和主要原则。作者将基本研究方法分为'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。文章首先讨论了词语搭配研究的基本方法, 之后概括了研究应遵循的主要原则。

卫乃兴的 《搭配研究50年: 概念的演变与方法的发展》[6] 概述了在过去半个世纪里, 搭配(collocation) 概念的演变和研究方法的发展。主要的理论体系包括弗思学派的概念和研究方法, 米切尔等人的

收稿日期: 2007-06-16

作者简介: 秦平新(1963-),男,河南扶沟人,平顶山工学院科研外事处副处长,副教授,主要从事应用语言学及语料库语言学研究。

综合法,韩礼德和哈桑的篇章衔接概念,博林杰等人的惯例化搭配研究,以及语料库研究方法。该文讨论了各研究体系的界定特点及其差异,并概括了概念演变的脉络与方法发展的趋势。

濮建忠的《英语词汇教学中的类联接、搭配及词块》[7]从类联接、搭配这两个与词汇知识深度相关的关键层面入手,利用中国英语学习者语料库,指出学习者在词汇知识深度上的问题和不足与未能充分掌握常用词在使用时的典型类联接和搭配直接相关,亦即未能充分掌握词块。作者进而提出:英语词汇教学的重点之一应置于词块教学。

周明亚的《词语搭配现象与大学英语词汇教学》[8]探讨了词语的习惯性搭配与大学英语中的词汇教学。英汉两种语言的词语搭配习惯受各自语言特征与文化因素的制约或影响。掌握目的语中词语的搭配习惯对外语学习者来说至关重要。教师在词汇教学过程中着重讲授词语的搭配知识,帮助学生掌握词语的正确习惯用法,词汇教学方能取得令人满意的效果。

邓耀臣的《词语搭配研究中的统计方法》[9]介绍了词语搭配研究中常用的三种统计方法的理念和实现方法,特别是对词语搭配研究中常见的MI值和T值的计算方法作了详细的介绍,并对每一种方法的优、缺点进行了比较研究。

其它相关的研究还有:海友尔,谢新卫(1991);李子云(1995);刘桂芳(1995);邵志洪(1994,1995);秦建栋(1996);练稳山(2000);杜凤兰,魏志中(2000);卫乃兴(2001);方艳(2002);鲍成莲(2002);张勇(2002);孙海燕(2004);李晓红(2004);缪海燕,孙蓝(2005);邓文英(2005);刘冬铃(2005);汤闻励(2005)等。

四、研究方法

基于语料库的方法(corpus- based approach)

(一)KWIC索引

KWIC索引是语料库语言学主要的技术手段和方法,也是基语料库于数据驱动学习(data- driven- learning,DDL)的方法。最为常见的索引形式为KWIC(key words in context),即"语境中的关键词"。而索引行(concordance line)则是指关键词及其语境的共现。语料库软件一般都带有索引功能。键入关键词后,语料库会自动检索出来包括关键词在内的一定数量的语境词并且关键词居中显示。以关键词为中心左右4- 5个词数构成关键词的"跨距"。(word span)跨距中的词语就构成了关键词的语境。该语境是连续的文本,可以围绕关键词,从关键词所在行、段落乃至语篇扩展显示。检索可广泛用于英语词汇、语法和语篇的学习。[10]

下面是以knowledge为关键词的语料库检索结果。

(1) lay an intensive and accurate knowledge of Latin grammar, logic- rheto

(2) reactions is that the added knowledge and increased familiarity wi

(3) for training and advanced knowledge. To meet this urgent need

(4) get of art and architectural knowledge- besides remembering that it

(5) der to a general background knowledge of the various types and capab

(6) I felt the need of a better knowledge of Hebrew and of archaeology,

(7) eness, a thorough if bookish knowledge of Asian lore, literature,

(8) thy, the worker's clinical knowledge must determine how these

(9) ts and birds; and her close knowledge of individual students.

(10) stems from the comfortable knowledge that every "volunteer" Demo

(11) the attainment of complete knowledge of the physical universe. For

(12) ses, in view of present - day knowledge of head growth, orthodonti

(13) depths, why not deeper? Knowledge gained from studying earthqu

(14) have little or no direct knowledge. In the autumn of 1959,

(15) t; of acquiring evidential knowledge of what happened in Athabas

(16) pped with everything except knowledge of the "outback" country.

(17) n, in the light of expanded knowledge of ancient manuscripts and

(18) nd on my supposedly expert knowledge of a trade of which they

(19) or the purpose of extending knowledge of the art of ballet in the

(20) lomat with an extraordinary knowledge of Russian language, histo

(21) almost anyone, with a fair knowledge of the English language, can

(22) kind of austere passion for knowledge; there had been lessons in

(23) erates with the target. Full knowledge of the edicbroadenacquiringrequisitegrammaticalthoroughintuitivepriorintimategaps science of oceanograp

(24) ck.This was done with full knowledge that

there would be no epid

(25) wledge.The need for greater knowledge is evident from their repl

(26) as in possession of guilty knowledge. Neither had any interest

(27) in 1904. He had first-hand knowledge of the patent wars which had

(28) rumored he was utilizing her knowledge of Constantinople as part of

(29) the acquisition of higher knowledge, but a positive handicap.

(30) had not planned to use his knowledge merely for war. David Cortla

(二) 搭配词统计

在语料库统计中有两种方法计算搭配词与搜索词的搭配强度或搭配力。一种是通过计算Z值或T值表示搭配的强度,Z值越高,搭配强度越强。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。再大样本的情况下,两种分值差别不大。Z值达到一定程度,搭配词即可视为显著搭配词(significant collocate),它与节点词组成的序列则是显著搭配(significant collocation)。[1]另一种是通过计算在跨距内每个词位的频数分布,根据峰值的显著性来确定搭配强度。如大学英语学习者语料库(CLEC)中knowledge一词的搭配词与搜索词的搭配强度可以用以上两种方法统计出来。在表1中,每个搭配词与搜索词的搭配强度用Z值表示,并从高到低排序。

表1 大学英语学习者语料库(COLEC)中knowledge的搭配词[12]

搭配词	z值
learn	37.776
learned	19.395
books	18.256
enrich	16.070
enwide	15.225
enlarge	14.489
broaden	13.560
acquire	11.494
specialized	11.354

李文中在《基于COLEC的中间语搭配及学习者策略分析》一文中认为,中国学生作文中VERB+NOUN类联接中搭配适当具有以下特征:第一,作为词组核心的名词所触发的动词搭配词自由而多变,缺乏英语本族人的搭配限制。高频名词,如knowledge等所触发的动词非常多,尤其是与名词knowledge搭配的动词多达19个,而这些动词大都具有相似的语义和用法特征。这些搭配不是违反了搭配中两个词的语义同现规则,就是使用了替代词语,造成语用失误。第二,学生过度使用有限的动词词群,并作为搭配词与名词交互重叠使用。同一组动词与某些名词交叉运用,且通过字面直接进入目的语运用,母语文化特征明显。如表2所示。相互搭配使用的动词和名词在语料库中都属于高频词,在意义上具有联想关系。这种特征反映了学生的表述母语文化涵义和交际需求。如在汉语中,"知识"一词与"学问"、"阅读"、"文化"、"识字"具有同义性,在某些语境中可以交替使用;"知识"可以"吸收"(absorb)、"消化"(digest)、"显示"(display)、"掌握"(grasp)、"积累"(accumulate),但这些词一旦被转化为英语,就显得缺乏意义的针对性并导致语用适当。第三,学生通过意义解释以及类比和推断选择搭配词,语用失误来自母语迁移和教学迁移的影响。

表2 knowledge在"VERB+NOUN"搭配中的典型特征(COLEC)[13]

stimulate,openwide, review, enlarge, study, deepen, grasp, know, enhance, gain, accumulatedigest, bring, enrich, learn, improve, expand,own, attain	knowledge

表3 词形knowledge有关搭配词的T值统计测量数据

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
gainint	2507	22	4.395361
imate	707	19	4.269362
acade	1885	19	4.120176
micmedic	6433	23	4.055357
alacquire	1029	17	3.985337
practical	2765	18	3.882875

表3显示的是COBUILD语料库与节点词knowledge共现的15个T值最高的搭配词及其相关信息(子语料库的词容为四千五百万,节点词的观察频数为3962)。由表中数据可以看出,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。[1]

表 4　词形 knowledge 有关搭配词的 MI 值统计测量数据

搭配词	总出现次数	与节点词共现的次数	T 值
encyclopedic	22	9	9.5034417
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
gramma	79	3	6.104764
ticalthorough	417	15	6.026561
intuitive	149	54	5.926324
prior	1445	2	5.719007
intimate	707	19	5.605893
gaps	338	9	5.592577

表4 显示的是COBUILD语料库与节点词knowledge共现的15个MI值最高的词及其相关信息。将表3与表4 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。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。[1]

(三)类联接(colligation)

类联接(colligation)是词语研究的一个重要概念, 它指的是文本中抽象的语法范畴间的结合,而搭配则是习惯性结伴使用的词汇。根据Mitchell的观点,类联接指的是语法范畴间的结合,是关于句法结构的表述,搭配则是类联接在词语层面上的具体实现;类联接不是与搭配平行的抽象,而是高一级的抽象。[1]一个类联接代表了一类搭配。

表5　Knowledge的类联接

类联接	搭配实例
V obj N*	e.g. have knowledge (884)
N* subj V	e.g. knowledge be (1134)
N* subj ADJ	e.g. knowledge useful (18)
N* subj N	e.g. knowledge power (16)
N* subj PREP	e.g. knowledge of (27)
ADJ N*	e.g. scientific knowledge (143)
N N*	e.g. background knowledge (68)
N* PREP	e.g. knowledge of (3950)

(注:N* 代表knowledge)

五、讨论

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, 甚至像Colorless green ideas sleep furiously (无色的绿色思想愤怒地入睡)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。[14]Sinclair说: There are virtually no impossible collocation, 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。统计一般有两种手段,Z值 (或T值)测量和MI测量。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。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。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。[1]

六、结语

基于语料库证据和语料库证据驱动的词语搭配研究为语言研究提供了一种全新的研究视角和研究手段。通过语料库检索,建立类联接,计算搭配词,以真实语料为基本依据,采用定量分析与定性分析相结合的方法,对常用词knowledge进行了基于语料库证据的实证研究,展示了语料库语言学方法在词语搭配研究方面的优势,揭示了词语搭配的一般规律,为语言研究和语言教学提供了科学、真实、可靠的手段和方法。同时,也为翻译和写作教学提供了一个可资借鉴的全新的思路与方法,为正确选用词语表达思想指出了一条有效的途径。

参考文献:

[1] 卫乃兴.词语搭配的界定与体系研究[M] 上海:上海交通大学出版社, 2002: 1- 2.
[2] Jones, S. and Sinclair, J. McH. 1974. English Lexical Collocations: a study in computational linguistics. Cashiers de Lexicologie 23: 2. 15- 16.
[3] Firth, J. R. 1957. Papers in Linguistics 1934- 1951. London: Oxford University Press.
[4] Susan Huston. 2002. Corpora in applied Linguistics.12- 13. London: Cambridge University Press.
[5] 卫乃兴. 基于语料库和语料库驱动的词语搭配研究[J] 北京:当代语言学, 2002,(2).
[6] 卫乃兴.搭配研究50年:概念的演变与方法的发展[J] .解放军外国语学院学报, 2003,(2) .
[7] 濮建忠.英语词汇教学中的类联接、搭配及词块[J] .外语教学与研究, 2003,(6) .
[8] 周明亚.词语搭配现象与大学英语词汇教学[J].外语界. 2003, (2) .
[9] 邓辉臣.词语搭配研究中的统计方法[J].大连海事大学学报, 2003, (4) .
[10] 甄凤超. 语料库数据驱动的外语学习:思想、方法和技术

[J].外语界, 2005, (4).
[11]李文中,濮建中.语料库索引在外语教学中的应用[J]解放军外国语学院学报, 2001,(2).
[12]卫乃兴,李文中,濮建忠.语料库应用研究[M]上海:上海外语教育出版社, 2005:130、138-139.
[13]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.
[14]卫乃兴.专业性搭配初探——语料库语言学方法[J].解放军外国语学院学报, 2001,(4).

A Corpus-Based Case Study of Collocation
—— "Knowledge" as an Example

QIN Pingxin

(Department of Scientific Research and Foreign Affairs,
Pindingshan Industrial Institute, Henan, Pingdingshan, 467001)

Abstract: It is Firth who first coined the term collocation and conducted research on it. He thinks that collocation is the way in which some words are often used together, or a particular combination of words used in this way. The present article takes "knowledge" as an example, conducting a corpus-based case study in order to find the rule of its collocation pattern. The ultimate purpose of the study is to provide some reference of the teaching of English vocabulary and translation.

Key Words: corpus; collocation; case study

[责任编辑:李 蕾]