语料库词典学

理论与方法探索

Corpus Lexicography:
Theory, Method and Application

李德俊 著

译林出版社

料库词典学与方法探索

Corpus Lexicography:

李德俊 著

译林出版社

图书在版编目(CIP)数据

语料库词典学：理论与方法探索 / 李德俊著. —南京：译林出版社，2015.1
ISBN 978-7-5447-5142-1

Ⅰ. ①语… Ⅱ. ①李… Ⅲ. ①语料库-词典学-研究 Ⅳ. ①H06

中国版本图书馆CIP数据核字（2014）第270083号

书　　名	语料库词典学：理论与方法探索
作　　者	李德俊
责任编辑	马绯璠
出版发行	凤凰出版传媒股份有限公司 译林出版社
出版社地址	南京市湖南路1号A楼，邮编：210009
电子邮箱	yilin@yilin.com
出版社网址	http://www.yilin.com
经　　销	凤凰出版传媒股份有限公司
印　　刷	江苏苏中印刷有限公司
开　　本	652毫米×960毫米　1/16
印　　张	15.75
字　　数	185千
版　　次	2015年1月第1版　2015年1月第1次印刷
书　　号	ISBN 978-7-5447-5142-1
定　　价	35.00元

译林版图书若有印装错误可向出版社调换
（电话：025-83658316）

前 言

词典研编与语料库的结合可谓历史悠久，早期词典编纂过程中使用的卡片可视为现代语料的前身。早在1747年，英语词典编纂的鼻祖塞缪尔·约翰逊(Samuel Johnson)就发表了《英语词典规划》(*Plan of an English Dictionary*)，将前人收集资料的好方法作了总结。他所编的英语词典所含例证和说明达15万条以上，可见其所收集的资料库规模已相当可观。《牛津英语词典》(*Oxford English Dictionary*, 简称为 *OED*)于1928年完成，所用的例证有400多万条，卡片1100多万张。《韦氏新国际英语词典》(*Webster's New International Dictionary of the English Language*)第二版的编写参照了100多万条例证，第三版于1961年付印时，新旧例证共达1000多万条。

利用真实语言资料进行研究，也一直是词汇学家和语法学家的传统做法。英语语法大师奥托·叶斯柏森(Otto Jesperson)在编写《英语语法要略》(*Essentials of English Grammar*)时，所使用的卡片数目多达30至40万张。20世纪40年代，美国的语言学家弗朗茨·博厄斯(Franz Boas)在研究美洲印第安语言时就使用了语料库的方法，后来的结构主义语言学家更是如此。

今天，语料库被视为现代语言学的三种主要研究方法之一，而在词典研编领域，不仅词典编纂离不开语料库，基于语料库的各项理论研究也正在如火如荼地进行。从国外的词典研编来看，语料库方法已经逐渐成为主流。

近年来出版的大型英语词典基本都采用了语料库辅助词典编纂(Corpus-aided Dictionary Compilation，简称为 CADIC)的手段。

基于语料库的词典编纂技术研究是语料库词典学的主要研究方向之一，基于语料库的词典学理论研究以及词典语料库建设的研究是语料库词典学的另外两个主要研究领域。

词典学是关于词汇的学问，词义的理解和重现是词典学的核心研究内容。汉语的词义研究从《尔雅》和《说文解字》开始，虽然汉语的词义研究已经有很长的历史，但是我国的词汇研究一直发展缓慢，直到上个世纪 50 年代，词汇研究还主要在“训诂”的范围内进行，真正意义上的汉语词汇(包括词义)研究从改革开放后才开始。改革开放以来，我国出版了一系列的词汇学和词汇语义学著作，这些著作对汉语词汇的系统性、词汇的社会性、词的内部形式、词的语义分类、词的搭配、词义和语素义的关系、语义场、义素分析、词汇应用等方面进行了研究。虽然可以说汉语的词汇研究进入了新的发展阶段，但除了词汇应用研究之外，在词汇研究的大多数领域，研究方法依然较为传统，基于大规模语料的定量分析方法还没有真正开始。

与词汇研究密切相关的汉语词典编纂也主要依靠内省的方法，这集中表现在释义和义项处理上。

内省法明显的缺点是主观性，因为个人的语感或直觉并不总是正确的，而且当自己的语感与其他人的语感发生冲突的时候，也难以找到一个判断正误的标准。由于内省法的缺点，国外有学者称这样的词典编纂为“扶手椅上的词典编纂”(armchair lexicography)。

与汉语单语词典相比，我国汉英词典编纂存在的问题更为严重。从目前出版的汉英词典来看，由于落后的编纂方法和对汉语词典的过分依赖，词典指导编码的功能普遍较低。

语料库语言学为词典编纂提供了全新的方法，基于语料库的方法(corpus-based method)和语料库驱动的方法(corpus-driven method)相互结合，使传统的演绎法和归纳法合二为一。语料库与词典研编在国外的成功证明了语料库的技术手段对于词典研编的重要性。语料库方法依靠计算机强大的储存、检索、索引和统计功能，在词义研究方面具有内省和人工检索无法比拟的优势。

词义与语境关系密切，语料库方法可以通过文本索引重现语境。语料库可以提供大量的释义选项和例证选项供词典编纂人员参考，这些基于大规模真实文本的语料经过词典编纂人员的合理加工后成为词典的现实释义和例证。由于它们来源于真实文本，因此在真实性、科学性和可插入性方面都具有很大的优势，有利于使用者在具体的语境里生成正确的目的语。

短语驱动词典学 (phraseology-driven lexicography) 的研究表明，由短语构成的小语境是最重要的言内语境。语料库词典学的释义和配例等都可以围绕短语展开。

统计功能是语料库的另一个强项，通过统计校验可以使英汉语之间在某一层面上的“联结模式” (association pattern) 凸显出来。统计手段还是研究搭配和用法的有效方法，通过互信息值、Z 值等可以衡量搭配词的搭配力。同时，统计手段还对义项的排序、常用词的常用度衡量等有不可或缺的作用。

对于汉英双语词典来说，平行语料库具有单语语料库无法比拟的优势。双语词典解决的是在具体的使用环境中该如何选择目的语进行表达。双语词典的释义其实就是从源语到目的语的翻译。基于平行语料库的英汉词汇对比研究对双语词典具有重要意义。

语料库词典学具有一定的跨学科性质和技术性，它与信息科学、计算语言学等具有互动关系。由于作者水平所限，书中难免有疏漏或浅薄之处，恳请广大同仁批评指正。

目　录

第一章　引论

◆ 1.1 什么是语料库词典学

语料库词典学可以简单定义为基于语料库的词典学理论研究和词典编纂技术的探讨。但对于词典学本身作为一门学科的地位都不甚牢固的今天，语料库词典学是否能作为一个学科来研究似乎更加令人怀疑。词典学通常被视为词汇学的分支，或者说词典学是将词汇学的理论运用于词典编纂的工作，其本身只能视为词汇学理论的应用。有人认为词典学只能算个应用学科，词汇学才是理论学科。据哈特曼（Hartmann，2006：9）介绍，国际上词典学的研究专刊只有几种：牛津大学出版社出版的《国际词典学学刊》（*International Journal of Lexicography*），北美词典学会（Dictionary Society of North America）出版的《词典学年鉴》（*Dictionaries*），尼迈耶（Niemeyer）出版的《国际词典学年刊》（*Lexicographica International Annual*），WAT 出版的《词典学研究年刊》（*Lexikos*）和上海辞书出版社出版的《辞书研究》等。国内外有影响的词典研究中心也为数不多。大部分词典和工具书也将词典学（lexicography）定义为“词典编写”（dictionary-making）的工艺，而不认为词典学是一个独立的学科。下面是一些权威工具书对词典学的定义：

词典学指的是词典编纂的实践。（Lexicography is “the practice of compiling dictionaries”.）（*NODE*，1998）

词典学指的是词典的撰写或编纂的过程或工作。（Lexicography is

"the process or work of writing or compiling a dictionary".)（*AHD*，1992）

词典学是撰写词典的活动或工作。（Lexicography is the activity or profession of writing dictionaries.）（《柯林斯COBUILD高级英汉双解词典》（电子版），2001）

词典学的地位尚且如此，那么语料库词典学还能作为一门学科来进行研究吗？

据作者掌握的资料，目前以"语料库词典学"为题名关键词出版的著作除了黄铭友（Ooi Beng Yeow）的《计算机语料库词典学》（*Computer Corpus Lexicography*）（Ooi，1998）之外，专门论述语料库词典学的著作在国内外都不多见。但这些都不能否定将语料库词典学作为一门学科来研究的意义。语料库与词典学的结合，不仅使传统词典学在方法论上发生了革命性的变化，语料库词典学关于意义的思考，特别是词义的形成和再现的研究拓展了词典学的理论研究内容。在信息化时代，语料库词典学具有跨学科的性质，它既是当代词典学最具前景的研究领域，也是计算语言学、自然语言处理等领域的重要研究内容。随着词典学学科地位的日益巩固以及计算机语料库技术的日臻成熟，语料库词典学必将受到广泛重视。

1.1.1 词典学学科地位日益巩固

大多数否定词典学是一门学科的人都认为词典学讨论的只是词典编纂的技术，它是一种实践，不是理论研究。这种观点貌似合理，其实简单的推理就能证明其片面性。

自然科学领域也有理学和工学的区分，但从来就没有人怀疑工学研究是一门学问。以计算机科学为例，它的许多领域都是技术性的，如软件开发、程序语言、数据库应用技术等。这些技术性的研究离不开理论，而它反对来又推动了计算机科学理论的发展。词典学不仅是对词典编纂技术的研究，也是词典理论的研究。经过众多学者的努力，词典学已

经具有了词典学本体研究和词典学跨学科研究的双重性质。

1.1.1.1 词典学的语言学传统

20世纪60年代以前，人们（包括词典编纂者）并没有从理论的高度来看待词典编纂，词典编纂与语言学研究长期脱节。马克瓦尔德（Marckwardt，1963：344）批评说："英语词典的编纂没有反映语言学的任何成果，在词典中，词类还是按传统的方式分为名词、形容词和动词等。词典编纂者也没有考虑是否依据形式或功能而给词典一个前后一致的结构。词典在释义的时候也看不到任何结构主义的影子。"

此时，词典编纂和词典学也长期被主流语言学家所忽视。这主要有两方面原因。

第一，语言学家们认为词典只是一种商品，从词典产生之日起，它就没有过任何的变化。词典与语言学理论无关，也没有科学性，不值得对其进行研究。词典充其量只能算是与语言学相关的一个不纯净的副产品（an impure by-product of linguistics）（Rey，1982：17）。

第二，因为词典是关于词汇的书，是语言中某些词的汇集，而在19世纪以及20世纪70年代之前，词汇一直被视为语言研究中无足轻重的成分。在当时，词汇学也没有被当作语言学的一个分支来看待。布龙菲尔德（Bloomfield）非常轻视词汇的研究，他认为"词汇表记录的只是一些没有规律的东西，词汇只能是语法的附庸，而语法反映的才是语言形式有意义的组合"（Bloomfield，1933：274）。除布龙菲尔德之外，20世纪最具影响力的转换生成语法学派在早期也对词汇研究持否定态度，该学派的两个代表人物乔姆斯基（Chomsky）和哈利（Halle）是这样定义词汇和认定词汇功能的：词汇如同一个垃圾桶，在语言研究中不规则的、没有共性的东西就会被扔到其中（Chomsky and Halle，1968：12）。

1961年《韦氏新国际英语词典第三版》（*Webster's Third New International Dictionary of the English Language*）的出版标志着词典编纂领

域一个全新时代的到来。它受结构主义语言学的影响，崇尚描写语言学而抛弃了规定主义的历史语言学。该词典的出版引起了语言学界的广泛争论，也引起了语言学家对词典学的关注。自 20 世纪 60 年代以来，词典学引起了越来越多的语言学家的注意，词典学家也意识到语言学理论对词典学的重要意义。

1960 年在美国印第安纳州的伯明顿（Bloomington）召开了第一次具有国际影响的词典学会议，参加会议的不仅有词典学家，还有一些有影响的语言学家。此后，许多语言学家不仅撰写词典学研究论文，而且还直接参加了词典的编写工作，如威廉·拉波夫（William Labov）、伦道夫·夸克（Randolph Quirk）、戴维·克里斯托尔（David Crystal）、约翰·辛克莱（John Sinclair）等。

早在上个世纪 40 和 50 年代，有些词典学家就意识到语言学对词典编纂的价值，从那时开始，词典学就不断从语言学研究中汲取营养。随着词典学研究的深入，语言学和词典学研究开始相互交织，互为补充，从单纯的给予和获利关系转变为互动关系（Bejoint，2002：177-178）。词典编纂工作的重要性和词典研究的价值也逐渐得到了语言学界的认可，词汇研究再次受到重视。

从 20 世纪 70 年代起，词汇研究也逐渐受到了语言学家的关注。转换生成学派发现用转换规则无法解释一些结构之间的关系，如 they destroyed Pompeii 和 their destruction of Pompeii。在这种情况下，乔姆斯基试图运用一种叫“词汇假设”（lexicalist hypothesis）的理论对词汇（像上面的 destroy 和 destruction）进行赋值。他认为这些问题靠语法（grammar）是解决不了的，应该开展词汇的研究（Chomsky，1970）。随乔姆斯基之后，杰肯诺夫（Jackendoff，1975）研究了与词汇相关的“冗余理论”（redundancy rule），后来他进一步认为词汇是语音结构（phonological structure，简称为 PS）、句法结构（syntactic structure，

简称为SS）和概念结构（conceptual structure，简称为CS）间联系的桥梁。哈德森（Hudson，1991：1-14）在《英语词汇语法》（*English Word Grammar*）中列举了当今语言学发展的8个趋势，其中“词汇主义”（lexicalism）被摆在首位（Hudson，1991：3-4）。词汇主义认为语言研究应该从语法中心转变为词汇中心。此外，词汇研究也受到了其他许多学者的重视，如韩礼德（Halliday，1985，1991，1992）等。

今天，词典学研究的内容和方法已经远远超出了语言学的范畴，词典学既不隶属于词汇学，也不是应用语言学的分支，它已经是一门独立的学科。

从元词典学（metalexicography）的角度看，哈特曼认为词典学的主要研究对象有5个：（1）词典史的研究；（2）词典评价的研究；（3）词典结构的研究；（4）词典类型的研究；（5）词典使用的研究。（Hartmann，2006：31）

哈特曼的分类其实不够全面，词典的主要功能之一是用语言（同一种或另一种）诠释语言，因此与词典诠释相关的理论和方法研究也应该是元词典学的主要研究对象，它与哈特曼所列举的5个领域共同构成了词典学理论研究的主要内容。

自兹古斯塔（Zgusta）之后，国内外涌现了许多词典学理论家，在理论词典学和应用词典学领域都出版了有影响力的著作，发表了大量词典学方面的论文。国内学者章宜华和雍和明认为“现代词典学已经形成一门自成一体、相对独立的学科”（2007：7）。他们还认为，将词典学简单视作一种技巧和艺术而不承认词典学是一门学问的观点是“对词典学的现代发展和词典学理论指导词典编纂的作用视而不见”（2007：4）。

1.1.1.2 词典学的跨学科研究

在计算语言学和机器翻译领域，有关自然语言的大部分知识都是以机器词典（又称“电子词典”、“自动词典”）的形式存储和利用的。机器