刘克强 著

语料库词典学与基于平行语料库的《三国演义》习语翻译词典的研编

Corpus Lexicography and Parallel Corpus-Based Compilation of Idiom Translation Dictionary of The Romance of Three Kingdoms

本课题尝试利用自建的《三国演义》汉英句对齐平行语料库编写《三国演义》习语翻译词典。全书共分为六章，第一章回顾并接受语料库词典学的历史；第二章详细分析利用语料库编撰词典的特点；第三章对平行语料库在双语词典编撰中的作用做了简要的介绍；第四章介绍《三国演义》汉英平行语料库系统的建设、标注等过程；第五章以自建的《三国演义》汉英平行语料库为基础，系统分析和探讨《三国演义》习语在Robert和Taylor两译本中的翻译情况；第六章是《三国演义》习语翻译词典的正文部分。

南大学出版社
University Press

语料库词典学与基于平行语料库的《三国演义》习语翻译词典的研编

Corpus Lexicography and Parallel Corpus-Based Compilation of Idiom Translation Dictionary of The Romance of Three Kingdoms

刘克强 著

云南大学出版社
Yunnan University Press

图书在版编目（CIP）数据

语料库词典学与基于平行语料库的《三国演义》习语翻译词典的研编 / 刘克强著. --昆明：云南大学出版社，2012

ISBN 978-7-5482-1219-5

Ⅰ.①语… Ⅱ.①刘… Ⅲ.①语料库—词典学 ②《三国演义》—英语—社会习惯语—翻译—词典 Ⅳ.①H06②H315.9-61

中国版本图书馆 CIP 数据核字（2012）第 216665 号

语料库词典学与基于平行语料库的《三国演义》习语翻译词典的研编

刘克强 著

责任编辑 石 可 刘 焰
封面设计 刘 雨
出版发行 云南大学出版社
印 装 昆明研汇印刷有限责任公司
开 本 787mm×1092mm 1/16
印 张 18.75
字 数 340 千
版 次 2012 年 9 月第 1 版
印 次 2012 年 9 月第 1 次印刷
书 号 ISBN 978-7-5482-1219-5
定 价 42.00 元

地 址：云南省昆明市翠湖北路 2 号云南大学英华园
邮 编：650091
网 址：http://www.ynup.com
E-mail：market@ynup.com

前 言

利用语料库编撰词典自古有之，且成为传统，中外概莫能外。不同的是，在计算机语料库问世之前，人们则完全依赖于手工搜集语料并记录在卡片上。James Murry 等编撰《牛津英语大词典》，从开始到结束的70年多年间曾经共制卡片500多万张；17卷本的《现代俄罗斯标准语词典》耗时17年，共制卡片高达600多万张。在我国，《辞源》主要编撰者吴泽炎先生亲手摘录的卡片达60万张。1964年，第一个计算机语料库布朗语料库在美国布朗大学诞生，规模达100万词次，开创了现代意义语料库的先河。但英国似乎有后来居上之势，先是 Sinclair 领导 COBUILD 项目，率先利用 KWIC 检索方法，编撰了柯林斯合作英语词典。该词典无论是词条选目及释义、例证选择、搭配及语法描写等都较以往的词典发生了革命性的变化；继而在国家战略的支持下，英国率先建立了世界上第一个国家语料库，即英国国家语料库（BNC），该语料库系一平衡语料库，库容达1亿词次，其中口语语料占1/10，且采用严格的统计抽样方法，较好地解决了语料的代表性问题。同时，值得注意的是，BNC 的建设过程也值得学习，即六大单位或公司通力合作，优势互补，不仅提高了工作效率，而且催生新的技术，如合作方之一的 Lancaster 大学负责语料词性标注，开发了基于概率与规则相结合的词性标注软件 CLAWS，至今准确率仍然居世界同类的前列。一方面，BNC 已成为其他国家建立国家语料库的模板，如已建成的俄国国家语料库（RNC 或 BOKR），至今仍然在建的美国国家语料库（ANC）。另一方面，BNC 发挥了强大的商业效用。1995年一年之间同时出现的《牛津高阶英语词典》、《朗文当代英语词典》、《柯林斯合作英语词典》、《剑桥高阶英语词典》及2002年问世的《麦克米伦高阶英语词典》等无不以此库为参照语料库。此外，国际上许多机构还订购了 BNC。目前，BNC 在互联网上也有多个检索引擎。而 BNC 也从原来最初的版本升级为现在的 BNC XML 版，使得标注更为规范，数据传输更加方便。

借助语料库也缩短了词典的升级换代的周期，并促进了词典的多样化、专业化。以《牛津高阶英语词典》为例，从第一版（1948 年）到第四版（1989 年）平均更新的周期为 11 年多，但从第五版（1995 年）到第八版（2010 年），更新周期则缩短至 5 年；柯林斯公司基于 BOE 于 20 世纪 90 年代推出系列词典，如《柯林斯最新英语词典》、《最新柯林斯英语学习词典》及《柯林斯精华英语词典》等，不胜枚举。

除基于单语语料库编撰词典外，利用平行语料库可以辅助编撰双语词典。平行语料库提供的对等翻译，对编撰双语词典发挥了巨大的作用。可以说，自 20 世纪末，不管是编撰新词典还是更新旧词典，不基于语料库数据的，几乎未闻。基于计算机语料库编撰的词典，开始已经成为英语词典出版业中的时尚与主流。这种做法已经被其他语种的词典编撰者学习并仿效。

我国对语料库的建设也十分重视，1979 年以来，多个专门研究机构或高校都建立了规模不等的机读语料库，从百万字次到几千万字次，有的也进行了深加工，但大多用于语言学研究，用于词典编撰的鲜有所闻。现有的文献也大多仅限于介绍语料库在词典编撰中的应用，利用语料库编撰词典的实践很少。

综上所述，我们认为有必要了解语料库词典学的历史，认识语料库在词典编撰中发挥的作用。需要说明的是，利用语料库编撰词典是一项实践性很强的活动，特别是利用平行语料库编写词典，建库工作量之大、任务之艰巨是不言而喻的。本课题尝试利用自建的《三国演义》汉英句对齐平行语料库编写《三国演义》习语翻译词典，以期在此方面有所突破。

本书分为 6 章，各章既相对独立，又共同组成一个有机整体。

第 1 章回顾并介绍语料库词典学的历史，分为四个阶段，其中每个阶段标志性的成果作为重点介绍对象。

第 2 章详细分析利用语料库编撰词典的特点，重点探讨语料库在词典编撰中具体环节上的应用，如义项划分与排序、例证选取等，特别是以语料库词典学第四阶段的标志性成果 Sketch Engine 为例，介绍了其在语料库与词典编撰之间的接口功能。

第 3 章对平行语料库在双语词典编撰中的作用作了简要地介绍，用

实例探索了平行语料库对双语词典词条释义方法的革新。

第4章介绍《三国演义》汉英平行语料库系统的建设、标注等过程，其中习语的标注为下面两章相关研究作好了铺垫。

第5章以自建的《三国演义》汉英平行语料库为基础，系统分析和探讨《三国演义》习语在Robert和Taylor两译本中的翻译情况。

第6章是《三国演义》习语翻译词典的正文部分，以笔画为序，以句子为单位，对比列举538个习语及其所在单位的翻译，其中大部分习语的出现频率超过一次，实际上包括习语译例达2158句次。

需要指出的是，《三国演义》是我国演义小说的开山之作，语言面貌呈历史演义体语言的特色，涵盖叙事性语言、对话性语言与引用语料。其中仅引用语料就包括诏令、疏表、奏章、策文、檄文、榜文、盟文、告示、书信、诗歌等。这些不同语料的共存使得《三国演义》变得复杂和特殊，同时给译者带来很大的挑战。不同的语料，译者翻译时须考虑采用不同的策略和方法。故归纳其中习语的翻译有着十分重要的意义：一方面，可以对比两译者的不同翻译；另一方面，更重要的是，对比同一习语在不同语料间的翻译，这些不同层次的对比不仅对深刻理解习语在不同语境下的翻译无疑具有巨大的帮助，而且对翻译鉴赏提供了真实、可靠的资源。此外，由于两译本问世时间相距达70年之久，从历时的角度亦可以探索译语的变化规律。

最后，笔者在写作本书的过程中，得到我的导师，上海外国语大学教授冯庆华先生的热情鼓励与支持，在此表示衷心的感谢；我的爱人曹淑芹女士牺牲自己的业余时间，帮我收集资料，在此也表示诚挚的感谢。此外，我的学生邢道远利用寒假，替我扫描《三国演义》的译本，对他付出的辛勤劳动也表示感谢。由于笔者水平有限，加之时间仓促，文中倘有错误之处，恳请方家批评指正。

刘克强

2012年6月于红河州蒙自苦研斋

目 录

第 1 章　语料库词典学的发展阶段

语料库在词典编撰中的应用问题是词典学界争论较多的话题之一。20 世纪 80 年代末的争论之一是词典编撰是否需要语料库，90 年代末的争论是语料库的规模及代表性的问题，此后研讨的便是如何从语料库资源中提取最有价值的信息（Kilgarriff & Rundell, 2002：1）。由此可见，语料库愈来愈应受到重视，已经成为学界共识。本章首先厘清语料库与词典学的概念，在此基础上梳理语料库词典学的历史进程。

尽管语料库的界定仍然存在争议，但基本可从广义和狭义两个层面理解，前者认为语料库是一些书面语或口语转写稿的集合，能为语言分析和描写提供帮助（Kennedy, 1998：1）；后者则认为，原则上，任何多于一个文本的集合都可以称作“语料库”，这仅仅相当于拉丁语中的“材料”。在当代语言学背景下，语料库更倾向包含更多的内涵，应主要从下面四个层次去理解：采样收集与代表性、规模有限、机读性及标准参考数据（McEnery, 2001：29）。显然，狭义的语料库是在当代语言学理论发展的基础上，结合概率论与计算机技术而界定的，具有显明的时代特征和现实的实践特质。实际上，现代意义上的语料库一般指狭义语料库。

语料库是语料库语言学的基础和重要组成部分。Leech（1992：105－106）认为语料库语言学不是指研究的领域或范围，而是指进行语言学研究的方法和基础。它研究自然语言文本的采集、存储、加工和统计分析，研究的焦点包括语言运用、语言描述、语言的定量和定性模型。目的是凭借大规模语料库提供的客观、翔实的语言证据来从事语言学研究和指导自然语言信息处理系统的开发。语料库语言学本质上是实证性的（empirical）。

词典学是计划和编写以字词为主的工具书的实际过程，此类工具书通常提供一种或数种语言的词汇信息。与此同时，词典学一词还指对上述过程及其产品的研究，如形式、内容、市场及使用等（Ilson, 1986：330）。显然，仅从词典学是计划和编写以字词为主的工具书的实际过程这一层含义来看，语料库与词典学就存在着必然联系。

基于 COBUILD（Collins Birmingham University International Language Database）词典编撰的经验，Sinclair（1985：81－94）和 Atkins（1991：167－

204）提出一种全新的方法论，旨在测评词汇言语行为所表现出来的实证结果，这种结果有可能产生比传统词库更全面、更连贯、更一致的效果。Atkins 将此研究方法称为语料库词典学。根据 Leech 的上述观点，我国学者章宜华（2004：15）将基于语料库的词典学理论研究和词典编撰工艺探讨定义为“语料库词典学”，并指出其研究范围包括三个方面：一是语料库的建立；二是语料库的管理；三是语料库的使用。

词典编撰的基础是必须具有反映词汇使用行为的材料，这些材料不管是纸质的还是存储于电脑中的，即机器可读的（即前文讨论的广义和狭义语料库），在词典编撰过程中发挥着十分重要的作用。实际上，语料库与词典的编撰的结合有着悠久的历史，按照时间顺序，综合计算机、语料库技术等的发展可以分为四个阶段。

1.1 第一阶段：计算机前时期

第一阶段称为计算机前时期，也就是计算机还没有问世的时期。这个时期出产的典型词典包括约翰逊的《英语词典》，这部被称为 18 世纪英语词典典范和具有里程碑意义的词典，是第一部由编者明示使用了语料库的词典（姚喜明、张霖欣，2008：51）。在序言中，约翰逊明确指出，此词典的引证均选自从菲利普·悉尼时代（1560 年）到王朝复辟时期（1660 年）100 年间的语言材料，全典收录例证 11400 条，大都出自著名作家作品，引用书证的目的是进一步阐明词义、说明用法、证实某些词的存在。默雷（Murry）等编撰的《牛津英语大词典》完全依赖于手工搜集语料并记录卡片，这些卡片组成传统意义上的语料库。从开始到结束的 70 年间曾用这种人工方式积累真实语言素材，共制卡片 500 多万张，其中，征集 1000 名志愿者及 800 多名英国学者和 400 多名美国学者投身于例证搜集工作中。可以说，利用真实的语言资料来编写词典是英国词典学家的传统。我国的词典编撰者历来也十分重视语料的搜集和整理，以《辞源》为例，主要的修订和编撰者吴泽炎先生亲手摘录的卡片就达 60 万张。总之，这一阶段的突出特点是人工工作量巨大、词典的编撰周期长、时效性差、资料缺乏客观性。随着计算机技术的发展，传统的卡片语料库被新型的基于计算机技术的语料库替代。

1.2　第二阶段：COBUILD

第二阶段以辛克莱（Sinclair）领导的柯林斯伯明翰大学国际语言数据库 COBUILD 项目为开始标志，始于 20 世纪 70 年代后期。当时辛克莱敏锐地发觉计算机蕴藏着巨大的潜力，可以实现诸如存储、分类、查询等功能，这些工作以前需用人工完成，而计算机可以将这些工作完成得更加准确、客观；此外，人工在搜集例证时往往找寻那些稀奇甚至古怪的例子，或者自认为是有趣的例子，因而将重点放在不常用的奇特的表达上，从而忽视了常用的表达。计算机则不同，既能显示常用的表达方式，又可以显示特例，从而比较客观地描述语言使用状况。从 COBUILD 开始，词典编撰者开始使用一种全新的方式来观察和审视词汇的使用行为，这就是后来语料库词典编撰学家必须使用的最基本的 KWIC（Key Word in Context），即关键词居中的索引工具。KWIC 逐渐成为语料库检索软件基本的功能。但由于当时计算机体积庞大、价格昂贵，词典编撰者无法在家中使用它们。《柯林斯 COBUILD 英语词典》的计算机专家 Jeremy Clear 曾写下这样一段话："1987 年初我写这篇文章时，微型计算机和网络通信技术仍然无法提供一个经济的、能使大批词典编撰者不用纸和笔进行词典编撰的系统。"（Clear，1987：41－61）事实上，第一代《柯林斯 COBUILD 英语词典》都是基于约 700 万词的语料库编写的，当时，检索输出需要花费很长时间，索引结果打印后分给词典编撰者。然后他们仍然采用传统的"彩笔法"，即词典编撰者手持彩笔，阅读这些索引行并判断词义，同时用同一颜色的彩笔标记同一词义的语言使用例证，直到阅读完所有索引行后再对该词的意项及句法结构加以归纳和定义。因而，词典编撰工作虽然较第一阶段有了进步，但仍然费时耗力、效率低下。但值得注意的是，这个阶段出现了许多语料库检索系统（Corpus Query System，CQS），利用这些系统软件，词条的用例可以得到检验，这些系统包括 Wordsmith Tools，MonoConc，the Stuttgart workbench 及 Manatee 等。这些系统工具现在随着时间的推移和计算机技术的发展而不断更新，目前仍然发挥着重要的作用。总而言之，这个阶段虽然计算机已经作为一种工具辅助编撰词典，但由于当时各方面条件所限，效率仍有待提高。

在语料库发展史上，第一代语料库以布朗语料库为模板，构成了布朗家族语料库。它们的共同特点是库容量是 100 万词且按照严格的抽样组成，而第二代语料库的库容增大了许多，达到 1 亿词（如 BNC）至 10 亿词（剑桥国际英语语

料库），甚至更多。这样语料库中的高频词出现的次数很多，一般来说，一个词（型）如果在索引中超过1000个，对词典编撰者来说已经是数据分析的极限了；如果超过此极限，人工将难以对付。但事实是，随着语料的增多，人们才会对词汇、语言进行更加完整和准确的描述和分析。随着语料库容量的急剧扩大，语言学家和词典编撰人员迫切要求有更高效的软件系统予以支持，从而从大量索引结果中获得最佳采样，并且对如何进行分析提供有效的技术指导。

1.3 第三阶段：词汇统计学

第三阶段以 Church 和 Hank（1989：76－83）开创的词汇统计学为标志，利用信息论中互信息值（Mutual Information）来度量词汇间的黏合度，从而判断语料库中有意义的词汇间搭配，包括实词与实词间的搭配，也包括实词与虚词（如动词与介词、形容与介词）间的搭配，甚至多个词组成词串间的黏合情况。下面以两词组合为例来说明其计算方法：将组成词对（也可以是任意两个被观察对象组成的同现对）的两个词 x 和 y 在一定跨距内的同现（一般是左右为5词），看作一个联合发生的随机事件，互信息值就是这个联合事件的概率 $P(x, y)$，除以这两个词分别出现的概率 $P(x)$ 和 $P(y)$，然后取其对数。

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

一般来说，两个词之间的联系越紧密，它们间的互信息值就越大；如果两个词是负相关（即一个出现时，另一个肯定不出现，反之亦然），那么它们的互信息值就是负数；如果两个词相互独立，即它们之间没有任何关系，则互信息值为零。结论是，互信息值较高的两个词很有可能组成一个有意义的搭配，而互信息值接近或小于零的词对则不可能构成搭配。表1.1及1.2是BNC中与动词“make”互信息值较大与较小的20个词的信息。

表1.1 BNC中与“make”的互信息较大的前10词

$I(x, y)$	$f(x, y)$	$f(x)$	x	$f(y)$	y
6.2874	6	77228	make	190	amends
5.1143	332	77228	make	23708	sure
5.0773	19	77228	make	1392	representations
4.8461	7	77228	make	602	donation

续表

$I(x, y)$	$f(x, y)$	$f(x)$	x	$f(y)$	y
4.8257	18	77228	make	1570	mistakes
4.8211	8	77228	make	700	adjustments
4.7334	57	77228	make	5300	contribution
4.4942	8	77228	make	878	judgements
4.491	15	77228	make	1650	enquiries
4.4467	7	77228	make	794	predictions

从上表可以发现，诸如 make amends，make sure，make donation，make mistakes 等都是常用的固定搭配。

相反，表 1.2 中的 world，times，months，research 等与“make”固定搭配则不可能构成固定搭配，因为它们与“make”间的互信息值均为负值。

表 1.2　BNC 中与“make”的互信息值为负的 10 词

$I(x, y)$	$f(x, y)$	$f(x)$	x	$f(y)$	y
-0.8371	13	77228	make	57447	world
-0.976	6	77228	make	29194	times
-0.9785	5	77228	make	24370	months
-1.1092	5	77228	make	26682	research
-1.2332	6	77228	make	34891	night
-1.2626	9	77228	make	53414	part
-1.2693	6	77228	make	35775	head
-1.3106	6	77228	make	36815	school
-1.3543	5	77228	make	31622	week
-1.9922	9	77228	make	88571	years

用互信息值识别出来的搭配词种类繁多，同时又呈现出一定的语义特征。这是互信息值的价值所在。但是，值得注意的是，当中心词与搭配词的共现频数 $f(x, y)$ 较小或搭配词在语料库中的总频数 $f(y)$ 相对较小时，互

信息值即使较大也是不太可靠的。实践发现，利用互信息值提取的搭配词往往会出现与节点词共现频率较低的情况，这样就不能得出它们是典型搭配的结论。因此，Church（1991：1－32）等利用T检验值消除互信息值存在的问题，剔除那些可能观测到的但不典型的词项组合，突出真正常用的有代表性的词项搭配。根据T检验统计理论的原理，假设两词 x 和 y，在某语料库中共现概率为 $P(x, y)$，各自单独出现的概率为 $P(x)$ 和 $P(y)$，那么所观测到的共现概率为 $P(x, y)$ 与随机共现的偶然概率 $P(x)$ 和 $P(y)$ 之间的T值为：

$$T(x,y) = \frac{P(x,y) - P(x)P(y)}{\sqrt{\frac{P(x,y)}{N}}}$$

根据 T 检验值的统计意义，两词项 x，y 的T值反映的是两词项间搭配强弱的相对差异。从统计学的角度来看，1.65 个均方差的差别是判断两词项搭配是否有意义的最低临界值。1.65 个均方差表明我们有 95% 的把握断言观测共现概率 $P(x, y)$，与偶然共现概率 $P(x)$、$P(y)$ 的差别是客观存在而非偶然巧合。同时，根据公式，我们可以发现，词汇间的共现概率 $P(x, y)$ 在获取T值中是十分重要的因素，所以正如上述所讨论的那样，T检验值避免了互信息值的缺陷。表1.3是BNC中与动词“make”共现且T值较高的词汇的相关信息。

表1.3　BNC中与“make”的T值较高的10词

$T(x, y)$	$f(x, y)$	$f(x)$	x	$f(y)$	y
8.9608	170	77228	make	207305	up
8.6911	105	77228	make	62165	use
8.0245	70	77228	make	11161	difference
7.84	72	77228	make	21349	sense
7.8084	149	77228	make	209333	more
7.4904	88	77228	make	69149	'll
7.3372	66	77228	make	24925	clear
7.3157	191	77228	make	350517	we
7.1369	436	77228	make	1118985	that
6.949	79	77228	make	67206	does

从上表可以观察到这些词与动词“make”的共现频率都较高，make up，make use（of），make difference，make sense，make clear 都是固定搭配，但也有 we'll 这些虚词，它们与“make”并不构成一般意义上的搭配，因此 T 值虽然克服了互信息值存在的缺陷，同时，本身也带来新的问题，即有些词汇间共现的几率高，导致 T 值增大，但实质是并不是固定搭配，这些共现词往往是代词、冠词、助动词等频率较高的虚词。

为了避免互信息值和 T 值出现的问题，研究者或是利用新的统计方法，如利用 Z 检验，从而获取词汇间的 Z 值；或是利用对数似然检验，获取 Log-Likelihood（LL）值；或是优化计算方法，如将上述的互信值中的词汇间共现的频率权重加大，于是产生平方互信息值 I2（MI2）和立方互信息值 I3（MI3）。必须指出的是，无论采用上述的任何一种统计方法或计算方法，得出的结果都存在一定的缺陷。在实践中，往往需要多种方法同时使用，共同考察词汇间的关系，才能得出较可靠的结论。这也是一些词汇分析与统计软件往往预置多种检验或计算词项间搭配力（强度）程序的原因。以牛津大学出版社使用的词汇分析软件 Wordsmith Tools 4.0 为例，此软件中使用六种方法来检验或计算词项间的黏合程度，分别是互信息值，立方互信息值及 Z 值、T 值、LL 值及 Dice 系数，分别用 MI、MI3 及 Z、T Score 和 Dice 表示（具体公式见软件附带的使用手册）。下图是利用此软件分析笔者建立的一文学语料库中“came”的搭配词结果。

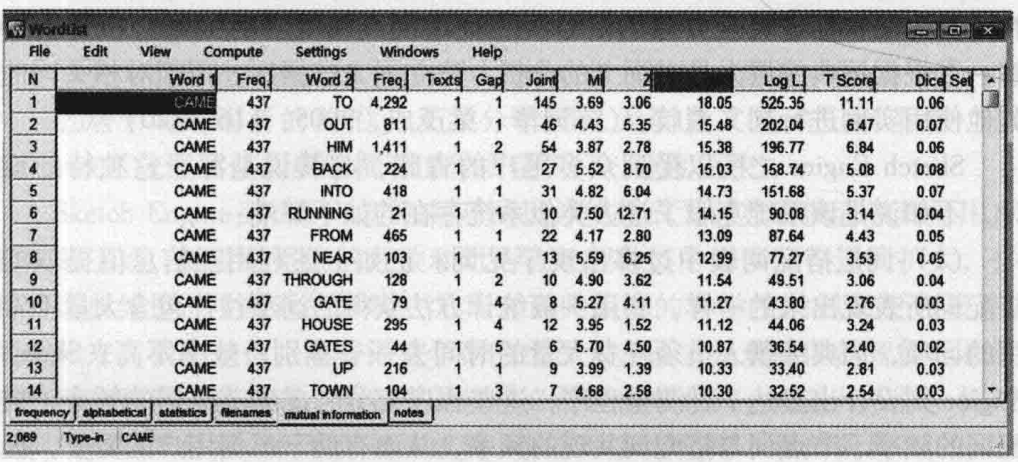

图 1.1　文学语料库中“came”搭配

综上所述，这一阶段的显著特点是引入了词汇统计学进入词典编撰的领

域，使得大量的词项的黏合力的计算变得简单化、科学化。同时，因为词项间的搭配是词典中词汇描写必不可少的一环，按辛克莱的观点，词汇间的搭配反映词汇的意义，不同搭配体现不同的意义，而词汇的义项分类及排列也是词典编撰人员必须考虑的。因此，通过计算词汇的搭配程度也有助于词汇义项的划分，从而可以提高词典编撰的效率。但是词汇的描写不仅仅是这些，还包括句法关系、同义词辨析及在不同语域的使用情况等等。随着这些问题的逐步解决，语料库词典学进入相对成熟的第四个阶段。

1.4 第四阶段：Sketch Engine

商用型词典编撰辅助系统 Sketch Engine（速描引擎）的投入使用，标志着语料库词典学进入第四阶段。该系统由 Lexical Computing Ltd. 研制，是基于网络的互动型平台，能够将大型语料库中的语言信息进行有效的汇总，充当语料库与词典编撰之间的接口。2002 年推出的《麦克米伦高价英语词典》，就采用了这一系统，该词典一经问世，便赢得“爱丁堡公爵英语联盟英语图书奖”和“英国文化委员会创新奖”，这两项大奖是 ELT 界最负盛名的奖项，其影响可见一斑。而且此系统已被牛津大学出版社、剑桥大学出版社、柯林斯公司等多家机构和用户使用，取得丰硕的经济效益。

Sketch Engine 是在对词语进行切分、词形还原、词性标注及句法分析的基础上对词项进行描写的，具体功能包括：①词语速描，即对特定词的搭配特征及语法特征进行全方位的汇总；②辅助词典编撰者对词的义项进行分解；③根据词典编撰人员对词义的分析，建立词义数据库，用于对相关词的其他使用实例进行词义消歧。（陈国华、梁茂成，2005：116 - 120）

Sketch Engine 之所以受到众多用户的青睐，与其说是源于它独特的设计，不如说是该系统克服了过去类似系统存在的如下弊端：

（1）词汇搭配词表中过多出现罕见词。正如上述利用互信息值提取的搭配词所表现出来的一样，利用一般统计方法获得的词项往往包含大量不常用的词项，词典编撰人员须耗费大量的时间去一一鉴别，效率不高；Sketch Engine 则设计出多达 7 种测量程序，这些程序充分考虑到节点词的频率、搭配词的频率、节点词与搭配词共现的频率，从而有助于根据用户的需要，提取适当的词项。如 log Dice 就可以帮助词典编撰者厘清词项的常用搭配词。词典编撰者根据搭配词进行聚类分析可确定节点词的义项，而义项的划分往往是词典编撰者最先考虑的因素，因此这个参数有助于解决义项划分问题。

（2）针对具体词形的分析。以往的软件或系统在分析时往往以一个具体词为对象，如针对 pigs 的分析；而 Sketch Engine 的分析范围更广，不仅包括具体词形，还包括这个词的词目，如可一次性分析 pig、pigs 两词，只要选择词目分析即可。这对词典编撰者来说是十分有用的。此外，还可以针对词性分析，因为英语中一个单词往往有多个词性，如 like，既是介词，又可以是动词，借助于词性分析，就可以解决词性问题。更重要的是，该系统设置了正则表达式检索分析，从而可以实施更加精确的分析，这也是一般系统所不具备的。

（3）节点词的语境呈现随意性。在语料库语言学中，分析节点词时，传统的方法是取其前后 4 到 5 个词的语境范围，这样靠近节点词的范围往往出现大量的语法词，远离节点词的范围则出现较多的实词，虽然从词性上可以区分搭配词，但噪音过大，让词典编撰无所适从；Sketch Engine 则充分利用词汇—语法模式，成功地解决了这个问题，从而丰富了词汇描写。

（4）静态的语法信息。以往的词典在描写词汇的语法信息时，名词标记可数还是不可数；动词标记及物还是不及物等，这些信息是静态的，而实际应用是动态的。Sketch Engine 通过建立数据库，可提供词项多达 27 种语法关系，极大地丰富了词汇的语法描写，提高了实用性。

由于采用全新的设计技术和理念，该系统的设计者认为他们的成就体现在以下几方面：（Kilgarriff & Koeling，2003：225 – 240）

①迈入最大的商业词典编撰领域，人类语言技术（尤其是词义排歧技术）研究领域，半商业半科研性质的机器翻译领域；

②史无前例地将基于语料库的词义排歧与人工输入融为一体；

③居于词典编撰的前沿，人工分析词义无疑是“艺术”而不是“科学”，该系统能够使词典编撰者工作质量更好、速度更快。

Sketch Engine 强大的功能和精巧的设计使其成为词典编撰者和其他用户青睐的工具，随着该系统的开放和对语言处理种类的增多及容量的扩大，必将对词典编撰产生深远的影响，尤其对词典编撰的科学性的提高更为显著。展望未来，我们坚信，随着语言学研究的深入，语言处理所需的智能化程度或自动化程度越来越高，语料库的规模必然也愈来愈大，愈具代表性。同时语料的标注软件和分析软件更加成熟，词典编撰必将迈入新的阶段。