

辞书研究文库
CISHU YANJIU WENKU

计算词典学

章宜华 著

JISUAN CIDIAN XUE

上海辞书出版社

辞书研究文库
CISHU YANJIU WENKU

计算词典学

章宜华 著

JISUÀNCÍDIǎNXUE



上海辞书出版社

图书在版编目(CIP)数据

计算词典学/章宜华著. —上海:上海辞书出版社,2013.2
(辞书研究文库)

ISBN 978-7-5326-3830-7

I. ①计… II. ①章… III. ①计算机技术—应用—词典学—研究 IV. ①H06-39

中国版本图书馆 CIP 数据核字(2012)第 297295 号

辞书研究文库

计算词典学

章宜华 著

上海世纪出版股份有限公司 出版、发行

上海辞书出版社
(上海市陕西北路 457 号 邮政编码 200040)

电话: 021-62472088

www.ewen.cc www.cishu.com.cn

上海展强印刷有限公司印刷

开本 890 毫米×1240 毫米 1/32 印张 15 插页 1 字数 389 000

2013 年 2 月第 1 版 2013 年 2 月第 1 次印刷

ISBN 978-7-5326-3830-7/H·544



定价: 48.00 元

如发生印刷、装订质量问题,读者可向工厂调换

联系电话: 021-66511611

目 录

“辞书研究文库”总序 曹先擢

前言 / 1

第一章 计算词典学概略 / 1

第一节 计算词典学的基础 / 1

一、计算词典学的源流 / 1

二、计算词典学的相关理论 / 5

三、计算词典学的相关方法 / 13

第二节 计算词典学的性质特征 / 16

一、计算词典学的理论特征 / 16

二、计算词典学的实践特征 / 20

三、计算词典学的学科交叉融合 / 22

第三节 计算词典学与机器词典 / 24

一、机器翻译的研究 / 24

二、机器词典的开发 / 30

三、机器词典与机读人用词典 / 34

第四节 计算词典学的国际背景 / 35

一、国际学术环境 / 35

二、与计算词典学相关的国际学术组织 / 41

三、语料和数据文本编码的国际标准 / 44

第二章 语料库与语料库词典学 / 46

第一节 语料库概说 / 46

第二节 语料库的基本特征 / 48

一、语料库的目的性 / 48

二、语料库语料的真实性 / 49

三、语料库语料的典型性 / 50

四、语料库语料的机读性 / 51

五、语料库语料的标准性 / 52

第三节 语料库的类型 / 56

一、按语言种类划分 / 56

二、按语言方式划分 / 57

三、按语料的载体划分 / 57

四、按语言的时域划分 / 58

五、按语言的用途划分 / 58

六、按语料的流通划分 / 59

七、按语料的分布划分 / 59

八、按语料的处理划分 / 60

第四节 语料库的基本功能 / 60

一、语料管理功能 / 61

二、语料索引功能 / 61

三、语料统计功能 / 62

四、语料标注功能 / 63

五、语音分析功能 / 65

第五节 语料库的建设与发展 / 66

一、语料库的历史背景 / 66

二、语料库的发展阶段 / 70

三、国内语料库的繁荣发展 / 76

四、语料库发展趋势 / 87

五、小结 / 90

第三章 电子词库的理论研究与实体构建 / 92

第一节 词库的理论概说 / 92

第二节 词库的相对关系 / 94

第三节 词库的基本内容和特色 / 96

第四节 主要词库的性能特征 / 100

一、词网 / 100

二、知网 / 109

第四章 词典语料的精加工与数据化 / 124

第一节 词汇属性速描 / 125

一、词汇属性速描的主要功能 / 126

二、词汇属性速描的描述方式 / 127

三、词汇速描的现状与问题 / 129

第二节 英语词汇数据库 / 130

一、英语词汇数据库的设计理念与方法 / 131

二、英语词汇数据库的容量与数据类型 / 132

三、英语词汇数据库的信息结构 / 133

四、DANTE 数据库的数据结构 / 137

第三节 语料库模式分析 / 143

一、语料库模式分析的理论构想 / 143

二、语料库模式的注释结构及构造 / 144

三、模式分析与模式表征 / 148

第四节 框架网——词汇数据库 / 152

一、框架网的设计思想 / 153

- 二、框架网的性质特点和构造 / 154
- 三、语义框架的结构及功能 / 156
- 四、框架元素的标注与属性的继承 / 160
- 五、框架的功能描述及其词典学意义 / 163

第五章 计算机技术与词典编纂的创新 / 169

第一节 概述 / 169

第二节 语料库在词典编纂中的应用 / 170

- 一、提供词典例证 / 173
- 二、辅助词典释义 / 173
- 三、辅助义项划分 / 173
- 四、辅助词典立目 / 174
- 五、辅助语用注释 / 174
- 六、提供句法信息 / 175
- 七、辅助语法注释 / 175
- 八、揭示文化信息 / 175
- 九、揭示同义词的分布与使用 / 176
- 十、词典自动生成 / 176

第三节 计算机辅助词典编纂平台 / 177

- 一、法国 IDM 的词典生成系统 DPS / 178
- 二、俄罗斯泰比词典编写系统 / 183
- 三、南非 TshwaneLex 词典编纂系统 / 188

第四节 计算机辅助编纂的新型词典 / 192

- 一、“牛津”系列词典 / 193
- 二、“朗文”系列词典 / 198
- 三、“柯林斯 COBUILD”系列词典 / 201

第六章 电子词典的类型与功能设置 / 207

第一节 电子词典概说 / 207

一、电子词典的发展简况 / 207

二、电子词典的定义 / 208

第二节 电子词典的类别 / 209

一、机器词典 / 209

二、微型芯片电子词典 / 213

三、光盘电子词典 / 217

四、网络电子词典 / 220

第三节 电子词典的主要功能 / 222

一、F1: 信息查询功能 / 222

二、F2: 媒体浏览功能 / 224

三、F3: 信息显示功能 / 225

四、F4: 发音功能 / 226

五、F5: 翻译功能 / 226

六、F6: 写作助理功能 / 227

七、F7: 汉化功能 / 228

八、F8: 系统设置功能 / 228

九、F9: 人机互动功能 / 230

十、F10—F15: 其他功能 / 231

第七章 电子词典数据库及其内部结构 / 233

第一节 电子词典数据库的特点 / 233

一、词典数据库的性质特征 / 234

二、词典数据库的理论基础 / 235

三、词典数据库的数据定义 / 237

第二节 电子词典的内部结构特点 / 239

- 一、一般电子词典的内部结构 / 239
- 二、机用电子词典的结构 / 244
- 第三节 电子词典词汇知识的表征形式 / 246
- 第四节 电子词典的知识信息 / 253
 - 一、词汇层次 / 254
 - 二、基本信息 / 255
 - 三、句法信息 / 265
 - 四、语义信息 / 270
- 第八章 电子词典编纂与制作的原则 / 281
 - 第一节 电子词典编纂的标准化原则 / 281
 - 一、国际词典编纂标准的背景 / 282
 - 二、国际词典编纂标准的特点 / 282
 - 三、国际词典编纂标准的内容 / 283
 - 四、国内词典编纂标准的状况 / 289
 - 第二节 电子词典制作的标准化原则 / 290
 - 一、标准通用标记语言 / 291
 - 二、可扩展标记语言 / 292
 - 三、可扩展样式表语言及其转换 / 293
 - 四、文件类型定义 / 294
 - 五、文件对象模型 / 295
 - 第三节 电子词典数据库的编码原则 / 297
 - 一、XML 的格式转换与文本编辑 / 297
 - 二、XML 格式的文本赋码方法 / 299
 - 三、LEXml 格式的文本赋码方法 / 305
 - 第四节 电子词典编纂的一般原则 / 311
 - 一、收词、立目与注音的原则 / 311

- 二、义项分立及释义的原则 / 313
- 三、括注和设例的原则 / 318
- 四、多媒体应用与附录设置的原则 / 320
- 五、综合、集合与孳生原则 / 323

第九章 光盘与芯片电子词典 / 325

第一节 牛津高级英语学习词典 / 327

- 一、主要功能特色 / 327
- 二、词典的界面结构 / 330
- 三、词典检索与查询 / 331
- 四、高级检索功能 / 332

第二节 朗文当代英语词典 / 334

- 一、主要功能特色 / 334
- 二、词典的界面与查询 / 338
- 三、高级检索功能 / 342

第三节 好易通电子词典 / 343

- 一、词典的基本配置 / 343
- 二、词典的功能特色 / 344
- 三、词典的查询功能 / 345
- 四、多媒体教学和扩展功能 / 346
- 五、牛津双解词典 / 346

第四节 诺亚舟电子词典 / 348

- 一、词典的基本配置 / 348
- 二、词典的功能特色 / 349
- 三、词典的界面与查询 / 350
- 四、新英汉词典的查询 / 352

第五节 卡西欧电子词典 / 353

- 一、词典的基本配置 / 354
- 二、词典的功能特色 / 354
- 三、词典的界面与查询 / 355
- 四、多媒体教学功能 / 356
- 五、新牛津英语词典 / 357

第十章 网络电子词典 / 359

第一节 Onelook[®] 词典网站 / 361

- 一、概述 / 361
- 二、主界面与主要功能 / 362
- 三、网站词典的类别 / 363
- 四、检索系统 / 368

第二节 在线英法品牌词典 / 373

- 一、在线牛津词典 / 374
- 二、在线牛津语言词典 / 379
- 三、在线拉鲁斯词典 / 383

第三节 在线品牌英语学习词典 / 391

- 一、在线牛津英语学习词典 / 391
- 二、在线朗文英语学习词典 / 395
- 三、在线剑桥英语学习词典 / 398

第四节 特大型在线词典 / 403

- 一、OEDO 的主要特性 / 403
- 二、OEDO 的界面及检索功能 / 404
- 三、OED 的修订特点 / 406
- 四、OED 类词典的发展趋势 / 407

第五节 维基百科全书 / 408

- 一、维基的源流与特点 / 408
- 二、维基百科 / 409

- 三、中文维基百科 / 410
- 四、维基百科的词条结构 / 412
- 五、大众百科 / 415

第十一章 电子词典的现状与发展趋势 / 417

第一节 电子词典的现状与问题 / 417

- 一、电子词典的基本现状 / 417
- 二、电子词典发展的策略问题 / 418
- 三、电子词典的技术策略问题 / 419
- 四、对问题的思考 / 420

第二节 问题的对策与发展策略 / 422

- 一、提高认识 加强合作 / 422
- 二、明确宗旨 突出特色 / 423
- 三、着眼未来 立足当前 / 423

第三节 电子词典的发展趋势 / 426

- 一、语言处理形式化 / 427
- 二、词典编纂自动化 / 427
- 三、小型词典情景化 / 428
- 四、专科词典集合化 / 428
- 五、大型词典综合化 / 429
- 六、综合词典网络化 / 429
- 七、常用词典芯片化 / 430
- 八、词典载体多元化 / 430
- 九、信息表述多媒体化 / 431
- 十、界面结构模块化 / 431
- 十一、界面窗口多样化 / 432
- 十二、窗口结构弹性化 / 432
- 十三、词典查询智能化 / 433

十四、信息显示任意化 / 433

十五、信息交流互动化 / 433

第四节 电子词典界面的结构模块 / 434

第五节 小结 / 440

主要参考文献 / 442

第一章

计算词典学概略

计算词典学是词典学与计算机技术结合的产物。尽管词典编纂有几千年的悠久历史,但它与计算机的结合直到 20 世纪 70—80 年代才初露端倪。

从 1946 年第一台电子计算机 ENIAC 诞生到 20 世纪 70 年代初期,计算机经历了从电子管、晶体管到集成电路和大规模集成电路的演变和发展;尽管技术不断更新,但其体积仍十分庞大、处理语言的能力极低,在词典语料库和词典编纂中的应用仅处于探索阶段,难以提供实质性的支持。到了 20 世纪 70 年代末期至 80 年代初期,计算机硬件和软件都有了飞跃发展,个人电脑出现并开始普及,信息处理和传播速度的提升加快,计算机辅助词典编纂有了实质性突破,有关应用方法和理论框架逐渐形成;自此计算词典学的研究开始走向繁荣时期,研究成果不断推出。历史悠久的词典乘上了信息化的快车,计算词典学的发展直接推动了辞书研究和编纂现代化的进程。正如李宇明(2006)所说:“信息时代的重要特征是:数字技术成为处理信息的主要手段,数据库成为信息的重要载体。辞书编纂手段现代化,其现实含义就是在辞书整个编纂过程中充分使用数字技术和各种数据库,具体说,就是开发功能强大的语料库、知识库和适用于辞书编纂的软件系统。”

第一节 计算词典学的基础

一、计算词典学的源流

人类有史可考的文明有五千多年了,词典编纂史若从词集开始算

起已有四千多年^[1]。我国字典和词典编纂史也有两千多年:《尔雅》是我国第一部按照词义系统和主题分类编纂的词典,而东汉许慎的《说文解字》则是我国最早的一部字典。长期以来,词典编纂的材料是以书页注释或卡片的形式积累的,而词典也都是以纸张为载体,蝇头小字黑鸦鸦一片,仅用排版形式、字体和元语言符号来区别词条中的义项和各类释义信息,词典查阅十分不便。而词典编纂更是一项十分浩繁的工作,词典编纂者要度过许多个寂寞的春秋,牺牲无数个夜晚和假日,成天被围困在书山字海之中,埋头搜集大量枯燥琐碎的语料——不停地去剪报、整理资料、摘录卡片,并对这些卡片进行分类、编排等。卡片积累多了,就成了一个“库”。一部四五百万字的中型词典需要上百万张卡片,可以堆满好几间库房。试想,上百万张卡片要一张张手工制作,数万例句要一句句抄录,数万条词目要一条条编写、誊写、校对、再誊写,加上全面的体例分析和综合文字处理,工作的繁杂和辛苦是可以想象的;特别是这些卡片语料的编排、存贮和查找,更是一件浩繁的工作。到20世纪中后期,随着全球经济建设的恢复、科学技术的快速发展,新事物、新语词、新术语不断增多,如何迅速搜集和有效管理这些与日俱增的新语词和新术语,成为一个异常棘手的问题。显然,手工制作卡片的传统方式已经远远满足不了科技发展的要求。这促使人们去思考,去寻求新的出路、新的途径——建立现代化的语料库,并用现代化手段进行词典编纂。

在1959年,英国伦敦大学学院英语语言文学系的 Quirk 带领10多名合作者开始建立容量为100万词的英国书面英语和口语语料库——“英语用法总览”(The Survey of English Usage)。其中,口语和书面语各占一半,口语语料记录在盘式录音带上,然后转写成书面文字,并以卡片的形式对这些语料进行了语调、语法和用法方面的详细标注。后来,这些语料的口语部分(51.06万词)全部电子化,由于电子化的合作者来自瑞典的隆德大学,后改名为“伦敦-隆德英语口语语料库

(*London-Lund Corpus of Spoken English*)”，近年来还在原标注的基础上增加了词类标注，成为研究英语口语使用最广的重要资源。

1961年至1964年间，美国布朗大学(Brown University)的 Nelson Francis 和 Henry Kucera 建立了一个具代表性的百万词计算机语料库。该语料库包含各种不同的文体，根据统计分析确定语料的文体平衡分布，根据这个分布搜集了100万词的语料，并进行了词类标注。与此同时，在卢森堡建立了世界上第一个服务于翻译的多语种机读术语库，包括法、德、意、英、荷、丹麦等6种语言，收词大约40余万条，是一个存储量大、检索方便的计算机术语数据库，大大地提高了术语翻译或处理的效率。

由于布朗语料库的电子化早于“伦敦-隆德英语口语语料库”，因此被称为第一个百万级的“大型”计算机语料库。它的建立标志着语料库语言学这门新兴学科的诞生。尽管布朗语料库的语料平衡和标注方式显得有些粗糙，但它创立了语料库建设的原则、方法和构架，因此一直被视为英语平衡语料库的标准和基础，后来出现的语料库基本都按照这个构架来建设。从布朗语料库开始，用于语言研究、翻译和词典编纂的现代大型语料库的开发与建设便拉开了序幕。

事实上，词典最直接与计算机结合，应始于20世纪40年代，始于机器翻译兴起之时(参见本章第三节)。可以说，词典与自然语言处理及机器翻译系统的研究有着不解之缘，因为翻译系统的核心之一就是双语或多语电子词典，也叫作机器词典(machine-translation dictionary)。这种词典在40年代开始开发，在80年代末期至90年代初期得到快速发展，其中部分词典被改造成供人查阅的电子词典，大量进入市场，使词典实现了从印刷本转向电子版的变革。从某种程度上讲，现代的电子词典起源于对机器词典的研究和开发。

迄今，词典与计算机等现代化技术手段的“联姻”已经历了半个世纪的磨合，相关理论的研究和实践也有三十多年的历史，无论是承载词

典信息的介质,还是知识内容的组织方式,都发生了巨大的变化。21世纪迎来了新经济时代的曙光,信息技术的发展更是一日千里,计算机网络的普及拉近了人与人的距离,一个网络环境下的数字地球村已逐渐形成。为了满足人们对语言信息与日俱增的需求,词典的知识信息表述形式和查阅方式都亟待变革,数字化多媒体词典应运而生。

集音频、视频和文本材料于一体的多媒体词典的诞生是当代词典学的重要标志之一。多媒体技术将词典的知识信息与计算机语言处理的现代化手段(如用计算机搜集、分析与处理词典语料)完美地结合在一起,词典编纂也开始实现计算机辅助运作,词典信息的承载形式以及传输和查阅方式实现了数字化。现代化处理手段和多媒体形式只是词典发生变化的外部特征,更重要的是其内部特征——词典编纂者对自然语言知识的认知及其在词典中的表述方法,都在发生着变化。

当代词汇学和词典学重要的理念之一表现为对心理词库(Mental Lexicon)本质特征的认识和描写。词汇学探讨词的起源和发展规律,研究词的构造、构成及规范,而词典则是关于词汇语言事实的记录或描写,与词汇学及语言学有着十分密切的关系。语言学的发展必然推动着词典学的发展,反之亦然。于是,词典学与语言学(尤其是与当代语言学)的接口问题始终撞击着各自的传统范式,不断迸发出新的思想火花,衍生出新的研究课题。传统语言学(包括词典学)首先受到质疑的,就是词汇与语法之间的人为界限问题。一般来讲,词汇和语法本是浑然一体的,语法的研究离不开词汇,词汇的研究也会涉及语法。而主流的传统语言学(结构主义语言学和生成语言学)对语法机制的研究则从句法入手,把词汇看作是无足轻重的成分。词汇论(Lexicalism)和功能语法则认为,语法和词汇处于同一个连续体,从词汇入手同样可以探求语法机制,甚至更方便、更直接——因为词汇是最精密的语法(lexis as most delicate grammar),说话人在语言系统网络中是以词汇为对象进行选择的(Halliday & Hasan 1976: 281; Halliday 1994: 15)。可见,所