辞书研究丛书
CISHU YANJIU CONGSHU

辞书研究与辞书发展论集
（第二辑）

王铁琨 李清山 亢世勇 主编

CISHU YANJIU YU
CISHU FAZHAN LUNJI

上海辞书出版社

辞书研究丛书

CISHU YANJIU CONGSHU

辞书研究与辞书发展论集

（第二辑）

王铁琨 李清山 亢世勇 主编

CISHU YANJIU YU
CISHU FAZHAN LUNJI

上海辞书出版社

图书在版编目(CIP)数据

辞书研究与辞书发展论集. 第二辑 / 王铁琨，李清山，亢世勇主编. —上海：上海辞书出版社，2014.11
（辞书研究丛书）
ISBN 978-7-5326-4241-0

Ⅰ.①辞… Ⅱ.①王… ②李… ③亢… Ⅲ.①汉语—辞书学—文集 Ⅳ.①H16-53

中国版本图书馆 CIP 数据核字(2014)第 163896 号

责任编辑 李潇潇
封面设计 姜 明

辞书研究与辞书发展论集(第二辑)
王铁琨 李清山 亢世勇 主编
上海世纪出版股份有限公司
上海辞书出版社 出版、发行
（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 9.5 插页 1 字数 243 000
2014 年 11 月第 1 版 2014 年 11 月第 1 次印刷
ISBN 978-7-5326-4241-0/H·592
定价：35.00 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

国家语委科研项目成果

我们对辞书的“五个关注”*

（代序）

尊敬的各位专家、学者，各位来宾：

大家上午好！

2007年教育部语言文字信息管理司和鲁东大学共同建立了“汉语辞书研究中心”，这是从国家语言文字事业和国家辞书研究领域的发展需要而做的战略性布局。几年来，“汉语辞书研究中心”在辞书领域的科学研究、人才培养和社会服务等方面取得了很多成果。昨天下午，辞书研究中心承担的国家语委的5项科研项目做了结题评审，辞书界的几位专家对辞书中心的研究成果给予了充分的肯定。这也说明辞书研究中心的工作方向正确，研究认真，搜集挖掘资料广泛，成果显现，我代表教育部语言文字信息管理司、代表国家语委对中心研究工作所取得的成效表示祝贺！同时也希望中心按照评审专家的意见建议精益求精，丰富充实研究成果。借此机会，也向评审专家以及语言文字工作的前辈、同行表示感谢！你们是国家语言文字工作的中流砥柱，尤其是老前辈老专家是国家语言文字工作的宝贵财富，你们对事业认真负责、不懈追求，白发苍苍仍在竭力奉献，值得学习。我深感敬佩！

今天，由“汉语辞书研究中心”主办的第三届汉语辞书高层论坛在鲁东大学召开，我谨代表教育部语言文字信息管理司对论坛的举办表示热烈的祝贺！对各位专家、学者表示热烈的欢迎！

---

* 本文原为教育部语信司司长张浩明在第三届汉语辞书高层论坛（2012年8月1—2日）开幕式上的讲话。

辞书是劳动智慧的结晶，是人类语言文字的归纳和总结，辞书不仅是人们在文明进步历程中答疑解惑的工具，同时又通过引导语言的应用、使用而具有语言教育的作用，通过帮助读者选择知识、过滤信息而具有知识教育的作用。

当前，汉语辞书的编纂、出版和研究蓬勃发展，日新月异，辞书产业正在快速成长。我们需要高度关注汉语辞书在国际经济发展、社会发展中发挥的重要作用，我认为在辞书研究和编纂中，要注意以下五个方面的关系或者是问题。

一、辞书与语言文字规范、社会语言生活的关系

语言文字规范同社会语言生活息息相关，而辞书是人对优秀知识的汇集，是社会的无冕教师，具有规范体系的其他成员代替不了的功能。辞书不但是落实贯彻国家制定的语言文字规范的最重要的途径，而且在规范覆盖不到的地方，可以通过专家权威释义等方式起到引导语言文字准确使用的作用，为民众语言交际、为各行业和各语言领域的专业语用提供规范和权威的答案。辞书与国家的语言文字规范互补，构成了一个全面的、带有一定引导性的广义的语言文字规范体系。因此，在辞书的研究编纂中，需要关注包括辞书如何调和语言文字规范中科学性与社会性之间的矛盾、辞书如何服务于社会语言生活等问题。

二、辞书与文化建设的关系

语言是文化的载体，辞书是语言文字的资源库，辞书收录的词语是文化的符号，因此也是文化资源的重要类型。党的十七届六中全会提出了建设社会主义文化强国的目标，突出了文化在建设和谐社会中的作用。中国拥有丰富的、渊源深厚的特色文化，这些文化都离不开语言文字载体，而辞书是收录、传承并辅助用户阅读、研习这些文化信息的最重要的工具。如何发挥辞书在我国文化建设、传承、传播中的作用，

这也是一个重要的研究课题。

三、辞书与在教育领域推广语言文字规范的作用

教育领域特别是基础教育领域是推广语言文字规范的重要阵地，语文辞书和语文教科书等一道通过引导的方式与规范文献相辅相成，在教育领域推广语言文字规范。我们需要特别关注辞书在教育领域中辅助推广语言文字规范作用的发挥。在这个方面，不但要重视规范性语文辞书、描写性语文辞书的编纂，还要高度重视学习性语文辞书的建设。目前汉语学习词典主要包括面向“汉语作为第二语言的教学”的外向型词典、面向母语教学的内向型学习词典两大类，其编纂和推广已经成为辞书界的新兴热点。学习词典是纯语文性的，是我们关注的重点，要针对不同教育对象编纂不同层次、不同特色的学习词典。另外，我们还需要有更多关于学习词典在有效推广语言文字规范方面的研究。

四、语言经济在辞书产业发展中的作用

辞书作为一种文化产业，其发展必须充分考虑文化产业的运作原理和机制，这方面西方的“六大词典家族”的以用户为中心、以市场为导向的运作机制带给我们很多的启发。我们要在借鉴的基础上，结合汉语汉字本体特点、汉语辞书用户特点，探索出一条有自己特色的辞书产业发展之路。这需要把辞书编纂者、出版者、管理者、研究者、教育者、用户等因素整合在一起来考虑。辞书产业的核心是语言文字内容，其本质是语言产业、语言经济，辞书只是载体和运作形式。因此，在整合过程中，需要关注“语言经济”在辞书产业中的作用，探索如何在辞书产业的发展过程中发挥好语言经济理论的引领作用。

五、要高度关注辞书理论的创新

理念演绎辞书，理念引领辞书，理念提升辞书。无论是面向哪一类

读者群的辞书都必须在一定的理论的基础上进行编纂，没有理论基础的辞书编纂就好比没有了设计图的建筑，因此一部辞书或是一系列辞书的编纂必须理念先行。我们一定要借鉴吸收西方发达国家的现代辞书编纂理念，结合我们的实际，及时吸收学术界的新理论，提出汉语辞书编纂的原创理论，指导辞书的编纂实践。要充分利用辞书编纂的现代化手段和电子辞书载体，利用语料库、知识库以及各种软件编纂、开发适合不同人群需要的汉语辞书，扩大汉语辞书在国际市场上的份额，推动辞书产业的发展。

我国虽然有古老而优秀的辞书传统，但在二十世纪上半叶却还只能算个辞书小国。今天，我国已经由辞书小国成长为辞书大国，但还不能称为辞书强国，要实现“辞书强国梦”，需要我们继续共同努力。“汉语辞书研究中心”的成绩是教育部语言文字信息管理司与鲁东大学一道为实现这一梦想搭建的一个基础工程。我们将会以更大的关心和支持来扶持中心的壮大与发展，也希望中心能练好内功，努力发展壮大，增强自己的实力。五年来，“汉语辞书研究中心”勇挑重担，加大了基础条件建设，发展有特色的研究方向，通过专、兼职结合，内外联合的方式在全国范围内组织了一支富有创新精神和创新能力的研究团队，承担了各类高层次项目30多项，取得了一批较好的研究成果，开发了一批辞书资源，编纂出版了一批高质量的辞书，收到了较好的经济效益和社会效益，为我国的辞书事业发展做出了积极的贡献。我希望汉语辞书研究中心能够进一步凝聚校内校外的力量，凝聚语言文字工作者、研究者的力量，以专业创新的理念和知识来推动辞书的发展，我相信语言文字工作在辞书产业中将大有可为，并期待着汉语辞书新的发展时期的到来！

最后，预祝第三届汉语辞书高层论坛圆满成功！

谢谢大家！

张浩明

2012年8月1日

目　录

词典学研究中的一门新兴学科——计算术语学

冯志伟

（杭州师范大学外国语学院）

近年来，在词典学和术语学的研究中，开始引进自然语言的计算机处理的方法和技术，出现了“计算术语学”（computational terminology）这样的学科。1998年的计算语言学国际会议COLING-ACL'98上，组织了世界上第一次计算术语学的讨论会（First Workshop on Computational Terminology），这次讨论会首次使用了“计算术语学”这个学科名称，并讨论了如下问题：

- 如何抽取术语以满足信息检索的需要；
- 如何抽取术语以便使用双语语料库来进行翻译；
- 如何进一步完善原有术语抽取的工作（例如，如何建立概念层级网络，如何搜索语义信息或概念信息）。

这次讨论会成为了计算术语学发展的催化剂，从此，计算术语学便成为一门新兴的术语学学科，活跃在当代科学技术的百花园中，并且一天天地成熟起来，初步具备了系统的理论和有效的方法，值得我们特别关注。

在“计算术语学”这个名称出现十年之前，笔者在1988年就注意到术语的自动处理问题，并在德国夫琅禾费研究院（Fraunhofer

Institute）使用计算机对汉语的词组型术语进行了自动结构分析，是国际上最早进行计算术语学研究的学者之一。计算术语学的研究主要包括术语结构的自动剖析、术语的自动发现、术语的自动标引等。本文主要介绍术语的自动发现和自动标引。

在自然语言的计算机处理的诸多领域中，都离不开术语，例如，机器翻译（machine translation）目前主要是翻译专业性的文献，术语的自动处理与机器翻译系统的译文质量有密切的关系；此外，信息检索（information retrieval）、信息抽取（information extraction）、文本分类（text classification）的运算的基本单位都是单词型术语或词组型术语，同时也离不开术语的自动处理。

术语是自然语言处理中一种特殊的词汇数据，与语言中一般的普通词汇不同，术语大多数都是由多个单词组成的词组型术语，它们对于科学技术的发展特别敏感，时时刻刻随着科学技术的发展而发展。在术语的发展过程中，它们不断地丰富，不断地充实，不断地变化，其语义也在不断地转移，旧的术语消失了，新的术语产生了。在这样的情况下，术语数据库需要经常地维护，不断地用新的术语充实原来的内容，有时甚至需要重建，以反映科学技术日新月异发展的要求。这样，术语的发现（term detection）或术语的获取（term acquisition）就成为了术语自动处理的一个重要内容。术语发现可以进一步分成两个类型：如果在术语发现中不依赖初始的术语数据，那么，这样的术语发现叫做“初始术语发现”（initial term acquisition）；如果在术语发现中要使用初始的术语数据，那么，这样的术语发现叫做“原有术语充实”（term enrichment）。

在文本自动处理中，术语的使用与术语的自动辨识（term recognition）是紧密联系在一起的。术语的自动辨识主要研究如何进行术语的自动标引（automatic indexing）。在自然语言处理中，为了便于信息的存取，文本文献总是要使用单词表或词组表，因此，有必要在

文本文献中进行术语的自动标引(automatic indexing of terms),然后根据自动标引的结果,使用计算机来自动地生成单词型术语表或词组型术语表。由于术语是科学技术知识在自然语言中的结晶,它能够浓缩地表示特定的科学技术领域中的主要概念,可以被看成是文本内容的抽象描述,文本文献经过术语的自动标引之后,就能大体上反映其内容。因此,在文本自动处理中,术语的自动标引是非常重要的。

根据在标引时是否依赖初始的术语数据,术语的自动标引也可以分为两个类型:如果在术语标引中不依赖初始的术语数据,那么,这样的术语标引叫做“自由标引”(free indexing);如果在术语标引中要使用初始的术语数据作为参照,那么,这样的术语标引叫做“受控标引”(controlled indexing)。总的来说,术语自动处理可以这样来分类,如表1所示:

表1　术语自动处理的四个主要领域

	不依赖于初始术语数据	依赖于初始术语数据
术语发现	初始术语发现	原有术语充实
术语辨识	自由标引	受控标引

下面我们介绍国外的术语发现和术语标引研究情况(Christian 2001)。

首先介绍“术语发现”的研究。发现候选术语的方法基本上分为符号法(symbolic approach)和统计法(statistical approach)两种。符号法根据术语(主要是名词词组)的句法描述来发现候选术语;统计法根据词组型术语中组成成分的互信息(mutual information)来发现术语,组成成分之间的互信息越大,组成术语的可能性也就越大。

(1)基于语法的术语发现方法:例如,在1994年,Lauriston (1995:147-170)在TERMINO系统中提出了一种基于语法的术语发现方法,这种方法要对文本进行剖析,利用文本中的单词和句法

线索（lexical and syntactic clues）来发现术语。剖析模型的操作顺序如下：

a. 预处理：首先对文本进行过滤，除去对术语发现无用的那些形式特征（虚词，停用词）；

b. 剖析并抽取术语：

■ 形态分析

■ 名词短语剖析

■ 术语生成

c. 交互式术语数据库的构建和管理：给用户提供友好的界面，用前面步骤中抽取出来的术语构建术语数据库。

（2）句法模式与机器学习到的选择限制相结合的方法：例如，D. Bourigault（1996：771-779）研制的术语自动处理工具LEXTER。LEXTER使用带标记的语料库，语料库中的标记有词汇特征的标记和句法模式的标记两种，这个工具有一个可视化界面，可用来确认并组织从带标记的语料库中抽取出来的术语。

a. 最大名词短语的分离：LEXTER可使用分离规则，从最大名词短语（maximal noun phrase）中把可能性最大的术语边界分离出来。例如，在法语的最大名词短语中，过去分词与介词结合而成的组合很可能是术语的边界，在法语最大名词短语 les clapets situés sur les tubes d'alimentation（位于进气管上的阀门）中，situés sur 是术语的边界，把整个名词短语分离为 les clapets（阀门）和 les tubes d'alimentation（进气管）两部分，这两部分分别是两个不同的术语。其中，“situés sur” 是句法模式，这个模式的使用取决于动词的选择限制，而动词的选择限制是通过内置的机器学习程序从语料库中自动学习得到的。

b. 把最大名词短语分解成候选术语：确定边界之后，最大名词短语被分离为两个部分，通过后处理，最后由人来判定这些候选

术语，并把确认后的术语加入到术语数据库中。例如，从最大名词短语 les clapets situés sur les tubes d'alimentation 中，把术语 les clapets 和术语 les tubes d'alimentation 自动抽取出来，作为候选术语，加入到术语数据库中。又如，在法语中，pylône à haute tension（高压电线架）的结构是：N + Prep + N + Adj，经过最大名词短语分离之后，把 haute tension（高压电）作为候选术语提取出来，加入到术语数据库中。

c. 最后，还可以根据这些候选术语在句法位置上的相似程度，把它们组织起来。例如，法语中的 vanne motorisés（电动门）、vanne pneumatique（气动门）、vanne d'alimentation（进气门）都有共同的中心词 vanne，就把它们组织起来，形成一组有关系的候选术语。

d. 这些进入术语数据库的候选术语，由专家做最后的审定，确定为正式的术语，充实了原有的术语。

（3）句法模式与统计过滤相结合的方法：例如，Daille（1996：49-66）研制的 ACABIT 是一个把句法模式与统计过滤结合起来的术语研究工具。ACABIT 获取候选术语的步骤如下：

a. 语言规则过滤（linguistic filtering）：根据术语结构的语言学规则，使用有限状态转移网络发现候选术语，在英语中，主要考虑三种模式的术语：Adj + N，N + N，N + Prep + N。由这三种模式扩展而形成的变体，也可以作为候选术语的筛选范围。例如，satellite transit network（N + N + N）可以看成是由 N + N 模式扩展而成的，multiple satellite links（Adj + N + N）可以看成是由 Adj + N 模式和 N + N 模式扩展而成的。

b. 统计排序（statistical ranking）：使用某些统计方法，对通过前面的步骤筛选出来的候选术语进行排序。例如，计算候选术语的“对数似然度”（log-likehood ratio），根据计算结果对候选术语

排序，得出在统计意义上可能性最大的术语。

（4）抽取搭配信息的方法：例如，Smadja（1993：143-177）研制的 Xtract 是一个专门用于抽取搭配关系的工具。Xtract 的重点不是关心术语本身，而是关心术语在意义上的可搭配性。只有那些在语义上可以搭配的词语才可以算做候选术语（例如，stock trader［存货商人］，last selloff［最后的存货］在语义上是可以搭配的）。

（5）非语言学的方法：例如，Enguehard & Pantera（1993）研制的术语提取工具 ANA。ANA 是独立于具体语言的术语自动抽取工具，它包括两个模块：

a. 预熟悉模块（familiarization module）：使用预熟悉模块来确定三类词语：

■ 停用词语表（stop list）：停用词通常是一些频度很高的词语，这些词语都不具有专业性。

■ 种子术语表（set of seed terms）：使用人工从语料库中选出反映专业概念的术语作为种子术语（seed term），构成种子术语表。

■ 结构词语表（set of scheme words）：这些结构词语一般是介词或限定词之类的虚词，它们在语料库中往往与种子术语一起出现。

b. 发现模块（discovery module）：使用机器自动学习中的“自举”（bootstrap）方法，一步一步地扩充从预熟悉模块中得到的种子术语的规模，从而发现更多的术语。

在用于术语发现的这五种方法中，前两种方法（TERMINO，LEXTER）不使用统计，假定文本中符合条件的全部词语都是候选术语，哪怕只出现一次的“罕用词语”（hapax legomenon），只要它们符合条件，也都在候选术语的考虑范围之内。这两种方法是非统计的方法。

使用这样的非统计方法时，术语的判定要由用户来进行，需要给用户提供交互工具，以便用户对候选术语进行选择。后面三种方法都要使用统计来进行过滤或排序，在这样的情况下，考虑候选术语出现的上下文环境就显得非常重要了，因为统计的数据需要在具体的文本或语料库中才可以计算出来，离开了具体的文本或语料库，不可能进行任何的统计，当然也就不可能发现术语了。

主题表（Thesaurus）是一种控制词汇的方式，它通过收集特定学科领域的词汇，并以特定的结构排列，以显示词汇之间的关系。主题表的编制主要包括准备工作、词汇选择、词汇整理、词汇分类、建立词间关系等步骤。传统主题表的构建主要由领域内专家手工完成，耗时较长，无法保证完全覆盖，也无法有效地进行自动更新，对于新词、组合词、外来词等的接收较慢。

最近，我们与中国科学院计算技术研究所合作，从维基百科中自动地获取关于“太空”领域的术语，得到了较好的效果。[1]

我们从维基百科上下载了66篇文档，随机按照比例10:1的文档集合，分别作为训练语料（航空航天语料wiki_Train）和测试语料（航空航天语料wiki_Test）。然后进行数据的清理，包括去掉html标签、生词位置标签等。然后，调用分词（work segmentation）程序，对净化后的文档进行分词，形成单个形符（token），然后对这些形符进行自动特征提取；最后对每一行特征向量进行人工的标注，形成训练集与测试集。原始的形符共有123150个，我们从这些形符中提取术语。

维基百科是一个由全社会参与、并且具有多种语言的百科全书协助计划，其目的是建立一个完整正确的百科全书。截止到2012年2月底其收录的英文词条超过386万，收录的中文词条接近40万。维基百科包含社会、经济、文化、教育、科技诸多领域的知识，由具有一定相关领域知识的社区积极分子进行维护与更新。因此，Web2.0带动下的全民织网，进行特定领域下的主题词抽取具有潜在的意义。