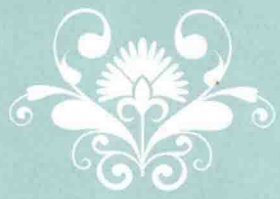
现代汉语未登录词
词类和语义类标注研究

邱立坤 著



科学出版社

现代汉语未登录词
词类和语义类标注研究

邱立坤 著

鲁东大学学科建设专项经费资助
国家自然科学基金（编号：61572245）资助

科学出版社
北京

内 容 简 介

本书是语言学与计算机科学相结合的产物。作者不仅用语言学理论来指导计算机算法的设计，而且通过计算机算法的实验结果反过来验证并丰富语言学的理论。在大量统计、算法的基础上，提出与目前主流的分布词类观相反的论点，并用实验数据证明：在自动判断汉语新词语类别时词语内部结构特征比上下文分布特征更有效，进而提出内外结合原则，即判断新词语类别时应同时使用内部结构特征和外部上下文特征。基于这一原则，设计了相应的算法，实验结果表明这些算法要明显优于单独使用内部特征或者外部特征的算法。

本书适用于语言学领域及计算语言学、语料库语言学领域的师生和研究人员。

图书在版编目(CIP)数据

现代汉语未登录词词类和语义类标注研究 / 邱立坤著. —北京：科学出版社，2016

ISBN 978-7-03-049180-0

I. ①现… II. ①邱… III. ①现代汉语-词类-研究 ②现代汉语-语义-研究 IV. ①H136 ②H146.2

中国版本图书馆 CIP 数据核字(2016)第 146881 号

责任编辑：石 悦 赵微微 / 责任校对：李 影
责任印制：张 伟 / 封面设计：华路天然工作室

科学出版社 出版
北京东黄城根北街 16 号
邮政编码: 100717
http://www. sciencep. com

北京京华虎彩印刷有限公司印刷

科学出版社发行 各地新华书店经销

*

2016 年 6 月第 一 版 开本：720×1000 B5
2016 年 6 月第一次印刷 印张：11 1/4
字数：230 000

定价：66.00 元

(如有印装质量问题，我社负责调换)

语素：汉语语法规则单位的重心（代序）

2000年我给北京大学中文系本科四年级学生开“理论语言学”课，邱立坤是班上最喜欢提问题和争论问题的同学之一。他也是一位痴迷语言学的年轻人，每次和同学或老师谈起语言学，眼睛就开始发亮。那时候我们经常讨论句法结构关系、语类问题和句法的初始概念问题。我一直指导他的本科学位论文、硕士学位论文和博士学位论文，句法理论的基本问题一直伴随着我们。学习期间，立坤还选修了大量计算机课程，人和机器的关系，自然语言理解的概率模型和规则模型也是我们讨论的重点。这本书是立坤在博士论文基础上扩展而成的，要解决的问题是未登录词的语类标注，包括语法的和语义的。这项工作的意义先得从单位和规则说起。

按照结构语言学的理论，先要确定词，再确定词类。词是最小的自由形式，比如“该校、该系、该所”等分别都是词。确定了词以后，再根据分布确定词类。“该校”这些词都分布在通常称为名词的环境中：

该校有问题，需要对该校进行调查
该系有问题，需要对该系进行调查
该所有问题，需要对该所进行调查

至于“该校、该系、该所”的构词语素“该、校、系、所”，尽管是最小的，但不自由，不是词，因此无法根据自由分布的理论对这些语素进行语法分类，也无法根据这些语素的分布信息来确定“该校、该系、该所”这些词的语类。概括地说，结构语言学确定“该校”的语类必须通过“该校”的分布。

但是，“该校、该系、该所”这些词的词类从内部构造看也有相当的规律，即在这样一种平行的条件下形成的很多组合，都具有名词的性质：

	左项语素的分布性质	右项语素的分布性质	语素组的分布性质
该校	?	?	N
该系	?	?	N
该所	?	?	N
该院	?	?	N
该厂	?	?	N

续表

	左项语素的分布性质	右项语素的分布性质	语素组的分布性质
该区	?	?	N
……	……	……	……

由于左项和右项都不自由，左项和右项的分布也不确定，但组合的语类则是确定的。在汉语中，通常所说的复合词与此类似，内部往往是有规则的，复合词的语类也往往可以通过规则推导出来，而不必依赖语境分布条件。事实上，“该 X”如此有规则，在很多人心目中甚至往往并不作为复合词看待，但两个成分都不自由，又不能当作词组。这是结构语言学分布理论蕴含的一种矛盾。

下面的实例通常是作为复合词看待的：

	左项语素的分布性质	右项语素的分布性质	语素组的分布性质
樟树	?	N	N
棕树	?	N	N
柞树	?	N	N
桦树	?	N	N
枞树	?	N	N
榆树	?	N	N
……	……	……	……

但都是有规则的组合，语类可推导。

汉语中大量复合词都是语素有规则的组合，这可以为徐通锵字本位理论提供一定的依据。字本位看到了语素字活动的规律性，起初希望把语法单位建立在字（语素字）上，但是，有些语素组是没有规则的组合，其语类无法预测。比较：

	左项语素的分布性质	右项语素的分布性质	语素组的分布性质
远视	A	?	N
近视	A	?	N/A
斜视	A	?	N/V
重视	A	?	V
轻视	A	?	V
弱视	A	?	N

可见，要把语法单位都建立在字或语素一种单位上是有困难的。徐通锵后来的字本位除了字，还有辞，辞相当于复合词，所以徐通锵后期的字本位已经承认双层单位论，只是字的地位更为重要。

双层语法单位应该是自然语言的基本属性。所不同的是，汉语中有大量黏着语素是规则活动的，词法和句法的推导规则有很大的相似性。这可能是汉语在类型上的一个重要特点。汉语中的规则语素比例比英语中的规则语素比例可能更高。英语中存在大量不规则语素，也就是存在大量不规则语素组。一个不规则语素组常常有一个且只有一个主重音，形成一个音系词（phonological word）。这个主重音先确定以后，词就可以确定，所以英语的词比较容易切分。与此不同，汉语缺少词重音或词形变化，确定词很困难。尽管双层语法单位是自然语言共有的，但从规则单位的类型看，重心可能不一样，汉语很可能是语素重心型的语言，而英语是词重心型的语言。字本位、语素本位，都属于语素重心论。而词本位则属于词重心论。

以上的分析最终会集中到一个瓶颈问题，到底汉语内部有规则的复合词比例有多高？有没有概率论意义上的显著性？如果我们只是举例分析，不会有明确的结论。邱立坤的这本书解决了这一难题。作者把计算机科学和语言学有机地结合起来，不仅用语言学理论来指导计算机算法的设计，而且通过计算机算法的实验结果反过来验证并丰富语言学的理论。作者通过实验数据证明：在自动判断汉语新词语类别时词语内部结构特征比上下文分布特征更有效。作者的大量统计、算法支持了汉语规则单位的语素重心论，这在语法单位的认识论上是一个相当大的突破。计算语言学如果只是停留在数据处理、语言识别和合成，而不能从认识论上对语言现象做出解释，不能算是真正的科学，只能算是技术应用。立坤的计算语言学研究超越了单纯的技术应用，很有科学研究的品位。

作者还在实际应用方面提出内外结合原则，即判断新词语类别时应同时使用内部结构特征和外部上下文特征。基于这一原则，设计了相应的算法，实验结果表明这些算法要明显优于单独使用内部特征或者外部特征的算法，为现代汉语未登录词的语类处理提供了极有价值的可操作手段。《计算机科学》期刊2010年第3期一篇综述文章认为该方法达到了“当前的最好水平”。在立坤的算法中，优先考虑内部特征的策略尤其值得关注和借鉴。譬如，分布词类观一般认为动词可以受“不”修饰，而名词不可以受“不”修饰。但是在“不边路进攻就无法取胜”这个句子中，“边路”语类的判定从分布语境入手难度很大，它虽然接在“不”后面，但却并不是受“不”修饰，因为它并没有与“不”产生直接联系。这里需要更复杂的直接成分分析。如果从内部特征去看，以“路”

结尾的词常常是名词，比如“财路、岔路、出路”等，以“边”开始的词也常常是名词，比如“边防、边关、边疆”等。因此，从内部结构去看就很容易得到正确的结果。

当然，立坤还有很多难点问题有待解决：哪些复合词的语类是不可预测的？汉语规则单位以语素为重心的特点，在语言类型学上处于什么样的地位？计算语言学在国内已经发展了许多年，但真正能同时具备计算机科学和语言学这两方面知识结构的学者很少。立坤属于这样的复合型人才，博士毕业后又跟计算语言学专家俞士汶教授做博士后，知识结构得到进一步提升。立坤已经主持了多项自然科学基金课题，在国际会议和刊物上发表了多篇有分量的论文。在计算语言学研究中，立坤常能超越纯技术而形成科学追问的眼光，我期待立坤在不远的将来有更大的突破。

陈保亚

2016 年 6 月于北街家园静山斋

序

邱立坤2010年获北京大学文学博士学位后，即进入北京大学计算机科学与技术博士后流动站继续进行计算语言学研究。作为他在博士后期间的合作教师，2011年底应北京大学中文系之约，我曾推荐他的博士论文《现代汉语未登录词词类和语义类标注研究》申报2012年度全国优秀博士学位论文，最终虽未能入选全国百优，但也获得了北京大学优秀博士论文荣誉。在毕业之后的这么多年中，邱立坤博士一直坚持科研实践，未曾懈怠对其博士论文的检验与推敲，并增补最新成果，终于锤炼成这部仍以原论文题目为书名的专著。约我作序，自当应允，尽管现在为专著写一篇有意义的序，对我来说，并不是一件轻松的事。

当年乐意写推荐信，现在又应允作序，是基于我对他的研究以及本书内容有如下认识。

第一，研究题目有意义。计算语言学的终极研究目标是让机器具有理解和运用人类语言的能力。为达到这个目标，需要让计算机逐步具备自动处理自然语言的能力。就书面汉语而言，一项最具基础性的工作就是将构成文本的汉字序列切分成词语的序列。这项看似简单的任务，中文信息处理学界已为此付出了30余年的智慧与辛劳，然而至今并未彻底解决。其中最难跨越的障碍之一就是未登录词的识别。未登录词指机器词典未收入的词，社会生活不断发展，未登录词也就不断出现。选择未登录词作为研究对象的意义是显见的，还表现了攻坚克难的勇气。

第二，研究内容有难度。未登录词的自动识别已是一个难题，邱立坤博士还要给未登录词自动标注词类和语义类，当然是难上加难了。不过，未登录词的自动识别与未登录词的词类和语义类的自动标注并非互不相关，恰恰相反，判断未登录词的词类和语义类有助于提高未登录词识别的性能。

第三，研究成果很丰硕。邱立坤博士深入考察未登录词的内部结构和外在关系，将获取的语言知识作为特征融入统计机器学习模型（如条件随机场）中，克服了统计机器学习模型通常只利用语言表层特征的缺点，提出新的改进模型，设计算法和程序，并采用真实的语言资源进行实验，还进一步探讨所提出的方法在机器词典版本升级中的应用。除博士论文的研究成果外，本书还吸收了应

用分布式词表示方法进行汉语词语相似和关系相似计算的最新研究成果。

邱立坤博士在语言学、计算语言学领域接受了全面系统的训练，基础扎实，功底很好，是难得的文理兼修的高端复合型人才。从本科生到博士生，邱立坤一直在北京大学中文系语言学专业就读。本科毕业后保送研究生，师从陈保亚教授，旋即在陈保亚和王洪君两位老师的推荐下进入北京大学计算语言学研究所兼修自然语言处理技术，参加词语切分和词性标注语料库的建设。博士后出站后，他到鲁东大学任教，在完成教学任务的同时，仍坚持进行计算语言学研究，做到了教学与研究相互促进。

邱立坤博士既勤奋、务实，又勇于创新。博士后期间申请到国家自然科学基金青年项目“基于自消歧模式的语法知识自动获取技术研究”、博士后基金项目“现代汉语未登录词语法属性自动标注研究”，并负责与几家大型 IT 公司的横向合作项目。近几年来，在主持国家自然科学基金面上项目“句法语义分析与开放域信息抽取技术研究”等多个研究项目的同时，还分担北京大学计算语言学研究所承接的 863、973 项目的任务。他的团结、协作精神也常为北京大学计算语言学研究所的老师们所称道。

邱立坤博士论著颇丰，已在自然语言处理和人工智能领域著名国际会议上发表了多篇论文，成为本领域活跃的年轻学者。

面向自然语言理解这个目标，语言知识库构建这项工作的基础性、重要性毋庸置疑，但需要长期投入，坚持不懈，不少人视为畏途，望而却步。邱立坤博士却投身其中，乐此不疲。我为有这样年轻、这样优秀的志同道合者感到十分幸福。科学研究永远是长江后浪推前浪。我也为青年学者取得远远超越自己的成就而感到十分高兴。

“路漫漫其修远兮，吾将上下而求索。”这是我的座右铭，也以此与邱立坤博士共勉。

俞士汶

2016 年 6 月于北京褐石园

目　录

表 目 录

图 目 录

第1章 绪 论

1.1 研究对象、背景、问题及应用价值

1.1.1 研究对象

本书以未登录词为研究对象。未登录词指词典未收录的词语，主要包括两大类：一类是普通新词语，如“网民、网吧、博客、甲流”等；一类是专有名词，如人名、地名、组织机构名等，第二类又称为命名实体。命名实体已成为自然语言处理中一个独立的研究对象，受到广泛关注，形成命名实体识别（named entity recognition）这一研究方向；命名实体所涉及的问题主要是识别，识别出来之后命名实体语法类和语义类的判断则相对简单。本书主要研究第一类未登录词。随着词典收词规则的变化，未登录词的外延差别也比较大。本书主要使用《北京大学现代汉语语法信息词典》（八万词版）和《同义词词林扩展版》作为基准词典。前者是语法词典，用作未登录词词类标注的基准词典；后者是语义词典，用作未登录词语义类标注的基准词典。

1.1.2 研究背景

关于未登录词词类和语义类的判断，一个角度是依据词语的内部成分来判断词语整体的词类和语义类。语言学本体研究对词语成分与词语整体的关系已有较多的研究。比如符淮青（1985）将词义同构成它的语素义之间的关系分成5个大类、9个小类；亢世勇（2004）发现仅有8.02%的词整体意义与成分意义没有任何关系；杨梅（2006）的统计则表明，39454个词中有90%以上的合成词为向心词。这些研究结果表明，无论是从语法角度还是从语义角度看，词语成分属性与词语整体属性都有着密切的关联。因此许多研究造词法和构词法的学者认为可以依据内部结构来猜测整体的属性（杨同用，2002；王洪君，2005）。上述研究分别以定性或定量的方法分析了现代汉语词语成分属性与词语整体属性之间的关系，这些研究都是从面向人的角度进行的。如果从面向计算机的角度来考虑上述问题，会碰到一些新的问题，其中最根本的一个是成分的歧义问

题。比如在“上边”和“上课”中“上”分别属于方位词和动词，要自动地从词语的成分属性判断词语的整体属性，首先需要对成分属性消歧。因此，相比于人的理解来说，让计算机理解未登录词难度要大得多。

另一个角度则认为可以从未登录词的用法来判断其词类。20 世纪 80 年代中国语言学研究的一个重要成果就是确定了词类判断的方法，即依据词语的分布来判断词类，包括与其他词语搭配的能力和充当特定句法成分的能力。换个角度说，这种方法就是依据一个词语的用法即上下文来判断其语法类别。从面向人的角度，语言学家总结了一系列依据分布判断词类的规则（郭锐，2002）。但是，将这些规则使用到计算机自动理解上时同样会遇到新的问题，主要是上下文结构歧义的问题。比如“咬死了猎人的狗”中，要判断“咬死”与“猎人”有没有直接搭配关系是一件很复杂的事情。至于从未登录词的用法来判断其语义类，语言学界讨论得就更少了。

关于汉语未登录词词类和语义类自动判断的研究主要是由美国学者和中国台湾学者进行的，中国大陆学者较少涉及。在判断未登录词词类时，内部特征（成分及成分的属性）和外部特征（上下文）以及两种特征的结合都有学者尝试过（Lu, 2005）。在判断未登录词语义类时，主要使用内部特征，外部特征有少量的尝试，但效果不佳（Lu, 2007）。

1.1.3 本书所要回答的问题及其难点

基于上述背景，本书在前人研究工作的基础上，构建了大规模的生语料库，分别使用基于内部特征和外部特征的方法以及两种特征相结合的方法来自动处理未登录词词类和语义类标注问题。所谓内部特征指未登录词的成分、成分的属性以及成分、成分属性的组合序列；所谓外部特征指未登录词在语料中的分布，通常用未登录词的上下文来表示。

本书尝试回答以下问题：

（1）从面向计算机的角度看，词的语法属性与内部特征关系更密切还是与外部特征关系更密切？

（2）从面向计算机的角度看，词的语义属性与内部特征关系更密切还是与外部特征关系更密切？

从应用的角度看，本书所要解决的主要问题是未登录词词类标注和语义类标注。以未登录词为对象的研究主要包括未登录词识别、词类标注和语义类标注。在中文信息处理研究中，未登录词识别研究得比较多，词类标注和语义类标注则研究得比较少，仅有的少量研究也主要是由北美和中国台湾的学者完成的，大陆学者罕有涉及。鉴于此，本书以汉语未登录词词类标注和语义类标注

为主要研究内容，从内部特征与上下文用法两个角度入手来解决未登录词词类和语义类标注的问题。

未登录词词类标注的难点是含有歧义成分或歧义结构的词。比如“V+N”式的双字词，前一个字是动词类的，后一个字是名词类的，整个词可能是名词类的，也可能是动词类的，因此这是一个歧义结构，而且能产性非常强。如“灌水”是动词，而“灌渠”是名词。对于内部结构有歧义的词，单纯依据内部特征很难正确地判断出词类来。

未登录词语义类标注问题的难点如下所示：

（1）有一些词语是由完全类推而生成的，遵循一定的生成规则，但这些规则都有许多反例。比如在《同义词词林扩展版》[①]中“橙子、柑子、桔子、梨子、李子……”这一组词符合规则“C（X）=Bh07，则C（X+子）=Bh07”[②]，其中Bh07指“水果、果品、鲜果”；这一规则有一个反例“瓜子”，“瓜”属于“水果、果品、鲜果”类，但“瓜子”却不属于此类。

（2）有一些词语是由创造性类推或者词汇化生成的，这些规则都会有核心成分，核心成分的语义类通常与词语整体的语义类一致，但是许多核心成分具有歧义，可能属于多个语义类，因此这一类词语的语义类标注就要面临核心成分歧义消解的难题。比如“头”在充当核心成分时，有“头目”“物体的顶端或末梢”“人身最上部或动物最前部长着口、鼻、眼等器官的部分”等义项，如“把头、工头”“车头、船头”“白头、牛头”这三组词中“头”分别属于这三个义项。因此，要根据内部特征来判断以“头”为核心成分的词语的语义类就是一个难度相当大的消歧工作。

（3）有一些词语属于缩略词、音译词、外来词或者是通过比喻、借代等方式产生的词语义项，内部成分与整体意义关系比较松散，这一类词语很难通过内部特征来判断其语义类。比如“八角”通常指一种调味香料，是由“八角茴香”缩略而来，在《同义词词林》中与“桂皮、茴香、芥末”同类，这些词在内部成分上没有任何共同之处。

1.1.4 应用价值

现有的语法词典（如《北京大学现代汉语语法信息词典》）和语义词典（如《同义词词林》《HowNet》）基本上是人工构建而成的，较少使用大规模语料库和自动计算的方式来辅助构建。当词典达到一定规模（如七八万词）时，要对词典进行扩充或者要进一步发现词典中存在的问题，单纯依靠人工是很难实现

① 参见 http://ir.hit.edu.cn/demo/ltp/，下文简称为《词林》。

② C（X）指词语X的语义类，下文同。