Jiyu Dongtai Liutong Yuliaoku de Xinciyu Jiance Yanjiu

基于动态流通语料库的新词语监测研究

刘长征 ◎著

世界图书出版公司

基于动态流通语料库的新词语监测研究

刘长征 著

世界图书出版公司
北京·

图书在版编目(CIP)数据

基于动态流通语料库的新词语监测研究/刘长征著. —北京：世界图书出版公司北京公司,2011.9
ISBN 978-7-5100-0334-9

Ⅰ. ①基… Ⅱ. ①刘… Ⅲ. ①汉语－新词语－研究 Ⅳ. ①H13

中国版本图书馆 CIP 数据核字（2011）第 106866 号

基于动态流通语料库的新词语监测研究

著　　者：刘长征
责任编辑：陈晓辉
装帧设计：春天·书装工作室

出　　版：世界图书出版公司北京公司
出 版 人：张跃明
发　　行：世界图书出版公司北京公司
（地址：北京朝内大街 137 号　邮编：100010
电话：64077922）
销　　售：各地新华书店和外文书店
印　　刷：北京博图彩色印刷有限公司

开　　本：787 mm × 1092 mm　1/16　　印张：19
字　　数：256 千
版　　次：2011 年 9 月第 1 版　2011 年 9 月第 1 次印刷

ISBN 978-7-5100-0334-9/H · 1228　　定　　价：35.00 元

序　言

冯志伟

德国著名人文学者洪堡特（W. Humboldt）曾经提出“语言绝不是产品（Ergon），而是一种创造性活动（Energeria）”，语言实际上是心智不断重复的活动，它使音节得以成为思想的表达。人类语言知识的本质就是语言知识如何构成的问题，其核心是洪堡特指出的“有限手段的无限使用”。由于采用有限的手段来生成无限数量的语言，使得语言总是处于不断的运动状态之中。

乔姆斯基（N. Chomsky）主张，语言知识的本质在于人类成员的心智/大脑（mind/brain）中，存在着一套语言认知系统，这样的认知系统表现为某种数量有限原则和规则体系；高度抽象的语法规则构成了语言应用所需要的语言知识，由于人们不能自觉地意识到这些抽象的语法规则，这些语言知识是一些不言而喻的或者无意识的知识。乔姆斯基主张把语言知识和语言的使用能力区分开来。两个人拥有同一语言的知识，他们在发音、词汇知识、对于句子结构的掌握等方面是一样的。但是，这两个人可能在语言使用的能力方面表现得非常不同。因此，语言知识和语言能力是两个不同的概念。语言能力可以改进，而语言知识则保持不变。语言能力可以损伤或者消失，而人们并不至于失去语言知识。所以，语言知识是内在于心智的特征和表现，语言能力是外在行为的表现。生成语法研究的是语言的心智知识，而不是语言的行为能力。语言知识体现为存在于心智/大脑中的认知系统。这种语言知识的无限使用，使得语言能力不断地变化，处于不停的运动状态之中。

尽管洪堡特和乔姆斯基的上述观点尚有可商榷之处，但是，他们都认为语言在本质上是处于不停的运动状态之中的，是动态的而不是静态的，这样的动态语言观是值得称道的。

20世纪末，北京语言大学张普教授提出了动态语言知识更新的理论，在原则上与洪堡德和乔姆斯基关于语言在本质上是动态的观点是一致的。动态语言知识更新理论采用控制论的调控机制、社会语言学的监测方法和计算语言学的处理手段，对语言现象进行动态的观测、描述、分析、归纳，从而推动整个信息传播、知识更新体系的循环发展。在这个理论的指导下建立动态流通语料库，对社会语言生活进行监测研究，是计算语言学的一个新的领域和研究方向。

2004年6月30日，“国家语言资源监测与研究中心（平面媒体）”在北京语言大学揭牌成立，对包括报纸、图书、期刊在内的平面媒体进行监测与研究，标志着我国对平面媒体语言资源开始实行动态管理。此后又陆续成立了有声媒体语言分中心、网络媒体语言分中心、教育教材语言分中心、海外华语研究分中心、少数民族语言分中心。从2006年开始，国家语言文字工作委员会以“语言生活绿皮书”的形式向社会公布语言监测研究的成果。

语言的动态变化在词汇中表现最为突出，刘长征博士的《基于动态流通语料库的新词语监测研究》一书，专门探讨新词语的监测，正是在这样的背景下展开研究的。

新词语是汉语词汇的重要组成部分，它们与社会生活息息相关，因为新词语凸显了社会的变化，所以，新词语已经成为语言生活中人们关注的焦点，也成为语言监测研究的重要内容。

刘长征的研究以《深圳特区报》1981—2009共29年的全部语料为考察对象，首先采用“自动分词＋参照词表筛滤”和“特征对比＋参照词表筛滤”两种互为补充的方法，自动获取候选新词语，同时统计并计算每个词语的频次、频率、出现文本数和出现年份数。把每一年作为一个监测时点，把29年作为监测时段，提出新词语“时点生命力”和“时段生命力”的概念和

计算公式，并据此对年度候选新词语进行分类排序，根据自动分词结果构建年度新词语档案库，根据“特征对比法”构建候选新词语档案库。

刘长征利用他所构建的这些档案库，进一步展开新词语监测研究，根据时点生命力值和时段生命力值对新词语进行排序与分类，探索新词语的词源和理据，进行新词语中类词缀和词族的历时定量研究以及新词语的历时生命力曲线类型研究。

这些研究成果对于汉语词汇本体研究、语言监测、语言规范、新词语词典编纂以及对外汉语教学等诸多领域都具有十分重要的理论意义和应用价值。

这项研究属于基于语言资源的新词语监测研究，利用动态流通语料库和现代信息技术，采用社会语言学的计量统计方法，与传统的新词语研究迥然不同。

刘长征把汉语词汇学的系统理论、相对时间观理论、动态语言知识更新理论和基于实态的语言监测等理论有机地结合起来，运用到大规模真实文本的考察研究之中，不但实践了这些理论，而且在具体的研究中进行了理论的深化和创新。

刘长征在技术路线的确定和研究方法上有许多创新。比如，他提出了新词语的“时点生命力”和“时段生命力”的概念和计算公式，构建了新词语档案库，并通过时点生命力值和时段生命力值分别对年度新词语表进行了排序、分类和分析，将新词语研究从“一表制”排序向“多表制”分类排序推进了一步。

虽然新词语生命力值的计算公式可以考虑纳入更多参数进一步改进与完善，但是在现有信息技术水平的“实态”条件下尚有一定的难度。尽管刘长征没有进一步研究和改进具体的计算公式，但是，他提出了一种研究新词语生命力值的创新性研究思路，这样的创新性思路或许可以为汉语词汇学和新词语监测研究开辟一片新的天地。

此外，刘长征还提出“与时俱进”与“回溯整理”相结合的监测研究方法，通过分析新词语的历时生命力曲线类型与社会生活的对应关系，探讨“计算语言”向“计算内容”和“计算

社会”推进的可能与方法；他还对命名实体进行了历时生命力考察实验，探索语言监测向社会服务转化的途径，取得了令人信服的实验结果，具有社会学意义和广阔的应用前景。

词汇信息的研究在计算语言学中起着举足轻重的作用，单词之间的相似度（similarity）的计算、词汇的搭配关系（lexical collocation）和词汇联想关系（lexical association）的自动获取、动词的次范畴框架（sub-categorization frame）的自动获取、词汇语义学（lexical semantics）等都是当前计算语言学研究的热点。在统计方法中引入了词汇信息，可以大大地提高统计分析的精确度，在句法分析中引入词汇信息，可以减少结构上歧义，提高句法分析的效率。机器可读词典和词汇知识库成为了当前国内外自然语言处理研究最关键、最重要的语言资源。刘长征的这本书，有助于我们进一步对汉语词汇进行深入的探讨，从而推动我国计算语言学研究的发展。

在本书出版之际，我写出了这个序言，算是对于作者的祝贺，希望刘长征再接再厉，在科学研究中再谱新篇。

2011 年 6 月于北京

目　录

第一章 绪 论

1.1 问题的提出

语言是由语音、词汇、语法三大要素构成的动态的、开放的、有生命的生态系统。一种语言的语音、词汇、语法又分别构成一个子系统。语言系统及其各个子系统都在随着社会政治、经济、文化、科技等各个方面的发展进步而不断地更新变化，但各个系统发展变化的速度是不一样的。从语言系统整体来讲，其发展变化的速度是最慢的，往往伴随着大规模的社会历史变迁。我们现在使用的现代汉语系统，在汉语发展的历史长河中，是伴随着“五四”运动中白话文运动的兴起而逐渐形成的。在语音、词汇、语法三个子系统中，语音、语法系统变化更新相对较慢，而词汇是语言系统中承载信息的基本载体，是语言系统中最活跃、最具生命力的元素。社会发展日新月异，新概念、新事物不断涌现，不同的语言在日益便捷的交流中相互借鉴、融合与发展，表现在社会语言生活中最为明显的就是新词语的不断产生，旧词语的不断消亡。

葛本仪（2004）指出，活的语言永远都是在运动和发展着的，同样，语言的词汇也永远都是一个运动着的整体，虽然词汇有静态和动态两种存在形式，但是词汇的静态形式是相对的，只有动态形式才是绝对的，所以词汇永远都是在运动中变化和发展着。

词语随着社会生活的发展变化而变化，其运动的轨迹往往是这样的：首先，一个新的个体成分的出现一般都呈现为个别的临时状态，随着它的应用，又会出现各种不同的情况，有的

可能只是昙花一现，再也没有踪影；有的在小的范围内使用后又很快消失了；有的使用范围逐渐扩大开来；还有的可能是在使用中几种形式并存。世界上只要是还在使用着的语言，莫不如此。

从下面的几则新闻报道中，我们可以窥见词语动态变化之一斑。

报道一：韩国网络用语“很卢武铉”被收为韩语新词汇[①]

2007-10-12 00:44:47 来源：中新网（北京）

中新网10月12日电 韩联社消息，韩国“国立国语院”在今年的韩文节里，将网络流通的和市面上的流行语收编为韩语新词汇，其中包括贬低卢武铉总统名字的词汇，成为热门话题。

报道称，“国语院”表示，对于2002至2006年期间在韩国社会流行或新组成的3 500多个新词汇，经筛选后，按其发音收编为韩语新词汇，并发行了一本名称为“词典里找不到的新用词”的词典。

据了解，除了在网络里常用的“火星语”之外，该词典还包括了社会的俗语、简语和新组合的造词等。其中还包括了拿当今韩国总统的名字“卢武铉”开玩笑，具有贬低含意的三个新词汇。

第一个最具有争议的新词是“很卢武铉”。该新词将“卢武铉”三个字缩读成两个音节，使其发出“很弄玄”的音，在含意上披上了“那家伙，那傻瓜”的意思。据“国立国语院”在词典里解释，“很卢武铉”是指让人难有期待，很让人失望的情况。

据悉，“国立国语院”在接受采访时表示，“青瓦台在得知‘很卢武铉’被编进韩语新词汇之后，曾就此提出抗议和回收所有词典的提议。”他们还表示，“当时据实回答了总统府，词典已印完并发布完。”

① http://news.163.com/07/1012/00/3QIJKQ440001121M.html

对此青瓦台通过发言人反驳称："对于污辱国家领导人的词汇，我们曾提醒国立国语院有必要慎重其事，但没有提出回收的要求。"

除了上述的"很卢武铉"之外，有关卢武铉的另两个新词汇是"卢哥"和"卢大"，其中前者是卢武铉的昵称，后者的意思是"姓卢的老大"，也普遍以此称呼卢武铉总统。

报道二："给力"登《人民日报》头条 曹景行评网络热词①

2010-11-12 08:11:00 来源：新浪网

网络热词"给力"登上《人民日报》头版头条。11 月 10 日头版头条为《江苏给力"文化强省"》，文章介绍了江苏从"文化大省"向"文化强省"的华丽转变，并用引题的方式总结了三条经验：改革攻坚迸发活力、政策创新激发活力和厚积薄发释放能力。对此，著名学者曹景行作如下点评：

语言实际上是一个活的东西，它会不断地更新、不断地淘汰，也会不断地补充。这些年来我们新的词汇最多产生的就是在互联网上，它不仅丰富了我们的用语，淘汰了那些陈旧的词汇，更重要的是它更能反映我们社会的变化和发展。

所以尽管网上会出现一些所谓火星文这样的新词，但是对我们的传统中文并没有造成破坏，反而在这几年当中我们的语言有了很大的丰富。不仅在网络上，在口语上，同样也在我们的传统媒体上。所以随着生活和社会的变化，网络的热词还是会不断地出现，这就是我们语言的强大的生命力。

报道三：西班牙作家学校呼吁人们"收养"过时词汇②

2007-04-13 05:51:19 来源：星辰在线

星辰在线 4 月 13 日报道 为保持西班牙语的丰富性，最近西班牙马德里的一个作家学校发出倡议，呼吁人们特别是名人来"收养"那些过时的濒危灭绝的西班牙词语。

---

① http://book.hebei.com.cn/wtylzx/dspd/sytj/201011/t20101112_2484401.shtml

② http://news.163.com/07/0413/05/3BUGR8QK0001121M.html

这个作家学校的负责人在解释他们提出这一创意的原因时表示，由于语言的“自然进化”，一些西班牙语词汇已经逐渐淡出人们的生活，这样导致的后果是人们使用的词汇日益贫瘠。因此，他们呼吁人们担任“养父”来“收养”这些濒危灭绝的词汇。“我们需要帮助来救救这些词语”，为此，他们开设了一个网站 http://www.escueladeescritores.com，凡是有意做“养父”的人只要在这个网站上做简单的登记，就可以认领到一个在西班牙皇家学院词典内出现和曾经出现的词语。不过，这名领导人表示对此有兴趣的人应该加快速度赶紧登录他们的网站，因为到本月21日他们就不再接受新的登记者了。

截至目前已经有超过5000个西班牙词语被来自42个国家的人认领，这些词语的“养父”主要来自西班牙和拉美国家，但是也有一些词语被中国人认领。他们当中有许多名人，其中最著名的就是现任首相罗德里格斯-萨帕特罗，他认领的词语是“andancio”，意思是一种传染病。这个词语曾经在他的故乡和古巴被使用过，但是现在几乎已经没有人知道它的意思了。

报道一中，在韩国网络流行的关于卢武铉的流行词语，虽然有官方的强烈抗议，但仍然阻挡不了人们的关注与使用，这种流行词语是规范的用法吗？随着时过境迁，总统易人，这些流行一时的热门词语命运又将如何呢？

报道二中的网络热词“给力”登上《人民日报》头版头条，使这一在网络中的流行词语扩大了使用范围，进入权威传统平面媒体，而权威传统媒体的引领必然提高其流通度和使用度。那么，它是不是就已经进入现代汉语词汇的稳态系统，成为一般词语了呢？

报道三中倡议人们“领养”濒临灭绝的西班牙词语，其初衷是为了保持西班牙语的丰富性，而这些“已经淡出人们生活”的词语，通过名人“领养”就能获得新的生命力吗？语言的丰富性可以通过“保存”这些词语不被人们忘记而得以保持吗？这些所谓“已经淡出人们生活”的词语真的会全部且永远地

“淡出人们的生活”吗?

以上问题,我们当然可以根据一般的经验和常识作出预测和判断,但这种预测和判断只是主观的。

通过对这三则报道的分析和思考,我们更深刻地认识到,语言和社会是一对“共生”、“共存”和“共变”的孪生子。社会的每一项细微的变化和发展都会在语言上以及时而适当的方式表达出来。(汤志祥,2004)中国改革开放以来,社会处于急速发展之中,新事物不断涌现,旧概念不断被更新,与这个时代的演进步伐相适应,汉语也产生了显著的变化。在语言的发展和变化之中,词汇和词义向来是反映社会变迁最敏感、最活跃的构成成分。改革开放以来汉语新词语的大量涌现,就是中国社会快速发展的一面镜子。有统计指出,改革开放30年以来,新词语以平均每年800个的速度产生,而且该数字还有不断增长的趋势。关于新词语的研究、新词语词典的编纂也日益引起语言学界、计算语言学界、社会语言学界甚至对外汉语教学界的重视。

新词语监测研究是新词语研究的基础性工作。只有在大规模动态流通语料库的基础上,以历时中含有共时、共时中含有历时的相对时间观为指导,通过对新词语的监测研究,才能相对准确和全面地掌握新词语在社会语言生活中的第一手材料,把握新词语使用和流通的脉搏,从而为进一步的新词语语言学、社会语言学、计算语言学、语言教学和词典编纂等研究和应用打下坚实的基础。

1.2 研究对象

取得深圳报业集团的授权,本文“基于动态流通语料库的新词语监测研究”以《深圳特区报》1981年至2009年全部语料为研究对象。其中1981年至2006年10月26日全部文本的电子文档由深圳报业集团直接提供,2006年10月27日至2009年12月31日文本由北京语言大学应用语言学研究所DCC博士研究室通