DONGTAI
YUYAN
ZHISHI
GENGXIN
YANJIU

# 动态语言知识更新研究

张普 著

创于1897 商務印書館 The Commercial Press

# 动态语言知识更新研究

张 普 著

商務印書館
2009年·北京

**图书在版编目(CIP)数据**

动态语言知识更新研究/张普著. —北京:商务印书馆,2009

ISBN 978-7-100-05925-1

Ⅰ. 动… Ⅱ. 张… Ⅲ. 语言—信息处理系统—研究
Ⅳ. TP391.1

中国版本图书馆 CIP 数据核字(2008)第 108635 号



DÒNGTÀI YǓYÁN ZHĪSHÍ GĒNGXĪN YÁNJIŪ

**动态语言知识更新研究**

张 普 著

商 务 印 书 馆 出 版
(北京王府井大街36号 邮政编码100710)
商 务 印 书 馆 发 行
北 京 龙 兴 印 刷 厂 印 刷
ISBN 978-7-100-05925-1

2009年5月第1版 开本 850×1168 1/32
2009年5月北京第1次印刷 印张 12⅞

定价:25.00元

# 内 容 简 介

本书是北京语言大学张普教授的第三部个人选集，收录了他的与“动态语言知识更新”研究有关的论文19篇，附录7篇。从第一篇发表到最后一篇写成，跨度10年。但是，如果从有想法算起，到付诸实践，到将阶段论文结集出版，前后却历经25年（参见前言）。

26篇论文分为5个部分：

第一部分 思考篇（5篇）

第二部分 控制论篇（4篇）

第三部分 理论篇（4篇）

第四部分 应用篇（6篇）

第五部分 附录篇（7篇）

五个部分涉及作者在总结上个世纪国内外语言资源建设的基础上，对新世纪的语言资源研究的深入思考，并分别论述了对“动态语言知识更新”研究的理论、方法论、实践和相关的研究与成果。在此期间作者负责建立了北京语言大学的应用语言学研究所，特别是建立了以“动态语言知识更新”研究为主要目标的“DCC博士研究室”，DCC即为英文Dynamic Circulating Corpus（动态流通语料库）的缩写。同时在此基础上，在教育部语信司和北京语言大学的领导与支持下，建立了“国家语言资源监测与研究中心”的第一家分中心——平面媒体分中心。

本书的创新点主要就在于作者力主建立“动态流通语料库”；在语言研究的时间观的争议中，强调历时研究与共时研究双重(音zhòng)，并基于现代大众传媒的发展，提出“共时中有历时和历时中有共时”的“相对时间观”；对于语感进行了一些基本分类，包括对于个人语感和大众语感、共同语感和差别语感、语感的空间特征和语感的时间特征等分类，提出通过测量大众传媒的流通度这个“白箱”来相似计算大众语感这个“黑箱”，提出计算机的“语感模拟”，把重要的语言规律“约定俗成”推向可计算。

尽管已经过去25年，但作者认为动态语言知识更新的思考、理论、方法、实践还是初步的，未臻成熟。不过，毕竟对语言的动态部分(例如新词新语新义、流行语、字母词语、术语等)和语言的稳态部分(例如通用词语、基本词汇、基本字集等)的监测和研究都已经开始，并引起了国内外的关注。

本书对于语言研究、应用语言学研究，特别是对于语言规划、语言规范化、社会语言学、计算语言学的研究，对词典编纂、语言教学、汉语海外传播、语言信息处理、自然语言理解等方面均有参考价值。

# 目　录

# 序

北京语言大学语言信息处理(后为应用语言学)研究所所长张普教授,早在上个世纪70年代就步入了中文信息处理工作的领域,当时他在武汉大学中文系任教,他与计算机系的一些老师一起利用计算机编纂文学名著语言资料索引,以供汉语研究和汉语辞书编纂之用。1983年第一本为计算机编纂的现代汉语语言资料索引《〈骆驼祥子〉逐字索引》问世,吕叔湘先生还专门为之写序,给予充分肯定。但是他真正专注动态语言更新研究,那是上个世纪80年代中期之后。引发他从事这一研究的,是日本语言学者引例看重所引文本的印数、发行量、销售排行榜上的位置之举。他开始意识到关注动态语言知识更新的重要。

1986年他由武汉大学转到北京语言学院(北京语言大学的前身)任教,由他实际主持北京语言学院刚成立的语言信息处理研究所的工作(研究所成立的第一年学校聘请北京大学马希文教授任所长),从事中文信息处理的研究教学工作。这使他能更多地了解国际上和中国国内自然语言处理的进展情况,也更有条件来思考动态语言知识更新的问题。上个世纪90年代初,他向学校提出建立"大规模现代汉语动态语料库"计划,以进行语言动态更新的研究。此计划得到北京语言学院领导和校内外专家学者的一致支持。但真要建设一个"大规模现代汉语动态语料库"谈何容易,必

须有人力、财力的支持，更要有理论的准备。为此，张普教授一方面注意物色人才，并努力争取校方的财力支持，另一方面广泛向业内专家学者虚心请教，并认真查阅有关文献资料，进行理论武装。进入 21 世纪，随着“DCC 博士研究室”(DCC 是英文 Dynamic Circulating Corpus 的缩写，意为“动态流通语料库”)的成立，随着他所带的一届届博士研究生的入学，通过网上下载的方式，逐步形成基于报纸语料的“现代汉语动态流通语料库”。所选取的报纸是发行排行榜上前 100 种里发行量最大的报纸(先选取了 10 种，后又增至 15 种)。这使动态语言知识更新的研究有了基本的保证。在有关方面的支持下，他们开展了“中国主流报纸十大流行语跟踪与发布研究”，从 2002 年开始每年发布一次，受到广泛的重视；2004 年，北京语言大学又和教育部语信司合作共建了“国家语言资源监测与研究中心平面媒体分中心”，这使动态语言知识更新的研究进入了良性发展的新时期。20 多年来，张普教授不仅自己，而且带领一批年轻学子，耕耘不止，撰写、发表了一篇又一篇有关动态语言知识更新的学术论文，为中文信息学界和现代汉语学界所瞩目。现在，呈现在读者面前的《动态语言知识更新研究》一书，就是他对自己所撰写、发表的有关这方面内容的文章精选、结集而成的。正文分“思考篇”、“控制论篇”、“理论篇”、“应用篇”四个部分，外加一个“附录”。这可以看做是他对动态语言知识更新研究的阶段性总结。他将自己的论文结集成书后，要我为书写序。我不好推辞，因为我是他在北大念书时的老师，给他们班讲授过“现代汉语”课，还跟他们班一起外出进行过田野调查。但是，也让我感到为难，因为我对动态语言知识更新问题，不要说没有任何研究，而且知之甚少。在这方面，应该说张普是我的老师。不过也好，借这个机会我

也可以学点儿新知识。于是，我把他给我的书稿的主要篇章都认真地阅读了一遍，还去查阅了有关文献资料。果然受益匪浅，长了不少见识。

讨论动态语言知识更新问题，必然会涉及这样一些理论问题：怎么看待规范问题？语感在动态语言知识更新研究中会占什么样的地位？怎么认识控制论与动态语言知识更新之间的关系？怎么正确认识索绪尔所提出的关于区分共时语言学与历时语言学的理论？

关于怎么认识语言规范问题，上个世纪八九十年代以及前两年都曾展开过广泛的讨论。本书好几篇文章都有所论及，特别是在《关于网络时代语言规划的思考》一文中有较多的论述。我赞同作者在书中所持的观点，这里不再赘述。只想简单地说两点意见，一是语言的变异是绝对的，规范是相对的，对语言的变异，说句实在话，语言学工作者没有什么规范的权利，有的是解释的义务；二是规范一定要有弹性，要允许不同意见的争论，过分的行政命令，只能事与愿违。

关于语感，本书也在多篇文章中有所论及。语感在不同研究领域，所占的位置可能不是一样的。在动态语言知识更新的研究中，正如作者所指出的，无论是个人语感还是公众语感都必须充分关注，甚至需要进行“语感量化”和“语感的计算机模拟”这样一些专项研究；在母语语文教学中，语感问题，特别是怎么培养学生良好的语感，也会放在重要的位置上来考虑；而在外语教学中，如在汉语作为第二语言/外语教学中，正如张旺熹教授所指出的，“语感是语言使用者对特定语言系统的形式、意义和功能之间所具有的特定内在联系性的高度自动化的判断意识，是语言使用者把握、使用这种特定内在联系性的纯熟的语言行为的表现”，因此，“语感培养是对外汉语教学的基本任务”。（《〈世界汉语教学〉创刊二十周

年笔谈会·〈语感培养是对外汉语教学的基本任务〉》,《世界汉语教学》2007年第3期第24—25页)可是在语言共时状态的研究中,如在语法研究或词汇研究中,我认为,语感只能成为我们研究、思考某个问题的诱因或出发点,不能作为最终立论的依据。

关于怎么认识控制论与动态语言知识更新之间的关系问题,书中《关于控制论与动态语言知识更新的思考》一文对此专门进行了讨论,特别强调控制论与动态语言知识的密切关系,论述了控制论对动态语言知识更新的理论支持或方法论上的支持。正如文章所指出的,当今信息传播的速度、领域、方式、效应均前所未有,人类有可能面临信息爆炸、信息泛滥、信息失控的局面,为防止这一不利局面的出现,急需建设基于社会传媒的网络语言规划模型,这个模型设想由四部分组成:语言自动控制体系、语言自动学习体系、语言知识自动反馈体系以及社会传媒之中的主页和文本自动检测体系。上述四大体系构成一个"学习—反馈—控制—检测"模型。这个模型可以说是信息时代语言信息处理的"自动机"。而这一模型显然既符合控制论的思想,也离不开控制论的指导。因此,我们必须从控制论的角度重新审视动态语言知识更新问题。结论是:动态语言知识更新与控制论紧密联系。张普教授这些观点,我想大家都会赞成。这里我想进一步补充的是,何止动态语言知识更新与控制论紧密联系,就是从相对静止的共时平面上看,语言的各个组成部分都存在着互相控制的关系,我们对语言作共时平面的研究时,也需要有控制论意识,也需要有控制论指导。举例来说,上个世纪80年代初,几乎同时,南北都提出了"三个平面"的思想,受到汉语语法学界的普遍关注,在汉语语法研究中广泛运用。但不少人只热衷于孤立地分别从句法、语义、语用三个平面来描写

说明自己所研究的某个词或某个句法格式，而不见有人深入思考句法、语义、语用这三者之间的联系或者说关系。其实，这三者之间就存在着互相制约、互相控制的关系。举例来说，“香蕉青的不买”，从理论上来说，可以有两种切分，可理解为两种意思：

A. 香蕉 青 的 不买 [意思大致是“不买青的香蕉”]

1 2 1—2 主谓关系

3 4 3—4 主谓关系

B. 香蕉 青 的 不买 [意思大致是“不买还挂着青的香蕉的香蕉树”][①]

1 2 1—2 主谓关系

3 4 3—4 “的”字结构

同样，“皮儿青的不买”，从理论上来说，也可以有两种切分，可理解为两种意思：

A. 皮儿 青 的 不买 [意思大致是“不买青的皮儿”]

1 2 1—2 主谓关系

3 4 3—4 主谓关系

B. 皮儿 青 的 不买 [意思大致是“不买皮儿青的那种水果”]

1 2 1—2 主谓关系

3 4 3—4 “的”字结构

但事实上，“香蕉青的不买”只能取(A)分析，而“皮儿青的不买”只能取(B)分析，而这完全是由语用因素决定的，甚至可以说

① 按朱德熙先生对“VP的”这类“的”字结构的研究(这里所说得VP涵盖动词、形容词)，“N+V+的”可以指称N所指的事物的领有者。如：“孩子游泳的”可以指称孩子的家长；“皮儿红的”可以指称某种水果。

是由社会生活决定的，因为在现实生活中，不存在香蕉树的买卖，也不存在水果皮儿的买卖。上面所举的可以认为是语用控制句法的典型例子。至于语义和句法之间的互相制约，语音和句法之间的互相制约，其例更是不胜枚举。总之，动态语言知识更新与控制论联系紧密，而就是从相对静止的共时平面上看，语言的各个组成部分之间的关系与控制论也联系紧密。

《论历时中包含有共时与共时中包含有历时》一文是专门讨论共时研究与历时研究的关系问题。文章在介绍了国外所说的两种时间说——物理学的时间和进化论生物学的时间，前者是可逆的，后者是不可逆的——的基础上，进一步论述了这样两个观点：一是“语言属于进化论生物学的时间”；二是“就语言的发展而言，历时中包含有共时，共时中包含有历时”。此外，对索绪尔所提出的区分共时语言学和历时语言学的观点进行了评论，强调语言的共时态是语言的空间态，语言的历时态是语言的时间态；“语言的历时研究与共时研究同等重要，不可偏废”；研究语言，“既要观察语言的共时状态，也要观察语言的历时状态，这样的观察才是全面的观察”。这些看法无疑都是正确的。但我在这里需要指出的是，索绪尔的下列观点还是有必要强调：“语言学应该分成共时语言学和历时语言学。共时语言学研究的是作为系统的语言，所以特别重要；历时语言学只研究个别语言要素的变异，不能构成系统，所以同共时语言学比起来，不如共时语言学那么重要。”这也就是说，历时研究与共时研究都很重要，不可偏废，但不等于二者可以不分主次。语言研究的事实告诉我们，语言的共时研究还是主要的，只有对语言共时状态作了充分的研究，才能很好地建立起历时语言学；而由于造成语言变异的因素很多，有语言自身的因素，有语言外在的因

素，所以历时语言学虽有助于语言共时状态的研究，但终究不能据此准确预测语言发展的走向和发展的必然趋势。我们现在强调要加强动态语言知识更新的研究，要加强语言历时状态的研究，只是因为过去这方面的研究太不注意了，而并不意味着对语言共时平面的研究和对语言历时状态的变异的研究可以不分主次。如果我们不这样来认识，又可能会走偏。据此，文章关于“索绪尔时间观”的提法还可以斟酌。按我理解，所谓“索绪尔时间观”里的“时间”跟上面所说的“物理学的时间”、“进化论生物学的时间”里的“时间”不是一个层次上的概念。

本书是个论文集，如果作者能根据已发表的论文的内容撰写成专著，可能出版效果更好一些，而且也可以避免一些前后行文上的重复。但话又得说回来，即使是一个论文集，也是“动态语言知识更新研究”方面的开山之作，很值得大家一读。是为序。

陆俭明

2007 年 7 月 28 日

# 前　言

这本《动态语言知识更新研究》结集出版是很不容易的。

光是这个前言，就写了5年，2003年写了第一稿，现在是第七稿。不断修改，不断计划增入新的文章，其间还遭遇了我一生中最严重的病患，可谓经历了死去活来。其实，这期间最重要的“更新”，就是我自己人生哲学的改变。我过去是透支生命、只争朝夕，以拼命三郎、工作狂为荣，现在变成更理性的“要拼搏，不要拼命”。命还在，才可以搏得更久，而命拼掉了，争到了朝夕，也不可持续发展。这一点，是我大病后的大彻大悟，幸而“重生”，得以有机会与诸君共勉。

不仅前言写了5年，本书中论文的写作时间更长。这是我的第三部个人选集，收入关于“动态语言知识更新”正编的研究论文19篇，附录7篇。从第一篇发表到最后一篇写成，跨度10年。但是，如果从1982年见到杉村博文先生的例句（下文有详述）有一点思考算起，到付诸实践，到将阶段论文结集出版，前后却历经25年。我自认为书中还有一点点创新，这就是：力主建立“动态流通语料库”；在语言研究的时间观的争议中，强调历时研究与共时研究双重（音ZHÒNG），并基于现代大众传媒的发展，提出“共时中有历时和历时中有共时”的“相对时间观”；对于语感进行了一些基本分类，包括对于个人语感和大众语感的分类，提出通过测量大众

传媒的流通度这个“白箱”来相似计算大众语感这个“黑箱”，把重要的语言规律“约定俗成”推向可计算。尽管“弹指一挥间”25 年已经过去，但动态语言知识更新的思考、理论、方法、实践还是初步的，未臻成熟。不过，毕竟对语言的动态部分（例如新词新语新义、流行语、字母词语、术语等）和语言的稳态部分（例如通用词语、基本词汇、基本字集等）的监测和研究都已经启动，并引起了国内外有关媒体和同行的关注。

25 年时间，仅此“一点点创新”，实在不值得炫耀。何况大量具体工作还都是博士、硕士们做的，我常常开玩笑和同学们说：“张老师现在是‘君子动口不动手’。”当然，我的文章还是自己动手写的。本前言的写法也想有一点儿改变，就是不想写成一般前言的本书内容简介与致谢。我想讲讲 25 年才有“一点点创新”的真实的故事，希望能比一般的前言好看和有益。

透过这个故事，我想告诉诸君，我可绝不是一个聪明人，不是“一不小心就玩上了语言研究”，玩出一本集子。我更不敢言“码字”或“写字”，我从有一点想法，到付诸实施，到形成一个集子，记录我的“学术心路历程”，实实的不容易。虽不敢言呕心沥血，一曝十寒，但是，岁月蹉跎，却也倏忽过了 25 个寒暑，最后竟然脑梗继而又“新生”。25 年，足以使一个初生的婴儿长大成人，然而，动态语言知识更新的学术创新还是只能算刚刚起步。因此，写写一个不聪明的甚至驽钝的人如何创新，就是想说：尽管现在技术进步了，有了“Google”和“百度”[①]的帮忙，但是创新谈何容易，学海无涯，万不可靠投机取巧、浅尝辄止。不过，创新又是人人都可以努

① “Google”和“百度”是常见的网络搜索引擎。

力争取的，“只要功夫深，铁杵磨成绣花针”的道理还管用。现代社会，不磨“铁杵”，不磨“绣花针”了，可以磨点别的，比如“动态语言知识更新”。

学术，常常是自以为是的，学界是否认同又另当别论。不过认同与否并不十分重要，如果都认同了，就都成了定论，学术也就没有了生命。学术，又常常是自以为非的，我也时常地自省，随着时间的推移，它还站得住吗？尤其是研究“语言知识动态更新”，难道这“动态更新”自己就不会被更新吗？

我深知自己的根基。做了30多年的“语言信息处理”方向研究，反而越做越觉得我们在交叉知识结构方面的欠缺。以我现在从事的语言知识动态更新研究而言，就已经汲取了理论语言学、社会语言学、计算语言学、认知科学、信息科学、传媒学、控制论等多学科的营养。我依然时时觉得自己功力不足，需要“充电”。我愿将这集子出版，一方面说明创新之树的生长，首先靠内因，当然自己要努力，根基重要、吸收重要、方向重要；一方面也想说明：此外，一要靠地帮忙，哪里有水、有土、有营养，根就往哪里扎，二要靠天帮忙，祈求风调雨顺，没有旱涝虫瘟。借助天地的力量，可以“化成万物”。[①] 天时、地利、人和，常常带有机会和运气。

我要借写《前言》的机会说明：也许我的机会和运气一直很好。

创新的故事，就从“天地化成万物”开始。

---

① 语出《说文解字》的首卷首字“一”的说解“惟出太始，道立于一，造分天地，化成万物”。更早还可见老子《道德经》三十六章：“道生一，一生二，二生三。三生万物。万物负阴而抱阳，冲气以为和。”