本文于 2010-04-20 发表于《中国社会科学报》，总 81 期

数学是语言学现代化的重要工具

-- 评介《语言学中的数学方法》[1]

冯志伟

冯志伟：教育部语言文字应用研究所研究员、博士生导师，长期从事语言学、数学和计算机科学的跨学科研究，主要研究领域为计算语言学、数理语言学、理论语言学。

Mathematical Methods in Linguistics《语言学中的数学方法》，世界图书出版公司，2009 年 3 月版

语言学和数学都是有相当长历史的古老学科。语言学历来被看成时典型的人文科学，数学则被许多人看成是最重要的自然科学。在学校的教育中，语文和数学被认为是两门最基础的学科，成为了任何一个受教育者的必修课。它们似乎成了学校教育的两个极点：一个极点是作为文科代表者的语文，一个极点是作为理科代表者的数学，在一般人看来，语文和数学似乎是两个风马牛不相及的学科，很少有人想到，这两门表面上如此不同的学科之间竟然还存在着深刻的内在联系。

可是，一些有远见卓识的学者却慧眼独具，敏锐地看出了语言和数学之间的联系。

例如，俄国数学家马尔可夫在 1913 年就采用概率论方法研究过《欧根·奥涅金》中的俄语元音和辅音字母序列的生成问题，提出可马尔可夫随机过程论，后来成了数学一个独立的分支，对现代数学产生了深远的影响。语言结构中所蕴藏的数学规律，成了马尔可夫创造性思想的源泉。

然而，这些构思巧妙的研究都没有对语言学本身发生显著的影响。这是由当时的社会实践的要求所决定的，因为当时的语言学，主要是为语言教学、文献翻译、文学创作和社会历史研究服务的，在这样的社会实践要求下，语言学还没有很大的必要与数学建立直接的联系。语言学仍然沿着自己传统的道路，孤立于数学之外，迟缓地发展着。

20 世纪以来，由于科学技术突飞猛进的发展，科技文献的数量与日俱增，世界各国每天出版的科技文献以数十万计，科技文献的这种增长情况被形容为“信息爆炸”。面对浩如烟海的科技文献，科技工作者为了了解外国的研究成果，取得科技信息，不得不花费大量的人力、物力来做难以数计的翻译工作，大大地影响了科研工作的效率。

1946 年，世界上第一台电子计算机研制成功，紧接着，在 20 世纪 50 年代初期，人们

---

[1] Mathematical Methods in Linguistics，《语言学中的数学方法》，世界图书出版公司，2009 年 3 月出版，书号：ISBN 978-7-5602-9287-0/H•1068

就开始考虑把这些工作交给电子计算机去做，利用电子计算机把一种形式的信息转换成另一种形式的信息，也就是将原始信息转换成为结果信息，这就提出了机器翻译、机器自动作文摘以及机器自动检索科技文献等信息加工问题，这样的研究叫做“自然语言处理”，这些都要求使用数学方法来研究语言。

随着信息技术的进步和网络的发展，互联网（Web）逐渐变成一个多语言的网络世界。目前，在互联网上除了使用英语之外，越来越多地使用汉语、西班牙语、德语、法语、日语、韩国语等英语之外的语言。英语在互联网上独霸天下的局面已经打破，互联网确实已经变成了多语言的网络世界，因此，互联网上的不同语言之间的自动翻译和处理也就越来越迫切了。而且，如何从包含海量信息的互联网上搜索到人们需要的信息，成为了网络时代的一个关键的技术问题，而这些信息大部分都是由语言文字来负荷的，也需要使用数学方法来研究语言。

在当今的信息时代，科学技术的发展日新月异，新的信息、新的知识如雨后春笋地不断增加，信息爆炸的情况更加严重。现在，世界上出版的科技刊物达 165000 种，平均每天有大约 2 万篇科技论文发表。专家估计，我们目前每天在互联网上传输的数据量之大，已经超过了整个 19 世纪的全部数据的总和；我们在新的 21 世纪所要处理的知识总量将要大大地超过我们在过去 2500 年历史长河中所积累起来的全部知识总量。随着知识突飞猛进的增长，机器翻译、信息检索、自动文摘、信息挖掘等自然语言处理的研究显得更加迫切，研究领域日益扩大，成为了当代语言学中最引人注意的一个新兴学科。

自然语言处理要求建立形式化、算法化、程序化、实用化的语言模型，这些都离不开数学，都必须使用数学方法来分析和描述语言，语言学与数学的结合已经到了迫在眉睫的地步了。

上面我们只是分析了语言学与数学结合的必要性，那么，语言学与数学的结合是否有可能呢？我们认为，不论从语言本身的性质来看，还是从当前科学技术发展的水平来看，都是有可能的。

语言本身的性质来看，正如索绪尔指出的，语言是一个符号系统，它可以同交通信号灯这样的符号系统相类比，只不过比交通信号灯复杂得多。每一种语言都是“能指”（即符号的物质表达）与“所指”（即概念或对象）的统一体，它为不同平面上的一定的结构规律制约着。因此，我们在研究语言时，可以只关注它的结构，至于这种语言是口说的或是手写的，还是用莫尔斯电码编了码的，对于研究者来说都是无关紧要的。这样，我们就可以把语言看成是一个抽象的符号系统，这种抽象的符号系统，当然可以用数学来加以研究。

从科学技术当前的发展水平来看，也为用数学来研究语言提供了理论和方法。现代数学日新月异地发展，20 世纪以来迅速发展着的概率论、数理统计、信息论、集合论、数理逻辑、图论、格论和抽象代数等数学部门，为用数学思想和方法研究语言提供了有力的武器。

现代语言学也逐渐向精密化方向发展，在传统语言学内，叶斯柏森提出了形式化的“分析句法”，在结构语言学内，布龙菲尔德、哈里斯等人提出了以替换和分布为手段，以辨别语素、分析层次为目标的一套严格的语言研究法。这些语言学派，在其语言观方面难免有片面之处，就是其具体方法本身，也有许多故弄玄虚、徒滋纷扰的地方。但是，由于采用了比过去的语言学更加严格的精密方法，在某些方面，对于用数学方法来研究语言也有一定的启示作用。

由此可见，使用数学方法来研究语言，不但是必要的，而且也是可能的。数学已经成为语言学现代化必不可少的重要工具。

随着我国自然语言处理研究的进一步发展，越来越多的学者开始关注语言学中数学方法的研究，数学方法在语言研究中的应用越来越广泛，就是在传统的语言学研究中，也开始采用数学的方法，不再认为使用数学方法来研究语言是一种离经叛道的古怪行为。在语言研究中采用数学方法，现在已经得到了我国语言学界的普遍认同。随着自然语言处理研究的发展，

数学已经成为语言学研究的最重要的一种工具。今天，现代语言学的研究，特别是面向计算机的语言学研究，离开了数学将寸步难行。生活在信息时代的语言学家，应当面对信息时代的需要，努力进行知识更新的再学习，改进自己的知识结构。

世界图书出版公司出版的《语言学中的数学方法》正好满足了这样的要求。

本书的三位作者都是语言学家。这本书是专门针对语言学的需要写的数学方法方面的著作，完全从语言学的角度来讲述数学，特别适合于那些想学习语言学中的数学方法的语言学专业的教师和学生阅读。

本书包括 A, B, C, D, E 五篇。A 篇讲述集合论，B 篇讲述逻辑和形式系统，C 篇讲述抽象代数，D 篇讲述作为形式语言的英语，E 篇讲述形式语言、形式语法和自动机。建议读者从 A 篇开始，一篇一篇地仔细阅读，反复推敲，认真做练习，逐步深入下去，就可以升堂入室，了解到语言学中使用的主要的数学方法。

我认为，本书具有以下特点：

第一，内容新颖，说理透彻

本书从数学的角度，深入地探讨了自然语言的形式特性，大大地丰富了我们对于自然语言特性的认识。

本书详细地介绍了语法的乔姆斯基层级。根据语法中重写规则的限制，乔姆斯基把形式语法分为四种：0 型语法，1 型语法，2 型语法，3 型语法。0 型语法又叫做递归可枚举语法，1 型语法又叫做上下文有关语法，2 型语法又叫做上下文无关语法，3 型语法又叫做有限状态语法。

这些语法之间有着明确的逻辑关系。每一个有限状态语法都是上下文无关的，每一个上下文无关语法都是上下文有关的，每一个上下文有关语法都是递归可枚举的。这样，可以把由 0 型语法生成的语言叫 0 型语言，把由上下文有关语法、上下文无关语法、有限状态语法生成的语言分别叫做上下文有关语言、上下文无关语言、有限状态语言，也可以分别叫做 1 型语言、2 型语言、3 型语言。由于从限制 1 到限制 3 的限制条件是逐渐增加的，因此，不论对于语法或对于语言来说，都有

$$0\text{ 型}\supseteq 1\text{ 型}\supseteq 2\text{ 型}\supseteq 3\text{ 型},$$

这就是关于形式语言和形式语法的乔姆斯基层级。

本书指出，从计算复杂性来看，自然语言处于上下文无关语言和上下文有关语言之间，这样的论述加深了我们对于自然语言的认识。这是当代语言学理论研究中的精粹，但在一般的语言学著作中是很难读到的。

此外，本书还介绍了乔姆斯基和库洛达等人关于语法和自动机相互关系的研究，他们把四种类型的语法分别与图灵机、线性有界自动机、下推自动机及有限自动机等四种类型的自动机相联系，并且证明了关于语言语法的生成能力和语言自动机的识别能力的等价性的四个重要结果，即：

1）若一语言 L 能用图灵机来识别，则它就能用 0 型语法生成，反之亦然；

2）若一语言 L 能用线性有界自动机识别，则它就能用 1 型(上下文有关)语法生成，反之亦然；

3）若一语言 L 能用下推自动机识别，则它就能用 2 型(上下文无关)语法生成，反之亦然；

4）若一语言 L 能用有限自动机识别，则它就能用 3 型(有限状态)语法生成，反之亦然。

上述理论提出了关于语言的生成过程和语言的识别过程的极为重要的见解，揭示了自然语言与计算机之间深刻的内在联系。

这些介绍使我们大开眼界，不但对于自然语言处理系统的开发具有指导意义，而且对于理论语言学的研究，也是很有启发的。

第二，深入浅出，通俗易懂

本书是专门为语言学工作者写的，讲数学问题时都紧紧扣住语言，通过恰如其分的语言实例来说明复杂的数学原理，深入浅出，具有高中数学知识的读者都不难理解本书的内容。文科背景的读者也可以通过本书学会如何在语言研究中使用数学方法。

第三，精选练习，注意实际

本书大多数章节的后面都有练习，这些练习都是经过作者精心挑选的，书末附有大部分练习的答案，可供读者做练习时参考。

第四，文献完善，编排精当

书末的参考文献几乎收集了与语言学中的数学方法有关的且全部英文文献，并且附有全书索引，便于读者进一步延伸阅读和检索。

作者的这些独具匠心的安排，不仅便于读者深入理解本书的内容，并可帮助读者把数学的概念应用到语言学的研究中去，正好满足了语言学工作者更新知识的迫切需要，是一本不可多得的优秀读物。

由于篇幅的限制，本书不可能把目前在语言学中所使用的全部数学方法都包揽无遗。例如，在统计自然语言处理中使用的概率和统计方法，在声学语音学中使用的信号处理和声波理论的数学方法，在机器翻译中使用的自动剖析方法等，本书都没有涉及。关于这些方面的内容，有兴趣的读者可以阅读另外的一些著作。