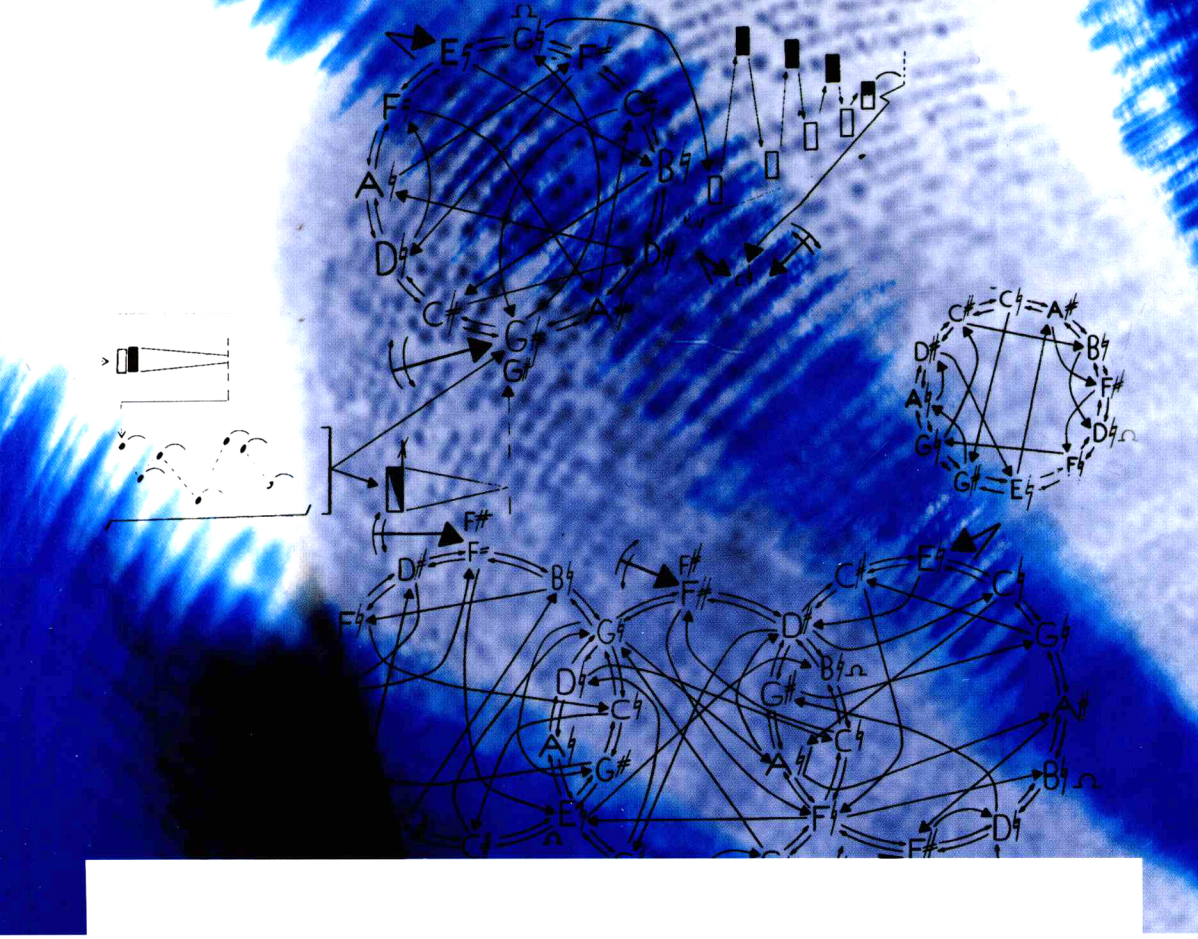

语言与数学

■冯志伟 著

世界图书出版公司

语言与数学

冯志伟　著

世界图书出版公司
北京·广州·上海·西安

图书在版编目(CIP)数据

语言与数学/冯志伟著. —北京：世界图书出版公司北京公司，2011.1
ISBN 978-7-5100-2806-9

Ⅰ. ①语… Ⅱ. ①冯… Ⅲ. ①数理语言学 Ⅳ. ①H087

中国版本图书馆 CIP 数据核字(2010)第 188787 号

语言与数学

著　　者：冯志伟
责任编辑：王晓燕
封面设计：然则设计公司

出　　版：世界图书出版公司北京公司
发　　行：世界图书出版公司北京公司
（北京朝内大街 137 号　邮编：100010　电话：64077922）
销　　售：各地新华书店及外文书店
印　　刷：北京高岭印刷有限公司

开　　本：787mm×1092mm　1/16　　**印张**：14.5
字　　数：240 千
版　　次：2011 年 1 月第 1 版　2011 年 1 月第 1 次印刷
书　　号：ISBN 978-7-5100-2806-9/H・1154
定　　价：38.00 元

信息时代语言学研究的基础

——读《语言与数学》有感

刘海涛
（中国传媒大学教授）

人是一种语言的动物，这是他与其他动物相比最大的不同之处。语言作为人类知识和信息的载体，对人类的发展和进步起到了无可比拟的作用。我们可以毫不夸张地说，没有语言就不会有人类的今天。正是语言的这种与人类的密切相关性和重要性，有史以来人类从未放弃过对语言本身的研究和探讨。综观人类对语言的研究可以发现，我们对于语言的认识和研究的深度是与社会的发展密切相关的，是与人类对整个世界的认识息息相关的。

众多的事实表明，我们目前处于一个信息和知识趋于“爆炸”的时代，大量信息的出现使得我们不得不寻找能够快速处理它们的技术和方法。由于对人而言，信息的主要载体可能就是语言，所以我们的研究重点也就成了寻求高效适宜的语言处理技术和装置。计算机的出现加快了定性信息和数据的处理。在语言信息处理方面，由于人类语言的模糊性、离散性及其他特性，而极大地限制了计算机在此领域中的应用。如果我们希望计算机能够进一步扩展与延伸人类的大脑，我们就必须研究怎样让计算机懂得人类的语言、能够处理人类的语言。于是研究语言不仅仅考虑人类，而且也应该顾及到机器，就成了信息时代语言研究的一大特点。这一点已被越来越多的语言学家和计算机专家所认同。

遗憾的是，语言作为人类所特有的现象具有许多特殊和繁复之处，其中最重要的就是语言的不规则性和模糊性。计算机作为一种定性的机器要处理语言材料则必须首先懂得语言的结构及其他特征，这就要求人们能把语言的结构和其他所需的材料精确地改写成计算机可以理解的程序和数据结构。数学是计算机科学，特别是软件理论和实践的重要基础。有鉴于此，研究语言学的数学化，或者说，从数学的观点去探索、研究自索绪尔

以来现代语言学的观点和理论应该是当代语言学家的一项重要任务。

近日读到冯志伟教授所著《语言与数学》一书，收获颇多，觉得有必要向语言学界引介。此书最早于1991年出版，原名《数学与语言》。由于当时这本书是作为《数学·我们·数学》丛书之一出版的，属于数学方面的著作，加之印量只有1400册，所以在语言学界很少有人读到此书。一个偶然的机会，笔者发现了此书。通读全书之后，我认为虽然本书是以“数学”丛书之一的面貌问世的，但实质是一位严肃的语言学家从数学的角度出发对索绪尔所建立的现代语言学理论的反思和发展，其目的在于建立一种信息时代的语言观。

作者冯志伟教授研究现代语言学和计算语言学多年，著述颇丰。《语言与数学》从索绪尔关于语言符号具有的两个重要特性（符号的任意性和能指的线条性）出发，进一步指出由于索绪尔所处时代的局限，他是无法提出那些只有在信息时代语言符号才能显现出的特性的。与任何事物一样，语言学也是不断发展的，信息时代的语言研究决不可能仅仅停留在索绪尔理论的框框里，而应该结合计算机处理语言的特点进一步发展索绪尔的理论，只有这样才能使语言学这门古老的学科焕发青春，成为真正意义上的“领先学科”。

本着发展索绪尔理论的思想，《语言与数学》作者结合计算语言学和现代数学等学科的新成果、新理论，重新审视语言这一极为复杂的符号系统，提出语言符号除了索绪尔提出的两大特征外，还具有以下几大特点：

① 语言符号的随机性：语言符号的出现和分布规律不是完全确定的，具有随机性，这一特性使得语言与统计数学发生了联系。

② 语言符号的冗余性：语言符号之间彼此制约，使得我们可以根据前后符号的关系来判断有关语言符号的性能，这样语言符号就显示出冗余性，这一特性使得语言与信息论发生了联系。

③ 语言符号的离散性：语言符号是由一些离散的单元组成的，具有离散性，这一特点使得语言与集合论发生了联系。

④ 语言符号的递归性：语言符号可以反复地使用有限的规则构成无限的句子，具有递归性，这一特点使得语言与公理化方法发生了联系。

⑤ 语言符号的层次性：语言的句子并不是由各个单词依前后的线性顺序排列而成的简单的线性序列，而是一个层次的立体性结构，具有层次性。每一个句子的线性序列的表层之下，都隐藏着一个层次分明的树形图。这一特点使得语言和图论发生了联系。

⑥ 语言符号的非单元性：语言符号并不是一个无结构的单元性符号，

而是一个有结构的、有多个复杂特征的非单元性符号，具有非单元性。这一特点使得语言与数理逻辑的许多演算方法发生了联系。

⑦ 语言符号的模糊性：语法符号中普遍存在着模糊现象。这一特点使得语言和模糊数学发生了联系。

作者围绕语言符号的这“七大”特点展开有理有据的分析、讨论，这些内容既有别人的一些研究成果，更多的则是作者多年来潜心研究的结果。

限于篇幅，本文不可能对作者提出的观点一一进行评介，下面就书中所提出的语言的公理化和语言的随机性谈一些自己的看法。

控制论的创立者维纳在其《人有人的用处》一书中指出：“人对语言的兴趣似乎是一种天生的对编码和译码的兴趣，它看来是在人的任何兴趣中最近乎人所独有的。言语是人的最大兴趣，也是人的最突出的成就。”维纳的这一段话，再次说明了语言对于人类的重要性和人类对于语言研究的浓厚兴趣。维纳将语言与编码、译码相提并论，说明了对语言用数学方法进行分析、处理的可能性。

德国著名语言学家洪堡特认为，语言是“用递归手段生成的系统，生成的法则是固定不变的，而生成的范围以及使用的具体办法则是完全没有一定的”，简言之，语言是“有限手段的无限应用”。正是在此基础上，乔姆斯基提出了著名的生成语法。如果洪堡特和乔姆斯基对于语言的理解是正确的，那么对于语言的描述就是尽可能抽取出一个符合内在规律的规则系统，选定了一种语言中特有的规则系统元素，就等于确定了这一种人类语言。通过规则来描述语言使得语言具有了一定的可计算性和可操作性，从而导致了数理语言学和计算语言学的产生。而数理语言学和计算语言学又可以通过数学方法去“仿真”人类的某些智力行为，这对于推动人类心智的研究有极大的意义。由此可以看出数学方法和理论对于语言研究的重要性。有鉴于此，《语言与数学》作者专辟一章“语言符号的递归性与公理化方法”较为详尽地研究了语言的这一重要特性，也可以说，语言的公理化特质是现代数理语言学和计算语言学得以发展的基础。

递归的定义与归纳证明具有相似的逻辑结构，它们均是由预先给定的有限数目的命题出发，反复运用一套特定的规则，推导出无限数目的外加命题。这种逻辑结构与数学中的公理方法极其相似，假定的起始命题为公理，外加命题叫定理，定理可由公理以及前面已推出的定理反复运用推理规则进一步推导出来。公理的集合、推理规则的集合，以及用来写这些集合的字母表，构成了一个公理系统。由此看来，递归定义很像一个公理系统，其中基底类似于公理，递归步骤类似于推理规则，这样的递归定义所

刻画出的集合的元，除了那些有基底给出的元之外，就相当于公理系统的定理。《语言与数学》作者认为："上下文无关文法与数学中的半图厄系统存在等价关系，而半图厄系统是一种特殊的扩展公理系统。"

既然如此，如何使用有限的手段来描述无限的语言，便成为理解与生成人类语言最基本的理论和实际问题。换言之，如何选定数目有限的规则，来生成无限的语言，是语言研究的一个基本问题。通过公理化的方法来研究人类语言，可以更精密地揭示人类处理语言的机制。公理化的语言描述也使得机器"仿真"人类的语言理解行为成为可能。就目前而言，得益于用公理化来分析语言的领域要算是计算语言学了，而数理语言学是计算语言学的理论基础。

我们认为，自然语言的这种可使用公理来描述的特点，体现了自然语言的可计算性。为使计算机可以懂得如何处理自然语言，必须赋予它足够的语言规则。所谓语言规则就是把语言由概念组合成判断和推理的规律和方法。语言规则可分为三大类：人类语言共有的规则，某种语言所特有的规则和某词所特有的规则。人类语言共有的规则可称作公理规则，它是由特定语言理论或从先验中归纳、概括出的描写、制约语言表达和语言结构的抽象原则的总和。公理规则是语言理论的指导思想，它制约着其他两类规则的研究和概括，又影响着算法的建立和优化。

当然，作者也指出公理化方法的不足，主要是由于任何公理系统都是一个封闭的自足的系统，一般而言，它的适用范围只是在一个句子内。毫无疑问的是，自然语言的公理化描述极大地有利于语言的形式化描述，而形式化又是语言自动处理的必要一步。理论上讲，公理化的句子生成技术和理论，也可进一步扩展到比句子更高的层次（如篇章）上。事实上，近年来已有学者提出基于篇章的自然语言生成理论和方法。这更说明递归和公理化特性是语言的基本特质之一。

语言的随机性和模糊性，是阻碍定性的计算机处理人类语言的最大障碍。关于语言的这两大特性，索绪尔也没有谈到。《语言与数学》作者根据自己多年来从事自然语言计算机处理的经验，对此问题作了深入的分析。更为可贵的是，采用数理方法的描述使得在计算机上处理语言的模糊性和随机性成为可能。

从信息论和控制论的角度看，语言是人类之间、人机之间传递信息的工具。理论语言学研究的重点为人类语言的一般特性，而计算语言学研究的是人机之间采用自然语言的交流问题。计算语言学也可看做是研究通过计算机等机器去实现人类语言处理机制的仿真科学。定性与不定性之间的

矛盾使得计算机在处理人的语言时遇到了难以跨越的障碍。

我们说语言在各个层次充满了模糊和不确定，指的正是语言语义方面的问题，因为语言的句法方面是不难通过形式化的方法来描述的，乔姆斯基的生成语法理论和泰尼埃的依存语法理论都是这一方面的杰出代表。就语言符号的理解和分析而言，内容是一种隐含的、模糊的东西，它只有通过形式才能被感知。不幸正在于此，一种形式可能表示多种内容，这就是计算机在处理自然语言时遇到的最大难题——歧义问题。

歧义问题是自然语言处理系统面临的最大难题，在自然语言的各个层次都存在着这个问题，由于层次的不同，歧义可以分为词汇歧义和结构歧义，词汇歧义属于词汇语义学的研究范畴，而结构歧义属于语法或句法语义学的课题。按照语言自动处理的观点看，所谓歧义就是一个字词串经过文法分析器后产生多个合格输出的现象，或者说歧义指的是同一句子可能有几个符合句法解释的现象。歧义作为影响语言正确理解的最大障碍，理所当然便成为计算语言学中语义研究的重点。对于语义的研究导致了计算语义学的产生，它的目的在于研究自然语言语义形式化的理论和方法，狭义说，计算语义学是将语义分析看做一种演算过程，它通过逻辑的方法处理语义问题，广义上讲，计算语义学研究利用计算机来处理和仿真人类语义处理机制的方法和理论，特别是歧义问题的处理和消解。我在分析了逻辑和自然语言的关系、计划语言学及众多其他相关学科已有的研究成果之后，认为计算语义学的广义概念可能更适宜于大规模、真实文本处理的应用。

对于有歧义的语句，理解的任务就是从多种结构中选出最适宜的和最可能的结构，注意我们在这里使用了“适宜”和“可能”这两个非绝对的词，目的在于说明在语言理解领域没有什么绝对的正确，而只有相对的“可能”。这些词汇的应用说明了语言的模糊性是多么的根深蒂固。

如果承认计算语义学研究是人类语义处理机制的仿真，那么分析人类对于语义的处理方式和消解过程可能是有益的。人类处理歧义问题的关键在于人的大脑中存有大量的知识，这些知识包括句法的、语义的和其他各类常识，利用这些知识人们可以很容易地理解对计算机来说有歧义的语句。与人一样，为了较完美地解决这个难题，计算机必然需要大量的各种知识。由于计算机与人有着极大的不同，知识需通过“显式”的方法表示出来，然而许多知识是模糊的、难于量化的，换言之，寻求适宜的、有效的知识表示方法是利用现有计算资源实现自然语言处理系统的唯一途径。理论上，我们不难把某些有关外部世界的知识授予计算机，难就难在世上

的知识是无穷尽的，而我们还不十分清楚，为了消除歧义，系统究竟需要什么样的知识。《语言与数学》一书所提出的语言符号的几大特质将有利于我们对语言进行知识表示和处理的研究。

语义的不可分解性和隐含性、歧义问题的复杂性、语言理解的无限性、语义的关联性、大规模真实文本处理的迫切性等，所有这些因素交织在一起，使得我们必须寻求新的语义处理方法和机制。

“歧义”是自然语言的特点之一，也是自然语言与其他人造符号体系的根本不同之处。但人人都有这样的经验，孤立地看有多个意义的词一旦被放到一个句子中，它的意义就很清楚了。有时为了正确地确定某个词的意义，人们甚至应将考虑的语境范围扩大到几个句子和段落。这说明在一篇文章中的某一个词（句子）和文章中的其他词（句子）有一种内在的联系，换言之，是一个词与句中其他元素的关系确定了它本身的意义。词在一定类型的上下文中只体现自己的某一种意义，它可以与具有一定意义的、数量不等的词构成伙伴关系，因此一个多义词有几种意义就可能有几种类型的上下文，有几种特点不同的组合联系。一个词具有的语境关系的总和便是我们所理解的著名哲学家维特根斯坦语言哲学中“意义即用法”的涵义。一个词的意义等于它的语境关系的总和，所谓“语境关系”就是一个词项在各种语境中遇到的全部正常关系。这一理解基本上是针对词汇语义而言的，它是我们建立非分解原则语义处理机制的语言哲学理据。

语言是人类用于交往的工具，在交往的过程中一般涉及接收者与发送者。接收者在交际过程中选择的词汇理解组合是在正常情况下最有可能的语境话语。在这里，“选择”和“可能”道出了语言理解的真谛。也就是说，在语义处理的过程中没有绝对的概念，人们理解一个句子的意义，只不过是因为它比其他意义更可能而已。而作出这种选择的主要根据就是语境或词在此时此地的用法。按照我们的理解，“词义”是蕴涵于它的语境之中的，词义是不能脱离它的语境关系来研究与讨论的。

自然语言处理本质上是一种人类语言处理能力的仿真。人类在处理语言的过程中更多地利用了类比和学习机制，我们根据维特根斯坦语言哲学“用法论”提出的“某一语言单位的意义即它的全部语境关系”的看法，是进一步建立类比语义处理机制的基础。语境关系是一个词在各种语境中所遇到的全部关系，具体来说，这些关系有句法关系和语义关系，如词的搭配关系、同现关系、支配关系等。我们认为，语义是隐含于它的所有语境关系中的。在语言理解和生成过程中，人们所利用的就是自己过去曾处理过的语例，并没有利用什么抽象的分析方法，他使用的只是一种基于类

比机制的方法来处理语义的。有时我们根据语境能判断出一个新词的意义，这说明利用语境关系确能推断出一个词的（可能）意义。表示知识和意义最好的工具就是自然语言，这是毫无疑问的。语言处理与别的计算机应用相比，有一个有利的条件，那就是存在大量的文献可供计算机作为语言处理的基础。

类比语义理论是目前计算语言学界基于经验的自然语言处理方法的基础，基于经验的语言处理方法（目前最有代表性的为“语料库语言学”）正是人们为了解决语言中的模糊性和随机性而诞生的，它的理论基础是统计数学。《语言与数学》作者在书中所提出的语言的随机性和模糊性，再次成为信息时代语言研究的基础。他所提出的许多数学方法极其有利于语言的计算机处理。

统观全书，我们认为《语言与数学》立论新颖、论述严谨，在语言符号的基本特性方面，进一步发展了自索绪尔以来现代语言学中的某些认识和看法，建立了一个以精密为特点的信息时代语言学研究的理论框架。这些理论和语言符号新特点的提出，对于普通语言学、计算语言学及语言学的其他分支都有积极的意义。

我们认为，此书探索、研究、发展了现代语言学中的一些基本问题，对于语言学诸分支都有较大的意义，值得每一位语言学工作者研读，希望本书能尽快再版，使书中所提出的重要思想为世人所知，并进一步发扬光大。

参考文献

1. 冯志伟. 1985. 数理语言学. 北京：知识出版社.
2. 冯志伟. 2004. 机器翻译研究. 北京：中国对外翻译出版公司.
3. 冯志伟. 1996. 自然语言的计算机处理. 上海：上海外语教育出版社.
4. 洪堡特. 1997. 论人类语言结构的差异及其对人类精神发展的影响. 北京：商务印书馆.
5. 刘海涛. 1993. 维特根斯坦语言哲学对计算语义学的影响. 见《计算语言学研究与应用》. 北京：北京语言学院出版社.
6. 刘海涛. 1997. 基于类比的计算语义处理机制. 见《语言工程》. 北京：清华大学出版社.
7. 乔姆斯基. 1992. 乔姆斯基语言哲学文选. 北京：商务印书馆.
8. N. 维纳. 1978. 人有人的用处. 北京：商务印书馆.
9. 伍铁平. 1994. 语言学是一门领先的科学. 北京：北京语言学院出版社.

再版前言

冯志伟

1989年，当时的北京大学校长丁石孙教授决定出版一套《数学·我们·数学》的丛书，内容包括“数学与经济”、“数学与军事”、“数学与教育”、“数学与语言”等许多方面，丁石孙校长委托北京大学数学系马希文教授找我，希望我写一本《数学与语言》。我是一个语言学工作者，当时正从事机器翻译和信息检索等应用系统的开发研究，接触到不少语言学中的数学方法问题，对于数学与语言之间关系的问题作过一些思考，因此，我欣然接受了丁石孙教授的这个任务，历时两年，写成了这本《数学与语言》专著，于1991年出版。

在这本专著中，我从数学的角度，对于自然语言的性质进行了深入的思考，明确地指出，除了索绪尔过去提出的语言符号的任意性之外，语言符号还具有另外7个明显的特性，它们是：语言符号的随机性、语言符号的冗余性、语言符号的离散性、语言符号的递归性、语言符号的层次性、语言符号的非单元性、语言符号的模糊性。我提出的语言符号的这7个新的特性，显然补充了索绪尔关于语言符号任意性的思想，使我们对于语言符号的特性有了更加深刻的认识。

我在本书中提出语言符号的这7个新的特性之后，引起了许多语言学家的关注，有的语言学家指出，语言符号的这7个特性，反映了“信息时代的语言观”。语言学家刘海涛在本书的书评中指出，语言符号的这7个特性应当成为“信息时代语言学研究的基础”，并指出，本书“值得每一位语言工作者研读”。19年过去了，我提出的语言符号这7个特性的“语言观”仍然显得很有生命力，越来越多的事实将会继续证明这种“语言观”的正确性。

本书在1991年出版时只印了1400册，早已销售一空。不少读者尽管听说过此书，但是，踏破铁鞋无觅处，根本买不到此书。有的读者直接写信给我，希望我能够帮助他们买到此书，由于市场上已经没有，我只好把

自己仅存的几本藏书送给他们，现在，我的藏书也送完了。世界图书出版公司敏锐地了解到这种情况，决定再版此书，以满足广大读者对于本书的需求。

本书是在19年前写成的，书中的个别例子反映了19年前的社会思潮和时代背景，本书再版时仍然保留了这些例子，以反映当时的真实情况和原书的风貌。

19年来，自然语言的数学研究有了很大的进展，统计方法在语言学中得到更多的应用，出现了隐马尔可夫模型、最大熵模型等新的统计模型，统计机器翻译得到了长足的发展，信息抽取、文本分类、文本数据挖掘、自动问答、人机自然语言接口等技术在信息化社会中发挥着越来越大的作用，汉字和汉语词语的统计也有了一些新的数据，对于这些新的发展，我们在世界图书出版公司将要出版的《自然语言处理简明教程》一书中详细论述，本书也就不再介绍了。

世界图书出版公司曾经在2009年出版过《语言学中的数学方法》一书的英文本，深入浅出地介绍了语言学中的各种数学方法，对于数学方法有兴趣的读者可以阅读此书。本书中对于这些数学方法也不作详细介绍。

感谢世界图书出版公司再版此书。希望此书的再版，能够吸引更多的语言学工作者来关心语言学中的数学问题。

2010年2月23日 于德国 Wiesloch 乡间

目　录

绪 言

——语言学是数学和人文科学之间的桥梁

法国数学家阿达玛（J. Hadamard）曾经说过："语言学是数学和人文科学之间的桥梁。"阿达玛不愧是一位有远见卓识的学者，他清楚地看出了语言学在人文科学中是最容易与数学建立联系的。

然而，在科学发展史上，人们是经过了相当长的过程才认识到语言学和数学之间的这种亲密关系的。

传统语言学的目的在于规定正确的读和写的种种规则，这样的语言学有点像法律。历史语言学用谱系树的方法研究语言的亲属关系，明显地受到进化论思想的影响，这样的语言学一如生物学。结构语言学着力于研究语言结构，力图找出语言中各种要素之间的结构规律，这样的语言学则似化学。

语言学和数学都是有着相当长历史的古老学科。语言学历来被看做典型的人文科学，数学则被许多人看做是最重要的自然科学。在学校教育中，语文和数学被认为是两门最基础的学科，成为任何一个受教育者的必修课。它们似乎成了学校教育中的两个极点：一个极点是作为文科代表者的语文，另一个极点是作为理科代表者的数学。很少有人会想到，这两门表面上如此不同的学科之间还有着深刻的内在联系。

直到 19 世纪中叶，才有人提出用数学方法来研究语言现象的想法。1847 年，俄国数学家布里亚柯夫斯基（В. Я. Буляковский）认为可以用概率论进行语法、词源及语言历史比较的研究。1894 年，瑞士语言学家索绪尔（De Saussure）指出，"在基本性质方面，语言中的量和量之间的关系可以用数学公式有规律地表达出来"，后来，他在其名著《普通语言学教程》（1916 年）中又指出，语言学好比一个几何系统，"它可以归结为一些待证的定理"。1904 年，波兰语言学家博杜恩·德·库尔特内（Baudouin de Courtenay）认为，语言学家不仅应该掌握初等数学，而且还有必要掌握高等数学。他表示坚信，语言学将日益接近精密科学，语言学将根据数

学的模式，一方面“更多地扩展量的概念”，一方面“将发展新的演绎思想的方法”。1933 年，美国语言学家布龙菲尔德（L. Bloomfield）提出了一个著名的论点：“数学不过是语言所能达到的最高境界。”

当时，学者们不仅仅只是提出这些颇具新意的想法，还有许多学者用数学方法对语言进行了实际的研究。1851 年，英国数学家德·摩根（A. de Morgan）曾把词长作为文章风格的一个特征进行过统计研究。1867 年，苏格兰学者坎贝尔（L. Campbell）用统计方法来确定柏拉图著作的执笔时期。1881 年，德国学者迪丁贝尔格（W. Dittinberger）进一步用统计方法把柏拉图著作的执笔时期分为前期、中期和后期三个阶段。1887 年，美国学者门登荷尔（T. C. Mendenhall）对不同时期的英国文学著作进行过统计分析，特别是研究了莎士比亚的作品。1898 年，德国学者凯定（F. W. Kaeding）编制了世界上第一部频度词典《德语频度词典》，用以改进速记的方法。1913 年，俄国数学家马尔可夫（A. A. Марков）研究了普希金叙事长诗《欧根·奥涅金》中俄语字母序列的生成问题，提出了马尔可夫随机过程论。1925 年，我国教育家陈鹤琴发表了第一部汉字频率统计的著作《语体文应用字汇》。1935 年，美国语文学家齐夫（G. K. Zipf）发表了齐夫定律。同年，加拿大学者贝克（E. Varder Beke）提出了词的分布率的概念，认为词典选词时，应以分布率为主要标准，频度为辅助标准。1944 年，英国数学家尤勒（G. U. Yule）发表了《文学词语的统计分析》一书，大规模地使用概率和统计方法来研究语言。

然而，上述各种用数学方法来研究语言的想法和具体的工作，都没有对当时的语言学研究发生显著的影响。这主要是由当时的社会实践的要求决定的。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。在这样的实践要求下，语言学没有多大的必要与数学接近。当然，上述各种研究中不乏卓越的工作。例如，马尔可夫在研究俄语字母序列的数学研究中，提出了马尔可夫随机过程论，后来成了一个独立的数学分支，对现代数学的发展产生了深远的影响。语言结构中所蕴藏着的数学规律，成了马尔可夫创造性思想的源泉。可惜的是，马尔可夫这一卓越的成就，在语言学界却鲜为人知。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。

第二次世界大战以来，由于科学技术突飞猛进的发展，科技文献的数量迅速增加，其增长速度十年翻一番。据联合国经济合作与发展组织估计，从 1960 年到 1985 年，世界情报量增加了 10～16 倍。全世界发行的图书总数是：1952 年约 25 万种，1962 年近 40 万种，1972 年约 56 万种，1980

年达到70万种。现在，世界上出版的科技刊物达16.5万种，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。科技文献的这种增长情况被形容为“情报爆炸”。面对浩如烟海的科技文献，研究人员为了取得全面而准确的科技情报，不得不花费大量的人力、物力、财力来做难以数计的翻译工作和检索工作，犹如大海捞针，严重地影响了科研工作的效率。

1946年第一台电子计算机问世后，人们开始考虑把这些繁重的工作交给计算机去做，这就提出了机器翻译、机器自动做文摘、机器自动检索科技文献等自然语言信息处理的问题。

在用计算机进行自动翻译的时候，必须进行原语词法、句法和语义的自动分析以及译语句法和词法的自动生成。这就首先要把这些问题用数学的语言加以描述，从而建立语言的数学模型。

在用计算机自动做文摘和检索时，要求把科技文献的信息储存在计算机中，建立数据库。数据库可以按照人们的要求，在其所储存的信息范围内，对人们提出的词题自动地作出回答。这种数据库中用以存储信息的语言，在内容上应该是严格的、精确的，在形式上应该适于数据库储存形式的要求，这当然也要求用精密的数学方法对自然语言进行描述。

由于自动化技术和计算技术的发展，人们正迅速地解决生产过程自动化问题，用自然语言来进行“人机对话”，让电子计算机理解自然语言，这就要用数学方法来研究句法结构和语义结构的形式化表达方式以及知识的形式表示技术。

目前微型计算机已经普及，在办公室的事务管理中得到了广泛的使用，这就是“办公室自动化”问题。自动化的办公室要用微型计算机来编辑和处理各种书面文件，这就要求对语言文字进行严格的形式化的描述。

另外，通讯技术的发展，要求为负荷信息的语言寻找最佳编码方法，要求提高信道的传输能力，以便在保持意义不变的前提下，最大限度地压缩所传输的文句，在单位时间内传输最多的信息，这就要求对语言的统计特性进行精密的研究。

在上述各种促使语言学与数学接近的因素中，最为关键的因素是电子计算机的出现。电子计算机是一种信息处理机，而自然语言是信息最主要的载体，电子计算机的研制和发展离不开自然语言的信息处理，而自然语言的信息处理离不开数学。语言学家必须采用数学思想和数学方法来研究自然语言，才能回答信息化时代对语言学提出的严峻挑战。