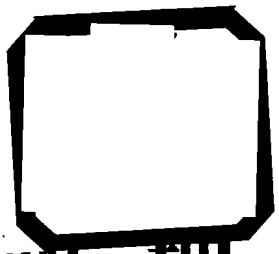
冯志伟 胡凤国 著

数理语言学

（增订本）

SHULI YUYANXUE

商務印書館



数理语言学

（增订本）

冯志伟 胡凤国 著

商務印書館 The Commercial Press
创于1897
2012年·北京

图书在版编目(CIP)数据

数理语言学/冯志伟，胡凤国著. —增订本. —
北京：商务印书馆，2012
ISBN 978-7-100-08391-1

I. ①数… II. ①冯…②胡… III. ①数理语言学
IV. ①H087

中国版本图书馆 CIP 数据核字(2011)第 104454 号

数 理 语 言 学

(增订本)

冯志伟 胡凤国 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
北京市松源印刷有限公司印刷
ISBN 978-7-100-08391-1

2012年4月第1版 开本 880×1230 1/32
2012年4月北京第1次印刷 印张 $16\frac{1}{8}$
定价：35.00元

内 容 提 要

数理语言学是用数学思想和数学方法研究语言现象的一门新兴的边缘性学科，它的出现使语言学与现代数学、计算机科学、控制论以及人工智能等学科有了关联，并逐渐地走上了现代化的道路。数理语言学主要包括代数语言学、统计语言学、应用数理语言学三个部分，本书系统地、全面地、深入浅出地介绍了这三个部分的基本知识和最新成就。为了便于文科读者透彻地理解本书内容，本书专门辟出一章讲述语言学中的离散数学方法。本书可作为数理语言学的入门教材，著者在写作时尽量考虑到跨学科读者的需要，既可供想了解这门新兴边缘学科而数学准备不够的语言学工作者和其他文科读者阅读，亦可供要求了解语言学方面的现代化知识的理工科读者阅读。

前　言

数理语言学(mathematical linguistics)是用数学思想和数学方法来研究语言现象的一门新兴的语言学科。这门新兴学科的出现，使得语言学的研究与现代数学、计算机科学、控制论以及人工智能等学科发生了密切的联系，并逐渐走上了现代化的道路。

语言学和数学都是有相当长历史的古老学科。语言学历来被看成是典型的人文科学，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，是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存在着深刻的内在联系。

但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。他们开始从计算的角度来研究语言现象，揭示语言的数学面貌。

1847 年，俄国数学家 В. Я. Буляковский(Buljakovski，布里亚柯夫斯基)认为可以用概率论方法来进行语法、词源和语言历史比较的研究。

1851 年，英国数学家 A. De Morgen(摩根)把词长作为文章风格的一个特征进行统计研究。

1894 年，瑞士语言学家 De Saussure(索绪尔)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，他在 1916 年出版的《普通语言学教程》中又指出，语言好比一个几何系统，它可以归结为一些待证的定理。

1898 年，德国学者 F. W. Kaeding(凯定)统计了德语词汇在文本中的出现频率，编制了世界上第一部频率词典《德语频率词典》。

1904 年，波兰语言学家 Baudouin de Courtenay(博顿·德·古尔特内)指出，语言学家不仅应当掌握初等数学，而且还要掌握高等数学，他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，更多地扩展量的概念，发展新的演绎思想的方法。

1913 年，俄国数学家 A. A. Марков(Markov，马尔可夫)用概率论方法研究过普希金叙事长诗《欧根·奥涅金》中的俄语元音和辅音字母的序列，并在此基础上提出了马尔可夫随机过程论。

1933 年，美国语言学家 L. Bloomfield(布龙菲尔德)提出了一个著名的论点："数学只不过是语言所能达到的最高境界。"

1935 年，美国语文学家 George Kingsley Zipf(齐夫)研究频率词典中单词的序号与频率的关系，提出了齐夫定律(Zipf's law)。

1935 年，加拿大学者 E. Varder Beke(贝克)提出了词的分布率(range)的概念，并以之作为词典选词的主要标准。

1944 年，英国数学家 G. U. Yule(尤勒)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的方法来研究词汇。

上述这些事实说明，关于语言计算的思想和研究是源远流长的。可惜的是，这些独具慧眼的思想和别具一格的研究，都没有对语言学本身产生显著的影响。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，语言学还没有与数学建立

直接的联系。

20世纪50年代以来，情况发生了巨大的、急剧的变化。1955年，美国哈佛大学首先创办了数理语言学讨论班，1957年正式开设了数理语言学课程。接着，麻省理工学院、密歇根大学、宾夕法尼亚大学、印第安纳大学、加利福尼亚大学都相继开设了数理语言学课程。同年，日本成立了计量语言学会，创办了数理语言学杂志《计量国语学》，德国的波恩大学也开设了数理语言学课程，前苏联在莫斯科大学、列宁格勒大学（现为圣彼得堡大学）及莫斯科国立第一外国语师范学院也进行了数理语言学的研究工作。1958年，莫斯科大学、高尔基大学、萨拉托夫大学、托姆斯克大学，分别给数学系及语文系的学生开设了数理语言学的选修课，并在列宁格勒大学（现为圣彼得堡大学）设置了数理语言学专业。

此外，罗马尼亚、匈牙利、捷克、英国、法国、挪威、波兰、瑞典等国，都先后开展了数理语言学的研究工作，有的国家还创办了专门的刊物，成立了专门的研究机构。

近年来，数理语言学成了语言学、数学、计算机科学、人工智能等学科所共同关注的重要领域。在有关上述学科的国际学术会议上，数理语言学经常是中心议题之一。

我国从20世纪50年代起便逐步开展了数理语言学的研究，在用数学方法研究汉语的句子结构、中文信息处理、言语统计等方面取得了一定成绩。

这一新兴学科不但引起了我国语言学界的重视，也引起了我国数学界的重视，在《数学辞海》中，有专门的一章是讲数理语言学的，可见我国数学界已经认识到数理语言学对于数学本身的价值。

为什么数理语言学会得到如此迅速的发展呢？我们可以从必要性和可能性两方面来分析这个问题。

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，在20世纪50年代初期，人们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(machine translation)、自动文摘(automatic summarization)以及自动信息检索(automatic information retrieval)等自然语言处理(natural language processing)问题。

在用计算机将一种语言A翻译为另一种语言B时，除了确定语言A中的每一个词在语言B中相应的等价物之外，还必须分析语言A的句子结构和语义结构，并把翻译出来的词作某种变化，按照语言B的结构把它们配置起来，这样，人们就得“教会”计算机自动地分析和综合句子。但是我们知道，任何一个问题要用计算机自动地来解决，首先就要使该问题所涉及的现象能够用数学语言来描述，也就是要把所考虑的问题“数学化”。所以，为了进行机器翻译，首先就要采用数学语言来描写语言现象，对传统语言学中的各种概念用数学的方法进行严格的分析，建立语言的数学模型。

在早期的机器自动文摘和自动信息检索中，要求把文献的信息储存在计算机中，计算机可以按照人们的要求，在其所储存的信息的范围内，对人们提出的问题自动地做出回答。在计算机中用以描述信息的语言，在内容上应该是严格的、精确的；在形式上应该适于计

算机储存形式的要求，这当然也要用精密的数学方法来研制。目前这些信息大部分是非结构化的，它们是以自然语言的形式存储的，语言是信息的主要载体，为了提高信息检索的查准率(precision)和查全率(recall)，需要使用数学方法对于负荷这些非结构化信息的语言进行形式化的描述。

由于自动化技术和计算技术的发展，人们正在迅速解决生产过程自动化问题，用自然语言来进行“人机对话”(man-machine dialogue)，让电子计算机能理解自然语言，这就要求将自然语言代码化和算法化，以便计算机自动地从自然语言的外部形态中，抽出它所表示的语义内容，并将计算机所理解到的语义内容，根据“人机对话”的要求，由计算机组织成相应的语句，回答人所提出的问题。

另外，由于通信技术的发展，要求给负载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，这就需要对语言的统计特性进行精密的研究。

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般地不断增加，出现了信息爆炸(information explosin)的情况。随着知识突飞猛进的增长，翻译市场供不应求的局面越来越严重了；由于无法消化大量从国际上传来的信息流，我们的信息不灵，就有可能使我们在国际竞争中失去大量的机会。在这种情况下，机器翻译、自动信息检索、自动文摘等自然语言处理的研究显得更加迫切，研究领域日益扩大，自然语言处理成为了当代语言学和计算机科学中最引人注意的新兴学科。

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迫在眉睫了。

在这些新的实践要求下，必须采用数学思想和数学方法来研究语言现象，在语言学中建立数理语言学这门新学科。

以上我们分析了建立数理语言学的必要性，那么建立这门新学科是否有可能呢？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。

从语言本身的性质来看，正如 De Saussure 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。音位学支配着语音的结合，形态学支配着构词和变词，句法学支配着词的组合，语义学支配着语义的组合。因此，我们在研究语言时，可以只管它的结构，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。这正如在下棋时，棋局的结构是重要的，而用木头的棋子或是用象牙的棋子则是无关紧要的一样。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。

科学技术当前的发展水平，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分支，为用数学的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，出现了 O. Jespersen(叶斯泊森)的“分析句法”，在结构语言学内，L. Bloomfield，Z. Harris(哈里斯)和 C. Hockett(霍凯特)等人提出了以“替换”(substitution)和“分布”(distribution)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，在其语言观方面可能有片面之处，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，由于采用了比过去语言学更加严格的精密方法，在某些方面，对于用数学思想和数学方法来研究语言也有一定的启示作用。

另外，20 世纪以来，控制论(cybernetics)逐渐成熟起来。控制论是研究机器与机器之间、人与人之间、人与机器之间的信息的传输、接收、储存、加工和利用的一门综合学科，而语言是人类最重要的交际工具，是信息的最主要的负荷者，对语言进行精密的研究，有助于控制论的发展，而控制论采用的一些方法，特别是模拟方法，也可以作为建立语言模型的借鉴。

近年来，计算机科学(computer science)发展迅速，语言学与计算机科学日益接近并互相渗透。计算机科学中使用的高级程序语言要尽量与人们的自然语言相接近，而其高级的程度，往往就是依这种程序语言与自然语言相接近的程度而定的，越接近自然语言就越高级。这样，计算机科学中对程序语言结构和编译技术的研究，就可以作为用数学思想和方法研究自然语言的参考。

目前，人工智能(artificiol intelligent)已经成为国内外科技界十分关注的一个领域。自然语言是人类最重要的一种智能，人工智能所探讨的有关人类智能活动的一般规律，对数理语言学的研究有着一般性的指导作用。

在上述各种因素的综合作用下，在 20 世纪 50 年代初期，作为一门独立学科的数理语言学便应运而生了。

数理语言学主要包括三个部门：(1)代数语言学(algebraic linguistics)，(2)统计语言学(statistical linguistics)，(3)应用数理语言学(applied mathematical linguistics)。

我们对数理语言学作这样分科的理论根据是瑞士著名语言学家 De Saussure 关于语言与言语区分的学说。De Saussure 在其名著

《普通语言学教程》中把语言现象分成言语行为(langage)、言语(parole)[①]和语言(langue)三样东西，他指出："语言是一种表示意念的符号系统"，而言语则是言语行为的过程(也就是交际过程)和言语行为的结果(也就是口头的或书面的言语作品)。"把语言和言语分开，一下子就把(1)什么是社会的，什么是个人的，(2)什么是主要的，什么是附属的和多少是偶然的分开来了。"

在 De Saussure 关于语言和言语的区分的理论影响下，美国语言学家 N. Chomsky(乔姆斯基)提出，必须把说具体语言的人对这种语言的内在知识和他具体使用语言的行为区别开来，并把前者叫作语言能力(competence)，把后者叫作语言运用(performance)。依我们看来，Chomsky 的语言能力，大致相当于 De Saussure 的语言，Chomsky 的语言运用，大致相当于 De Saussure 的言语。

正是从这样的观点出发，我们认为数理语言学的研究应该从语言的内部结构和语言的交际活动两方面来进行，也就是说，我们应该把数理语言学的研究首先分为对作为符号系统的语言的数学性质的研究和对作为交际活动的过程及结果的言语的数学性质的研究两个部分。

作为符号系统的语言，本质上是由一些离散的单元构成的，我们可以采用集合论、数理逻辑、算法理论、图论、格论等离散的、代数的方法来研究它，这方面的研究就叫作代数语言学(algebraic linguistics)。

在言语中，在用语言进行交际的活动中，有的语言成分使用得多些，有的语言成分使用得少些，各语言成分的使用有一定的随机性，而交际过程本身，又是一个信息传输的过程，我们可以使用概率论、

---

① 我国有的语言学家把 parole 译为"言谈"，我们这里把它译为"言语"，是按大多数语言学家的译法。

数理统计和信息论等统计数学的方法来研究它，这方面的研究就叫作统计语言学(statistical linguistics)。

当然，在语言与言语、语言能力与语言运用之间也是有联系的。正如 De Saussure 所指出的："无疑地，这两个对象是紧密地联系着的，而且是互为前提的：要使言语让人听得懂，产生它的效果，必须有语言；要使语言能够建立起来，也必须有言语。"因此，在代数语言学和统计语言学之间也是有联系的：我们要研究作为符号系统的语言的数学性质，就要注意到各语言成分的统计特征，而在对言语作统计研究时，也必须考虑到整个语言符号系统的总体。

代数语言学和统计语言学都是数理语言学中的理论性部门，把这两个部门的研究成果应用于自动翻译、人机对话以及自动信息检索的实际工作中，还有许多十分具体的技巧和方法需要人们进行深入的研究，这方面的研究，便构成了数理语言学的第三个部门——应用数理语言学(applied mathematical linguistics)。

应该承认，数理语言学，正如其他语言学方法一样，并不是万能的，它还在发展之中，还不很完备。近年来，美国数学家 L. A. Zadeh (查德)认真地研究了语言中的模糊现象，并在此基础上提出了模糊数学(fuzzy mathematics)的基本理论。或许模糊数学的发展和完善能为数理语言学提供一些新的工具，因为语言现象是复杂的，不可能处处都作精细入微的描述。尽管如此，数理语言学的重要意义是不容忽视的。本书的内容只限于代数语言学、统计语言学和应用数理语言学三个分支，至于模糊数学方法在语言研究中的应用，由于本书篇幅的限制，只好略而不谈了[①]。

---

① 参看钱锋、冯志伟，"试论模糊数学在方言研究中的应用"，《华东师范大学学报》，1983 年第 4 期 p. 27—33。

本书将分章介绍数理语言学的这三个部门。为了便于数学准备不够的广大语文工作者理解本书内容，我们专门辟出一章来讲语言学中的离散数学方法。这样一来，只要具备中学数学程度的读者，在数学方面就不会再有什么困难了。

数理语言学和计量语言学(Quantitative Linguistics)有着密切的关系。计量语言学这个术语在英国统计学家 Gustav Herdan(赫丹)1964 年出版的《计量语言学》(Quantitative Linguistics，1964)一书中就出现了，而计算语言学(Computational Linguistics)这个术语是 1965 年才正式在正规的学术出版物中出现的，计量语言学在语言学中的资格比计算语言学还老①。Gustav Herdan(赫丹)还出版过《语言作为选择和机会的理论》(Advanced Theory of Language as Choice and Chance，1966)一书，这是计量语言学的奠基性著作。

近年来，计量语言学得到了快速的发展，其主要代表人物都是来自德国和东欧地区的，其中最著名的是已退休的德国波鸿大学(Bochum University)Gabriel Altmann(阿尔特曼)教授，他在计量语言学的诸多领域均有重要贡献，被誉为 Zipf 之后最重要的计量语言学家。另外一位著名的计量语言学家是德国特里尔大学(Trier University)的 Reinhard Köhler(柯勒)教授，他对计量语言学最大的贡献是提出了“协同语言学”(synergetic linguistics)的理论。

我们认为，计量语言学与本书所说的“数理语言学”有着密切的联系，它是“统计语言学”的进一步拓展。我们在研究数理语言学的时候，应该密切关注计量语言学的发展。

本书中使用了一些必需的数学符号，希望读者不要讨厌这些

---

① 1965 年，美国的 Machine Translation 杂志改名为 Machine Translation and Computational Linguistics，在杂志封面上首次出现了 Computational Linguistics 这个术语。

符号，只要你细心揣摩它，体会它的细微含义，你就会发现，这样的数学符号比之于文字叙述要简明得多。作者在1982年给北京大学中文系汉语专业的学生讲授《语言学中的数学问题》这门新开的选修课时，曾讲过本书的部分内容，同学们产生了很大的兴趣，成绩良好，足见学习语言学的文科学生是完全有可能掌握数理语言学知识的。

本书第一版在编写过程中，曾得到国家语言文字工作委员会倪海曙、中国社会科学院语言研究所刘涌泉、上海人民出版社胡道静、北京大学中文系叶蜚声、中国银行电脑部姚兆炜、中国科学院上海冶金研究所谢雷鸣、华东师范大学钱锋、汉语大词典出版社徐文堪等同志的关心和帮助。初稿完成后，又承刘涌泉、姚兆炜、谢雷鸣等同志通读全稿，提出了很好的修改意见。作者谨向他们表示衷心的谢意！本书参考过多种国内外时贤著作，在每章末已列出了主要的参考文献，特在这里一并致谢。

本书第一版是在1985年出版的，这是我国第一本数理语言学的专著，由于它内容新颖、深入浅出，受到了读者的欢迎，现在我国自然语言处理界的一些专家，正是在20多年之前读了本书以后，对于自然语言的计算机处理发生了兴趣，开始走上了自然语言处理的道路，而今，他们中的很多人，已经成为我国自然处理领域的中坚力量了。

本书第一版出版后的20多年中，科学技术的发展更加迅速，新的信息、新的知识更加迅猛地增长，信息爆炸。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。

随着信息技术的突飞猛进的进步和网络日新月异的发展，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。如何从包含海量信息的互联网上搜索到人们需要的信息，就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，需要我们使用数学方法来研究语言。

在互联网(web)出现之后，网络上的信息检索(information retrieval)、信息抽取(information extraction)、文本数据挖掘(text data mining)、统计机器翻译(statistical machine translation)成为网络信息处理的重要内容，为了从网络的海量数据中获取有用的信息和知识，统计方法有了长足的进步，数理语言学中的统计语言学这个分支更加成熟，数理语言学与互联网的联系也更加密切，数理语言学对于社会进步的重要性也更加明显了。

在这种情况下，本书的第一版已经难以满足社会的需求了。

商务印书馆周洪波先生敏锐地察觉到这种重大变化，建议我进一步增订本书的第一版，并且表示商务印书馆愿意出版本书。我欣然接受了周洪波先生的建议，在保持原书基本面貌的前提下，增加了树邻接语法、词汇增幅率研究、基于语料库的语言研究、信息检索中的文档处理、信息检索中的统计匹配技术、基于语料库的机器翻译等内容，形成了现在的这个增订本。

本书增订本曾在中国传媒大学应用语言学专业的“数理语言学”课程中作为教材试用，同学们反映良好，希望这个增订本的出版，能够继续受到广大读者的欢迎。

在本书增订过程中，中国传媒大学胡凤国老师根据同学们对于该校“数理语言学”课程教学的反映，提出了不少建设性的意见，事实上他成为了本书增订本的作者之一。

这是一本关于数理语言学基本理论和应用的著作，涉及的问题很多，作者水平有限，错误在所难免，敬希广大读者不吝赐教！

冯志伟

2010年6月15日于北京