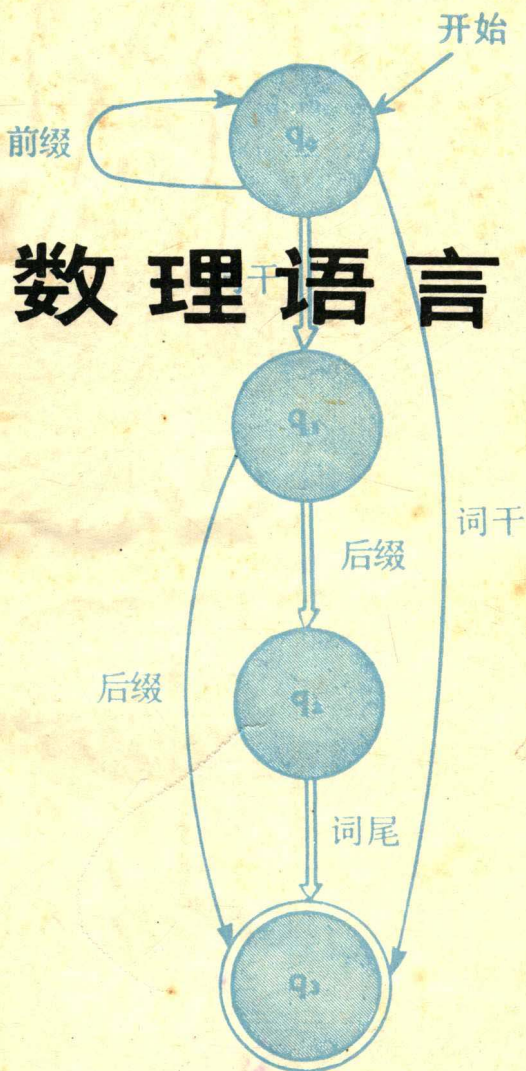


现代化知识文库



知 识 出 版 社



现代化知识文库

倪海曙 主编

数 理 语 言 学

冯志伟 编著

知 识 出 版 社

1985·8·上海



20-011
数学语言学
装帧设计 张苏予

现代化知识文库
数 理 语 言 学

Shuli Yuyanxue

冯志伟 编著

知识出版社出版发行

(上海古北路650号)

(沪 版)

新华书店上海发行所经销 上海东方印刷厂印刷

开本 850×1035 毫米 1/32 印张 8.375 插页 2 字数 270,000

1985年8月第1版 1987年10月第2次印刷

印数: 10,001—13,000

ISBN 7-5015-5177-4 / H · 21

定价: 1.70 元

总 序

社会主义现代化建设需要知识，需要在不断更新中的现代化知识。

人类的知识是不断发展、不断更新的。现代的社会，文化科学突飞猛进，人类知识的更新速度空前加快；假定 19 世纪的知识更新周期是 80~90 年，现在已缩短为 15 年，而某些领先学科更缩短为 5~10 年。知识体系不断更新，人的知识结构也必须不断更新，进学校求得适用一辈子的知识的“一次教育”已经成为陈旧的观念。这样，不断地进行更新知识的再学习，也就成为现代人生活和工作的需要。“活到老，学到老”这句格言有了新的含义。现在，好些国家已经在研究和推行“终身教育”，又称为“知识更新教育”，它的主要方法是提供对最新知识的深入浅出的介绍，以便自学。现代化的人才要由实行全面的终身教育来造就。

人类认识日新月异，各门科学的新分支层出不穷，边缘性、交叉性学科随着发展，形成了人类知识结构的综合化和整体化的新趋向。因此，现代化社会不仅需要“专才”，而更需要“通才”，也就是具有新的知识结构的科学人才。现在许多成就卓著的科学家，极少是只限于一门专业的，他们往往在边缘性、交叉性学科领域中以博识多才取胜。当然，一个人不可能通晓一切知识的细节；但是，如果知识深广，视野开

阔，就可以具有融会贯通、触类旁通的创造能力。我国的现代化事业正需要成千上万这样的通才。

《现代化知识文库》就是为了提供知识更新的学习材料而出版的。它将系统地、全面地、通俗地介绍从自然科学到社会科学各个部门的最新成就，特别是边缘性、交叉性学科的新进展以及它的难题和解决的方向。《文库》的有些内容在国内还是第一次作系统介绍，希望它的出版对正在探索科学文化新境界的读者有所帮助。

这套文库将不断补充新的选题，分辑出版，每辑 10 本。编著者大多是中年科研人员，由老一辈的著名科学家担任编审。从内容到文体都将按照客观情况的发展不断更新。

知识就是力量，我们的工作希望得到大家的支持和帮助。

《现代化知识文库》编辑部

1982 年 5 月

前 言

数理语言学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挂上了钩,逐渐走上了现代化的道路。

早在 19 世纪中叶,就有人提出过用数学来研究语言现象的想法。例如,1847 年,俄国数学家 В.Я.Буляковский (Buljakovski, 布里亚柯夫斯基)认为可以用概率论来进行语法、词源及语言历史比较的研究。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”,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 Baudouin de Courtenay (博顿·德·古尔特内)认为,语言学家不仅应该掌握初等数学,而且还要掌握高等数学。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 L.Bloomfield (布龙菲尔德)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,人们不仅仅提出了这些想法,并且还有人用数学方法对语言进行了实际的研究。例如,俄国数学家 А.А.Марков (Markov, 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的序列。不过,这种研究只是为了说明他的数学理论(马尔可夫随机过程论),与语言学本身并无直接关系,只是一种纯粹的数学研究。

所以,不论是 Буляковский, Saussure, Baudouin 和 Bloomfield 的想法和信念也好;还是 Марков 的实际研究也好,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,语言学还没有与数学建立直接的联系。

最近二十多年来,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》,西德的波恩大学也开设了数理语言学课程,苏联在莫斯科大学、列宁格勒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,并在列宁格勒大学设置了数理语言学专业。

此外,罗马尼亚、匈牙利、捷克斯洛伐克、英国、法国、挪威、德意志民主共和国、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50年代起便逐步开展了数理语言学的研究^①,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一定成绩。这一新兴学科已引起了我国语言和数学工作者的重视。

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。

本世纪以来,由于科学技术突飞猛进的发展,科技文献的数量与日俱增,其增长速度十年翻一番,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“情报爆炸”。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,取得科技情报,不得不花费大量的人力、物力来作难以数计的翻译工作,大大地影响了科研工作的效率。

电子计算机出现之后,在本世纪50年代初期,人们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,也就是将原始数据(原始信息)转换成结果数据(结果信息),这就提出了机器翻译、机器自动文摘以及机器自动检索科技文献等信息加工问题。

^① 早在1959年,刘涌泉同志就在《中国语文》1959年第6期上,发表了“谈数理语言学”一文,首次在国内介绍这一学科。据刘涌泉同志讲,在50年代初期,我国著名数理逻辑专家胡世华先生就注意到数理语言学的问题了。

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。

用机器自动做文摘和检索时,要求把科技文献的信息储存在机器中,这种储存信息的机器叫“信息机”。信息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作出回答。在这种信息机中用以储存信息的语言,在内容上应该是严格的、精确的,在形式上应该适于信息机储存形式的要求,这当然也要用精密的数学方法来加以描述。

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,争取在不远的将来,用自然语言来进行“人机对话”,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,变为电子计算机所能理解的形式,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,并将电子计算机所理解到的语义内容,根据“人机对话”的要求,由电子计算机组织成相应的语句,回答人所提出的问题。

另外,由于通讯技术的发展,要求对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,这就需要对话言的统计特性进行精密的研究。

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。

且是互为前提的：要使言谈让人听得懂，产生它的效果，必须有语言；要使语言能够建立起来，也必须有言谈。”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，而在对言谈作统计研究时，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、人机对话以及自动情报检索的实际工作中，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。

应该承认，数理语言学，正如其它语言学方法一样，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部门，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。

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。

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。但愿本书的出版，能吸引更多的语文工作者投入到数理语言学的研究中来，使语言学这门古老的学科尽快地现代化起来。

^① 有兴趣的读者可参看：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。

本书在编写过程中,曾得到中国文字改革委员会倪海曙、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、汉语大词典编纂处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感谢!本书参考过国内外时贤著作多种,在每章末已列出了主要的参考文献,特在这里一并致谢。

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

1983年元月1日于北京

目 录

前言	i
----------	---

第一章 离散数学与语言	1
-------------------	---

1. 逻辑与集合(1)

- 1.1 逻辑语言(1) 1.2 英语辅音和集合论的基本概念(9) 1.3 集合的几种记法,空集,通集(11) 1.4 对立与集合运算(13)

2. 函数与关系(17)

- 2.1 函数与关系的基本概念(17) 2.2 关系的性质(22) 2.3 等价关系(26) 2.4 偏序关系和全序关系(28)

3. 成分结构树 (30)

- 3.1 支配关系(32) 3.2 前于关系(33) 3.3 标记(34) 3.4 成分结构树的几个性质(35) 3.5 属于,句友和统率(35)

4. 代数结构 (37)

- 4.1 运算(37) 4.2 运算的性质(38) 4.3 包含运算的数学构形(42) 4.4 群和子群(42) 4.5 独异点和子独异点(45) 4.6 半群(47) 4.7 符号串(47) 4.8 毗连(48) 4.9 语言是词的集合上的自由独异点(49)

5. 递归 (50)

- 5.1 递归定义(50) 5.2 归纳(53) 5.3 公理系统(54) 5.4 扩展公理系统(56) 5.5 半 Thue 系统(58)

第二章 代数语言学	60
-----------------	----

1. 形式文法和转换文法 (60)

- 1.1 形式文法的概念(60) 1.2 形式文法的类型(64) 1.3 有限状态文法(66) 1.4 上下文自由文法(72) 1.5 上下文敏感文法和 0 型文法(85) 1.6 上下文自由文法在描写自然语言方面的缺陷(88) 1.7 转换文法的数学模型(89)

2. 自动机 (92)

- 2.1 自动机与文法的关系(92) 2.2 有限自动机(93) 2.3 图灵机(95) 2.4 后进先出自动机(98) 2.5 线性有界自动机(102) 2.6 自动机在自然语言处理中的作用及其局限(105)

3. 语言的集合论模型 (106)

- 3.1 集合论模型的三个初始概念(106) 3.2 B -划分(107) 3.3 B -结构(108) 3.4 B -等价(109) 3.5 型和词类(109) 3.6 简单语言(110) 3.7 B -构形(112) 3.8 名词的语法性(115) 3.9 语言集合论模型的效用(118)

4. 句法类型演算 (120)

- 4.1 句法类型(120) 4.2 演算规则(121) 4.3 动词短语的句法类型演算(125)

5. 蒙太格文法 (131)

- 5.1 蒙太格文法的两个来源(131) 5.2 深层结构与表层结构(131) 5.3 深层结构的内涵逻辑表达式(133) 5.4 蒙太格文法的语义理论(138)

第三章 统计语言学 140

1. 统计语言学的基本概念 (140)

- 1.1 必然事件和随机事件(140) 1.2 频率和概率(141) 1.3 抽样(144)

2. 语言成分的统计研究 (149)

- 2.1 频率词典的编写(149) 2.2 词的序号分布规律(151) 2.3 语言使用中的某些统计规律(157) 2.4 用计算机作词语统计研究(160)

3. 语言与信息 (164)

- 3.1 语言的使用是一个随机过程(164) 3.2 熵和多余度(169)

第四章 应用数理语言学 175

1. 信息语的设计 (175)

- 1.1 信息语的概念(175) 1.2 信息语的设计方法(175)

2. 语言自动处理方法的研究 (177)

- 2.1 预示分析法(177) 2.2 扩充转移网络(187) 2.3 通用句法生成器(191) 2.4 文句的自动生成方法(194)

3. 机器翻译 (199)

- 3.1 机器翻译概述(199) 3.2 英汉机器翻译系统(201) 3.3 汉一法、英、日、俄、德多语言机器翻译系统(204)

4. 自然语言理解 (214)

- 4.1 SAD-SAM 系统(214) 4.2 BASEBALL 系统(215) 4.3 SIR 系统(216) 4.4 STUDENT 系统(218) 4.5 LUNAR 系统(220) 4.6 SHRDLU 系统(221) 4.7 MARGIE 系统(224) 4.8 SAM 系统和 PAM 系统(229) 4.9 RJD-80 汉语人机对话系统(237)

结语	240
专门名词索引	243
外国人名索引	249

第一章 离散数学与语言

本章从语言学的角度,介绍了数理逻辑、集合论、函数、关系、树、代数结构、递归等离散数学的基本知识,每个数学问题都尽量结合语言学的例子来讲解。最后引出了公理系统、半 Thue 系统等概念,为第二章的主要内容形式文法作了准备。本章是学习以后各章的数学基础。

1. 逻辑与集合

1.1 逻辑语言

逻辑的基本单位是命题。命题在语言学上就是一个有意义的陈述句。

下面的句子都是命题:

“一切人都是要死的。”

“如果 $a < b$, 则 $b > a$ 。”

“所有的狗都是动物。”

“P 是一个双唇塞音。”

“在标准英语中没有喉塞音。”

“所有的猪都有翅膀。”

一个命题不一定是真的。它或为真,或为假,但不能又真又假。为简单起见,命题可用符号 $P, Q, R, \dots$ 来表示,也可以加下标。

我们可以用逻辑连词把两个命题构成更复杂的命题:

“P 和 Q”写为 $P \wedge Q$, 表示命题 P 和命题 Q 的合取。

“P 或 Q”写为 $P \vee Q$, 表示命题 P 以及命题 Q 的析取。

“如果 P 则 Q”或“P 蕴涵 Q”, 写为 $P \Rightarrow Q$, 表示命题 P 是命题 Q 的条件, P 叫做假设, Q 叫做结论。

“P 为假”或“非 P”, 写为 $\neg P$, 表示命题 P 的否定。

我们可采用括号来避免在一般的书面语言中容易产生的二义性。例如:

“桃花红或棉花白和油菜花蓝”

这一命题,或为真,或为假。

如果我们插入这样的括号

“桃花红或(棉花白和油菜花蓝)”,

则命题为真,因为第一部分为真,而如果 P 为真或 Q 为真,则 $P \vee Q$ 为真。

如果我们这样插入括号

“(桃花红或棉花白)和油菜花蓝”,

则命题为假,因为油菜花不是蓝的,而仅当P和Q二者同时为真时, $P \wedge Q$ 才为真。

一般说来, $P \vee Q \wedge R$ 是有二义性的,它不是命题,因为命题不能既为真又为假。但如果我们使用括号,得到 $(P \vee Q) \wedge R$ 或 $P \vee (Q \wedge R)$,则二者均为命题。

上面我们讨论了逻辑的句法,下面我们来讨论逻辑的语义,也就是命题的真假值。

我们可对每一个逻辑连词列出其真值表,这样,我们便可根据组成复合命题的各个命题的真假值,来确定复合命题的真假值。在下面的真值表中,我们用T表示命题为真,用F表示命题为假。

否定的真值表:

表 1-1

P	$\neg P$
T	F
F	T

析取的真值表:

表 1-2

P	Q	$P \vee Q$
T	T	T
T	F	T
F	T	T
F	F	F

合取的真值表:

表 1-3

P	Q	$P \wedge Q$
T	T	T
T	F	F
F	T	F
F	F	F

条件的真值表:

表 1-4

P	Q	$P \Rightarrow Q$
T	T	T
T	F	F
F	T	T
F	F	T

从否定的真值表中可看出, 当 P 为真时, $\neg P$ 为假; 当 P 为假时, $\neg P$ 为真. P 与 $\neg P$ 的真假值总是相反的.

从析取的真值表中可看出, 当 P 以及 Q 的真假值都为假时, $P \vee Q$ 的真假值为假; 否则 $P \vee Q$ 的真假值为真.

从合取的真值表中可看出, 当 P 以及 Q 的真假值都为真时, $P \wedge Q$ 的真假值为真; 否则, $P \wedge Q$ 的真假值为假.

从条件的真值表中可看出, 当 P 为真 Q 为假时, $P \Rightarrow Q$ 的真假值为假, 否则, $P \Rightarrow Q$ 的真假值为真.

因此, 如果 P 为真, Q 为假, 则 $\neg P$ 为假, $\neg Q$ 为真, $P \wedge Q$ 为假, $P \vee Q$ 为真, $P \Rightarrow Q$ 为假, $Q \Rightarrow P$ 为真, $\neg P \Rightarrow Q$ 为真.

有时, 我们可用一个命题 y 来替换一个复合命题 x , 当且仅当 y 为真时, x 为真, 这时, y 与 x 在逻辑上等价. 这可通过计算这两个命题的真值表来证明.

例如, 为了证明 $P \wedge (P \vee Q)$ 与 P 在逻辑上等价, 可借助于表 1-2 和表 1-3 来建立表 1-5, 从表 1-5 可看出, $P \wedge (P \vee Q)$ 为真时, P 也为真.

表 1-5

P	Q	$P \vee Q$	$P \wedge (P \vee Q)$
T	T	T	T
T	F	T	T
F	T	T	F
F	F	F	F

一般地说, 在任何一个逻辑证明中, 一个命题可以用一个在逻辑上与之等价的命题来替换.

间接证法是基于这样的原则：如果我们希望证明 $P \Rightarrow Q$ ，我们就先证明 $P \Rightarrow Q$ 与 $\neg Q \Rightarrow \neg P$ 是否在逻辑上等价，如果它们在逻辑上等价，那么，我们可通过证明 $\neg Q \Rightarrow \neg P$ 为真，来间接地证明 $P \Rightarrow Q$ 也为真。

另一种证法是归谬法。如果我们想证明 $P \Rightarrow Q$ ，可先假定 P 为真， Q 为假，并导出矛盾，即一个命题又为真，又为假，从而证明 $P \Rightarrow Q$ 为真。

我们再来考虑下述命题：

(1) “英语中的一切喉辅音都是清辅音。”

(2) “英语中有双唇塞音。”

(3) “英语中无送气边音。”

(4) “并非一切的塞擦音都是浊辅音。”

对英语的语音作了初步的分析之后，我们会承认上述命题均为真。

下面，我们来研究这些命题的内部结构。

命题(1)和(2)可改写如下：

(1') “对于一切的英语辅音 x ，如果 x 是喉音，则 x 就是清辅音。”

(2') “存在着一个英语辅音 x ，使得 x 是双唇塞音。”

或者这样说更好一些：

(2'') “对于某一个英语辅音 x ， x 是双唇塞音。”

命题(3)和(4)也可以用类似的方法加以改写。

但是，象下面这样的陈述(5)和(6)却不是命题：

(5) “如果 x 是喉音，那么 x 是清辅音。”

(6) “ x 是双唇塞音。”

这是由于在(5)(6)中，对于 x 没有作说明，它们既不能为真，也不能为假。

例如，我们如果用辅音音位 $|P|$ 来替换(6)中的 x ，我们得到“ $|P|$ 是一个双唇塞音”为真，但我们也可用 $|h|$ 来替换 x ，我们得到“ $|h|$ 是一个双唇塞音”为假。

(5)和(6)中仅包含一个变元 x ，有时，还可以包含多个变元，这种包含一个或多个变元的表达方式，称为“谓词”。我们用形式为 $P(x)$ 或 $P(x, y)$ 或 $P(x_1, x_2, \dots, x_n)$ 这样的式子来表示含有一个变元、两个变元或多个变元的谓词。

谓词也可以用逻辑连词 $\neg$ ， $\wedge$ ， $\vee$ ， $\Rightarrow$ 连接起来。例如，如果用 $P(x)$ 表示“ x 是喉音”， $Q(x)$ 表示“ x 是清辅音”，那么，(5)可写成 $P(x) \Rightarrow Q(x)$ ，它是有一个变元的谓词。