



语言

王
迈
著

形式化原理

上海大学出版社

语言形式化原理

王 迈 著



上海大学出版社

· 上海 ·

图书在版编目(CIP)数据

语言形式化原理/王迈著. —上海: 上海大学出版社,
2015.10

ISBN 978-7-5671-1918-5

I. ①语... II. ①王... III. ①语言学-研究
IV. ①H0

中国版本图书馆 CIP 数据核字(2015)第 242224 号

责任编辑 黄晓彦 石伟丽

封面设计 施羲雯

语言形式化原理

王 迈 著

上海大学出版社出版发行

(上海市上大路 99 号 邮政编码 200444)

(<http://www.press.shu.edu.cn> 发行热线 021-66135112)

出版人: 郭纯生

*

上海教育出版社经营有限公司排版

上海上大印刷有限公司印刷 各地新华书店经销

开本 787×960 1/16 印张 15.5 字数 270 000

2016 年 1 月第 1 版 2016 年 1 月第 1 次印刷

ISBN 978-7-5671-1918-5/H·317 定价: 30.00 元

前 言

语言是音义结合的词汇和语法的体系。语言的既有形式是自然形成的,是语言经历了千万年发展变化的结果,也是纷繁的言语变体通过长期竞争优胜劣汰的结果。因此,发展至今的各种语言都是一个相对完善的体系,它可以高效地满足本民族传统交际方式的需要,高效地满足社会信息传统传递方式的需要。但是,随着科学技术的进步,各种现代化的信息处理方式和传送方式不断出现,语言作为人们最重要的信息载体,不再局限于传统使用领域,它必须顺应时代的发展,适应现代化应用的需要。然而,语言固有的形式并不能很好地满足这一需求,它不是为现代化应用而生的,日益表现出局限性。特别是最近几十年,新型媒介的出现、计算机网络的广泛应用,以及人工智能的提上日程,使自然语言的不足和局限性日益突出,成为阻碍众多学科发展的瓶颈之一。我们对语言进行形式化研究,就是使用各种物质形式分析并重塑自然语言,拓展自然语言的适应性,使其满足各种现代应用领域的需要,更好地为其他学科服务。

本书(上海外国语大学 2013 年度校级重大科研项目“汉语国际教育和中国文化传播的对策研究”的阶段性研究成果,项目编号 KX161052)将从语言学和信息科学的视角,探讨语言形式化的一般原理和方法,内容涉及语音、语义、语法、语用修辞、文字等方面,下面选取若干问题作简要介绍。

1. 语音的属性及语音信息的获取

语音具有物理、生理和社会属性,这是对语音现象不同角度的分析和描述,合理利用这些特征,是实现语音形式化的基础,也是设计各种语音编码方案及压缩方案的重要参考。语音数字化的基本过程是采样、量化、编码;方案包括不均匀量化、差分量化、矢量量化、频域波形编码、参数编码等。语音实验可以根据语音的三种属性对数据作出不同层次的分析 and 解释。

2. 符号主义和联结主义

符号主义和联结主义是认知心理学的两种主要研究范式。符号主义把人脑比作图灵机,具有“有限”的性质,即有限的输入符号、有限的状态、有限的转移规则。

语义系统具有无限性,为适应符号主义处理机,必须进行有限化改写,但是任何对语义系统的有限化改写(义素分析、逻辑语义分析、语义格分析、词性分析等)都将造成语义缺失,破坏语言系统的完整性。

与此相对,联结主义从人的自然生理结构出发,把人脑看成由众多节点联结而成的开放式关系网络,具有并行处理、容错、自学习、遗忘、规则浮现等特征,这与人脑中的概念网络结构十分相似,是词汇语义形式化的理想模型。计算机语言作为典型的符号主义描写工具,伴随其智能化处理能力的不足,业已表现出明显的联结主义转向。

3. 语言的模糊性

模糊性是语义形式化面临的基本问题,它的来源不是语言单位的有限性,也不是客观世界的模糊性,而是人脑对客观世界的认知方式。其中,比较和概念化过程是模糊性产生的关键节点,而模糊性的产生又促进了人脑认知效率的大幅提升。符号主义范式对语义进行有限化改写的过程中所摒弃的主要内容,正是语义的模糊性,而联结主义范式可以实现对语义清晰与模糊的全覆盖。

4. 语法的抽象度

概念意义是明示的、开放的,语法意义是暗示的、封闭的,概念意义抽象为语法意义的过程,就是从明示的到暗示、从无限到有限的过程,这一过程受到语言发展的经济规律的制约。概念关系是多维的、普遍联系的,从深层概念结构到表层句法结构,是一个降维的线性化过程,语法就是作为多维信息损失的补偿机制而产生的。语法单位的有限性决定了其较词汇语义更易于形式化,符号主义范式可以胜任这一工作。

5. 语境形式化方法

语境形式化要素包括参与者信息、客观环境、上下文、语言知识、常识性知识、社会文化背景知识六类。基于实用性考虑,形式化语境的构成不再区分语言性和知识性,而是影响意义表达和意义理解的一切因素的总和。我们尝试用C++建构了一个常规语境类,用以模拟言语活动的一般过程:随着交谈的进行,新信息不断产生,旧信息不断被修正,交谈者的信息结构和心理状态也随之改变,语境就是这样一个处于变化中的动态平衡,话语的建构和理解实时依赖语境,语境义也实时发生着变化。

6. 通感的认知心理机制

通感既是五种感觉之间的相通,同时也是内省的情绪、情感之间的交融。通感

与比喻、比拟等传统辞格具有相同的认知心理基础,都是处在心智连续统上的不同区域间的彼此联通,因此可以把它们共同纳入广义的通感范畴,心智连续统也成为修辞形式化的重要参考模型。通感不仅大量存在于言语中,更多的通感现象已经经过社会的约定俗成,词汇化后进入语言,成为语言事实。通感是词汇系统丰富化和表达方式丰富化的重要源泉。

7. 语音修辞与象似性

语音的物理特征与意义之间具有某种程度的象似性,它不单体现在拟声词对声音的直接模拟,也可以表达非听觉的概念意义和情感意义,这是语音修辞的一种类型。象似性包括生理共鸣腔的大小、语音的时长、基频和强度等。本书尝试构建宏量指数模型,对随机抽取的35组单音节反义词进行测算,获取的宏量指数可以反映语音的真实面貌,并与心理感知相一致,揭示了语音修辞存在的物质基础。

8. 汉字的熵

熵是度量信息量的概念,汉字是高熵系统,其诸多特点,如字形复杂、数量庞大、区别度高、信息承载量大等,都可以从信息熵的角度得到合理解释。人类对信息序列的位数的瞬间感知能力极其有限,而对信息序列每一位所包含信息的瞬间区别能力却异常强大。这一显著倾向告诉我们,适合人类学习记忆的信息序列类型是:位数趋少;每一位所蕴含的信息量趋多。这正是高熵系统的特点,也是汉字体系的特点。信息熵同汉字编码位数的关系可以形象地描述为不同数制间的特征差异。信息熵是度量语言共性和特性的标尺之一,衡量语言发达程度的标准是该语言的词汇系统所表达的概念体系的复杂程度,或许可以称之为概念熵。

9. 同位汉字的编码

汉字体系具有跨语种跨地区的同位性特征(如“图、圖、囙”等),却在字符集中为语种所隔断,形成了同位异码的不利局面,汉字跨语种传递信息的能力受到限制。国际汉字单位(ICC)是跨语种同位汉字的集合,它建立在自顶向下的汉字系统观之上,可以有效解决汉字同位异码问题,促进汉字在世界范围内的整理和规范化,适应现代化信息交换的需要。从历时和共时角度对汉字的形、音、义进行考察,可以为同位汉字的判别制定标准。

以上为本书的主要内容。语言的复杂性及其广泛的学科基础决定了语言形式化研究是一项浩大的系统工程,绝非一人一时之功。因此,本书的研究当然不可能面面俱到,确切地说,它只是蜻蜓点水般的一些尝试,尝试通过分析语言的本质特征探寻语言形式化的一般原理和方法,以及语言形式化领域的主要方面。本书的

研究也不是穷尽性或辞书性的,某些具体问题也许不能直接从书中找到答案,却可能从中获得解决问题的理论和方法论的探讨,这是笔者的努力方向。

限于笔者的水平和经验,书中错漏之处在所难免,敬请读者批评指正。

作 者

2015年9月28日

目 录

1 语言的形式与意义	1
1.1 语言	1
1.2 语言是形式与意义的辩证统一	3
1.3 语言形式化研究的目的	4
1.4 语言形式化研究的学科支持	6
1.5 语言形式化研究的原则	6
1.6 语言的结构层次与形式化	8
2 语音与形式化	11
2.1 语音的属性	11
2.2 语流分解与离散化处理	14
2.3 语音数字化编码与 PCM	15
2.3.1 采样	15
2.3.2 量化	16
2.3.3 编码	17
2.3.4 波形编码的文件结构	18
2.4 语音编码技术的发展	20
2.4.1 不均匀量化和 A 律、 μ 律	21
2.4.2 信号相关性与 DPCM、ADPCM	22
2.4.3 频域波形编码与 SBC	22
2.4.4 参数编码与 LPC	24
2.4.5 混合编码	25

2.5	语音数据的压缩	25
2.5.1	利用语音的物理属性进行数据压缩	25
2.5.2	利用语音的生理属性进行数据压缩	26
2.5.3	利用语音的社会属性进行数据压缩	27
2.5.4	矢量量化技术	28
2.6	语音合成	29
2.6.1	波形拼接	30
2.6.2	参数合成	31
2.6.3	关于语音的自然度	31
2.7	语音识别	33
2.7.1	原理与过程	34
2.7.2	隐马尔科夫模型	36
2.8	语音的实验分析	38
2.8.1	频谱分析	38
2.8.2	基频分析与汉语的声调	41
2.8.3	音强分析	48
2.8.4	共振峰分析与元音的音色	51
3	语义与形式化	57
3.1	语义及其分类	57
3.2	语言结构的符号主义范式	59
3.2.1	从图灵机说起	59
3.2.2	有限状态自动机与正则表达式	60
3.2.3	符号主义范式的基本架构和特点	63
3.2.4	义素分析与语义的有限化	65
3.2.5	经典逻辑与语义二分	66
3.3	语言结构的联结主义范式	69
3.3.1	联结主义范式的基本架构和特点	69

3.3.2	语言结构的再分析	72
3.3.3	两种范式的互补性	78
3.3.4	再谈百科全书式的知识系统	79
3.3.5	语义形式化实例	80
3.3.6	语义网络的进一步说明	82
3.3.7	计算机语言对联结主义机制的支持	83
3.4	关于语义的模糊性	85
3.4.1	模糊性产生根源的两种观点	85
3.4.2	语言的模糊性源于人脑对客观世界的认知机制	88
3.4.3	模糊与精确的相互制约	90
3.4.4	语言模糊性的形式化表达	91
4	语法与形式化	94
4.1	语法及其分类	94
4.2	基本语法单位的形式化	96
4.2.1	从无限的概念意义到有限的语法意义	96
4.2.2	语法手段的有限性	98
4.2.3	语法单位是语言线性化过程的补偿机制	101
4.3	语法分析方法简述	103
4.3.1	以词法和词类为主要研究对象的欧洲传统语法学	103
4.3.2	历史比较语言学	105
4.3.3	索绪尔对现代语言学的贡献	106
4.3.4	结构主义语言学	107
4.3.5	转换生成语法	109
4.4	语法形式化的部分具体问题	111
4.4.1	上下文无关语法及 N 元语法	111
4.4.2	汉语分词技术	113
4.4.3	词类划分与标注	117

4.4.4	词类标记集分析	119
4.5	“把”字结构的句型意义	123
4.5.1	语义原型	124
4.5.2	“把”字结构不能出现光杆动词的深层语义原因	126
4.5.3	“把”字结构的形式化处理	127
4.6	再议汉语述补结构	128
4.6.1	述补结构难点分析	129
4.6.2	汉语述补结构在日语中的对应形式	130
5	语用修辞与形式化	135
5.1	语用与修辞	135
5.2	语境的形式化表达	137
5.2.1	语境类	138
5.2.2	参与者类	141
5.2.3	客观环境类	145
5.2.4	上下文类	149
5.2.5	话语理解和建构的一般过程	151
5.3	从通感看修辞格产生的心理基础	154
5.3.1	通感的三种类型	155
5.3.2	语言通感和言语通感	158
5.3.3	辞格的共性及感觉·情绪连续统	159
5.3.4	辞格的形式化表述	161
5.4	语音象似性修辞的量化分析	161
5.4.1	语音特征的基本量度	162
5.4.2	语音宏量指数模型	164
5.4.3	汉语常见对等范畴的语音宏值	168
5.4.4	再议语音象似性修辞	172

6 文字与形式化	174
6.1 文字的性质与分类	174
6.2 汉字的三个维度	175
6.3 本位观与“字本位”	178
6.4 语言单位的信息量	183
6.4.1 汉字的熵	184
6.4.2 熵与形式化	186
6.4.3 隐藏在信息熵下的语言共性	187
6.5 字符的内码、外码与形码	190
6.5.1 字符内码与标准字符集	190
6.5.2 字形码与字形描述	196
6.5.3 字符外码与输入法	200
6.6 同位汉字编码问题	203
6.6.1 同位汉字字形关系的类型	203
6.6.2 国际汉字单位	206
6.6.3 同位汉字的判定标准	209
6.7 汉日同位汉字辅音系统的对应规律	211
6.7.1 汉语中古音的声母系统	212
6.7.2 三十六字母所辖汉字的日语读音	213
6.7.3 中古汉语声母系统在现代日语中留存状况	215
6.7.4 归纳总结	219
6.8 汉字识别技术	220
参考文献	223
附录一 正则表达式元字符及使用方法	230
附录二 FrameNet:Commerce_buy	234

1 语言的形式与意义

1.1 语言

在现代科技的帮助下,人们的触手已经延展到几十亿千米外的太阳系边缘,可是我们脚下仅几千千米之遥的地球核心依然是难以企及的探索盲区;人们对赖以生存的外界环境的研究开发已是硕果累累,可是对自身生命奥秘的认识却刚刚起步。也许是控制物种进化的那只无形之手,把人类的感官千篇一律地设计成指向身体外部,我们对世界的认知也就有了关注外部先于关注内部的不自觉倾向。

我们对语言的认知也是如此。语言自产生之初已经陪伴我们走过了数万年的发展历程,我们借此思考问题、交流思想、存储信息。一个智力健全的人,用6~10年就可以基本掌握母语并熟练运用;一个受过基本训练的语言学学者,可以对可感知的语言单位(语音、词汇、语法等)进行细致入微的形式化描写和分析。可是,当我们试图对语言的产生机制、习得内化机制、言语生成和外化机制进行深入研究时,却遇到了难以克服的困难,以至于现有的理论大都停留在猜想阶段,无法证实或证伪^①,也无法很好地指导实践。这个环节的研究滞后突出表现在对语言内化机制和言语生成机制的人工模拟上;突出表现在计算机对人类语言的理解、掌握和话语生成上;突出表现在自然语言人机对话和人工智能上。

现代语言学研究领域不断扩展,从语言体系到言语活动再到言语机制,每一次扩展都是一次新大陆的发现,从拓荒到收获,使我们对语言本质的认识里程碑式地加深。语言的形式化研究也在沿着这条主线发展着,不再仅满足于对语言体系的

^① 科学命题应该可以通过经验或形式推理加以证明,波普尔(K. R. Popper)认为可证伪性也是区分科学和非科学的标准之一。在言语脑机制的研究中,经验和形式推理往往显得力不从心,既不能证实,也不能证伪。例如乔姆斯基(A. N. Chomsky)基于天赋论的“语言习得装置”,其真实性就无法通过个体经验或推理来验证。这固然有验证的技术手段落后于理论猜想等现实原因,却也不能排除这个理论原本就缺乏科学性。

形式化描写,开始越来越注重对言语活动、言语环境、语用和修辞进行定量定性分析,如今更把注意力转到对语言的脑机制和心理机制的模拟、对言语生成和理解过程的模拟、对信息通过语言在大脑中存储方式的模拟等领域中来了。

王德春(2009)将现代语言学研究领域归纳为一个建构性的循环网络:“在这个网络中,通过语言和言语的循环建构语言体系;通过言语活动和话语的循环建构话语;通过交际双方在言语活动中的循环建构话语意义的共同理解;通过言语活动和言语机制的循环建构内部言语;通过语言体系和言语机制的循环建构个体语言体系,提供建构内部言语的语料;通过语言体系和文字体系的循环建构文字体系;通过文字体系与言语活动的循环建构书面语;通过主管言语机制的人脑和处理语言信息的电脑的循环建构机器语言体系和话语,达到人机对话。”

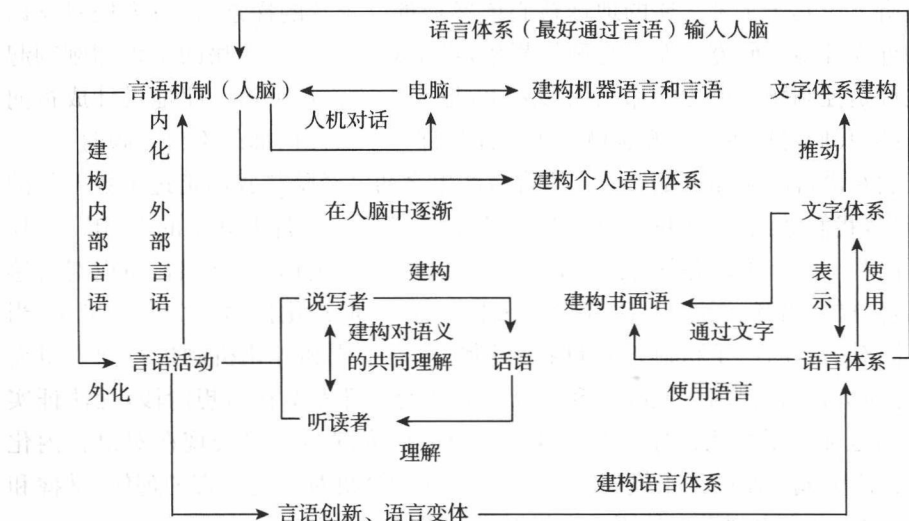


图 1.1 现代语言学研究的建构性循环网络

图 1.1 中的语言体系、言语活动和言语机制是四处连线密集的枢纽,以它们为主要支点,联系各个循环,构成完整的语言学研究体系。本书的形式化研究,大致相当于“人机对话”这个环节,有其相对独立性,但在研究过程中不可避免地涉及其他各分支,并且以整个循环网络为理论基础。我们需要首先讨论语言体系各要素的形式化,在此基础上讨论语境和言语活动的形式化,进而探索电脑对人脑机制的模拟,即人脑言语机制的形式化,这是实现形式化研究最终目的——人工智能的基础。我们可以把“人机对话”看成一个包容整个语言学大体系的小分支,这个分支

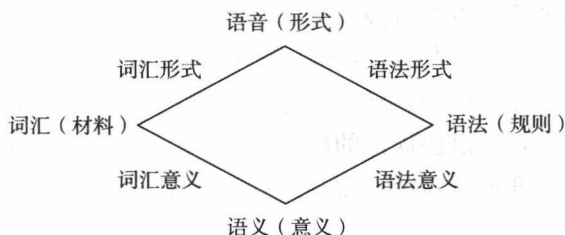
专门研究语言体系、言语过程、言语机制的形式化和人工机器模拟。

语言学研究是一个普遍联系的整体,各部分相互制约、相互补充,牵一发而动全身。在这个体系中,与形式化研究如影随形、密切相关的,是语言的意义。脱离了语言的意义,语言的形式化研究就成了无水之源、无本之木。因此,我们有必要讨论语言形式与意义的关系问题。

1.2 语言是形式与意义的辩证统一

语言是词汇和语法的体系。词汇是语言的建筑材料,是表达客体的主要手段。语法是语言的组织规律,是表达客体关系的主要手段。人们利用语法规则把词汇材料组织成表达思想的话语,同时依据语法规则分析话语结构以获取话语信息。

词汇和语法都是音义结合的。语音是语言的物质外壳,是记录语言的主要形式。语义是语言的意义内容,是语言所承载的思维信息。人们利用语音形式记录和表达思想,同时依靠语音形式传递和交流思想。语言的语音和语义就如同一张纸的两面,结合紧密而又缺一不可,是形式与意义的统一(图 1.2)。



例词(形)	students	问问
词汇形式	[ˈstu:dənts]	[ʋənʋən]
词汇意义	在学校学习的人	不明白的事请人解答
语法形式	[-s]	重叠
语法意义	名词、复数	动词、短时、尝试

图 1.2 语言的四要素及其相互关系

形式与意义从语言诞生之初就是不可分割的两个方面,它们都是语言固有的。

马克思主义的语言起源观^①认为,语言是人们在劳动中为满足交际的需要而产生的,人类语言从一开始就是有声语言。动物的嚎叫与嘶鸣虽然有其声音形式,但它并不表达思维意识,不能完成社会交际的任务,因此,它不是语言;结绳记事与划痕计时虽然表达一定的意义,但它不具备口耳相传的有声形式,因此,它也不是语言。原始人类在劳动中经过生理和心理的长期进化,逐渐尝试将承载着思维意识的声音呼号出口,形式才最终与意义相结合,那一刻起,语言才得以诞生。因此,从历时的角度看,音与义的统一始于语言诞生之初,是语言的安身立命之本。

言语作为思维和语言相互作用的产物,也是音义结合的,并且,音义结合始于言语建构的底层,即始于人的思维。言语生成过程不是在既有的语音外壳上捆绑思维意识,也不是给现成的思维意识包裹一个语音外壳。思维过程必须借助语言的参与,语言也必须在思维过程中实现话语交际意义,语言和思维是共时共存、不可分割的。语言是思维的工具,思维产生新的意识,经约定俗成后固化为具有语音形式的语义,不断充实语言体系,促进语言体系的新陈代谢和发展变化。

言语生成的心理过程可以分为表述动机、表述意图、内部言语、扩展的言语表述等几个环节(王德春,1997)。其中,内部言语是从体性的压缩的语义初迹向线性的扩展的言语表述转化的重要阶段。它虽然没有外显的语音形式,但心理存储方式仍然不能脱离语音。儿童在语言习得过程中,会经历一个内部言语外显的阶段,例如在做动作前,时常会低声细语“吃苹果吧”“宝宝想妈妈了”等。人类的梦中呓语经常是一些不符合语法规则的碎句,甚至是语义不连贯的离散单词。这些例子让我们从不同角度窥见,内部言语是“无声的说话”,它在外化之前就已经是音义结合的了。

——认知语法在分析语言时更加彻底,只允许三类单位:语音单位、语义单位、象征单位。象征单位是语音单位和语义单位的结合,这种限定符合语言本质特征,使语言分析方法回归到音义结合的道路上来,“它排除了任何描写上的随意性,排除任意设立的空范畴、虚迹和各种有名无实的附加符号”(沈家煊,1994)。

1.3 语言形式化研究的目的

以上从共时、历时以及言语生成机制等方面探讨了语言形式与意义的关系。

^① 关于语言起源,有很多猜想和假说,如神授论、名称固有论、自然声音模仿论、情绪表达论、劳动号子论、手势决定论、社会契约论、劳动创造论等。

这是语言形式化研究的理论基础,同时也引出了几个需要思考的问题:

第一,语言是音义结合的,既然语言已经具有其形式,对语言的形式化研究是否还有必要?回答是肯定的。语言的既有形式是自然形成的,是语言经历了千百年发展变化的结果,也是纷繁的言语变体通过长期竞争优胜劣汰的结果。因此,发展至今的各种语言都是一个相对完善的体系,它可以高效地满足该民族的传统交际方式的需要,高效地满足社会信息的传统传递方式的需要。但是,随着人类科学技术的进步,各种现代化的信息处理方式和传送方式层出不穷。语言作为人类最重要的信息载体,不再局限于传统使用领域,它必须顺应时代的发展,适应现代化应用的需要。然而,语言固有的形式并不能很好地满足这一需要,它不是为现代化应用而生的,日益表现出其局限性。特别是最近几十年,声、光、电、微波等媒介的广泛应用,计算机网络的出现以及人工智能的提上日程,使自然语言的不足和局限性日益突出,成为阻碍众多学科发展的主要瓶颈之一。我们对语言进行形式化研究,就是要使用各种物质形式分析并重塑自然语言,拓展自然语言的适应性,使其满足各种现代应用领域的需要,以促进多学科均衡发展。语言形式化的最终目标是实现人工智能,让机器像人一样使用自然语言进行思维和交流,虽然这个目标还显得遥不可及。

第二,语言已经具有其形式,将形式进一步形式化,其可行性如何?我们在编纂词典时,以词释词是经常采用的方法之一;我们在进行言语表述时,也经常“换言之”“简言之”,这些其实也是一种形式化,是采用语言中既有的形式来“形式化”语言本身,说明语言在某种程度上具有自足性(自解释力)。但是这种形式化在逻辑上是一种没有入口的循环定义,它也许可以满足文学、艺术的需要,却并不适用于科学研究。为避免循环定义,必须采用语言之外的物质形式来形式化语言,这种物质形式是严格意义上的元语言,即语言的语言。并且,元语言还可以进一步形式化,也就是说,语言的形式化是可以无限叠加的。例如,通过国际互联网进行语音传输的过程就是一个语音形式化的多次叠加过程和反向过程。

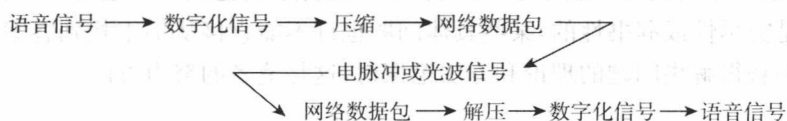


图 1.3 语音传输示意图

图 1.3 显示,语言的形式化不但可行,而且可以是一个多次叠加的过程,并且