

The Exploration and Practice of
Language Information Processing

语言信息处理 探索与实践

(第1辑)

朱小健 主 编



NLIC 2970677108



北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
北京师范大学出版社

The Exploration and Practice of
Language Information Processing

语言信息处理 探索与实践

(第1辑)



朱小健 主编



NLIC 2970677108

北京师范大学出版集团

北京师范出版社

图书在版编目(CIP)数据

语言信息处理探索与实践.第1辑/朱小健主编.—北京:
北京师范大学出版社,2011.1
ISBN 978-7-303-10772-8

I. ①语… II. ①朱… III. ①汉字信息处理—文集
IV. ①TP391.12-53

中国版本图书馆 CIP 数据核字(2010)第 014511 号

营销中心电话 010-58802181 58808006
北师大出版社高等教育分社网 <http://gaojiao.bnup.com.cn>
电子信箱 beishida168@126.com

出版发行:北京师范大学出版社 www.bnup.com.cn
北京新街口外大街19号
邮政编码:100875

印 刷:北京嘉实印刷有限公司
经 销:全国新华书店
开 本:184 mm × 260 mm
印 张:24.25
字 数:551千字
版 次:2011年1月第1版
印 次:2011年1月第1次印刷
定 价:48.00元

策划编辑:杨帆	责任编辑:杨帆
美术编辑:毛佳	装帧设计:天泽润
责任校对:李茵	责任印制:李啸

版权所有 侵权必究

反盗版、侵权举报电话:010-58800697

北京读者服务部电话:010-58808104

外埠邮购电话:010-58808083

本书如有印装质量问题,请与印制管理部联系调换。

印制管理部电话:010-58800825

序

许嘉璐

北京师范大学中文信息处理研究所恐怕是中国高校里同类机构中年龄最轻的，到现在才五岁。从她诞生那天起，就承担着两项艰难的任务：（1）以 HNC 理论为主体，探索中文信息自动化处理的技术实现，出成果，出人才；（2）和相关企业合作，尝试着走产—学—研相结合的路子。在自然科学基金和有关部委以及学校的支持下，研究所总算没有辜负各方面的期望，研究工作达到了预期目标^①，硕、博士生的培养也如期地、较好地完成了任务。这本论文集就是学生们向学校、向社会所作的一次汇报。

本书中所收论文涉及中文信息处理的许多方面：词、短语、句、句群；语义、识别、检索、语境；中外对译、繁简转换，等等。文章或许显得比较稚嫩，但是却有着共同的特点：都是和产业化的需求紧密结合的，也都是比较前沿的课题。

研究所希望能培养出一批“两栖人”，即在语言学和计算语言学两个方面都有相当基础的人才，这是中文信息处理界感到缺憾的老问题，发现、提出多年了，情况依然没有大的改变。我们决定试一试，招收的学生即来自两个专业。学生入学后立即进入项目组，在实践中各自补充另一方面的不足，课程也是两边兼顾，也要结合中文信息处理的实际。与产业结合的教学与科研除了能及时发现基础理论、应用基础理论的不足与偏误，同时还经常遇到具体的、局部的技术问题。学生们的论文题目就从这些问题里选择，这样就自然站在“前沿”了。我们心中的“前沿”，并不全着眼于国内外学术界所提出的新理论、新热点，而是更强调在共同攻关过程中所遇到的新问题、新思路；换言之，我们要让学生始终站在“应用”的第一线上。

一个教学单位把硕、博士生的毕业论文结集出版，这在我国似乎还并不多见。我们之所以这样做，就是想请社会，特别是业内专家、同行通过学生们的“作业”帮助我们判定：我们这样建设中文信息处理这一应用学科的得失与长短。当然，书的出版也是对学生们的鼓励。作为论文的作者们，自然希望听到对他们作品的批评，而我却更想得到关于如何进一步改进教学模式和方法的批评和建议。

书中论文的作者有的还在所里学习或工作，有的已经在祖国各地就业。我希望他们之所学能够在现在和未来的工作中派上用场。

衷心地祝福他们！

2010 年 10 月 12 日于日读一卷书屋灯下

^① 参见许嘉璐、朱小健：《中文信息处理研究工作的新进展》，载《云南师范大学学报》（哲学社会科学版），2010（4），转载于《新华文摘》，2010（18）。



目 录

CONTENTS

训诂学的系统性对 HNC 语境理论的启示	田 涛 (1)
汉语理解处理中的动态词研究	唐兴全 (68)
基于 HNC 理论的汉英逻辑组合变换研究	李千驹 (88)
现代汉语语料库例句检索的后处理系统	谭晓平 (97)
中文繁简转换研究与系统实现	冯 霞 (141)
现代汉语感谢和道歉类言语行为的识别及其语义信息的提取	徐 圆 (169)
现代汉语文本中离合词扩展形式的自动识别	冯向华 (210)
基于 HNC 的句群领域判定研究	袁义春 (240)
中文时间短语的自动识别	刘 奎 (272)
现代汉语文本中人与机构间任职关系的信息抽取	陶 莹 (308)
农业知识查询中广义工具语义关系的识别	张竹村 (342)
专利文献中机具名的自动识别	宋冬雪 (375)

训诂学的系统性对 HNC 语境理论的启示

田 涛

摘 要：语境可消解语言模糊，在人类的言语交际中起着非常重要的作用。目前的计算机之所以无法像人类一样理解自然语言，基本原因之一就是它不能有效获得和利用语境，因而无法有效消解语言的模糊。因此，研究语境形式化，让计算机“掌握”语境，是实现自然语言理解的一个关键。但是，由于语境是无限的、动态的，实现语境的形式化困难重重。HNC（Hierarchical Network of Concepts，概念层次网络）理论吸收中国传统训诂学的思想精髓，实现了语境的形式化，并将其应用于语句和篇章的理解处理，取得了重大进展。对 HNC 语境理论的训诂学依据进行系统总结，对促进训诂学和自然语言理解的交融具有重要意义。

系统性是训诂学的重要特点之一，HNC 语境理论继承发展了这一优良传统。本文对此进行了研究和总结：《尔雅》作为我国第一部概念分类词典，开创性地地区分了具体概念和抽象概念，同时在概念分类中贯彻了层次性和有序性原则，形成了一个相对完备的概念系统，这实际上是对人脑在使用语言中形成的概念体系的模拟。HNC 理论借鉴了《尔雅》的系统性原则，构建了模拟人类大脑的语言概念空间，并针对计算机理解自然语言的需要，设计了表达概念关联性的形式化方法。

关键词：语境 HNC 理论 训诂学 中文信息处理 系统性

一、引言

自然语言理解的一个基本目标是使计算机能够像人一样理解自然语言，从 20 世纪 50 年代英国数学家图灵（Turing）提出著名的图灵测试至今，经过众多语言学家、逻辑学家以及计算机工作者的努力，自然语言理解技术虽然取得了长足进步，但是这一目标依然是一个看似无望解决的难题。要解决这一问题必须对自然语言、人脑以及计算机的特点及其关系进行深入的研究。模糊性是自然语言的最大特点之一，它普遍存在于自然语言的各个层面。HNC（Hierarchical Network of Concepts，概念层次网络）理论创始人，中国科学院黄曾阳（1998：3）研究员把它们概括为五重模糊：发音模糊、音词转换模糊、词的多义模糊、语义块构成的分合模糊、指代冗缺模糊（书面语只有后三重模糊）。对于语言模糊性的原因，黄曾阳先生也进行过深入的分析，其中的关键原因之一是：在人脑中相互关联的概念体系在自然语言空间里表现为彼此孤立的语言符号，它们的关联性在交际中无法从形式上表现出来，这就造成了歧义、含混、指代不明等种种现象，给交际造成了重重障碍。但是人的交流并没有因为语言的模糊性而无法进行，大多数交流

都能顺利展开，其原因在于人的交际并不是语言符号的孤立传递，而是能够通过言内信息和言外信息的耦合生成交际语境，大多数的模糊现象都可以在语境中消除。关于语境的功能众多国内外学者曾进行过研究，日本学者西楨光正（1992：27—44）总结出了语境的八种功能：绝对功能、制约功能、解释功能、设计功能、滤补功能、生成功能、转化功能、习得功能。语境的这些功能在很大程度上消解了语言模糊，使交流得以实现。对于目前的自然语言理解而言，在计算机面对一段自然语言时，它所具有的只是关于这段语言的言内信息，而文本以外的信息以及人脑中与之相关的信息计算机是没有的，因此语境也就无法生成，这就导致了目前的人机交互无法克服自然语言的模糊性。

信息处理学界已经认识到语境对于自然语言理解的重要作用，并且开始尝试将语境知识应用于自然语言理解，具体说来就是要实现语境的形式化。冯志伟（1999：215）先生指出，计算机对自然语言的研究和处理应该经过如下三个方面的过程：

（1）把需要研究的问题在语言学加以形式化，使之能以一定的数学形式，严密而规整地表示出来；

（2）把这种严密而规整的数学形式表示为算法，使之在计算上形式化；

（3）根据算法编写计算机程序，使之在计算机上加以实现。

实际上，语境的形式化是一项非常困难的工作。首先，语境是一个包罗万象的范畴，交际中几乎所有与话题相关的内容，都可以归入语境。并且语境在交际过程中并非一成不变，而是随着时间、空间等各种条件的变化不断变化。相对于外部世界的信息，人脑中的相关信息更是难以把握。其次，现有的语境理论发展尚不完善，对于繁杂的语境信息只是进行了大概的分类而无法系统把握，而且所关注主要是人与人之间的交际语境，而对人机交互语境研究甚少。此外，现在的认知心理学所研究的内容仅仅局限于认知的某一领域，对人脑中的语境信息进行系统表征也就无从谈起。所以，语境的系统形式化困难重重。目前语境在信息处理中的应用大多限于利用文本的上下文信息进行歧义消解等方面的工作，语境在信息处理中的更广泛应用还有待于语境理论以及认知心理学等研究的进一步深入展开。

中文信息处理在经历了字处理、词处理阶段以后，在句处理和篇章处理阶段遇到了难以突破的困境，中文信息处理的瓶颈在于语义，而突破这个瓶颈的关键就是语境和语义的结合。长期以来，统计概率和基于西方语法理论的规则是中文信息处理的主要方式。但是，语言现象是复杂多变的，单靠统计概率无法统摄。而基于西方语法理论的规则体系是以有着丰富形态标记的西方语言为基础的，并不适用于中文信息处理；而且立足于自然语言空间对语言进行语法逻辑分析，也无法完全克服自然语言关联性缺失的问题。近年来自然语言处理学界虽然认识到了语境的重要性，但主要还是试图通过逻辑推理的方式来实现语境的形式化，而逻辑方式难以从根本上把握繁杂多变的语境信息。中文信息处理的语义研究在传统的道路上难以取得突破，因此有必要从汉语的实际特点出发，从中国传统的语言研究中寻找出路。许嘉璐（2003）先生指出：“中国自古重视语义的解释和对其演变规律的探索，其成果集中体现在训诂学中。在训诂学对语义的解释中，语言环境和语义分类受到极大关注，这或许是训诂学区别于西方语义研究的特色。这一特色，是历代学者徜徉于语言事实的海洋中体味到的，换言之，是基于语言实际的。古人

的这些经验（其中也有不少反面经验），是我们今天解开现代汉语语义之谜的重要参考。我们在陆俭明等先生的论述中已经明显地感到了传统学术的影响和对前人成果的关注。如何看待中国独有的训诂学，能否利用之，这或许又是解决中文信息处理应该注意的问题之一。”

正如许先生所言，训诂学历来重视语义和语境的作用。以训诂学为代表的中国传统语言文字学面对的是汉语的实际语言，训诂学家们为了使当时的读者能够理解古文，进行了大量深入细致的训诂工作，积累了汉语语言理解的宝贵经验，其中最重要的就是重语义以及对语言环境的运用，这也正是解决汉语缺乏狭义形态所带来的种种难题的关键方法。训诂工作是为了帮助后人理解前人的文献，实际上可以视为训诂学家对古代文献语境的复原与营造，在这一点上训诂学与信息处理有着相通之处：都是一门语言应用之学，目的都是为了理解交流。因此两者完全有着相互借鉴的条件和价值。此外，中国历代训诂学家面对的是中国的语言文字，更能深刻地体悟到汉语的实际特点，因此从中国传统语言研究中寻找中文信息处理的新思路，不失为一条可行之路。

黄曾阳先生立足于汉语言的实际，继承和发展了训诂学的优良传统，创立了 HNC（概念层次网络）理论。HNC 理论认为交际中语境生成的关键在于人的大脑中存在着语言概念空间，其中的语言概念彼此关联，复原了自然语言空间缺失的关联性。人类的语言概念空间是相通的，在交际过程中能够生成可以相互交流的语境，计算机正是因为不具有这样的语言概念空间，才无法生成人机交互的语境。HNC 理论对语言概念空间进行了深入的研究并建立了完备的数学物理表示式，为计算机语言概念空间的建立创造了条件。针对语境的无限性，HNC 把语境单元定义为领域（DOM）、情景（SIT）、背景（BAC）的三维结构体，从而以有限的语境单元三要素控制了无限的语境。HNC 语境理论已经应用于中文信息处理并取得了很大的进展。HNC 语境理论的思路就是借鉴了训诂学的语境观，如以有限语境要素来控制无限语境就是受到了训诂学方法的启示。

训诂学在训诂实践中积累了语言理解与应用的宝贵经验，形成了科学的理论体系，但是，训诂学家却没有把这些经验加以整理概括，没有把他们心中存在的理论体系明确陈述出来。对此许嘉璐（1999：28，31）先生曾有过专门的论述，指出中国古代语言学的弱点在于“缺乏宏观的观察和归纳”，“忽略方法论的研究”。近代的章太炎和黄侃先生倡导在传统语言学的基础上建立现代语言文字之学，但《马氏文通》问世之后，近代语言学的研究集中于西方语言学理论，中国语言学的传统也就随之隔断了。中文信息处理要想从训诂学那里获得有益的启示，首先必须对前人的经验进行整理，在此基础上形成系统实用的理论体系，唯有如此，才能走出一条适合汉语特点的信息处理发展道路。对 HNC 语境理论的训诂学依据进行全面总结，对于促进训诂学和自然语言理解之间的交融具有重要意义。

中国传统语言学历来重视语言的系统性，古代汉语的形音义都建立了比较严整的体系，如《说文》的形义系统，清代小学家构建的音韵体系等。训诂学很早就有了概念分类和语义系统的观念，以《尔雅》为代表的雅学系列著作最具代表性。《尔雅》中具体概念和抽象概念的区分对 HNC 概念基元体系的设计产生了直接的启示，黄曾阳先生称之

为“尔雅原则”。《尔雅》义类和类义的系统观对中文信息处理的语义分类有着重要的借鉴意义，HNC 构建的具有层次网络特性的语言概念空间，在很大程度上是向训诂学传统的复归。以系统的观念构建概念和语义体系，与对语言进行逻辑化结构分析的思路有着根本不同，也是 HNC 语境理论区别于传统中文信息处理方法的一大特点。

本文选取《尔雅》为材料，对《尔雅》的分类体系、训诂方式和方法进行分析，由此进一步辨析古人的义类和类义观念，概念系统等。在此基础上进一步分析训诂学的系统性思想对 HNC 语境理论以及中文信息处理的借鉴意义。

二、自然语言空间的局限性

（一）“言意之辩”与中西方语言观的差异

“书不尽言，言不尽意”（《易传·系辞传》）是我国古代著名的语言哲学论断，反映了中国传统的语言观，也是与西方语言哲学的基本差异。中西方语言哲学对自然语言历来存在不同的观点，主要差别是中国语言哲学认为自然语言存在着局限性，仅仅依靠语言无法对意义、情感等进行充分表达，也就是“言不尽意”。西方语言哲学则认为语言是表达的必然工具和手段，一切事物都可以通过语言得到充分的表达和诠释。这种观点也成为西方重视对语言本体结构进行深入分析的原因之一。

中国很多思想家都曾谈到过语言的局限性。老子认为：“道，可道，非常道；名，可名，非常名。”（《老子》第一章）孔子则说：“予欲无言。”（《论语·阳货》）庄子对“言不尽意”的思想进行了深入的阐发，他说：“世之所贵道者，书也。书不过语，语有贵也。语之所贵者，意也。意有所随。意之所随也，不可以言传也，而世因贵言传书。世虽贵之，我犹不足贵也，为其贵非其贵也。”（《庄子·天道》）在庄子看来意义才是人真正要追求的，语言只是传达意义的工具，语言因为负载了意义才具有了价值。意义是有所指向的，意义的指向却是不能用语言来表达的。但是人们往往过分相信自己的感官体验，殊不知这样无法获得真正的意义实情，“故视而可见者，形与色也；听而可闻者，名与声也。悲夫！世人以形色名声为足以得彼之情。夫形色名声，果不足以得彼之情，则知者不言，言者不知，而世岂识之哉！”（《庄子·天道》）庄子在《秋水》篇中，进一步指出了语言的局限性，“可以言论者，物之粗也；可以意致者，物之精也；言之所不能论，意之所不能察致者，不期精粗焉。”事物外在的形体表象，也就是“粗”，是可以用来语言表达的，而事物的内在本质，也就是“精”则只能通过心意来体察。然而，还有无法用语言表达和心意体察的，那就是“道”，圣人之意，“道”无象无形，也就是“不期精粗”，是不可传达的。所以，《庄子》讥讽所谓记载古代圣人言论的书其实是“古人之糟粕”（《庄子·天道》）。在“言不尽意”的基础上，庄子提出了另一个著名的论断“得意忘言”：“荃者所以在鱼，得鱼而忘荃；蹄者所以在兔，得兔而忘蹄；言者所以在意，得意而忘言。吾安得忘言之人而与之言哉！”（《庄子·外物》）人们使用荃、蹄是为了捕捉鱼和兔，与此类比，运用语言的目的是为了理解和表达思想，也就是得到“意”，而语言如同荃、蹄一样，仅仅是得“意”的工具。

以庄子为代表的中国思想家普遍表现出重意轻言的倾向，而西方学者则大多重视语

言的作用，德国神学家、哲学家施莱尔马赫说：“诠释学的一切前提无非就是语言。”^①认为一切事物都可以通过语言进行诠释。洪堡特也说：“语言不只是表述已知真理的手段；更重要的是，它是揭启未知真理的手段。”^②与中国“言不尽意”的观念相反，洪堡特认为真理可以通过语言得到。表现在学术实践上，西方学者的诠释工作通常以一套具有严密逻辑和语法结构的语言形式表现出来，而中国的诠释语言在形式上则往往具有不确定性，中国古代的学者更重视心灵的观照和体悟，以直觉思维的方式来体验和传达圣人的本意。

对于中西方的这种差异，可以从语言 and 思维的特点入手进行分析。“言不尽意”并不是否认语言的表达功能，而是认为语言在表达上具有局限性，无法充分地表达意义。庄子所说的“意”主要指的是圣人之意，也就是古代先贤的思想和经验。首先，人的思想和经验是在特定的语境中形成的，既包括当时外部环境的信息，也包括人脑中已经具备的信息，各种信息彼此关联形成语境，人的思想经验的形成是与当时特定的语境紧密关联在一起的，离开了当时的语境，其中的思想经验是无法准确体验传达的。但是语境是不断动态变化的，组成语境的各种信息随时会缺失改变。这些信息形式多样，数量众多，而以语音和形体为表现形式的语言文字，无法完整地记录还原这些信息，语境也就无法完整地还原。同时，语言文字符号体系由于缺失了人脑中概念体系所具有的关联性，由它们记录的东西不可避免地具有模糊性，产生歧义含混等种种现象。在语境无法完整还原的前提下，这些模糊性也就无法完全消解，还原前人的确切意识。后人依据文本来理解前人的思想，往往会产生不同的解读，就是这个原因。其次，人的心理体验非常复杂细微，相当一部分没有形成明确的概念，而语言是经过形式化、规范化的产物，具有抽象、概括的特点，舍弃了思维中很多细节的部分，对意识中没经过抽象化、概念化的部分是无法表达的。故此“言”是不能“尽意”的。

由以上可知，通过对语言的结构分析来寻求完全的意义和真理是无法做到的。西方对于语言的看法在一定程度上夸大了语言的作用，有的西方学者也认识到了语言的局限性。黑格尔就曾说过：“语言实质上只表达普遍的东西；但人们所想的却是特殊的东西，个别的東西，因此不能用语言表达人们所想的東西。”

既然言不能尽意，那么应当通过何种方式来“尽意”？古人提出了解决的方案：“然则，圣人之意，其不可见乎？子曰：圣人立象以尽意，设卦以尽情伪，系辞焉以尽其言，变而通之以尽利，鼓之舞之以尽神。”（《易传·系辞传》）对于言不尽意的矛盾，《周易》提出了立象尽意的办法。具体说来，就是建立一套代表宇宙万物的卦爻符号体系以及对其进行解释的卦辞，通过推演进行信息预测，揭示事物的运行规律，补充语言所不能“尽”的意。古人创立《易经》的最初目的是为了占筮，即通过卦象的推演来预测吉凶，指导自己的行动。所以，创立卦象的目的在于寻求象所蕴含的“意”，设立卦象是得到意的方法和手段。“夫象，圣人有以见天下之赜，而拟诸其形容，象其物宜，是故谓之象。”（《易传·系辞上》）也就是圣人发现了幽深复杂的事理，把它譬拟成具体的形象，用来象

① 转引自洪汉鼎：《诠释学》，北京，人民出版社，2001。

② [德] 威廉·冯·洪堡特：《洪堡特语言哲学文集》，姚小平编辑译注，长沙，湖南教育出版社，2001。

征事物适宜的意义。

《易经》创立的卦符系统其初始符号是阴阳两爻，首先由阴阳两爻推演组合而成八经卦，即乾、坤、震、巽、坎、离、艮、兑，八经卦两两叠加组成六十四别卦，每别卦由六爻组成，共三百八十四爻。阴阳两爻和每一卦都代表一定的事物，以此通过象征说明事理。《易经》的卦爻符号可以象征多种事物，例如，乾可以象征天、圆、君、父、玉、金、寒、冰、大赤、良马等；坤可以象征地、母、布、釜、吝啬、均、大舆等（《易经·说卦》）。六十四卦的排列也不是任意的，而是按照各卦所代表事物之间的关系有序排列。对此，《序卦辞》进行了说明：

有天地，然后万物生焉。盈天地之间者，唯万物，故受之以屯。屯者盈也，屯者物之始生也；物生必蒙，故受之以蒙。蒙者蒙也，物之稚也；物稚不可不养也，故受之以需。需者饮食之道也；饮食必有讼，故受之以讼。讼必有众起，故受之以师……

卦象从天地生化万物起，经过生长、蒙昧的阶段，发展到和顺、安乐的状态。从某种角度说《易经》的卦象创造了一个具有关联性的概念体系，依据这个体系进行占卜预测，传情达意，在一定程度上克服了自然语言关联性、系统性不明确的局限。同时，通过具体的感性物象对宇宙万物的变化及其规律进行象征性地表达，可以使人生发直觉形象性思维，体验到语言文字所无法传达的感受。较之语言思维，“象”的思维更具整体性。

通过卦象传情达意虽然在一定程度上克服了语言的局限性，但是也存在一定问题，通过有限的卦象无法穷尽表示宇宙万物，这样在进行推演解释时必然会存在不确定性。因此《易经》的符号系统除了卦象以外还有对卦象进行解说的卦辞，也就是所谓的“系辞焉而明吉凶”（《系辞上》），对通过卦象无法说明的部分进行补充解说。象与辞的相互配合，完善了《易经》的占筮体系，《易经》通过“设卦”“观象”“系辞”最终达到“得意”的目的。《易经》的占筮通常是针对特定的事情，也就是在特定的语境中进行的，包括问卦者，所问的事情，事件的背景，等等。因此，占筮者在推演的过程中，也会充分考虑这些语境因素，陆宗明（1997：50—52）结合具体事例对这些语境因素在占筮中的作用及占筮者对它们的运用进行了分析。所以，依据《易经》进行占筮可以视作语用推理。综观其过程，是对言、意、象、卦、辞的综合运用，实现了语形、语义、语用、语境的统一。

魏晋南北朝时期，玄学家王弼进一步阐释了言、象、意三者间的相互关系，最著名的是在“得意忘言”的基础上提出了“得意忘象”说。

王弼首先肯定了象可尽意，言能表象，语言具有积极的一面：“夫象者，出意者也。言者，明象者也。尽意莫若象，尽象莫若言。言生于象，故可寻言以观象；象生于意，故可寻象以观意。意以象尽，象以言著。”（《周易略例·明象》）

同时王弼又从“名”的角度指出了语言的局限性，即语言表意不能周全穷尽。“名也者，定彼者也；称也者，从谓者也。名生乎彼，称出乎我。”（《老子指略》）事物的命名是人根据事物某个方面的外在形态进行的，具有主观性，无法反映事物的一切特征，尤

其是内在特征。因此，名所反映的事物与事物的本质状态可能大相径庭，甚至可能误导人对事物的认识。故王弼认为形具有局限性，而名是根据形来命的，所以名也具有局限。“名必有所分，称必有所由，有分则有不兼，有由则有不尽；不兼则大殊其真，不尽则不可以名。”这也是语言不能尽意的表现。王弼同时也指出了人的感知的局限性：“视之不足见，则不足以悦其目；听之不足闻，则不足以娱其耳。”（《老子注 35 章》）

在此基础上王弼认为言对于象，象对于意都是次要的，得到了象就不必拘守于言，得到了意就不必拘守于象，曰：

故言者所以明象，得象而忘言；象者所以存意，得意而忘象……是故，存言者，非得象者也；存象者，非得意者也。象生于意而存象焉，则所存乃非其象也；言生于象而存言焉，则所存乃非其言也。然则，忘象者，乃得意者也，忘言者，乃得象者也。得意在忘象，得象在忘言。故立象以尽意，而象可忘也；重画以尽情，而画可忘也。

王弼的“忘象”，并非是说不需要言与象，而是说不要执着于言与象，意才是真正要追求的目的。

在言和象的基础上，唐代诗学引入了“境”的概念，来表示传达内心的体验。“境”本是佛学的概念，包括声、色、香、味、触、法等六种感受，称为“六境”。相对于具体物象的“象”，“境”是主观性的内部体验。王昌龄的《诗格》最早从诗的角度提出了境：

诗有三境，一曰物境：欲为山水诗，则张泉石云峰之境，极丽绝秀者，神之于心，处身于境，视境于心，莹然掌中，然后用思，了然境象，故得形似；二曰情境：娱乐愁怨，皆张于意而处于身，然后驰思，深得其情；三曰意境：亦张之于意而思之于心，则得其真矣。

王昌龄所论的“境”，是从文艺学角度提出的心理认识对象空间，与后来的语言学语境概念存在很大的差别，但是境的物、情、意三分对语言学语境的分析有启示意义。境创造的是一种立体化的心理体验环境，以境传意较之以视觉感受为表达方式的立“象”尽意，更能全方位地传达人的内心体验，尤其是那些无法借助语言 and 具体形象表达的部分。

晚唐诗人司空图总结了唐代的诗境理论，他的《二十四诗品》将诗歌意境分成雄浑、冲淡、高古、典雅等 24 种，每种各以 12 句形象化的韵语来比喻概括，同时也阐明了意境创造的原理和过程。司空图的“象外之象”“景外之景”“韵外之致”“味外之旨”“超以象外，得其环中”展现了“境”的超越性，境能够传达出那些言和象所无法传达的细致入微的体验，正如他对“含蓄”品格所描述的“不著一字，尽得风流”。

发端于老庄和《周易》，成熟于魏晋的“得意忘言”说，是中国哲学的基本思想之一，对中国传统思维方式的形成产生了深远的影响。“得意忘言”强调不要过多纠缠于语言本身，而应把重点放在文本之外的东西即“言外之意”“言外之旨”，无论是“言”“象”还是“境”都是得到“意”的方法手段。“象”和“境”的提出是建立在自然语言的局限性基础之上的，是对自然语言的一种有效补充。

(二) 自然语言局限性的表现

自然语言是人类思维概念的外化,黄曾阳先生在 HNC 理论中将两者所属的领域分别命名为自然语言空间和语言概念空间,自然语言是人类交际不可替代的工具,但是相对于思维概念,它也有其自身的局限性。这些局限性给人类的交际造成了一定的消极影响,中国古代的立象尽意,以境言意就是试图克服语言在交际中的不利因素。在计算机的自然语言理解中,由于计算机自身客观条件的限制,语言的局限性就表现得更为突出。自然语言的局限性主要表现在以下几个方面:

1. 自然语言无法准确完整地描述还原人的思维和感受。人类的思维和感受是与特定的语境关联在一起的,非常复杂细微并且不断变化,难以把握。人类把自己的认知和感受进行抽象概括,形成体系,以概念的形式存储在人脑中。人类的概念系统在外界是与自然语言空间相对应的,也就是说人类语言是以概念为中介对思维和感受的抽象概括,省略了很多特定语境中具体细微的部分,只是对其主要框架和要素的描述。但是,哲人和诗人思想感情中最精妙的地方往往蕴含在自然语言无法传达的部分。中国古人早就意识到了这一点,所以才提出了通过“象”和“境”来弥补自然语言的局限,而不是致力于自然语言的结构分析。

2. 关联性的缺失。自然界和人类社会的万事万物,彼此之间是相互关联的,这是宇宙的普遍法则,反映在人脑的概念系统中,各类概念也是彼此关联的,这种关联性是人类认识世界的成果,并将其固化到概念系统之中,成为系统的有机组成部分。概念关联的类型是多种多样的,黄曾阳先生对此进行过系统的总结,有定义关联、强关联、交互关联、因果关联、果因关联、互为因果关联、属于关联、包含关联、虚实设置关联、等同关联和对应关联。但是这些关联在自然语言中是表现不出来的,自然语言以字形和语音的形式存在,彼此之间相互独立,它们之间的关联无法从形式上观察到。人类在使用自然语言的时候,获得它们关联性的途径是先把自然语言和人脑中的概念对应,然后通过概念之间的关联复原语言之间的关联。但是,如果人的认知存在缺陷,如语言和概念的对应存在问题,人脑中概念关联知识储备不足等,也会影响人的理解和交际。

3. 自然语言中语形和语义对应的不确定性。在自然语言空间,语形和语义并不是一一对应的,往往是一个语形对应着数种语义,而同一种语义也可以用几种不同的语形来表达。以词汇为例,由于词义的引申,原先单个义项的词语,后来往往具有多个义项。还有,在古代汉语中,假借也是常见的语言现象,对于假借,许慎言道:“假借者,本无其字,依声托事,令长是也。”也就是说没有为某个词专门造字,而是使用一个已有的字形来记录它,这样同一个字形就负载了两个意义毫无关联的词。这是最基本的假借现象,实际的情况还要复杂得多。此外,自然语言空间中同义关系是普遍存在的语言现象,其中一部分是意义完全相同的绝对同义词,如“西红柿一番茄”,“公分一厘米”等,这样同一个意义就对应着不同的词形。这种语形和语义对应的不确定性也在一定程度上加大了语言理解和记忆的难度。

4. 自然语言空间的不稳定性。语言概念空间是人类认知和思维的基础,它虽然也随着社会的发展和人类认识能力的提高而不断完善,但稳定性是其主要方面,这也是语

言文化得以传承交流的基本条件。相对于语言概念空间,自然语言空间的不稳定性十分显著,语言的语音、词汇、语法等各种要素都处于不断的变化之中。比如汉语的语音系统经历了一个由简到繁再到简的发展历程,汉语词汇也逐渐由单音节向双音节发展,新词不断出现,旧词不断消亡,如《尔雅·释诂》同一词条中就有很多古语词。自然语言的这种不稳定性有其积极的一面,能够促进语言的不断发展,使之更适应人类交际的需要;但也有其消极的一面,比如在某一时期未经规范的新词大量出现,会在一定程度上造成交际和理解的混乱。

自然语言的局限性造成了语言交际中的种种模糊现象,人类依靠头脑中的语言概念空间和自然语言空间进行对应,还原了自然语言的关联性,在很大程度上消解了语言的模糊。但对于缺失概念空间的计算机而言,语言的模糊性是难以克服的。为计算机建立一个模拟人脑的语言概念空间是解决这一难题的根本出路。对此,中国传统训诂学的系统语义观具有重要的启示意义。

三、训诂学的系统观分析——以《尔雅》为例

(一)《尔雅》的性质

《尔雅》作为我国第一部训诂专著,反映了我国早期的训诂观念以及对语义、概念等方面的认识水平,具有重要研究价值。对于《尔雅》的性质,学术界历来存在不同的观点。

以苏宝荣(2000:148)先生为代表的观点认为《尔雅》是一部同义词词典,《尔雅》把同义的词语集中到一起,用一个通俗的词语加以解释,实际上是起到了同义词类聚的作用。

而以管锡华先生为代表的观点认为《尔雅》是一部义类词典,《尔雅》聚合的是同一类别的词,而非同义词语,从这个角度《尔雅》可以视为汉语类义词典的早期形式。

我们认为,相对而言类义词典的说法更接近《尔雅》的实际。目前词汇通常的语义分类大致可分为两种:一是在同一层次的语义分类,如分为同义词、反义词等;二是按照语义上下层级的类别进行分类。《尔雅》十九篇,前三篇收录的是表达抽象意义的词语,后十六篇收录的是表示具体事物的词语,其中又可分为表示与人的社会活动相关的词语(释亲、释言、释器、释乐、释畜)和表示自然界各种事物的词语(释天、释地、释丘、释山、释水、释草、释木、释虫、释鱼、释鸟、释兽)。《尔雅》的后16篇每篇都是一个独立的语义类别,16大类下再分39小类,共1580种事物。《尔雅》编排最核心的原则是分类,以类分篇,聚类成条是《尔雅》贯穿始终的编排方法。所以《尔雅》是一部类义词典。语义类别从根本上反映的是概念的类别,同一语义类别的词表达的是共同的概念,《尔雅》实际上是将词语按照相应的概念从属关系分门别类进行类聚,表示同一概念的词被编为一个词群,因此它也是一部概念类别词典。正如何九盈(1984)先生所说:“《尔雅》作为一部‘正名之书’,既是在辨析事物的概念(如属概念、种概念等),也是对名词进行释义,这在战国时代是一件很重要的事情。”《尔雅》的内容和编排本质上是当时的人对各种语言现象及世界万物的认识,在某种程度上是当时人脑中概念体系

的反映。

与按照形序和音序编排的词典不同,类义词典中的词语不会因为词形的不同而编排在不同地方,而是按照义项的类属不同出现在不同的类中。同一概念的词语编排在一起,它们所反映的是意义的共同之处,而对其相异之处往往不做辨识,这也是与重在考察词语意义差别的同义词词典的不同之处。

概念类别词典具有概念类别直观性强的优点,其主要作用在于能够给人们集中提供表达关于某一概念的词语。人类大脑中的概念空间是与自然语言空间相对应的,人类概念的表达需要借助自然语言中的词语。人脑中积累的与某一概念相对应的词语往往是有限的,所以在想表达某一概念时却一时找不到与之相对应的确切的词语,而概念类别词典把表达同一概念的词语类聚到一起,当人需要表达某一概念时就可以从中进行挑选,这也是它与一般的解释词义的词典最大的不同之处。这一点对于自然语言理解有着重要的借鉴意义,人能够理解自然语言的重要因素之一就是人脑中的概念体系和自然语言中的词语对应关系的建立,为计算机建立这样一部概念类别词典,是实现计算机智能化的必要手段。此外,概念类别词典把相同概念的词汇类聚到一起,为语言研究尤其是词汇研究提供了丰富的材料,可以由此观照这些词汇意义的异同,分布等,如多词一义、一词多义等语言现象。

《尔雅》按照概念类别进行编排的特点也有检索不便的缺点,这与汉语汉字自身的特点有关,同时也是中国传统重视整体的思维方式的体现。“只按书写形式或读音形式排列,将这个‘精神世界’或‘体系’弄得支离破碎。”(刘叔新,1984:86)而《尔雅》的编排方式,则把人的概念体系完整地展现出来。西方语言学早期的词典基本上是以音序排列,直到19世纪中叶英国皇家学会会员罗捷特(Roget)才编写出第一部类义词典《英语词库》(*Thesaurus of English Words*),其后概念分类的思想日益受到语言学界和词典编纂界的重视,这在某种意义上是向传统的回归。

《尔雅》首创的按概念类别编排的体例,对后代辞书的发展有很大影响,历代出现了一系列以“雅”为名的著作,直至今日仍对类义词典以及百科全书的编纂具有重要的借鉴意义。

(二)《尔雅》的分类

分类是逻辑学基本的工作之一,对于概念分类词典而言,最根本的问题是怎样对概念范畴进行科学的分类。直至今日,语言学界和词典编纂学界仍然没有确立一个科学统一的标准,各家的分类有着很大的差异。

罗捷特的《英语词库》把全部词语分为六大类:抽象关系、空间、物质、理智、意志、感情。类下分纲、目,共一千目。德国的《杜登词典》(*Duden Woerterbuecher*)对概念范畴进行了更为详细的分类,共分为18类:

- 一、人物;二、品德;三、智能;四、情感;五、感觉;六、仪容;七、处境;
- 八、交往;九、争斗;十、语言;十一、信息;十二、物质;十三、运动;十四、
- 时间;十五、空间;十六、数量;十七、程度;十八、性质。

梅家驹等编著的《同义词词林》以“同义词”冠名,实则也是一部概念分类词典。全部词语首先分为四个范畴:(1)人类及一切事物概念的本体名称;(2)人和事物的属性、运动及状态;(3)人以及事物之间的相互关系;(4)语言本身的特殊表达方式。以下又分为12大类(人、物、时间和空间、抽象事物、特征、动作、心理活动、活动、现象与状态、关联、助语、敬语)、94中类、1938个小类、3933个标题。

由以上可以看出,各家对概念的分类是存在着很大差别的,这些分类是否属于同一层级,能否完备地涵盖各种自然和社会现象,这些问题都有待于进一步研究。如《杜登词典》的分类,“人物”和“物质”属于具体概念,而其余的则属于抽象概念,“品德”“智能”“情感”和“感觉”属于人类活动范畴,而“时间”“空间”“数量”等则属于自然的基本属性范畴。如何建立一个科学的概念分类系统,是建立概念分类词典的关键。

中国对“类”的研究有着悠久的历史。“类”的研究成果主要体现在先秦诸子对“名实”的研究中。首先,尹文子将名分为“命物之名”“毁誉之名”和“况谓之名”三类。其后,墨子对“名”进行了更为细致的分类,按外延大小划分为“达名”“类名”和“私名”三类,以具体和抽象为标准划分为“以形貌分者”和“不可以形貌分者”两类,又以种属关系分为“兼名”和“别名”两类,同时《墨子·小取》中提出“以类取”的方法。墨子对“类”的研究达到了相当的高度。“由于《墨辩》提出了关于‘类同’‘类异’和‘划分’的思想,从而使‘类’概念在中国逻辑史上成了一个比较科学的概念,成了逻辑推论的基础和条件。”(周云之,刘培育,1984:151)

《尔雅》的分类体系是先秦逻辑分类思想在训诂实践中的应用,其中前3类的抽象概念和后16类具体概念的区分更是直接体现了墨子“以形貌分者”和“不可以形貌分者”的分类标准。李开(1999)说:“(《尔雅》)把人间天地万事万物统分成19类,和后来的《说文》据文字形体分成540部首一样,是先秦就有的类逻辑思想的体现。……《尔雅》是先秦训诂资料的汇编,它的做法,搜罗天地万物,汇集故训,分类纲统,类取类予,比物相附,释以雅言,其办法和精神仿佛就是《小取》逻辑的具体化。”

1. 《尔雅》的分篇

《尔雅》的分类思想首先体现在它的分篇上,其中表达具体概念的后16篇是按照具体概念的种属类别进行分篇的,其分类标准比较明显。而关于前3篇的分篇标准,学术界存在一定的争议。孔颖达以为:“《释诂》《释言》,通古今之字,古与今异言也。”(《诗经·周南·关雎》疏)其后,众多学者都对此提出过自己的见解。其中朱骏声的观点得到了学术界较为普遍的认同,“《释诂》者,释古言也;《释言》者,释方言也;《释训》者,释双声迭韵连语及单辞、重辞与发声助词之辞也。”(《说文通训定声·豫部·第九》)黄侃从另一个角度指出:“《广雅》:‘诂,言也。’然则诂与言对言有别,通言不殊。”“《释言》专释六艺成言。《释诂》顺衍而下,如云:古所谓初;今所谓始也。《释言》逆溯而上,如云:今所谓中;古所谓殷也。又《释诂》字大抵本义;《释言》字大抵非本义而引申。”(《尔雅音训》)黄侃认为《释诂》主要解释词的本义,《释言》主要解释词的引申义。各位学者提出的分类标准虽然各异,但是从中可以看出,《尔雅》前三篇的分类是基于语言形式的分类,而不是概念类别的分类。对这些反映抽象概念的词语依据语言形式进行分类,主要出于训诂实际的需要,但也在一定程度上反映了当时对抽象概念进行

更高层次概括能力的不足。

2. 《尔雅》各篇之下的分类

《尔雅》对各篇之中的事物也进行了分类，形成了不同的概念类别。在此《尔雅》主要运用了类聚之后进行求同和辨异的方法。

(1) 《尔雅》概念的类聚

类聚就是把表达相同概念的词语聚集在一起，形成词条。类聚在《尔雅》各篇都有运用，其中在《释诂》中运用最为广泛，如：

靖、惟、漠、图、询、度、咨、諏、究、如、虑、谟、猷、肇、基、访，谋也。

以上16个被训释词以“谋”作训释，共同具有的抽象概念是“谋划”，在HNC语义网络的概念节点中属于8行“思维”中的“831 策划”。《释言》和《释训》中的多词类聚相对较少。

后16篇概念类别的分类依据非常明确，它们依据各自反映的具体概念类属加以类聚，在各篇中大量运用，例如：

父为考，母为妣。父之考为王父，父之妣为王母。王父之考为曾祖王父，王父之妣为曾祖王母。曾祖王父之考为高祖王父，曾祖王父之妣为高祖王母。父之世父、叔父为从祖祖父，父之世母、叔母为从祖祖母。（《释亲》）

宫谓之室，室谓之宫。牖户之间谓之扂，其内谓之家，东西墙谓之序，西南隅谓之奥，西北隅谓之屋漏，东北隅谓之宦，东南隅谓之窔。（《释宫》）

木豆谓之豆，竹豆谓之筴，瓦豆谓之登。（《释器》）

宫谓之重，商谓之敏，角谓之经，徵谓之迭，羽谓之柳。（《释乐》）

春为苍天，夏为昊天，秋为旻天，冬为上天。（《释天》）

两河间曰冀州，河南曰豫州，河西曰雍州，汉南曰荊州，江南曰扬州，济河间曰兖州，济东曰徐州，燕曰幽州，齐曰营州。（《释地》）

丘一成为敦丘，再成为陶丘，再成锐上为融丘，三成为昆仑丘。（《释丘》）

河南，华；河西，岳；河东，岱；河北，恒；江南，衡。（《释山》）

水中可居者曰洲，小洲曰渚，小渚曰沚，小沚曰坻，人所为为渚。（《释水》）

藿，山韭。茗，山葱。薺，山薺。蒿，山蒜。薜，山蕮。（《释草》）

枣，壶枣；边，要枣；檇，白枣；楸，酸枣；杨彻，齐枣；遵，羊枣；洗，大枣；煮，填枣；蹶泄，苦枣；皙，无实枣；还味，棣枣。（《释木》）

次蠹，蜘蛛；蜘蛛，蛛蝥。土蜘蛛。草蜘蛛。（《释虫》）

贝，居陆贖，在水者蛎。大者鮓，小者贔。玄贝，贻贝。余贐，黄白文。余泉，白黄文。（《释鱼》）

鸛雉。鸛雉。鸛雉。鸛雉。秩秩，海雉。鸛，山雉。（《释鸟》）

麋，牡麋，牝麋，其子麋，其迹麋，绝有力狄。鹿，牡鹿，牝鹿，其子鹿，其迹速，绝有力麋。（《释兽》）

羊，牡羊，牝羊。夏羊，牡羊，牝羊。角不齐，觝。角三觝，羴。（《释畜》）

把表达相同概念的词语形成类聚是进一步建立概念分类体系的基础。对于这些类聚