◆迈向 圆世纪的语言学◆

主摇编摇吴国华
副主编摇王铭玉

计 算 语 言 学

易绵竹摇南振兴摇编

上海外语教育出版社

摇图书在版编目(悦陨孕)数据
摇计算语言学辕易绵竹,南振兴编郾—上海:上海外语教育出版社,圆园园缘
摇(迈向圆员世纪的语言学丛书 辕桂诗春 宁春岩 钱冠连 王初明 沈家煊 林克难 姜望琪 刘润清主编)
摇摇摇陨杂月晕苑 原缘愿员园原员员远远原缘
摇Ⅰ郾计…摇Ⅱ郾①易…②南…摇Ⅲ郾计算语言学—文集摇Ⅳ郾匀园怨园原缘猿
摇中国版本图书馆 悦陨孕数据核字(圆园园缘)第 园猿缘远远猿号

摇摇摇

摇摇摇摇出版发行:上海外语教育出版社
摇摇摇摇摇(上海外国语大学内)摇邮编:圆园园园愿猿
电摇摇话:园圆员原远缘源圆缘猿园园(总机),远缘源圆圆愿猿缘(发行部)
电子邮箱:遭蚤藻蚤泽澡赃遭怎泽蚤灶藻泽泽岳泽澡枣造藻责援糟燥皂
网摇摇址:澡贼贼责:辕辕憎憎憎援泽澡枣造藻责援糟燥皂摇摇澡贼贼责:辕辕憎憎憎援泽澡枣造藻责援糟燥皂援糟灶
责任编辑:冯玉柱

印摇摇刷:摇摇摇摇摇摇摇摇摇摇印刷厂
经摇摇销:新华书店上海发行所
开摇摇本:摇摇 伊摇摇 员辕猿圆摇摇印张摇摇摇字数摇摇千字
版摇摇次:圆园园缘年 圆月第 员版摇圆园园缘年 圆月第 员次印刷
印摇摇数:摇摇册

书摇摇号:陨杂月晕苑 原缘愿员园原员员远远原缘 辕匀·员园远
定摇摇价:圆员郾园园元
摇本版图书如有印装质量问题,可向本社调换

总摇序

摇摇我在“对中国英语教育的若干思考”(《外语研究》圆猿猿圆年第 猿期)一文的附注中曾有这么一段话:“感谢社会主义的市场经济,由于上外出版社和北外出版社的竞争,近两年它们分别引进出版了大批语言和文学原版书,受到广大导师和研究生的欢迎。”人们可以从不同视角去解读这两家权威出版社互动的深远意义,如这是中国改革开放在外语出版界的反映,这是中国参加 宰耦的必然结果,这是中国高等教育走向现代化、国际化、全球化的标志,不一而足。我们教师是讲实际的,有书看比没有书看好,有许多许多好书比一门课只念一本书好。但是,人们也清醒地认识到,事物总是两方面的:现代化、国际化、全球化不能光抓引进一面。无论是老师,还是学生,都想了解国内的情况,知己知彼,才能学得深入,学以致用。何况我们也应立足于培养高水平的人才,拿出自己的具有国际水平的科研成果。所幸解放军外国语学院的老师们早看到了这一点,近几年来立项研究国内语言学界的研究动向和成果,这套《迈向 圆员世纪的语言学》系列丛书便是他们深邃远见所凝聚的心血。

语言学的领域广,分支多,要在短时间内摸透所有的学科,未免强人所难,不合实际。幸好本书的编者们不愧是有素养的军人,能集中兵力,围攻主要战略目标,这主要战略目标便是当代语言学的最具代表性的学科和具有时代意义的前沿学科。在本丛书八个分册中:《普通语言学》和《应用语言学》是有关语言学的基本理论和实践的当家学科,当在首列;语言作为人们交际的符号之一,

而且是最主要的符号工具，说明《符号语言学》的重要性；语言是用来表达思想和做事的，《语义学》和《语用学》分别从语言本身和语境讨论意义的表述和理解；语言是人在所生活的社团中交际的工具，在这个意义上，《社会语言学》和《语言国情学》把我们引入探讨语言和使用者之间的关系，语言使用者和民族文化、习俗等关系，以及对语言使用的影响；最后，上世纪中叶启动的《计算语言学》必然在新世纪得到飞速的发展，它在多媒体教学、机器翻译、语料库建设、信息技术等领域的作用将关系到一个民族的存亡。

[illegible]年，当我和一些老师合作编写《语言学教程》一书时，王宗炎先生和许国璋先生在“第一版序”中曾给过一针见血的但又是善意的批评和帮助。不妨转录如下：

[illegible]引进的理论，能用汉语说得清，讲得懂；能用汉语的例证加以测验。

[illegible]凡有可能，不妨采用现场工作法。我国社会语言学、心理语言学和测试学研究者已做出榜样，值得学习。

[illegible]凡在汉语诸范畴中验证外国某一理论，其有解释力者肯定之，其解释力太弱或不具解释力者指出之，其主观臆测者直言之，不以权威而护短，不以宗师而慑服。

[illegible]尊重我所不懂或不明白价值所在的理论，不以有用无用、正统邪说为取与舍的标准。对理论有矢志不渝的精神，理解深，教得熟，力求贯通、比较、自创。

[illegible]汉语研究者中的前辈已经做出的自创，外语系出身的研究者应该认真读，读懂，直至应用到自己的研究工作。

十五年过去了，上述情况是否有所改观呢？我认为是有的。本丛书所遴选的论文可以佐证。这些论文印下了同行们前进的足迹。有不少论文对国外的理论比过去说得更清楚，有不少论文注重现场调查，有不少论文选用汉语例证，有不少论文试图提出自创的理论，更令人高兴的是有许多论文出自汉语界研究者，为我们提供了学习的机会。如果考虑到本丛书作者群中有各个语种的研究

者,也有计算机科学的研究者,在我们眼前顿时涌现了一支生机勃勃的来自四面八方的语言学大军,我国的语言学研究自当会有更蓬勃的发展!

还应指出,解放军外国语学院的老师们的工作不仅仅是遴选论文,而是为每一分册整理了该学科的概观、发展和趋势的引论,并对每篇文章进行客观分析与评论。这一编辑思想保证了论文集走上一个新的台阶,也是对解放军外国语学院师资力量和水平的检阅。愿他们今后取得更大成就。

胡壮麟

二〇〇三年 元月

北京大学蓝旗营

前摇言

摇摇近几十年来,国外语言学研究的进展迅速,国内学者的研究成果更是汗牛充栋。挖掘学科成果,提炼学术思想,形成系列研究,巩固学科发展,弘扬民族精神,是编撰本丛书的主要目的。

本丛书以八大语言学前沿学科为线索,以我国具有代表性的研究成果为基础,对各学科在我国的发展进行系统的总结,对我国语言学家的重要思想观点进行评析,对 圆员世纪语言学的发展做出展望。丛书共有八个分册,分别为《普通语言学》、《符号语言学》、《应用语言学》、《语义学》、《语用学》、《社会语言学》、《计算语言学》、《语言国情学》。各部分的主体结构如下:(员 总序;(圆 引论(包括学科概观、学科在我国的发展、学科在 圆员世纪的发展趋势等内容);(猿 圆员世纪(主要是 员怨愿园年以后)我国学者的代表性研究成果;(源 评析(对每篇学术文章进行客观分析与评价);(缘 主要文献索引。

应该说,中国的语言学研究成果并不算少,但各自为政、零散分布是主要问题。在进入 圆员世纪之际,系统研究、合力共现,将对中国的语言学研究与发展具有重要的现实意义。

另外,我国高校的语言理论教学对国内外学术思想和科研成果的追踪和发展往往重视不够,重复性劳动普遍存在,造成严重的人才和资源上的浪费。本丛书将为展示前沿科研成果、促进语言理论的研究和教学提供较为完备的案头参考。

鉴于编者涉猎面有限,在遴选论文的过程中,难免会挂一漏万,“引论”中的一些观点和对每篇论文的“评析”也会有不够准

确、客观之处；另外，本丛书所收录的论文基本上选自国内的一些学术期刊，为便于读者查阅原作，编者未对论文的体例作大的调整和统一，尽量保持原来的样式，敬请同行专家批评指正。

需要说明的是，编者在选定论文之后，向每篇论文作者发去了联系函，并得到了他们的积极响应和大力支持。在此，我代表课题组向各位专家、学者表示衷心感谢。遗憾的是，个别原作者由于地址变更等原因，至今未能联系上。在此，敬请有关论文作者或其亲朋好友见书后及时将有关信息通知我们或出版社，以便取得联系。

吴国华

二〇〇五年五月

目录

引摇论

易绵竹摇南振兴

近半个世纪以来,特别是改革开放 圆园多年来,我国计算语言学的基础理论研究和应用开发研究取得了辉煌成就,但同时也存在诸多难题和瓶颈亟待突破。不言而喻,中国计算语言学研究的首要目标应当定位在关注与提升中文信息处理的理论发现和技术发明水平。回顾并展望该学科在我国的发展历程、研究现状及未来趋势,必将有助于锤炼广大计算语言学研究者的原则性思维与技术性思维,积极探索针对中文信息处理领域若干语言学保障与程序保障问题的有效解决方案和突破路径,从而增强我国新世纪 隐息产业的国际竞争力。

一、中国计算语言学的学科背景及发展历程

任何科学研究都有其学术背景。追溯中国计算语言学的发展历程,其主要目的就是为了厘清和把握该学科的起源、特色、分期等有关学科发展背景方面的线索与脉络。

员.计算语言学是时代发展和科技进步的必然产物

中国工程院院士、中国中文信息学会理事长倪光南先生(圆园园缘)曾撰文畅谈计算机的过去、现在和将来,文章开宗明义指出,计算机无疑是人类历史上最重大的发明之一,计算机的发明导致了信息革命,使人类社会进入了信息社会。计算机不仅推动了经济领域的变革,而且还推动了文化、科技和生活等领域的变革。

因此可以说,计算机对人类社会影响的深度和广度为其他任何发明所不及。为了加深对信息社会本质的认识,我们援引中国工程院院士汪成为先生([illegible])的一段话:“集成电路是信息社会的细胞,通讯网络是信息社会的神经,计算机是信息社会的大脑,信息资源是信息社会的血浆”。

随着计算机技术的迅猛发展和日益普及,使人造工具由过去只是人手的延长与扩展演化为人脑的模拟与延伸,正是在这个意义上,计算机又被通俗地称作“电脑”([illegible])。人机交流对话系统成为新的认识主体,人的认识能力得到极大的提高,从而使人对外在物质世界、内在精神世界和由抽象概念或观念所形成的语言世界的感知与认识水平提到空前高度,其影响力已超出科技层面,渗入到社会生活的各个领域,诸如行政管理、生产经营、教学科研、新闻资讯、信息检索等行业和部门都普遍使用电脑,电脑文化蔚然成风。正是电脑文化使当今社会呈现出[illegible]种基本形态:以数字化为特征,以网络化为基础,以信息化为目的。美国当代著名未来学家尼葛洛庞蒂([illegible])认为:“计算不再只和计算机有关,它决定我们的生存”,这就是所谓“数字化生存”([illegible])。要了解数字化生存方式的实质,就应清楚地认识到,作为“信息[illegible]”的比特([illegible])正迅速取代原子([illegible])而成为人类社会的基本要素。“信息高速公路”([illegible])正是以光速在全球传输没有重量的比特,在国际互联网上互传比特的数字一族,逐渐营造出一种崭新的生活方式。另一位美国著名未来学家托夫勒([illegible])曾预言:“谁掌握了信息,控制了网络,谁就将拥有整个世界”。我国著名信息科学家钟义信([illegible])在全面审视信息化的时代特征之后,从现实国情和国家利益的高处着眼,提出必须采用“超越战略”,高速度、高质量地实现工业、农业、科学技术和国防的全面现代化,并充分论述了信息化是生产力发展的现实要求,强调构建作为信息化的社会基础结构——信息网及其传输交换平台——通信

网的重要性，积极倡导开展一项名为“中国人的和谐信息化国家行动”（[illegible]）。

众所周知，信息作为客观世界存在的三大要素之一（另两个要素是物质和能量），其最主要的载体是自然语言。在信息时代，属于人文科学传统学科的语言学在现代科学体系中的地位发生了急剧变化，它和属于现代科学技术前沿的电脑、控制论、系统论、信息论及自动化技术发生牵丝攀藤的交叉性联系。特别是圆园世纪源园年代，由计算机科学和语言学结合而产生的计算语言学，不仅成了计算机进入人文科学领域一个重要的里程碑，而且发展为应用语言学学科群中最具现代化色彩的一门学科，在心理语言学、社会语言学以及语言教学等众多学科领域均积极采用计算机模拟实验方法，语言研究的计算机化特色日趋明显。从国际学术发展的大背景来看，早期计算语言学研究的中心课题一直是机器翻译。圆园世纪远园年代，美国率先成立计算语言学学会和国际计算语言学委员会，世界各国的相关研究于是逐步走上有组织阶段，由此形成一门以计算语言学理论为基础、具有广泛应用价值的语言工程学。中国的计算语言学研究，如果从员怨缘苑年开始的俄汉机器翻译试验算起，与其他国家相比，起步并不晚。但是，我国计算语言学的相关研究在愿园年代中后期才全面展开并形成规模，其标志性事件就是成立中国中文信息学会和隶属于该学会的自然语言处理专业委员会和计算语言学专业委员会。员怨愿愿年，我国台湾省也成立了计算语言学学会。

圆援计算语言学的学科定位与学科特色辨析

如何确定计算语言学在现代语言学中的学科位置？我国学者刘涌泉、乔毅（员怨怨员：员—员缘）把应用语言学区分为一般应用语言学和机器应用语言学，将工程语言学、机器语言学、计算语言学列为后者的研究内容，其特点是利用电子计算机等先进工具来处理自然语言。两位学者在综述该学科的各项研究课题之后，构拟了一

个涉及范围相当广泛的中文信息处理系统。显然，计算语言学应当归属于应用语言学的研究范畴。

计算语言学的研究对象是什么？美国学者格里什曼（[illegible]）将该学科定义为“对自然语言理解与生成的计算机系统之研究”（[illegible]）。实际上，计算语言学与自然语言处理（[illegible]）、自然语言理解（[illegible]）等术语有着密不可分的联系，前者是从理论和方法角度而言，后者是从技术和应用角度而言。为了更准确地把握该学科的研究内容和特色，在此我们不妨援引《[illegible]认知科学百科全书》（英文）对词条“[illegible]”的界定（试译成中文）：“计算语言学（[illegible]）亦称自然语言处理，[illegible]涉及：（1）研究语言结构和功能、语言使用和语言习得的计算模型；（2）设计、开发和实施各种系统，如语音识别、语言理解和自然语言生成等。[illegible]的应用领域包括：与数据库、文本处理和消息理解的接口，外文函电、网页和受限领域中语音翻译（[illegible]）等方面的辅助性多语言接口。在理论上，计算语言学利用计算模型对语言的句法、语义、语用（指说话人与受话人之间关系的某些方面，而在计算语言学系统情形中是指用户与系统之间的特定关系）以及语言的话语（[illegible]）诸方面进行研究。这些研究具有学科交叉性（[illegible]），并从人工智能（[illegible]）、语言学、逻辑学和心理学中借用一些概念。由于计算机科学和语言学的密切联系，计算语言学在认知科学中起着关键作用。”（[illegible]）由此可见，在某种程度上，计算语言学与自然语言处理是同义术语，它作为认知科学及人工智能的一个分支，其主要特色是学科交叉性。

我国著名计算机专家刘开瑛与郭炳炎（[illegible]）较早从事中文信息处理的研究工作，他们将自然语言处理划归于高技术学科之列，认为它是知识信息处理中的核心课题，指出目前自然语言处

理的研究已打破传统的语言学、心理学、数学以及计算机科学的界限，通过这些有关学科之间的互相渗透、互相影响，并已形成了具有新概念、新理论、新技术的交叉学科——计算语言学。我国著名计算语言学家俞士汶（员怨怨缘：猿–愿）从汉语信息处理基础研究相当薄弱的现实出发，着重阐明了计算语言学具有基础学科性质的特点，将现代汉语语法电子词典、语料库标注、汉语短语结构体系及其形式化确定为先期开发的基础工程。我国计算语言学研究先驱者之一冯志伟（员怨愿员：Ⅲ–员–圆）在简明论述由自然语言处理技术所形成的边缘性交叉性学科之基础上，指出自然语言处理既是电子计算机模拟人类智能的一个重要方面，也是研制智能化电子计算机的一项基础性工作，清晰地描述了自然语言计算机处理需要经过的猿个过程，并强调该学科同时涉及文科、理科和工科三大知识领域。此外，我国汉语语言学研究者姚亚平（员怨怨缘：缘猿–远愿）用“新兴学科”、“应用学科”和“交叉学科”来概述中国计算语言学的性质和特征。我国两位留学海外的学者翁富良与王野翊（员怨愿：员–猿）把计算语言学定义为一门边缘学科，详细列举了与该学科研究密切相关的众多学科。

我国台湾著名学者黄居仁（员怨怨苑）从科际整合与整合科技的角度论述计算语言学与语料库语言学之角色与发展，指出这两门学科一向被认为是标准的科际整合，进而断定它们势将成为人文社会科学中最关键的整合性科技。在学术研究方面，这两项研究将提供认知科学、自然语言处理以及所有文本研究的基础架构。在应用方面，信息时代的来临与信息高速公路的普及，意味着计算语言学将提供驾驭信息的基本工具，而语料库语言学则提供信息的架构与内容。根据科际整合与整合科技的理念，黄先生（员怨怨园）把计算语言学看作是个整合学科，它是人工智能、语言学、认知科学的结合，无论采用何种定义，计算语言学都应当包括：构建处理语言的计算机系统以及自然语言规范模式的研究。台湾另一位著名学者谢清俊（员怨怨园）认为，“计算语言学”就字面看来，仍是以“语

言学”为主角，它是语言学的一支，是在运用计算机的环境下，为解决语言学上的问题所产生的学问。由于它必须在运用计算机的环境下发展，所以它为解决问题所用的方法、工具，甚至于观念等都和传统的语言学大不相同；更由于计算机在储存、记忆、计算，甚至推理和判断方面，做起事来和人的能力与性质不尽相同，于是利用计算机做研究所表现出的功能和结果也和传统人做的计算语言学研究迥异。因此计算语言学能产生新的语言学理论和模式，这种影响力直接震撼了语言学的根本。对于计算机科学而言，计算语言学研究的成果直接提升了计算机处理语言的能力，不仅是对计算机使用的人工语言如此，更使计算机具有处理自然语言的能力，进而使计算机由数据（[illegible]）处理提升至信息（[illegible]）处理，进而能够进行知识（[illegible]）处理。这种计算机能力的跃升情形也对计算机科学的发展和应用带来巨大的冲击，因此计算语言学的发展引起了计算机科学与语言学相互的激荡，不仅在应用上相辅相成，在基本理论上也迭有新猷。作者换一个角度来说，正因为有新的理论产生，才构成了成立新学问——计算语言学的条件，这是计算机科学和语言学科际整合的结果。作者指出，计算语言学（[illegible]）、社会语言学（[illegible]）和心理语言学（[illegible]）都是语言学和另一门学问经整合而产生的新学问，它们位居两种原有的学问之间，扮演着桥梁的角色，而这个桥梁的关系则可经由理论上和应用上两个角度加以说明。从理论研究范畴和主要应用领域这两个方面，作者阐述了中文计算语言学发展的构思。

特别值得一提的是，当前语言研究更趋明显的技术化色彩引起了国际学术界的普遍关注。俄罗斯著名语言学家博格达诺夫（В.В.Богданов，[illegible]）指出，[illegible]世纪中期以来，人类社会信息化的步伐明显加快，电子计算机和国际互联网技术日益渗透到人们的生活领域和学术领域，由此引发了所谓的“技术语言学革命”（технолингвистическая революция），信息、意思、意义、内容

等成为研究者的关注焦点，因而“语义中心论”（семантикоцентризм）在当代语言学中开始占据主导地位。我国理论语言学研究者李葆嘉（圆圆圆：圆圆圆—圆圆圆）在阐述自己提出的语言科学与语言技术新思维时宣称，当代语言学已经凸显“语言科学”与“语言技术”的二分互补格局，语言技术包括语言文本处理技术和语言系统模拟技术。作者指出，界定计算语言学要考虑到工具性、技术性和理论性这三个方面的特点。因此，作者把计算语言学界定为利用计算机作为语言研究工具和探索语言的可计算性问题、同时建构基于计算机科学等相关学科基础之上的语言理论的新学科，并将面向语言系统模拟的计算语言学研究比喻为“电脑模拟语言能力工程”或“语言能力移植电脑工程”，认为语义性是人类语言的本质共性，而语义结构网络研究乃是语言能力移植电脑工程的基础研究，这一研究可比喻为“语言基因图谱分析工程”，强调研究对象的语义性转移和研究成果的技术性趋势将成为圆圆世纪的语言学精神。通过进一步论证，作者提出了“语言科技”这一新概念，其内涵是以语言学研究为枢纽，沟通计算机应用、数学、认知科学和现代教育技术等相关学科以实现语言信息处理的技术化，而其外延表现为语言信息科技、语言教学科技和语言研究科技。语言信息科技的目标是实现自然语言的计算机理解和生成，研究领域是人—机对话；语言教学科技的目标是实现教学的媒体多维化和网络远程化，研究领域是人—人对话；语言研究科技的目标是实现语言研究自身的计算机程序化，研究领域是语言研究工具。应当承认，这一新思维既突出了当代科技发展所要求的“语言技术论”，又体现了以语言学为本而沟通文理工相关学科的研究旨趣。

此外，我国具有不同学术背景的研究者对与计算语言学相关的基础理论及工程开发也表现出浓厚兴趣。比如刘莎在《计算机世界》报发表的“破解人类语言的数字基因密码”（圆圆圆：圆圆圆—圆圆圆），邹晓辉在《系统科学

之窗》发表的“融智学新范式”（[illegible]）和在潜科学网站发表的“语言及语义信息的统一参照系”（[illegible]）等论文，对语义信息编码及处理技术做了有益的探索。

中国计算语言学的发展时期及主要业绩

如前所述，计算语言学研究肇始于机器翻译（下称机译）实验。可以说，中国计算语言学的发展史实际上就是一部机译研究史。下面我们主要根据发布在“[illegible]”网站（[illegible]）冯志伟的论文“面向计算机的语言研究（三）”，对我国机译研究的历史作一简要回顾（我们对个别文字略作改动）。

冯先生指出，我国是继美国、（前）苏联、英国之后，世界上第四个开展机译研究工作的国家。但是与国外机译的发展情况相比，我国除了有草创期、复苏期和繁荣期之外，由于文化大革命的影响，还有一个非常特别的时期——停滞期。并且由于我国在机译理论和方法上以及设备上的底子都很薄，因此每一个时期又都比同时期的国外机译稍微滞后。这是作者对我国机译研究状况和特点的总体评估。

1. 草创期（[illegible]）：在这个时期，我国学者对机译进行了初步的探索和试验。[illegible]年，国家便把机译研究纳入我国科学工作的发展规划，成为其中的一个课题，课题名称是“机器翻译、自然语言翻译规则的建立和自然语言的数学理论”。[illegible]年中国科学院语言研究所与计算技术研究所合作开始俄汉机译试验，[illegible]年国庆十周年前夕在我国[illegible]大型电子计算机上该项试验获得了成功，一共翻译了[illegible]个不同类型的较为复杂的句子。输入是利用穿孔纸带，输出的是代码而不是汉字，这在当时已是了不起的成果。[illegible]年由中国科学技术情报研究所编辑、科学技术文献出版社出版的《机器翻译论文选辑》，集中反映了以刘涌泉先生为首的课题组在俄汉机译中所取得的开创性研究成果。在草创期，北