外研社学术文库·当代国外语言学与应用语言学

The Oxford Handbook of Computational Linguistics

牛津计算语言学手册

Ruslan Mitkov 编

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS
牛津大学出版社
OXFORD UNIVERSITY PRESS

外研社学术文库·当代国外语言学与应用语言学

The Oxford Handbook of Computational Linguistics

牛津计算语言学手册

Ruslan Mitkov 编

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS
牛津大学出版社
OXFORD UNIVERSITY PRESS
北京 BEIJING

京权图字：01-2009-5082

图书在版编目(CIP)数据

牛津计算语言学手册：英文/（英）米特科夫（Mitkov，R.）著. — 北京：外语教学与研究出版社，2009.8
（当代国外语言学与应用语言学文库）
书名原文：The Oxford Handbook of Computational Linguistics
ISBN 978-7-5600-8529-6

Ⅰ. 牛… Ⅱ. 米… Ⅲ. 数理语言学—手册—英文 Ⅳ. H087-62

中国版本图书馆CIP数据核字（2009）第147796号

出 版 人：蔡剑峰
责任编辑：官亚平
封面设计：牛茜茜
出版发行：外语教学与研究出版社
社　　址：北京市西三环北路19号（100089）
网　　址：http://www.fltrp.com
印　　刷：北京京科印刷有限公司
开　　本：650×980　1/16
印　　张：52.25
版　　次：2012年8月第1版　2012年8月第1次印刷
书　　号：ISBN 978-7-5600-8529-6

*　　*　　*

购书咨询：(010)88819929　　电子邮箱：club@fltrp.com
如有印刷、装订质量问题，请与出版社联系
联系电话：(010)61207896　　电子邮箱：zhijian@fltrp.com
制售盗版必究 举报查实奖励
版权保护办公室举报电话：(010)88817519
物料号：185290001

导读

⊙ 冯志伟

一　计算语言学的发展历史与现状

计算语言学（Computational Linguistics）是当代语言学中的一门新兴学科，在这门学科的发展过程中，人们曾经从计算机科学、电子工程、语言学、心理学、认知科学等不同的学科角度分别进行过研究。之所以出现这种情况，是由于计算语言学涵盖了一系列性质不同而又彼此交叉的学科。这里，我们简要介绍计算语言学的萌芽期、发展期、繁荣期，并分析计算语言学当前的一些特点。

计算语言学的萌芽期

从 20 世纪 40 年代到 50 年代末这个时期是计算语言学的萌芽期。

在“计算语言学”这一术语出现之前，关于语言与计算的研究早就开始了。有 4 项基础性的研究特别值得注意：

——一项是关于马尔可夫模型的研究；

——一项是关于可计算性理论和图灵机模型的研究；

——一项是关于概率和信息论模型的研究；

——一项是关于形式语言理论的研究。

早在 1913 年，俄罗斯著名数学家 A. A. Markov 就注意到俄罗斯诗人普希金的叙事长诗《欧根 · 奥涅金》（Eugene Onegin）中语言符号的出现概率之间的相互影响，他试图以语言符号的出现概率为实例，来研究随机过程的数学理论，提出了“马尔可夫链”（Markov chain）的思想，

他将这一开创性的成果用法文发表在俄罗斯皇家科学院的通报上[1]。后来Markov的这一思想发展成为在计算语言学中广为使用的“马尔可夫模型”(Markov model)，是当代计算语言学最重要的理论支柱之一。

在计算机出现以前，英国数学家A. M. Turing就预见到未来的计算机将会对自然语言研究提出新的问题。

1936年，Turing向伦敦一家权威的数学杂志投了一篇论文，题为《论可计算数及其在判定问题中的应用》。在这篇开创性的论文中，Turing给“可计算性”下了一个严格的数学定义，并提出了著名的“图灵机”(Turing machine)的数学模型。“图灵机”不是一种具体的机器，而是一种抽象的数学模型，使用这样的数学模型可以构造一种十分简单但运算能力极强的计算装置，用来计算所有能想象得到的可计算函数。1950年10月，Turing在《机器能思维吗》一文中指出：“我们可以期待，总有一天机器会同人在一切的智能领域里竞争起来。但是，以哪一点作为竞争的出发点呢？这是一个很难决定的问题。许多人以为可以把下棋之类的极为抽象的活动作为最好的出发点，不过，我更倾向于支持另一种主张，这种主张认为，最好的出发点是制造出一种具有智能的、可用钱买到的机器，然后，教这种机器理解英语并且说英语。这个过程可以仿效教小孩子学说话的那种办法来进行。”Turing提出，检验计算机智能高低的最好办法是让计算机来讲英语和理解英语，他天才般地预见到计算机和自然语言将会结下不解之缘。

20世纪50年代提出的自动机理论来源于Turing在1936年提出的可计算性理论和图灵机模型，Turing划时代的研究工作被认为是现代计算机科学的基础。Turing的工作首先导致了McCulloch-Pitts的神经元(neuron)理论。一个简单的神经元模型就是一个计算的单元，它可以用命题逻辑来描述。接着，Turing的工作还激发了Kleene关于有限自动机和正则表达式的研究。

1948年，美国学者C. E. Shannon使用离散马尔可夫过程的概率模型来描述语言的自动机。

Shannon的另一个贡献是创立了“信息论”(information theory)。他

[1] A. A. Markov, Essai d'une recherche statistique sur le texte du roman “Eugene Onegin” illustrant la liaison des epreuve en chain, Bulletin de l'Academie Impériale des Sciences de St-Pétersbourg, 7, 153-162.

把通过诸如通信信道或声学语音这样的媒介传输语言的行为比喻为“噪声信道”（noisy channel）或者“解码”（decoding）。Shannon 还借用热力学的术语“熵”（entropy）来作为测量信道的信息能力或者语言的信息量的一种方法，并且他用概率技术首次测定了英语的熵[2]。

1956 年，美国语言学家 N. Chomsky 从 Shannon 的工作中吸取了有限状态马尔可夫过程的思想，首先把有限状态自动机作为一种工具来刻画语言的语法，并且把有限状态语言定义为由有限状态语法生成的语言。这些早期的研究工作产生了“形式语言理论”（formal language theory）这样的研究领域，采用代数和集合论把形式语言定义为符号的序列。Chomsky 在研究自然语言的时候首先提出了“上下文无关语法”（context-free grammar），后来，Backus 和 Naur 等在描述 ALGOL 程序语言的工作中，分别于 1959 年和 1960 年也独立地发现了这种上下文无关语法。这些研究都把数学、计算机科学与语言学巧妙地结合了起来。

Chomsky 在计算机出现的初期把计算机程序设计语言与自然语言置于相同的平面上，用统一的观点进行研究和界说。他在《自然语言形式分析导论》一文中，从数学的角度对语言提出了新的定义，指出：“这个定义既适用于自然语言，又适用于逻辑和计算机程序设计理论中的人造语言”。在《语法的形式特性》[3]一文中，他专门用了一节的篇幅来论述程序设计语言，讨论了有关程序设计语言的编译程序问题，这些问题是作为“组成成分结构的语法的形式研究”从数学的角度提出来，并从计算机科学理论的角度来探讨的。他在《上下文无关语言的代数理论》一文中提出：“我们这里要考虑的是各种生成句子的装置，它们又以各种各样的方式，同自然语言的语法和各种人造语言的语法二者都有着密切的联系。我们将把语言直接地看成在符号的某一有限集合 V 中的符号串的集合，而 V 就叫做该语言的词汇……我们把语法看成是对程序设计语言的详细说明，而把符号串看成是程序。”在这里乔姆斯基把自然语言和程序设计语言放在同一平面上，从数学和计算机科学的角度，用统一的观点来加以考察，对“语言”、“词汇”等语言学中的基本概念，获得了高度

[2] C. E Shannon. “A Mathematical Theory of Communication” [J]. *Bell System Technical Journal*. 27 (1948): 379-423.

[3] N. Chomsky. “Formal Properties of Grammars”. In *Handbook of Mathematical Psychology*, Vol. 2. Ed. R. D. Luce, R. R. Bush, and E. Galanter. New York: Wiley and Sons, 1963.

抽象化的认识。

Markov, Turing, Shannon 和 Chomsky 这四位著名学者对于语言和计算关系的探讨，是计算语言学萌芽期最重要的研究成果。

在应用研究中，计算语言学首先在语音的计算方面取得了令人振奋的成绩。1946 年，König 等研究了声谱，为尔后语音识别奠定了基础。20 世纪 50 年代，第一台机器语音识别器研制成功。1952 年，贝尔实验室的研究人员研制的语音识别系统，可以识别由单独一个说话人说出的 10 个任意的数目字。该系统存储了 10 个依赖于说话人的模型，它们粗略地代表了数目字的头两个元音的共振峰。贝尔实验室的研究人员采用选择与输入具有最高相关系数模式的方法来进行语音识别，达到了 97% –99% 的准确率。

在 20 世纪 50 年代末期到 60 年代中期，处于萌芽期的计算语言学明显地分成两个阵营：一个是符号派 (symbolic)，一个是随机派 (stochastic)。

符号派的工作可分为两个方面。

一方面是 20 世纪 50 年代后期以及 60 年代初期和中期 Chomsky 等的形式语言理论和生成句法研究，很多语言学家和计算机科学家热衷于研究剖析算法。1960 年，J. Cocke 提出使用二分的上下文无关规则来分析自然语言的 Cocke 算法，接着，Younger 和 Kasami 等分别进行这种算法的研究，形成了 Cocke-Younger-Kasami 算法（简称 CYK 算法），同时提出的分析算法还有自顶向下分析算法、自底向上分析算法、动态规划算法。这样一来，形式语法理论便成为了一种可以计算的理论，被直接应用到自然语言的计算机处理中，成为自然语言自动剖析的有力工具。美国语言学家 Z. Harris 研制了最早的完整的英语自动剖析系统“转换与话语分析课题”（transformation and discourse analysis project，简称 TDAP），这个剖析系统于 1958 年 6 月至 1959 年 7 月在美国宾夕法尼亚大学研制成功。

符号派另一方面的工作是人工智能的研究。在 1956 年夏天，J. McCarthy，M. Minsky，C. Shannon 和 N. Rochester 等学者汇聚到一起，组成了一个为期两个月的研究组，讨论关于他们称之为“人工智能”（artificial intelligence，简称 AI）的问题。尽管有少数的 AI 研究者着重于研究随机算法和统计算法（包括概率模型和神经网络），但是大多数的 AI 研究者着重研究推理和逻辑问题。Newell 和 Simon 研制了“逻辑理论家”（logic theorist）和“通用问题解答器”（general problem solver）等可以自

动进行逻辑推理的系统。早期的自然语言理解系统几乎都是按照他们的观点建立起来的。这些简单的系统把模式匹配和关键词搜索与简单试探的方法结合起来进行推理和自动问答，它们都只能在某一个领域内使用。在 20 世纪 60 年代末期，学者们又研制了更多的形式逻辑系统。

随机派主要是一些来自统计学专业和电子学专业的研究人员。在 20 世纪 50 年代后期，他们使用“贝叶斯方法”（Bayesian method）来解决最优字符识别的问题。1959 年，Bledsoe 和 Browning 建立了用于文本识别的贝叶斯系统，该系统使用了一部大词典，首先计算出词典的单词中所观察的字母系列的似然度，然后把单词中每一个字母的似然度相乘，就可以求出整个字母系列的似然度来。1964 年，Mosteller 和 Wallace 用贝叶斯方法解决了在《联邦主义者》（The Federalist）一文中的原作者的分布问题。

20 世纪 50 年代还出现了基于转换语法的第一个人类语言计算机处理的可严格测定的心理模型；并且还出现了第一个联机语料库：布朗美国英语语料库（Brown corpus），该语料库包含 100 万单词的语料，样本来自不同文体的五百多篇书面文本，涉及的文体有新闻、中篇小说、写实小说、科技文章等。这些语料是布朗大学（Brown University）在 1963—1964 年收集的。

计算语言学萌芽期的这些出色的基础性研究和应用性研究，为计算语言学的理论和技术奠定了坚实的基础。计算语言学从萌芽期一开始，就把不同的学科紧密地结合起来，带有明显的边缘性交叉学科的特点。可以说，计算语言学就是在各个相关学科的交融和协作中萌芽成长起来的。

机器翻译是计算语言学最重要的应用领域。在计算语言学的萌芽期，机器翻译研究取得长足进展。

1946 年，美国宾夕法尼亚大学的 J. P. Eckert 和 J. W. Mauchly 设计并制造出了世界上第一台电子计算机 ENIAC，电子计算机惊人的运算速度，启示着人们考虑翻译技术的革新问题。因此，在电子计算机问世的同一年，英国工程师 A. D. Booth 和美国洛克菲勒基金会副总裁 W. Weaver 在讨论电子计算机的应用范围时，就提出了利用计算机进行语言自动翻译的想法。1947 年 3 月 6 日，Booth 与 Weaver 在纽约的洛克菲勒中心会面，Weaver 提出，“如果将计算机用在非数值计算方面，是比较有希望的。”

在 Weaver 与 Booth 会面之前，Weaver 在 1947 年 3 月 4 日给控制论学者 N. Wiener 写信，讨论了机器翻译的问题，Weaver 说："我怀疑是否真的制造不出一台能够作翻译的计算机。即使只能翻译科学性的文章（在语义上问题较少），或是翻译出来的结果不怎么优雅（但能够理解），对我而言都值得一试。"可是，Wiener 给 Weaver 泼了一瓢冷水，他在同年 4 月 30 日给 Weaver 的回信中写道："老实说，恐怕每一种语言的词汇，范围都相当模糊；而其中表达的感情和言外之意，要以类似机器翻译的方法来处理，恐怕不是很乐观。"不过 Weaver 仍然坚持自己的意见。1949 年，Weaver 发表了一份以《翻译》为题的备忘录，正式提出了机器翻译问题。在这份备忘录中，他除了提出各种语言都有许多共同的特征这一论点之外，还有两点值得我们注意。

第一，他认为翻译类似于解读密码的过程。他说："当我阅读一篇用俄语写的文章的时候，我可以说，这篇文章实际上是用英语写的，只不过它是用另外一种奇怪的符号编了码而已，当我在阅读时，我是在进行解码。"

在这段话中，Weaver 首先提出了用解读密码的方法进行机器翻译的想法，这种想法成为后来"噪声信道理论"的滥觞，是统计机器翻译的重要的理论依据。

备忘录中还记载了一个有趣的故事，布朗大学数学系的 R. E. Gilman 曾经解读了一篇长约一百个词的土耳其文的密码，而他既不懂土耳其文，也不知道这篇密码是用土耳其文写的。

Weaver 认为，Gilman 的成功足以证明解读密码的技巧和能力不受语言的影响，因而可以用解读密码的办法来进行机器翻译。

第二，他认为原文与译文"说的是同样的事情"，因此，当把语言 A 翻译为语言 B 时，就意味着，从语言 A 出发，经过某一"通用语言"（universal language）或"中间语言"（interlingua），然后转换为语言 B，这种"通用语言"或"中间语言"，可以假定是全人类共通的。

可以看出，Weaver 把机器翻译仅仅看成一种机械的解读密码的过程，他还远远没有看到机器翻译在词法分析、句法分析以及语义分析等方面的复杂性。

早期机器翻译系统的研制受到 Weaver 上述思想的很大影响，许多机器翻译研究者都把机器翻译的过程与解读密码的过程相类比，试图通过

查询词典的方法来实现词对词的机器翻译，因而译文的可读性很差，难于付诸实用。

由于学者的热心倡导，实业界的大力支持，美国的机器翻译研究一时兴盛起来。1954 年，美国乔治敦大学在国际商用机器公司（IBM 公司）的协同下，用 IBM-701 计算机，进行了世界上第一次机器翻译试验，把几个简单的俄语句子翻译成英语，接着，苏联、英国、日本也进行了机器翻译试验，机器翻译出现热潮。

1952 年，在美国的麻省理工学院召开了第一次机器翻译会议。1954 年，出版了第一本机器翻译的杂志，该杂志的名称就叫做 *Machine Translation*（《机器翻译》）。尽管人们在自然语言的计算方面开展了很多研究工作，但是，直到 20 世纪 60 年代中期，才出现了“computational linguistics”（计算语言学）这一术语，而且，在刚开始的时候，这一术语是偷偷摸摸、羞羞涩涩地出现的。

1965 年，*Machine Translation* 杂志改名为 *Machine Translation and Computational Linguistics*（《机器翻译和计算语言学》），在杂志的封面上，首次出现了“Computational Linguistics”这样的字眼，但是，“and Computational Linguistics”这三个单词是用特别小号的字母排印的。这说明，人们对于“计算语言学”是否能够算得上一门真正独立的学科还没有把握。“计算语言学”刚刚登上学术这一庄严的殿堂时，还带有“千呼万唤始出来，犹抱琵琶半遮面”般的羞涩，以致于人们不敢用与 Machine Translation 同样大小的字母来排印它。当时 *Machine Translation* 杂志之所以改名，是因为在 1962 年美国成立了“机器翻译和计算语言学学会”（Association for Machine Translation and Computational Linguistics），通过改名可以使杂志的名称与学会的名称保持一致。

根据这些史料，我们认为，早在 1962 年，就出现了“计算语言学”这门学科了，尽管它在刚出现时还是偷偷摸摸的，显示出少女般的羞涩。但是，无论如何，“计算语言学”这门新兴的学科终于萌芽了，她破土而出，悄悄地登上了学术的殿堂。

1964 年，美国科学院成立了语言自动处理咨询委员会（Automatic Language Processing Advisory Committee，简称 ALPAC 委员会），调查机器翻译的研究情况，并于 1966 年 11 月公布了一份题为《语言与机器》

的报告，简称 ALPAC 报告[4]，这份报告对机器翻译采取了否定的态度，报告宣称“在目前给机器翻译以大力支持理由还不充分”；这份报告还指出，机器翻译研究遇到了难以克服的“语义障碍”（semantic barrier）。在 ALPAC 报告的影响下，许多国家的机器翻译研究陷入低潮，许多已经建立起来的机器翻译研究单位遇到了行政上和经费上的困难，在世界范围内，机器翻译的热潮突然消失了，出现了空前萧条的局面。

美国语言学家 D. Hays 是 ALPAC 委员会的成员之一，他参与起草了 ALPAC 报告，在 ALPAC 报告中，他建议，在放弃机器翻译这个短期的工程项目的时候，应当加强语言和自然语言计算机处理的基础研究，可以把原来用于机器翻译研制的经费使用到自然语言处理的基础研究方面，Hays 把这样的基础研究正式命名为“Computational Linguistics”（计算语言学）。所以，我们可以说，“计算语言学”这个学科名称最早出现于 1962 年，而在 1966 年才在美国科学院的 ALPAC 报告中正式得到学术界的承认。

计算语言学的发展期

20 世纪 60 年代中期到 80 年代末期是计算语言学的发展期。

在计算语言学的发展期，各个相关学科彼此协作，联合攻关，取得了一些令人振奋的成绩。

统计方法在语音识别算法的研制中取得成功。其中特别重要的是“隐马尔可夫模型”（hidden Markov model）和“噪声信道与解码模型”（noisy channel model and decoding model）。这些模型是由两支队伍分别独立地研制的。一支由 Jelinek、Bahl、Mercer 和 IBM 的华生研究中心的研究人员组成，另一支由卡内基梅隆大学（Carnegie Mellon University）的 Baker 等组成，Baker 受到普林斯顿防护分析研究所的 Baum 和他的同事们的工作的影响。AT&T 的贝尔实验室（Bell laboratories）也是语音识别和语音合成的中心之一。

逻辑方法在计算语言学中取得了很好的成绩。1970 年，Colmerauer 和他的同事们使用逻辑方法研制了 Q 系统（Q-system）和“变形语法”（metamorphosis grammar）并在机器翻译中得到应用，Colmerauer 还是

[4] “ALPAC, Language and Machines: Computer in Translation and Linguistics”. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Publication 1416, Washington.

Prolog 语言的先驱者，他使用逻辑程序设计的思想设计了 Prolog 语言。1980 年 Pereira 和 Warren 提出的“定子句语法”（definite clause grammar）也是在计算语言学中使用逻辑方法的成功范例之一。1979 年 Kay 对于“功能语法”（functional grammar）的研究，1982 年 Bresnan 和 Kaplan 在“词汇功能语法”（lexical function grammar，简称 LFG）方面的工作，都是“特征结构合一”（feature structure unification）研究方面的重要成果，他们的研究引入了“复杂特征”（complex feature）的概念。与此同时，我国学者冯志伟提出了“多叉多标记树形图模型”（multiple-branched multiple-labeled tree model，简称 MMT 模型），在他设计的多语言机器翻译 FAJRA 中采用了“多标记”（multiple label）的概念。“多标记”的概念与“复杂特征”的概念实质上是一致的，这些关于自然语言特征结构的研究成果，都有效地克服了 Chomsky 短语结构语法的生成能力过强的缺陷。

在这一时期，自然语言理解（natural language understanding）也取得了明显的成绩。自然语言理解肇始于 T. Winograd 在 1972 年研制的 SHRDLU 系统，这个系统能够模拟一个嵌入玩具积木世界的机器人的行为。该系统的程序能够接受自然语言的书面指令（例如 Move the red block on top of the smaller green one [请把红色的积木块移动到绿色的小积木块的上端]），从而指挥机器人摆弄玩具积木块。这是一个非常复杂而精妙的系统。这个系统还首次尝试建立基于 Halliday“系统语法”（systemic grammar）的全面的英语语法。Winograd 的模型还清楚地说明，句法剖析也应该重视语义和话语的模型。1977 年，R. Schank 和他在耶鲁大学的同事和学生们建立了一些语言理解程序，这些程序构成一个系列，他们重点研究诸如脚本、计划和目的这样的人类的概念知识以及人类的记忆机制。他们的工作经常使用基于网络的语义学理论，并且在他们的表达方式中开始引进 Fillmore 在 1968 年提出的关于“深层格”（deep case）的概念。

在自然语言理解研究中也使用过逻辑学的方法，例如 1967 年 Woods 在他研制的 LUNAR 问答系统中，就使用谓词逻辑来进行语义解释。

计算语言学在“话语分析”（discourse analysis）方面也取得了很大的成绩。基于计算的话语分析集中探讨了话语研究中的 4 个关键领域：话语子结构的研究、话语焦点的研究、自动参照消解的研究、基于逻辑的言语行为的研究。1977 年，Crosz 和她的同事们研究了话语中的“子结构”

(substructure) 和话语焦点；1972 年，Hobbs 开始研究“自动参照消解”(automatic reference resolution)。在基于逻辑的言语行为研究中，Perrault 和 Allen 在 1980 年建立了“信念－愿望－意图”(belief-desire-intention，简称 BDI) 的框架。

在 1983—1993 年的 10 年中，计算语言学研究者对于过去的研究历史进行了反思，发现过去被否定的有限状态模型和经验主义方法仍然有其合理的内核。在这 10 年中，计算语言学的研究又回到了 20 世纪 50 年代末期到 60 年代初期几乎被否定的有限状态模型和经验主义方法上去，之所以出现这样的复苏，其部分原因在于 1959 年 Chomsky 对于 Skinner 的“言语行为”(verbal behavior) 极其影响的评论在 20 世纪 80 年代和 90 年代之际遭到了理论上的反对。

这种反思的第一个倾向是重新评价有限状态模型，由于 Kaplan 和 Kay 在有限状态音系学和形态学方面的工作，以及 Church 在句法的有限状态模型方面的工作，有限状态模型仍然显示出强大的功能，因此，这种模型又重新得到计算语言学界的注意。

这种反思的第二个倾向是所谓的“重新回到经验主义”。这里值得特别注意的是语音和语言处理的概率模型的提出，这样的模型受到 IBM 公司华生研究中心的语音识别概率模型的强烈影响。这些概率模型和其他数据驱动的方法还运用到了词类标注、句法剖析、名词短语附着歧义的判定以及从语音识别到语义学的联接主义方法的研究中。

此外，在这一时期，自然语言的生成研究也取得了令人瞩目的成绩。

计算语言学的繁荣期

从 20 世纪 90 年代开始，计算语言学进入了繁荣期。1993 年 7 月在日本神户召开的第四届机器翻译高层会议上，英国著名学者 J. Hutchins 在他的特约报告中指出，自 1989 年以来，机器翻译的发展进入了一个新纪元。这个新纪元的重要标志是，在基于规则的技术中引入了语料库方法，其中包括统计方法、基于实例的方法、通过语料加工手段使语料库转化为语言知识库的方法，等等。这种建立在大规模真实文本处理基础上的机器翻译，是机器翻译研究史上的一场革命，它将会把计算语言学推向一个崭新的阶段。随着机器翻译新纪元的开始，计算语言学进入了繁荣期。

在20世纪90年代的最后5年（1994—1999），计算语言学的研究发生了很大变化，出现了空前繁荣的局面。这主要表现在以下三个方面。

第一，概率和数据驱动的方法几乎成为了计算语言学的标准方法。句法剖析、词类标注、参照消解、话语处理、机器翻译的算法全都开始引入概率，并且采用从语音识别和信息检索中借鉴的基于概率和数据驱动的评测方法。

第二，计算语言学的应用研究日新月异。由于计算机的速度和存储量的增加，使得在计算语言学的一些应用领域，特别是在语音合成、语音识别、文字识别、拼写检查、语法检查这些应用领域，有可能进行商品化的开发。自然语言处理的算法开始被应用于“增强交替通信”（augmentative and alternative communication，简称AAC）中，语音合成、语音识别和文字识别的技术被应用于“移动通信”（mobile communication）中。除了传统的机器翻译和信息检索等应用研究进一步得到发展之外，“信息抽取”（information extraction）、“问答系统”（question answering system）、“自动文摘”（text summarization）、“术语的自动抽取和标引”（term extraction and automatic indexing）、“文本数据挖掘”（text data mining）、“自然语言接口”（natural language interaction）、“计算机辅助语言教学”（computer-assisted language learning）等新兴的应用研究都有了长足的进展。此外，自然语言处理技术在“多媒体系统”（multimedia system）和“多模态系统”（multimodal system）中也得到了应用。计算语言学的应用研究呈现出日新月异的局面。

第三，多语言的在线自然语言处理技术迅猛发展。随着网络技术的发展，因特网逐渐变成一个多语言的网络世界，因特网上的机器翻译、信息检索和信息抽取的需要变得更加迫切。目前，在因特网上除了使用英语之外，越来越多地使用汉语、西班牙语、葡萄牙语、德语、法语、俄语、日语、韩语等英语之外的语言。从2000年到2005年，因特网上使用英语的人数仅仅增加了126.9%，而在此期间，因特网上使用俄语的人数增加了664.5%，使用葡萄牙语的人数增加了327.3%，使用汉语的人数增加了309.6%，使用法语的人数增加了235.9%。因特网上使用英语之外的其他语言的人数增加得越来越多，英语在因特网上独霸天下的局面已经被打破，因特网确实已经变成了多语言的网络世界。因此，网络上的不同自然语言之间的计算机自动处理也就变得越来越迫切了。网络上

多语言的机器翻译、信息检索、信息抽取正在迅猛地发展。“语言辨别”(language identification)、“跨语言信息检索”(cross-language information retrieval)、“双语言术语对齐”(bilingual terminology alignment)和“语言理解助手”(comprehension aids)等计算语言学的“多语言的在线处理技术”(multilingual on-line processing)已经成为了互联网技术的重要支柱。

在信息时代，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不断增加，出现了“信息爆炸”(information explosion)的局面。现在，世界上出版的科技刊物达 165,000 种，平均每天有约 2 万篇科技论文发表。专家估计，我们目前每天在因特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的全部知识总量。而所有的这些信息主要都是以语言文字作为载体的，也就是说，网络世界主要是由语言文字构成的。

为了说明计算语言学的重要性，我们可以把它与物理学作如下的类比：我们说物理学之所以重要，是因为物质世界是由物质构成的，而物理学恰恰是研究物质运动的学科；我们说计算语言学之所以重要，是因为网络世界主要是由语言文字构成的，而计算语言学恰恰是研究语言文字自动处理的学科。

可以预见，知识日新月异的增长和网络技术突飞猛进的进步，一定会把计算语言学的研究推向一个崭新的阶段。计算语言学有可能成为当代语言学中最有发展潜力的学科，计算语言学已经给有着悠久传统的古老的语言学注入了新的生命力，在计算语言学的推动下，语言学有可能真正成为当代科学百花园中一门名副其实的领先学科。

当前计算语言学发展的四个特点

21 世纪以来，由于互联网的普及，自然语言的计算机处理成为了从互联网上获取知识的重要手段，生活在信息网络时代的现代人，几乎都要与互联网打交道，都要或多或少地使用计算语言学的研究成果来帮助他们获取或挖掘广阔无边的互联网上的各种知识和信息。因此，世界各国都非常重视计算语言学的研究，投入了大量的人力、物力和财力。

当前国外计算语言学研究有四个显著的特点：

第一，随着语料库建设和语料库语言学的崛起，大规模真实文本的处理成为计算语言学的主要战略目标：在过去的四十多年中，从事计算语言学系统开发的绝大多数学者，都把自己的研究目标局限于某个十分狭窄的专业领域中，他们采用的主流技术是基于规则的句法—语义分析，尽管这些应用系统在某些受限的“子语言”（sublanguage）中也曾经获得一定程度的成功，但是，要想进一步扩大这些系统的覆盖面，用它们来处理大规模的真实文本，仍然有很大困难。因为从自然语言系统所需要装备的语言知识来看，其数量之浩大和颗粒度之精细，都是以往的任何系统所远远不及的；而且随着系统拥有的知识在数量上和程度上发生的巨大变化，系统在如何获取、表示和管理知识等基本问题上，不得不另辟蹊径。这样，就提出了大规模真实文本的自动处理问题。1990 年 8 月在芬兰赫尔辛基举行的第 13 届国际计算语言学会议为会前讲座确定的主题是“处理大规模真实文本的理论、方法和工具”，这说明实现大规模真实文本的处理将是计算语言学在今后一个相当长的时期内的战略目标。为了实现战略目标的转移，需要在理论、方法和工具等方面进行重大的革新。1992 年 6 月在加拿大蒙特利尔举行的第四届机器翻译的理论与方法国际会议上，会议宣布的主题是“机器翻译中的经验主义和理性主义的方法”。所谓“理性主义”，就是指以生成语言学为基础的方法，所谓“经验主义”，就是指以大规模语料库的分析为基础的方法。从中可以看出当前计算语言学关注的焦点。当前语料库的建设和语料库语言学的崛起，正是计算语言学战略目标转移的一个重要标志。随着人们对大规模真实文本处理的日益关注，越来越多的学者认识到，基于语料库的分析方法（即经验主义的方法）至少是对基于规则的分析方法（即理性主义的方法）的一个重要补充，因为从“大规模”和“真实”这两个因素来考察，语料库才是最理想的语言知识资源。但是，要想使语料库名副其实地成为自然语言的知识库，就有必要首先对语料库中的语料进行自动标注，使之由“生语料”变成“熟语料”，以便于人们从中提取丰富的语言知识。

第二，计算语言学中越来越多地使用机器自动学习的方法来获取语言知识。传统语言学基本上是通过语言学家归纳总结语言现象的手工方法来获取语言知识的，由于人的记忆能力有限，任何语言学家，哪怕是语言学界的权威泰斗，都不可能记忆和处理浩如烟海的全部的语言数据。因此，使用传统的手工方法来获取语言知识，犹如以管窥豹，以蠡测海，

这种获取语言知识的方法带有很大的主观性。传统语言学中啧啧称道的所谓“例不过十不立，反例不过十不破”的朴学精神，貌似严格，实际上，在浩如烟海的语言数据中，以十个正例或十个反例就轻而易举地来决定语言规则的取舍，难道就能够万无一失地保证这些规则是可靠的吗？这是大大地值得怀疑的。当前的计算语言学研究提倡建立语料库，使用机器学习的方法，让计算机自动地从浩如烟海的语料库中获取准确的语言知识。机器词典和大规模语料库的建设，成为了当前计算语言学的热点。这是语言学获取语言知识方式的巨大变化，作为21世纪的语言学工作者，应该注意到这样的变化，逐渐改变传统的获取语言知识的手段。

第三，计算语言学中越来越多地使用统计数学方法来分析语言数据。使用人工观察和内省的方法，显然不可能从浩如烟海的语料库中获取精确可靠的语言知识，必须使用统计数学的方法。目前，计算语言学中的统计数学方法已经相当成熟，如果我们认真学会并努力掌握了统计数学，就会使我们在获取语言知识的过程中如虎添翼。目前，在机器翻译中使用统计方法获得了很好的成绩，统计机器翻译（statistical machine translation，简称SMT）成为了机器翻译的主流技术。

2003年7月，在美国马里兰州巴尔的摩（Baltimore, Maryland）由美国商业部国家标准与技术研究所（National Institute of Standards and Technology）主持的评比中，来自德国亚琛大学（Achen University）的一位年轻的博士研究生F. J. Och获最好成绩。他使用统计方法，在很短的时间之内就构造了阿拉伯语和汉语到英语的若干个机器翻译系统。伟大的希腊科学家阿基米德说过：“给我一个支点，我就能撬动地球。”（Give me a place to stand on, and I will move the world.）而这次评比中，Och也模仿着阿基米德说：“只要给我充分的平行语料，那么，对于任何的两种语言，我就可以在几小时之内给你构造出一个机器翻译系统。”（Give me enough parallel data, and you can have translation system for any two languages in a matter of hours.）这反映了新一代的机器翻译研究者大无畏的探索精神和继往开来的豪情壮志。看来，Och似乎已经找到了机器翻译的有效方法，至少按照他的路子走下去，也许有可能开创出机器翻译研究的一片新天地，使我们在探索真理的曲折道路上看到了灿烂的曙光。过去我们研制一个机器翻译系统往往需要几年的时间，而现在采用Och的方法构造机器翻译系统只要几个小时就可以了，研制机器翻译系统的