ZIRAN YUYAN JIQI FANYI XINLUN
自然语言机器翻译新论

冯志伟 著

YUWEN CHUBANSHE
语文出版社

ZIRAN YUYAN JIQI FANYI XINLUN

自 然 语 言 机 器 翻 译 新 论

冯志伟 著

*

YUWEN CHUBANSHE CHUBAN

语 文 出 版 社 出 版

北京朝阳门南小街51号 邮政编码:100010

新华书店经销 北京密云胶印厂印刷

*

787×1092毫米 1/16 16.625印张 426千字

印数:1-10000 定价:39.00元

ISBN 7-80006-744-0/H·160

内容提要

本书是一本关于机器翻译的专著，着重讲述近年来国内外机器翻译研究中的新情况、新理论、新方法。全书共分十章，主要内容有：机器翻译的新发展，国外的机器翻译系统、我国的机器翻译研究、机器翻译与现代语言学理论、机器翻译与逻辑、语言的自动分析和生成技术、自然语言的复杂特征描述与运算、形态分析器和结构分析生成器、词汇转换器和形态生成器、机器翻译的工程化。书中对于广义短语结构语法、词汇功能语法、功能合一语法、中心语驱动的短语结构语法、孟德鸠语法、定子句语法等最新的语言学理论，都作了详细的介绍。本书内容丰富、观点新颖、深入浅出、通俗易懂，可供从事语言学、计算机科学、数学、逻辑学、人工智能、知识工程、计算语言学的高等学校师生和科研工作者阅读，也可以作为计算语言学和机器翻译初学者的入门书。本书写作时充分地考虑到跨学科读者的需要，只要具备高中以上文化水平的文科和理科的读者，都不难理解本书的内容。

目 录

序言

机器翻译是应用电子计算机进行语言之间的自动翻译的一门边缘学科，它需要语言学、数学、计算机科学方面的专业人员的合作。这门学科的研究难度很大，可是，它越是难，就越有不怕困难的人来攻它，就越能引起探索者对它的兴趣。机器翻译的研究者就像侦察兵，没有道路的路，对于他们来说，才是最好的路。这条艰险的荆棘之路一旦被机器翻译的侦察兵开通了，前面就是一马平川的坦途。正是这种对未来的坚定信念，机器翻译这门新兴学科吸引了来自不同专业的有志之士，许多人为之贡献了自己的聪明和才智，甚至生命。

1984 年 1 月，我与杨平同志曾合写过一本《自动翻译》的书，此书于 1987 年 11 月由知识出版社出版。由于国内外机器翻译研究发展很快，在 1984 年 1 月完稿的《自动翻译》一书中，许多重要的新情况未能及时地反映出来，因此我曾希望，如果有机会，再写一本书来弥补《自动翻译》一书的不足。可是，当今学术著作出版困难，是否能有这样的机会，我感到很茫然。

近年来，机器翻译的发展日新月异，新的理论和方法层出不穷，其中不乏深刻的理论和巧妙的方法。由于机器翻译研究对社会经济发展有着潜在的价值，各国对机器翻译更加重视，纷纷投资。欧洲共同体为了把 EUROTRA 多语言机器翻译系统实用化，五年内投资 2800 万美元。法国制定了一个 ESOPE 机器翻译计划，用于 ARIANE 系统的实用化，投资 5600 万法郎。日本对机器翻译的专项投资为 140 亿日元（约相当于 1 亿美元），其中，仅 CICC 计划的投资就达 62 亿日元，京都大学的 μ 计划投资 1 亿日元。而日本新一代计算技术研究所的第五代计算机系统，计划十年之内投资 5 亿美元，其中包括研制一个实用的日英机助翻译系统，编制 10 万词的日英机器词典，机助翻译正确率要求达到 90%。我国对机器翻译也进行了巨额的投资。七五计划期间，投资人民币 260 万元。这个数字当然不能同欧洲和日本的投资数额相比，但在我国目前经济实力还不很强的条件下，这已经是一笔相当大的投资了。

中山大学英语系王宗炎教授非常关心机器翻译的发展，他建议我再写一本关于机器翻译的书，把近几年机器翻译发展的新情况、新理论、新方法反映出来。我意识到机器翻译这一学科对于经济发展的巨大价值，同时，我早就有再写一本书反映机器翻译新情况的宿愿，在王宗炎教授的鼓励之下，我终于写成了这本《自然语言机器翻译新论》。语文出版社社长李行健教授热心支持学术著作出版，热情给我提供出版此书的机会，使得本书得以问世。北京大学中文系叶蜚声教授在百忙中通读了全书，并请北京大学数学系的专家审阅本书与数学有关的内容，提出了许多中肯的建议。语文出版社的有关编辑及出版人员为本书的出版付出了辛勤的劳动。在本书出版之际，特向他们表示衷心的感谢。

1990 年 9 月，我应联邦德国莱茵-法尔茨州教育部的邀请，前往特里尔大学担任客座教授，曾向德国学生们讲过本书的部分内容，引起了他们对于机器翻译领域中的新理论和新方法的兴趣，有的德国学生甚至主动选择了机器翻译作为自己的学位论文题目，立志从事机器翻译的工作。实践证明，采用此书作为大学本科或研究生的机器翻译课程或计算语言学课程的教材是成功的。目前我国不少大学的计算机系都开设了计算语言学的课程，有些大学的中文系或外语系也准备开设计算语言学课程，均可选用此书作为教学的参考。

我原是学中文的，对于计算机、数学和外语都是外行，60 年代末期开始进行更新知识的再学

习，学了一点计算机，学了一点数学，学了一点外语，但是，由于天资不高，底子太薄，所知仍十分有限。1979 年至 1981 年，我有机会到法国格勒诺布尔理科医科大学应用数学研究所的自动翻译中心深造，我的导师就是国际上著名的机器翻译专家和数学家、国际计算语言学学会主席沃古瓦（B. Vauquois）教授，在他的精心指导下，我才跨进了机器翻译这一新领域，完成了汉-法/英/日/俄/德多语言机器翻译系统 FAJRA 的设计，并在 IBM-4341 计算机上顺利地通过了试验，使我对于机器翻译开始有了较为全面的认识。1985 年 9 月，年仅 55 岁的沃古瓦教授不幸英年早逝，噩耗传来，使我感到无比的悲痛。沃古瓦教授对中国人民始终怀着友好的感情，他曾经多次来中国访问，我国许多机器翻译研究者都是他的老朋友。在本书完稿之际，恰值沃古瓦教授逝世 9 周年的纪念日，谨以此书，寄托我们的哀思，献给沃古瓦教授，以表达我们对这位为机器翻译事业奋斗终生的学者的深切的怀念之情。

机器翻译是一门综合性边缘学科，涉及多个科学部门，作者孤陋寡闻，水平有限，如有不当之处，敬请读者批评指正。

本书写作时，参考过国内外时贤著作多种，已在参考文献中注明，谨对他们表示感谢。

冯志伟

1994 年 9 月于北京

第一章　机器翻译的新发展

圣经《创世纪》中说，古代人说的原是一种统一的语言，交流思想非常方便，劳动效率也很高，他们曾经想建造一座高达天庭的通天塔，叫做巴比塔，来显示他们的丰功伟绩。建造巴比塔的壮举震惊了上帝，上帝便施展伎俩，让不同的人说不同的语言，使人们难于交流思想，无法协调工作，以此来惩罚异想天开的巴比塔建造者。结果，巴比塔没有建成，而语言的不同，却成为了人们相互交往的极大障碍。这样的传说当然不足为训，但是，语言的障碍却时时刻刻在困扰着人们。

现在我们已经进入了信息化的时代，语言是信息的最主要的负荷者，因此，如何有效地使用现代化手段来突破人们之间的语言障碍，成为了全人类面临的共同问题，机器翻译便是解决这个问题有力手段之一，它有可能成为消除人们语言障碍的真正的通往理想境界的巴比塔。本章首先对机器翻译的历史作一回顾，然后着重讨论 70 年代以来机器翻译发展的新特点。

第一节　机器翻译发展的曲折道路

在自然语言处理中，机器翻译是一个最早研究的课题。远在 1946 年电子计算机问世之时，英国工程师布斯（A. D. Booth）和美国工程师韦弗（W. Weaver）就提出了利用计算机进行机器翻译的想法。1954 年，美国乔治敦大学在国际商用机器公司（即 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，首次用计算机把俄语译成了英语。这是计算机最早的在非数值处理方面的应用，一时吸引了人们的注意，许多人认为这是一个大有可为的计算机应用领域。50 年代，在世界范围内出现了机器翻译的热潮。之所以出现这种热潮，是因为：

第一，随着科学技术日新月异的发展，国际科技交流日趋频繁，世界进入了所谓“情报爆炸”的时代，翻译成为一种十分重要的而且必不可少的事业，为了提高翻译工作的效率，迫切需要把传统的手工式翻译工作自动化。

第二，翻译活动是人们经常进行的一种普通活动，由于人们对于翻译活动习焉不察，不少人误以为它的机制是容易用计算机来模拟的，因而对于机器翻译寄以过高的希望，以为可以一蹴而就。

第三，在翻译工作中，词典的找查是最费时间的，许多人认为，如果在计算机中建立一部机器词典，词典的找查将会变得十分简易。

基于这些原因，人们以为用计算机代替人来做翻译，将是一件并不十分困难的工作。在这个机器翻译的热潮中，机器翻译实验室像雨后春笋般地建立了起来，并出现了首批的机器翻译实用系统，如美国乔治敦大学经过进一步扩充的俄英机器翻译系统，美国橡树岭国家原子能实验室的俄英机器翻译系统，欧洲原子能联营（EURATOM）设在意大利瓦雷泽的俄英机器翻译系统，美国空军国外技术部的俄英机器翻译系统等等。

但是，刚刚兴起的机器翻译很快就遇到了困难，译文质量的低劣引起了用户们无休止的抱怨，甚至一些专家也对机器翻译颇有微词，认为他们受了不学无术而冒充内行的江湖骗子的损害。因

此，在1963年10月，美国国家基金会向美国科学院提出如下要求：建议国防部、中央情报局及国家基金会对在用机器来翻译外语的一般领域内的研究和发展情况进行调查。根据这个建议，美国科学院于1964年4月成立了一个语言自动处理咨询委员会（Automatic Language Processing Advisory Committee，简称ALPAC）。

ALPAC委员会在下述三个方面进行了调查：

第一，政府部门和科学界对翻译工作的需要情况。

第二，现有的翻译人员和设备能否满足这种需要。

第三，机器翻译的优点、缺点及其前景。

1966年11月，ALPAC委员会发表了《语言与机器》的黑皮书（又称ALPAC报告），否认机器翻译实用的可能性，指出“一般科技文章的机器翻译是不现实的”，散布悲观主义的论调，致使许多国家的机器翻译研究工作遇到了财政上和组织上的困难，在世界范围内，机器翻译转入低潮。

尽管如此，法国、加拿大、日本和前苏联等国，仍然坚持机器翻译的研究，并取得了一定的成绩。而且，就是在机器翻译空前萧条的美国，乔治敦大学的俄英机器翻译系统，仍然在极端艰难的条件下运行着。不少研究者认识到机器翻译的困难性，开始采取比较现实的态度来对待机器翻译，不再像早期那样急于求成，而是认真地进行了语言结构的研究以及翻译时知识背景的研究，这样，机器翻译的研究工作就更加深入了。

70年代以来，由于翻译需求量的增加、计算机技术的进步、语言学理论的发展以及人工智能研究中自然语言理解模型研究的进展，机器翻译又复苏了。在美国和日本，一些机器翻译系统已经从实验研究进入实用研究，个别机器翻译系统已经商品化。加拿大蒙特利尔大学的TAUM－METEO天气预报机器翻译系统、法国格勒诺布尔大学的ARIANE－78机器翻译专用软件系统、前苏联的ФР法俄机器翻译系统都在踏踏实实进行研究，取得了长足的进展。80年代以来，日本成为了世界上机器翻译最为兴盛的国家。

机器翻译由盲目的热潮，到骤然的萧条，又到重新的复苏，走过了一条曲折的发展道路。尽管在这条曲折的道路上遇到了重重的困难，但是，那些为追求科学真理而斗争的艰辛的机器翻译研究者们，终于迎来了机器翻译事业在全世界范围内的空前繁荣。

第二节　机器翻译发展的新特点

70年代以来的机器翻译研究具有如下的新特点：

一、独立分析与独立生成

1957年美国著名机器翻译学者英格维（V. Yngve）提出了独立分析和独立生成的思想。他认为，在机器翻译的过程中，应该首先在不受译语影响的情况下进行原语分析，接着进行原语译语的转换，最后进行译语文句的生成。这样，机器翻译过程可分为三个阶段：

1. 以代码化的结构标志来表示原语的结构；

2. 把原语的结构标志转换成译语的结构标志；

3. 构成译语的输出文句。

第一阶段只涉及原语，不受译语的影响，第三阶段只涉及译语，不受原语的影响，只有第二阶段才涉及原语和译语两者。

可惜英格维的这种卓越思想并没有受到当时机器翻译界的足够重视。尽管有极个别的机器翻

译系统曾采用独立分析和独立生成的办法，但很快就改弦易辙了。在绝大多数的机器翻译系统中，只有原语形态分析是独立的，其他阶段几乎都是相关的。原语句法语义分析的目的，在于确定译语的信息，而不是确定原语本身的结构。这样，从原语到译语的翻译过程，就混合成一个从输入文句的形态转换为输出文句的形态的复杂过程。俄国著名数理语言学家库拉金娜（О. С. Кулагина）指出，这实质上是一种“相关句法分析方法”。

70年代以来，不少机器翻译工作者批评这种相关句法分析方法的缺点。他们指出，这种原语与译语混杂在一起的方法，不可能对原语的句法和语义作深入的分析，也不可能对译语的句法和语义作有效的生成，这正是机器翻译的译文质量多年来始终提不高的症结所在。法国学者沃古瓦（B. Vauquois）发展了英格维的思想。他尖锐地批评了相关句法分析方法，并且提出，在机器翻译中，不仅要对原语的形态作独立的分析，而且对原语的句法和语义也要独立于译语来进行分析，译语的句法生成和语义生成，也要独立于原语来进行，而在原语与译语的接口处，进行原语和译语的词汇转换和结构转换，以处理原语和译语之间在词汇和结构方面的差别。这样，一个完整的机器翻译过程便可分为六个阶段来进行：原语形态分析→原语句法语义分析→原语译语词汇转换→原语译语结构转换→译语句法语义生成→译语形态生成。

本书作者于1981年在法国格勒诺布尔大学自动翻译研究中心采用这种独立分析独立生成的方法，进行了汉－法/英/日/俄/德多语言机器翻译试验。这个试验把汉语的句法语义分析与各种译语的句法语义生成完全分开，各种译语的句法语义生成也独立地进行，分析汉语时完全不考虑译语，生成译语时完全不考虑汉语，这样，便可从同一个汉语分析的结果出发，分别生成法语、英语、日语、俄语、德语等译语。如果采用相关句法分析方法，把汉语的句法语义分析与法、英、日、俄、德等译语的句法语义生成合在一起来考虑，就得针对每一种译语分别进行汉语的句法语义分析，单是汉语的句法语义分析方案就得制定5套，而且，由于不能对汉语的句法和语义进行透彻深入的研究，制定出的方案也不会是高水平的。本文作者于1985年又采用这种独立分析与独立生成的方法，在北京遥感技术研究所进行了德汉和法汉机器翻译试验，把汉语的生成独立于德语和法语来进行，用同一个汉语生成程序接受来自德语和法语的分析和转换结果，得到了通顺的汉语译文。实践证明，这种独立分析独立生成的方法，不论对于汉外机器翻译还是对于外汉机器翻译都是行之有效的。

近年来，国外不少机器翻译单位都放弃了相关句法分析方法而采用独立分析和独立生成的方法。例如，苏联科学院应用数学研究所1954年制定的法俄机器翻译 ФР－I 系统，对法语的句法分析，采用的就是相关句法分析方法。但是，这个研究所在70年代新设计的法俄机器翻译 ФР－II 系统中，则明确地采用独立分析独立生成的方法。该系统的主要设计人员库拉金娜（О. С. Кулагина）说，在这个系统中，“翻译的总图式是：形态分析，句法分析，转换，句法生成，形态生成。这时，分析和生成都脱离另一种语言来进行”。由于把法语句法分析与俄语句法生成分开进行，大大地提高译文的质量。

采用独立分析独立生成的方法，为在机器翻译领域进行广泛的国际合作提供可能性。例如，欧洲共同体最近制定的多语言机器翻译的“莱布尼茨计划”规定，由参与该计划的法、德、意、英等国分别制定本国语言的分析与生成方案，由有关国家共同制定双语言的转换方案，从而建立法、德、意、英四种语言互译的实用性机器翻译系统。由于参加国着重研究他们十分熟悉的本国语言，这就为提高机器翻译的研究水平提供了有利的条件。

一般说来，这种独立分析和独立生成的方法，对于一对多、多对多的机器翻译是特别适合的。但是，对于多对一的机器翻译则未必是一种简便的办法。在当前机器翻译的研究中，是不是一定

要采用这种独立分析独立生成的方法，学者们还有不同的看法。

二、语言和程序分开

早期的机器翻译系统，语言和程序是分不开的，进行语言分析和生成的每一步，都必须考虑它在程序上实现的具体细节，这样制定出来的机器翻译方案必然是十分庞杂和琐碎的。因此，在60年代初期，就有人提出语言和程序分开。但是，由于人们当时还没有找到强有力的技术手段，尽管有些人主观意图是要把语言和程序分开，而实际上仍然分不开。其中在技术上的一个重要原因，是因为大多数的机器翻译系统，对查明的语言现象（例如，词汇的歧义，句法功能的歧义，词与词之间的句法语义关系等）的加工，往往是使用一般的高级程序语言（如FORTRAN，COBOL，SNOBOL，LISP，BOL，..... 等等）来实现的。而这些高级语言的格式，一般并不与自然语言的形式结构相适应，其表达方式与自然语言句法语义表达方式也不协调，所以，当采用这样的高级语言来描述自然语言的形式结构时，仍然不可避免地要考虑程序上的种种细微末节。在这种情况下，程序对于语言的分析有很大的依赖性，它只能按语言学家所制定的机器翻译方案来安排，程序人员的作用，只是把语言学家的语言学思想，用程序的形式表达出来而已，本身并没有多少主动性。用这样的办法，语言的语法就是程序，语言的语法是以程序的形式来起作用的。

70年代以来，人们提出了不少适于描写自然语言的机器翻译专用软件。例如，美国兰德公司马丁．凯依（M. Kay）的MIND系统（1970），B. B. N. 公司伍兹（W. Wood）的ATN系统（1971），加拿大蒙特利尔大学科尔迈洛埃（A. Colmerauer）的Q系统（1971），法国格勒诺布尔大学硕塞（J. Chauché）的ATEF系统（1972）和CETA系统（1974），日本京都大学长尾真的PLATON系统（1974），加拿大蒙特利尔大学斯特瓦德（G. Stewart）的REZO系统（1975），法国格勒诺布尔大学基尧姆（P. Guillaume）和盖采尔－安布律拉兹（M. Quezel－Ambrunaz）的TRANSF系统（1976），布瓦戴（Ch. Boitet）的ROBRA系统（1978），饶叶尔（D. Jaeger）的SYGMOR系统（1978）。近来又出现了GRADE系统、LINGOL系统、扩充的LINGOL系统、BUP系统、PATR－II系统、ESPER系统、AGLAI系统等。这些机器翻译专用软件，一般都以某种自动机为逻辑模型。采用这样的方法，语言的语法不再是程序而是数据，它以数据的形式起作用。由于程序的书写格式与自然语言的结构相适应，因此，可以把自然语言的规则以数据的形式直接填写到程序的格式中去，不必再考虑程序实现上的各种细微末节。这样，就可以在算法不变的条件下，进行各种语言数据的代换。在选好了程序的书写格式之后，语言现象的书写和修改，可以完全不考虑算法。

语言和程序分开，是机器翻译研究中的一大进步。当然，语言和程序分开，并不意味着语言学家可以不管程序，程序人员可以不管语言。一个理想的机器翻译研究者，应该既懂得语言知识，又精于程序技巧，或者至少精通自己的本行，而对于另一行比较熟悉。如果研究人员的知识结构不完善，是适应不了机器翻译的要求的，因此，我们提倡机器翻译研究者不断地进行更新知识的再学习。

三、语言研究更加深入

当代机器翻译发展的一个重要特点是，人们向语言的更深的层次——语义进行探索，广泛地探讨了机器翻译中的各种语义问题。俄国机器翻译界提出了“意义-文句”（МЫСЛЬ-ТЕКСТ）模型，用它来改进英俄机器翻译系统（Кулагина，1971）和法俄机器翻译系统（Апресян，1987）。在Апресян的俄法机器翻译系统中，提出了深层句法平面，并把它当作机器翻译过程的一个独立的阶段。

美国斯坦福大学的威尔克斯（Y. A. Wilks）于1973年采用优先语义学的方法，设计了一个英法机器翻译模型，用600个英语单词组成英语日常用语输入计算机，能译成通顺的法语。这个系统的语义分析比较细致，能解决句法分析中难以解决的歧义、指代等疑难问题。

日本京都大学的长尾真提出了利用语义和上下文进行句子分析的方法，根据他的方法，较好地处理了句子之内或句子与句子之间的代词的指代问题。此外，他还提出了利用语义信息进行句子的随机生成的方法。

美国逻辑学家孟德鸠（R. Montague）于70年代前后，把内涵逻辑学应用于自然语言研究，提出了孟德鸠语法（Montague grammar）。日本京都大学的西田丰明等人，用孟德鸠语法来进行英日机器翻译．在机器作句法分析的同时，把相应的句法结构转化为内涵逻辑表达式，从而使句法分析和逻辑语义分析巧妙地结合起来。

此外，在机器翻译的研究工作中，还采用了语义网络理论（semantic network）、格语法（case grammar）、框架理论（frame theory）、概念依存理论（conceptual dependency）等来处理自然语言分析和生成中的语义问题。

引入了语义平面之后，有必要在语言的描写方面作一些实质性的改变。在一般的机器翻译系统中，最小的翻译单位是词，最大的翻译单位是单个的句子，而不考虑属于不同句子的词与词之间的关系，由于引入了语义平面，就必须超出句子范围来考虑问题，从最小的研究单位——义素出发，逐渐地建立起由这些义素形成的更大的单位来（如词、词组、句子、句段、文章等）。这样，语言处理的范围就从“词--句子”扩大为“义素--文章”了。

由于语言研究的深入，机器词典的内容也更加丰富了。词典的书写形式与语法的书写形式越来越接近，词典与语法的形式描述方法日趋统一。而且，在词典中，特别注意词条的语义特性的描述。

例如，俄国莫斯科国立外国语师范学院机器翻译实验室编写的自动词典 APMAC（Англо－русский многоаспектрый автоматический словарь）有英语和俄语两部分，词典正文由词汇单元、形态单元、句法信息、语义信息、词汇信息等五个方面组成，内容十分丰富，可以给机器翻译提供尽可能多的信息。显而易见，要编写这样的包含多方面信息的词典，没有对语言作过全面深入的研究是根本做不到的。

目前，电子词典的研究日趋兴盛。日本专门成立了电子词典研究所，以电子词典的研究为中心，来带动自然语言处理的其它研究课题。

语言研究的深入还表现在多种句法分析方法的提出。自动句法分析是机器翻译的关键。在目前语义分析还不十分完善的情况下，机器翻译中主要还是靠句法分析。迄今人们提出的自动句法分析方法可以分为两类：一类是弱面向模型的语法（weakly model－oriented grammar），它着重于处理方法本身，是描写性的；另一类是强面向模型的语法（strong model－orientedgrammar），它着重于理论，是规范性的。

弱面向模型的语法主要有：英格维（V. Yngve）提出的结构格式分析法，爱. 罗德斯（Ida Rhodes）和埃丁格尔（A. G. Oettinger）提出的预示分析法（predicative analysis），加尔文（P. Garvin）和提出的支点分析法（fulcrum approach），威尔斯（K. Wells）提出的直接成分分析法（immediate constituent analysis），帕. 罗德斯（A. F. Parker－Rhodes）和李德汉（R. Needham）提出的团块理论（clump theory），赛卡托（S. Ceccato）提出的关联语法（correlational grammar）等。

1. 结构格式分析法

所谓结构格式分析法，就是首先根据词的分布状态把它们分成若干个类别，再根据这些类别确定词组类型，并构成词组类型清单。在进行机器翻译时，首先借助机器词典把词的类别特征记录在文句中所有的词上，然后，计算机把所要分析的文句与词组类型清单进行比较，从而在文句中找到相应的词组类型。这样，就可以确定文句中词与词之间的句法关系。

2. 预示分析法

所谓预示分析法，就是当计算机在从左到右“阅读”句子时，后面每一个词的句法作用，可以由前面一些词的句法作用预示出来。例如，有A，B，C三个词，对词B进行分析时，可用下列方式完成：

(1) 辨明词A的哪种信息和词B本身的特性一致；

(2) 根据对词B分析的结果，为词C提供一个或几个信息。

如此连续进行下去，直到完成原语句子的整个分析为止。

3. 支点分析法

支点分析法中的所谓“支点”(fulcrum)，就是一种决定句法单位性质的成分。例如，在俄语中，句子的支点是谓语，而谓语的支点是人称形式的动词或短尾形容词。支点分析法先由句子的最小成分开始分析，然后把这些成分逐渐组成越来越大的单位。

4. 直接成分分析法

直接成分分析法把句子作为最大的语言分析单位，经过一系列的分割和再分割，产生出越来越小的句法成分。使用这一方法来作句法分析时，对于所分析的句子，必须来回地进行多次的扫描。

5. 团块理论

团块理论采用格论方法建立语言的格论模型，从而研制出一种形式化的句法语段系统。这种形式化的句法语段，称之为“团块”(clump)。用团块理论来作自动句法分析时，也要对分析的句子进行多次的扫描，并根据从属关系和支配关系把有关的句子成分组织起来。

6. 关联语法

关联语法要分析每一个词汇单元与语言中其它各成分之间的关联关系，作出一份数量有限的关系表，从而建立起自动句法分析的规则系统。

强面向模型的语法主要有：海斯（D. Hays）的从属理论（dependency theory），兰姆（S. Lamb）的层级理论（stratificational theory），哈里斯（Z. Harris）的转换理论（transformational theory），塞内沙勒（D. Senechalle）的组成理论（formational theory）等。

1. 从属理论

从属理论的目的在于建立句子各成分之间的从属关系，采用从属关系树形图把这种关系表示出来。

2. 层级理论

层级理论认为，语言的结构可以分为若干个等级严格的平面（或层级），最低的平面是语音平面，最高的平面是语义平面，语言就是一个严格的层级系统。

3. 转换理论

转换理论认为，一种语言中有一定数量的基本的句子类型，叫做核心句，该语言中其它较为复杂的句子可以利用转换规则从核心句转换而来，因此，一种语言就可以看成是由核心句及其转换式组成的系统。

4. 组成理论

组成理论是在转换语法理论的基础上，把符号行的数学理论加以形式化而提出来的，这种理论采用元语言（metalanguage）来描写对象语言（object language）。

70 年代以来，自动句法分析的方法在原有基础上得到了进一步的发展。例如，依尔里（Earley）于 1970 年在直接成分分析法和转换理论的基础上，提出了“依尔里分析法”。这种分析法把自顶向下分析与自底向上分析两种方法结合起来，从树形图的顶点开始，逐级进行试探，最后由计算机造出能表示句子结构的成分结构树来，这种方法大大地提高了自动句法分析的效率，使用起来也十分灵活。

值得我们注意的是，自从进入 80 年代以来，在现代语言学中，继转换生成语法、扩充转移网络语法、格语法和系统语法之后，又陆续出现了一些新的语法理论。其中比较著名的有：布列斯南（J. W. Bresnan）和卡普兰（R. Kaplan）的词汇功能语法（lexical functional grammar，简称 LFG），盖兹达（G. Gazdar）等的广义短语结构语法（generalized phrase structure grammar，简称 GPSG），马丁. 凯依（M. Kay）的功能合一语法（functional unification grammar，简称 FUG），佩瑞拉（F. Pereira）等的定子句语法（definite clause grammar）。这些新出现的语法理论有三个共同的特点：

第一、这些语法都采用了复杂特征集来对词、短语和句子各级语言单位进行描写，而不像早先的许多语法理论那样，只采用语法范畴（如词类或短语）这种单一的标记来描写，这就提高了语法的描述能力和对语言现象的解释能力，特别适合于在机器翻译中对语言进行多方面的描述。

第二、这些语法在分析语言的过程中，广泛地采用了“合一”（unification）运算，而不只单纯地依靠简单的模式匹配技术，这就大大地扩充了短语结构语法的生成能力，同时又保持了短语结构语法表达清晰、处理效率高等优点。

第三、这些语法理论都带有浓烈的算法色彩，具有可操作性，有着一种自然科学式的严谨，叙述比较形式化，逻辑上也很有条理，它们的在算法上的复杂性曾有人作过严格的数学证明。

这些新的语法理论无疑会对机器翻译的语言分析技术产生重大的影响，并对传统的语法研究提出大量的新课题。

四、数学研究更加精细

在早期的机器翻译系统中，数据加工的形式化方法是通过把这些数据改变为机器内部语言的方式来实现的，而规则是通过标准规则和算子来建立的。后来，由于语言与程序分开，程序的书写格式要反映自然语言的结构，这就要从数学的角度探讨描述句法结构的形式化方法。为此，学者们采用了成分结构树和从属树来描述句法结构。

美国语言学家乔姆斯基提出了形式语言理论。他认为，语言的语法在于利用有限的规则生成语言中无限的句子。这样的语法用 G 来表示，可定义为四元组，即

$$G=(V_N, V_T, S, P)$$

其中，V_N 是非终极符号的集合；

V_T 是终极符号的集合；

S 是 V_N 中的初始符号，也就是句子的标示；

P 是重写规则，其一般形式为

$$\alpha A\beta \rightarrow \alpha\omega\beta,$$

$A \in V_N$，α，β，ω 是 V 上的符号串。

在重写规则中，当 α，β 为空符号串时，它的形式就可写为

$$A \rightarrow \omega,$$

这时，乔姆斯基就把这种语法叫做上下文自由语法（context free grammar）。

后来证明，成分结构树和从属树都可以用上下文自由语法来统一地加以描述。

为了把上下文自由语法应用到机器翻译的程序中去，可采用科克算法（Cocke algorithm）。

科克算法的规则形式为

$$W => XY$$

或 $$W => x$$

其中，W，X，Y 是非终极符号，x 是终极符号，相应地把前一规则叫做非终极规则，后一规则叫做终极规则，它们显然符合于上下文自由语法的重写规则的形式。

分析开始时，要用终极规则把全部的终极符号规约为非终极符号，这样得到的句法单位叫做第一级句法单位。接着，用非终极规则对所分析的句子建立第二级句法单位，它们是由一对一对的第一级句法单位两两相邻而构成的。然后建立第三级句法单位。第一个第三级句法单位或者由一个第一级句法单位后面跟着一个第二级句法单位组成，或者由一个第二级句法单位后面跟着一个第一级句法单位组成。换言之，如有三个词毗连，它们或者可以表示为 3＝2＋1，或者可以表示为 3＝1＋2，其中，1 表示第一级句法单位，它只包含一个词，2 表示第二级句法单位，它包含两个词，3 表示第三级句法单位，它包含三个词。

设 n 为所分析句子的词数，则第 n 级句法单位（即标有符号 S 的这一层，它是树的根）的情况，可标示出句子分析的结果是否可靠。如果在第 n 级句法单位只得到一个 S，则说明该句子的分析是无歧义的，分析的结果正确；如果在第 n 级句法单位得不到 S，则说明所分析的句子与规则不匹配，分析失败。

科克算法的表示方法过于简单，在分析过程中会产生大量的句法单位，特别在分析的中间阶段尤为突出，因而这种算法执行起来效率不高。

法国格勒诺布尔大学沃古瓦（B. Vauquois）教授改进了科克算法，提高了这种算法的工作效率，使之便于书写，易于修改。在该大学设计的第一个俄法机器翻译系统中，句法分析分两个阶段进行。在第一个阶段，采用经过修改的科克算法，建立结构成分树；在第二个阶段，把建立好的成分结构树改造成从属树，并使用筛选的办法，除去寄生结构。

随着机器翻译研究的深入，人们发现，各种机器翻译系统在程序上进行的操作，在逻辑上几乎都有着共同的结构。为了从数学的角度描述机器翻译的过程，数学家们把自动机理论中的自动机的概念加以扩展，提出了别具一格的转录机（traducteur）的概念。

在自动机理论中，自动机是形式语言的抽象的识别机，它能对输入的语言素材进行判断，从而决定它们是不是成立句子，如果是成立句子就接受，如果是不成立句子就不接受。但是，机器翻译的目的不在于识别输入的句子是不是成立，而在于对句子进行转换，也就是把输入带子上的语言数据，按一定的规则转录到输出带子上。为此，法国数学家硕塞（J. Chauché）提出了转录机的理论。

一个转录机可以定义为六元组

$$T = (V_E, V_S, Q, \delta, Q_O, F)$$

其中，V_E 是输入词汇的有限集合，

V_S 是输出词汇的有限集合，

Q 是转录机状态的有限集合，

Q_0 是初始状态的集合，

F 是最后状态的集合，

δ 是转录函数，是如下定义的一个映射

$$V_E \times Q \to P\ (Q \times V_s^*)$$

这里，×是卡氏积符号，P（A）表示 A 的有限部分的集合，V_S 是由V_S^*及其毗连运算“.”产生的自由幺半群，转换函数 δ 在状态 Q 时，把输入词汇 V_E 中的元素转录到输出词汇 V_S 中去，从而达到了自动地转录语言信息的目的。

转录机由状态控制器、输入带子和输出带子组成，其结构原理如下图所示：

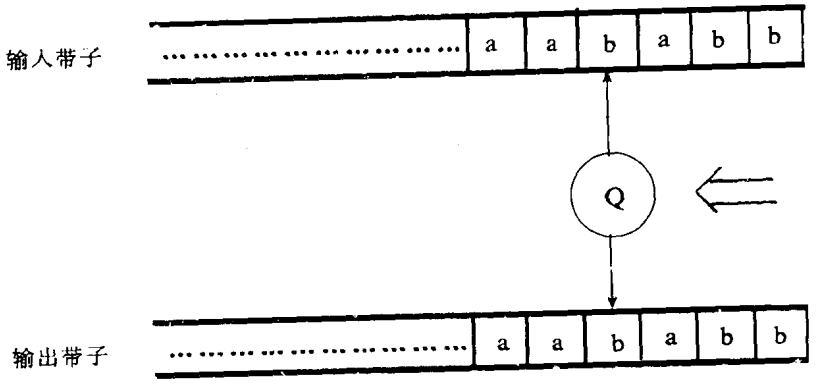

图 1. 1 转录机

输入带子分为若干单元，每个单元记录一个符号，这样，便可把输入符号记录在输入带子上。读数头从最右边的符号开始扫描，每扫描一个输入符号就向左移动一个单元。控制器中有 K 个有限的状态，当转录机 T 处于最右边的符号的读数头开始工作时，状态控制器处于初始状态 q_0，每扫描一个符号向左移动一个单元，同时把扫描到的符号及其所带的语言信息记录在输出带子上，状态控制器也就改变到一个新的状态。这样，通过有限个状态之后，便可把输入带子上的符号及其所带的语言信息转录到输出带子上。

转录机能近似地模拟机器翻译的过程，为机器翻译提供了数学模型。

五、程序技术更加成熟

早期的机器翻译系统程序设计的目的只是建立一套实现算法的执行程序。70 年代以后，机器翻译的程序设计要求从机器翻译的总体上考虑问题，除了建立实现算法的执行程序之外，还必须加上一些辅助程序。目前的发展趋势是力图研究控制整个机器翻译过程的程序系统。这样的程序系统应该包括如下的内容：

1. 实现机器翻译算法的执行程序；
2. 控制机器翻译过程的控制程序；
3. 实现外部语言到机器内部语言转写的编译程序；
4. 在上述三种程序的工作过程中，实现人机联作的程序。

由于语言数据与执行程序完全分开，这样的程序系统可以灵活而有效地控制整个机器翻译的过程。例如，借助于这样的程序系统，语言工作者可以自如地准备语言数据，根据自己的需要，既

能执行整个机器翻译过程，也能执行机器翻译中的某个阶段，如果在执行过程中发现问题，还可以随时追踪，找出问题的症结所在，一步一步地把机器翻译程序调试好，使机器翻译系统的研制与开发，有了一个良好的软件环境。

法国格勒诺布尔大学研制的 ARIANE－78 系统，就是这种机器翻译专用程序系统的典型。该系统由四部分组成：ATEF，ROBRA，TRANSF 和 SYGMOR，它们是程序的生成器，可以连续地执行机器翻译过程的六个阶段。

ATEF 是一个不确定的有限状态转录机。它的程序接收原语文句作为输入，并提供出该文句的形态解释作为输出。

ROBRA 是一个树形图转录机。它的程序接收形态分析的结果作为输入，借助于文法规则进行各种运算，输出能够表示文句结构的有标记树形图。ROBRA 还可按同样的方式实现原语-译语结构转换和译语句法生成。

TRANSF 可借助于双语言词典进行原语－译语词汇转换。

SYGMOR 是一个确定的树－链转录机，它接收译语句法生成的结果作为输入，进行译语形态生成，并以字符链的形式提供出译文。

ARIANE－78 系统的控制程序、编译程序、人机联作程序配合上述的执行程序进行工作，使得整个的机器翻译过程有条不紊。

利用 ARIANE－78 系统进行机器翻译的过程如下图所示。

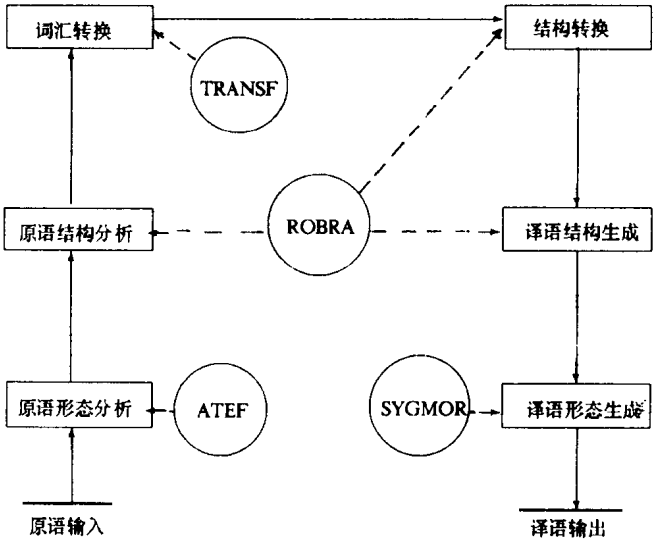

图 1．2　ARIANE－78 系统进行机器翻译的过程

从图中可看出，原语形态分析由 ATEF 执行，原语句法分析、原语－译语结构转换、译语句法生成都由 ROBRA 执行，词汇转换由 TRANSF 执行，译语形态生成由 SYGMOR 执行。输入的原语文句是字符链，执行过程中变为有标记树形图，经过树形图转换，最后变为译语文句的字符链作为输出。

ARIANE－78 系统中，语言和程序是完全分开的。语言工作者把按照程序规定编好的语言数据送入计算机，ARIANE－78 系统就开始工作。首先由编译程序对人编的语言数据进行处理，使之成为由计算机执行的语言数据，然后由执行程序进行语言数据的解释和转换，也就是执行机器翻译过程的六个阶段。语言数据的修改可在人机联作的情况下进行，在机器翻译的执行过程中，操