语言和计算机

《语言和计算机》编辑组编

1

LANGUAGE AND COMPUTER

中国社会科学出版社

语言和计算机

（1）

《语言和计算机》编辑组编

中国社会科学出版社

语 言 和 计 算 机

（1）

*

中国社会科学出版社出版

新华书店北京发行所发行

八九九二〇印刷厂印刷

787×1092毫米 16开本 9印张 214千字

1982年8月第1版 1982年8月第1次印刷

印数1—8,000册

统一书号：9190·025 定价：0.97元

目　　录

LANGUAGE AND COMPUTER

Main Articles

应用语言学的新发展

——自然语言的计算机处理

编　　者

“语言是人类最重要的交际工具”。语言的发展总是与社会的发展密切相关的。人们在运用语言的过程中不断改善和扩大它的交际职能。

最初的语言，只有有声语言的形式，后来出现了文字，产生了书面语言，但是严格地说，书面语言只是有声语言的记录。自从有了书面语言之后，语言的应用扩大了。从时间上说，它能流传到后代，从空间上说，它能传达到远方。后来又出现了印刷机，语言的应用得到进一步扩大，不仅能流传到后代和远方，而且能比较快速地大量印刷，广为传播。录音机和电话的出现，又是一次飞跃，从此，语言不但可以保留其书面形式，而且也可以保留其语音形式，不但能以书面形式，而且能以语音形式迅速传递到远方。电视、传真、复印、模式识别、各种语言分析合成仪器、电子计算机等先进技术和设备的出现，更把语言的应用提高到一个新水平。在“计算机文化”到来的社会里，语言已不仅是人与人之间的交际工具，而且是人机对话的基础。

语言应用范围每扩大一步，都会出现一些新问题。为了解决这些新问题，常常又必须采用一些新的方法和新的工具，而新方法和新工具的应用又反过来促进语言学的发展。

在整个科学技术大发展的年代，语言学也得到了巨大的发展。这种发展表现在很多方面，如：数理语言学作为语言学的一个分支已站稳了脚跟；结构语言学从只注重形式研究业已发展为注重语义研究；继转换生成语法学派之后，最近又兴起了文句语言学(Text Linguistics)。值得强调指出的是，应用语言学① 的长足进展，则是这种发展的一个最显著的特征。

应用语言学是语言学的一个重要方面，它不同于理论语言学。前者着重解决现实当中的实际问题，一般不接触语言的历史状态，而后者则着重探讨理论问题，在总结规律时常常涉及语言的历史状态。当然，两者之间的关系是密不可分的。

最初，语言研究的应用方面和理论方面没有划分开来，那是由于语言学不够发达，还没有多少必要作这种区分。十九世纪初，理论方面和应用方面的研究开始分化，例如，作为应用语言学一个分支的语言教学同当时着重探讨历史问题的语言学开始分手。十九世纪末叶，有的语言学家就提出了“应用语言学”这个概念，但是没有得到广泛的注意。二十世纪以来，语言科学得到了进一步的发展。特别是最近二三十年，各方面向语言科学提出了一系列与科学技术、国民经济发展有密切关系的新任务，语言的应用范围空前扩大，这才使得语言应用

① 应用语言学 (Applied Linguistics) 这个术语目前在国内外语言学界大体上有两种用法：（1）狭义的用法主要指语言教学；（2）广义的用法，除语言教学外，还包括其他许多内容（详见正文）。这里采用的是后一种。

方面的研究和理论方面的研究明确地区别开来，“应用语言学”这个名词开始广泛地运用。

另外，语言学由于应用范围的扩大，同其他学科的联系也增多了。过去谈到语言学同其他学科的联系时，只能说，它同文学、人类学、历史学、考古学、心理学、哲学和文化史等有联系。而今天这样说，就远远不够了。除了这些之外，还得说，同数学、信息论、控制论、生理学、物理学、电子学、医学、符号学、情报学、计算技术、通讯技术、自动化技术等方面都有联系。语言学领域如此扩展，无疑地带来了新的分工和一些部门的分化。有的人进行历史比较，有的人钻研理论，有的人去解决实际问题，还有的人则去研究一些最新的、对国民经济有重大作用、而且与尖端学科相联系的边缘问题。所有这一切，进一步扩大了应用语言学与理论语言学的分化。

应用语言学是一个较大的概念，与语言应用有关的问题可以说都属于应用语言学的研究范围。但是，根据近年来的发展趋势看，有必要区分一般应用语言学和机器应用语言学。前者处理语言应用的一些老问题，例如：语言教学，文字的创制和改革，翻译研究和标音、转写问题研究，人名、地名研究，科技语言研究和术语问题研究，言语矫正学和舞台语言研究，国际辅助语的建立，速记系统的制定，聋哑盲教育学等等。后者处理语言应用的一些新问题，而主要是跟机器打交道的一些新问题。例如：所有工程语言学、机器语言学、计算语言学等方面的语言学问题均可属此类。下面着重谈谈后一方面的问题。

电子计算机的出现，引起了科学技术的巨大变化，同时也为语言学开辟了新的发展途径。计算机一方面对语言学提出了一系列新要求，要求武装它的“头脑”以发展它的智力（如赋予它检索能力、翻译能力），给它增添“翅膀”以赋予它听觉（识别口语）、更强的视觉（识别文字），说话能力（言语合成）和听写能力（语音打字）①。另一方面它又充当语言学工作者的得力助手，帮助语言学工作者对语言素材进行分类、演算、控制和模拟。机器应用语言学是研究如何利用计算机来处理自然语言的一个新部门。目前研究得比较多的课题有以下几个：

1. 机器翻译

克服语言障碍，是一个老问题。人们曾想通过翻译、通过设计国际辅助语等途径来解决它，总未见有多大成效。由于科学技术的日新月异，运输工具的日益发达，各民族间文化交流越来越频繁，语言障碍问题相对说来也就越来越严重。电子计算机的出现，为解决这个问题带来了新的曙光。还在计算机刚刚诞生的1946年，人们便开始讨论用它作翻译的问题了。经过一段时间的探索和研究，到了1954年，终于在电子计算机上第一次完成了把俄文译成英文的试验。

机器翻译最主要的问题和最大的难关是语言问题。对于这一点，人们逐渐加深了认识。

（1）机器翻译所需要的技术条件目前已基本具备：a.运算速度高达每秒一亿五千万次的计算机已经问世。1959年，我们进行俄汉机器翻译试验时所用的计算机是当时运算速度比较高的，每秒达一万次。如果说当时机器的翻译速度与人相等，那么，现在就等于人的一万五千倍了。b.大存储量的问题也已解决，内存2048兆字节或更大的计算机已不止一种。c.各种类型高速输出装置也已具备。d.理想的输入装置②虽然还不具备，但是一种字体的光学自动阅读器已经研制成功。目前各国正在大力研制多种字体的阅读器，估计不久的将来定会突破。

① 当然，这些问题要在其他学科的积极配合下才能解决，其中有些主要还是其他学科的任务。

② 由于出现了磁带形式的书本，这些磁带又可以作为计算机的直接输入资料，因而这个问题也可算部分解决了。

（2）整个机器翻译的历史可以说是语言研究不断加深的历史。从词的对译发展到注重语法分析，又从语法分析发展到语义分析。这个过程，正好反映了对翻译问题的复杂性逐步加深认识的过程。

（3）在目前的条件下，由于技术上已有相当的保证，语言研究方面也做了不少工作，机器翻译在一定范围内已经达到某种初步实用的程度。但是，要使机器翻译达到比较高的质量，还需要在语言研究上花相当大的力气。

目前机器翻译研究已处于实际应用的前夕，世界上已经有十来个初步实用的机器翻译系统。据学者们估计，到1980—1990年，机器翻译的产品将流通于世。

2. 情报检索

科技情报的数量每八年到十年就翻一番，在大量的资料中查找某些需要的东西，犹如大海捞针。如不采用新技术，真是不堪设想。手工检索（包括穿孔卡片）和机械检索（机电检索系统、光电检索系统）已进行多年，成效有限。只是由于采用了电子计算机，情报检索才出现崭新的面貌，整个情报工作才得以进入一个“情报—计算机—电讯”三位一体的新时期。

情报检索系统中的关键问题是情报检索语言的建立。这种语言应具备能精确表达文献主题和提问主题所需的词汇语法手段，不应产生歧义，不受用户主观因素的影响，并且便于用程序运算方式进行检索。为了提高情报检索系统的效能（具有较高的查全率和查准率），在词汇方面如何消除术语的同义性和多义性，在语法方面如何求得既经济又足够的语法手段来表达必要的语法关系，这些都还需要进行认真的研究。语言学工作者在这方面可以而且应该发挥自己的作用。

3. 言语识别和言语合成

言语识别和言语合成技术的研究，目的在于提高通讯效能，解决计算机的语音输入、输出问题，实现语音控制和提高计算机的人工智能。

言语识别的研究已经进行多年，提出过不少方案，进行过不少实验，但是始终未能突破。只是由于采用了电子计算机，这方面的研究才取得了较大的进展。例如，日本的“基础口语一号”（“Spoken Basic I”）言语识别系统已能从一百四十二个句子（四个发音人均为男性）中正确识别一百一十六句（81.7%）①。这里不想具体地谈论言语识别技术和各方案的短长，而只想指出一点：言语识别单纯靠识别语音特征是行不通的，还必须把声学信息同语言信息结合起来，即利用语言的概率性辅助识别，才能得到正确结论。为此，在言语识别器中除具有辨识声学信息的机制外，还必须存有大量的语言信息或“语言知识”。上述日本的言语识别系统就由四个部分组成：(1)声学加工器，(2)词汇匹配程序，(3)句法加工器，(4)语义加工器。

言语合成的研究基本上可以说是言语识别的一个反过程，即根据一定规则（生理模型或声学模型）把言语分析所得的各种物理参数（音色、音高、音强、音长等）加以综合处理的过程。在这个过程中，语言信息或“语言知识”，同样也是必要的。

由此可见，语言学工作者除了提供言语识别和言语合成所需要的各种物理参数外，还必须提供机器所需要的“语言知识”。

4. 汉字信息处理

随着计算机非数值应用范围的不断扩大，汉字信息处理问题越来越尖锐。这个问题不解

① 日本《情报处理》杂志，1977年第5期。

决，什么情报工作自动化、什么经营管理现代化、什么印刷排版自动化、都将落为空谈。汉字信息处理系统中一个关键问题是汉字编码。汉字字形繁复，字数庞杂，而且存在大量一音多字、一字多音现象。这给编码输入带来很大麻烦。如何使编码做到简单易学、操作方便、输入迅速、没有重码，这是目前急需解决的问题。为了合理解决这个问题，有必要对汉字进行多方面的（数理的，语言学的，工程心理学的）分析研究。从语言学方面说，把语言作为一个符号系统进行研究，已经做了不少工作，而把文字作为一个符号系统来研究的还不多见。加强汉字基础理论的研究不仅对汉字信息处理是必需的，而且也是语言文字学的迫切任务。书面语的统计研究（词、字、字元、偏旁、部首、笔划）和口语的统计研究（多音节词、音节、音素、声调）不仅有利于编码方案的设计，而且也有利于输入输出装置的设制（例如键盘的设计和字库的安排）。利用计算机辅助设计编码方案或比较编码方案的优劣，同样也是一些重要课题。

5. 语言分析自动化

这里面包含的内容非常广泛。凡是利用计算机对语言素材进行加工，包括以上谈的几个方面的内容，均可谓之语言分析自动化。尽管内容多种多样，但是根据加工的复杂程度，通常可以把它分为以下三类：

（1）自动编排。这是最简单易行的一种语言分析自动化。利用计算机编辑各种类型的索引，为编词典搜集词汇并制成词表，进行各种素材的频率统计，编辑逆序词典① 等等，均属此类。这里边没有什么语言研究工作，只是一种技术性工作。例如，我们要给某一翻译规则系统编词典，就可以让计算机从我们选好的材料中把不同的词抽出来，进行频率统计，同时还可让它注出每个词的出处。计算机根据一定程序进行加工后，可以给出按频率高低编排、按字母顺序编排、按字母个数编排和带有出处的几种词表，并且同时能告诉我们词的总数。

（2）自动分析。这是一种较复杂的语言分析自动化。这种自动分析系统是根据事先存入计算机内的特定语言信息进行工作，目的在于得到预先规定的结论。例如让机器查词典，或根据词类信息把句子分解为词组等，均属此类。若结论有误，就证明该系统不够完善，需要对原先的数据或规则加以修订或补充。

（3）自动研究。这是一种更复杂的语言分析自动化。这种自动研究系统是根据计算机内存储的一般语言信息进行工作，借助统计对比等手段，得出自己推断的结论。目前，这方面还没有什么比较成熟的研究成果。

除了以上几个方面以外，计量风格学、程序教学、人工智能中的自然语言理解、以及各种数据库的建立等，也都是从不同角度和针对不同目的来处理自然语言的一些重要课题，它们也都属于机器应用语言学的研究范围。

* * *

检验语言学或语文研究工作搞得好坏，唯一的标准是实践，看它是否能满足社会的需要，是否有助于掌握语文知识，是否能提供人们足够的参考书，是否能满足语言自动加工的要求，是否能在建立“信息化社会”方面发挥作用，等等。根据这个标准来衡量，我们的工作还做得很不够。我们必须大力加强语言学研究，尤其是机器应用语言学的研究，使我国语言学在四个现代化中做出自己应有的贡献。

① 逆序词典与一般词典不同，词条是按字母顺序从词尾向词首排列。这样，词尾相同的词（如“社会”、“协会”、“学会”、“议会”）就排在了一起。人们可以利用这种词典挑选用词、检查收词情况以及研究语音、语法和词汇问题。

机　器　翻　译*

刘　涌　泉

一、引　言

机器翻译，早在计算机问世之初，就成了有关学者研究的课题，至今已有三十来年的历史了。它的发展经历了一个马鞍形的过程。经过60年代中期以后的低落，如今又开始蒸蒸日上。这个事实说明，采用现代化电子技术来克服语言障碍，更广泛地进行国际间的科学技术文化交流是人类社会发展的必然趋势，而机器翻译则是适应这一潮流而崛起的一门新兴学科。在这个历史进程中，它必将发挥越来越大的作用。

在我国需要充分借鉴外国先进科学技术文化实现四个现代化的今天，我们应当加速机器翻译的研究，加速实现科技情报工作的现代化，为四化做出贡献。

本文将简要地介绍机器翻译的基本原理，以英汉翻译为例，比较具体地谈谈它的实现过程，并且概略地阐述一下机器翻译与科技情报工作现代化的关系，最后再谈谈机器翻译的过去、现状及其未来前景。

二、机器翻译是一门边缘学科

研究翻译自动化问题的一门新学科叫机器翻译。机器翻译是语言学、数学、计算技术、自动化等科学部门相结合的产物。语言学工作者提供适合于电子计算机进行加工的词典和语法。数学工作者把语言学工作者提供的材料代码化和程序化，即变成机器语言。计算技术工作者研制便于进行翻译的计算机。自动化工作者解决外部设备问题，如光电输入装置等等。机器翻译的实现有赖于这几方面的成就和共同努力。

三、机器翻译的原理

机器翻译，是由电子计算机根据一定程序进行的翻译，它的原理没有什么奥秘。简单地说，就是让机器模拟人的翻译过程。人在进行翻译之前，必须掌握两种语言的词汇和语法。机器也是这样，它在进行翻译之前，在它的存储器（等于人脑）中已存储（或“记忆”）了语言学工作者编好的并由数学工作者加工过的机器词典和机器语法。人进行翻译时所经过的过程，机器也同样遵照执行：先查词典得到词的意义和一些基本的语法特征（如词类等），如果查到的词不止一个意义，那么就要根据上下文选取所需要的意义。在词汇意义和基本语法

---

* 本文原是1977年年底开办的“机器翻译训练班”的部分讲稿。收入本书时，一些地方作了增删。

特征弄清楚之后，就要进一步明确各个词之间的关系。在这之后，根据译语的要求组成译文（包括改变词序、翻译原文词的一些形态特征及修辞）。

四、机器翻译的过程

机器翻译的过程一般包括四个阶段：原文输入、原文分析（包括查词典和语法分析）、译文综合（包括调整词序、修辞和从译文词典中取词）和译文输出。下面以英汉机器翻译为例，简要地说明一下机器翻译的整个过程。

1. 原文输入：

由于计算机只能接受数字，原文字母和符号必须按照一定的编码法转换成二进制数字，通过穿孔卡片或穿孔纸带输入计算机①。例如What are computers 这三个词就要变为下面这样三大串二进制代码：

What	110110	100111	100000	110011					
are	100000	110001	110100						
computers	100010	101110	101100	101111	110100	110011	100100	110001	110010

2. 原文分析：

原文分析包括两个阶段：查词典和语法分析。

（1）查词典。通过查词典，给出词或词组的译文代码和语法信息，为以后的语法分析及译文的输出提供条件。机器翻译中的词典按其任务不同而分成以下几种：

a. 综合词典：它是机器所能翻译的文献的词汇大全。一般包括：原文词及其语法特征（如词类）、语义特征和译文代码。对其中某些词，还要给出进一步加工的指示信息（如同形词特征、多义词特征等），以便对这些词进行专门处理。

综合词典加工结果举例如下：Professor——名词，有生命，职业名（我们用1·4·1表示），译文代码××××××；close——“同形词”，因为它既可是动词又可是形容词，所以要到同形词典进一步加工。

b. 成语词典：二词以上连贯使用的固定词组，我们称作成语。我们这个系统与众不同，为了提高翻译速度和质量，把它放到了综合词典前面（详见本刊王广义文）。例如，at the same time，不必经过综合词典得到每个词的信息后再到成语词典去找，即可直接得“副词状语”特征和“同时”的译文。

c. 同形词典：专门用来区分英语中有语法同形现象的词。例如close一词，经过综合词典加工未得到任何具体的词类，而只能得到该词是形/动同形词的指示信息，转到这里后，按照同形词典所提供的检验方法，来确定它在句中到底是用作形容词还是动词。同形词典是根据语言中各类词的形态特征和分布规律构成的。例如，动词、形容词同形的图示中，就有这样的规则：close后有er，est为形容词；处于“冠词＋close＋名词”和“形容词＋close＋名词”等环境时也为形容词，……

d. （分离）结构词典：某些词在语言中与其他词可构成一种可嵌套的固定格式，我们给这类词定为分离结构词。根据这种固定搭配关系，可以简便而又切实地给出一些词的词义和

① 理想的原文输入，是通过光电阅读装置自动识别出英文字母并自动转换成二进制代码。

语法特征（尤其是介词），从而减轻了语法分析部分的负担。例如：effect of.....on（××对××的影响）这一分离结构可给出：

effect＝偏谓，“影响”

of＝前介主 B，“零义”

on＝前介宾 C，“对…的”

e. 多义词典：语言中一词多义现象很普遍。造成多义的原因是非常复杂的。因此，要确定一个词究竟用作哪个义项，往往是很困难的，但是这又是必须解决的问题。

多义词的义项选择也主要是靠上下文解决。每一个词，不管其本身有几个义项，但在具体上下文中，一般必是只用一个义项。为了解决多义词问题，我们必须把源语的各个词划分为一定的类属组。例如：名词就要细分为专有名词、物体类名词、不可数物质名词、抽象名词、方式方法类名词、时间类名词、地点类名词等等。利用这样的语义类别来区分多义现象，是一种比较普遍的方法。例如：Effect 一词，当它前面是专有名词（例如人名）时，要选择“效应”为其词义：Barret effect “巴勒特效应”、Peltier effect “佩尔蒂尔效应”、Joule effect “焦耳效应”，等等。当它处在表示“过程”意义的动名词之后时就要译为“作用”，如：deoxidizing effect “脱氧作用”、extrusion effect “挤压作用”，等等。这种利用语义搭配的办法，并非万能，但总还是能解决相当一部分问题。

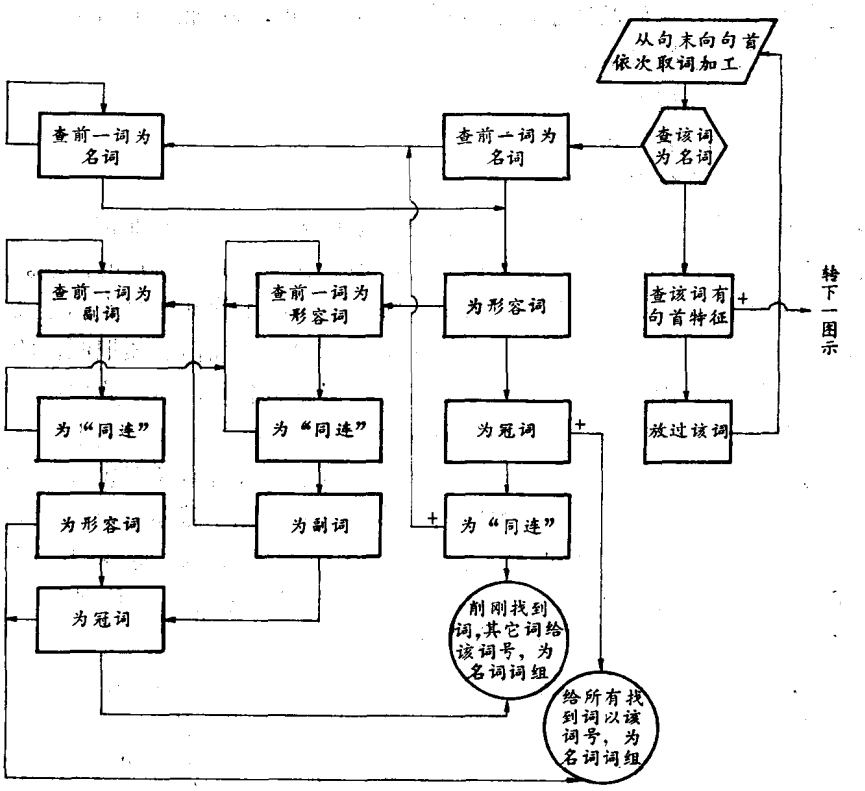
通过查词典，原文句中的词在语法类别上便可成为单功能的词，在词义上成为单义词（某些介词和连词除外）。这样就给下一步语法分析创造了有利条件。

（2）语法分析。在词典加工之后，输入句就进入语法分析阶段。语法分析的任务是：①进一步明确某些词的形态特征，②切分句子，③找出词与词之间句法上的联系，同时得出英汉语的中介成分，一句话，为下一步译文综合做好充分准备。

根据英汉语对比研究，我们发现，翻译英语句子（其他外语译成汉语，基本上也是如此），除了翻译各个词的意义之外，主要是调整词序和翻译一些形态成分。为了调整词序，首先必须弄清需要调整什么，即找出调整的对象。根据我们的分析，英语句子一般可以分为这样一些词组（或称词团）：(1) 动词词组，(2) 名词词组，(3) 介词词组，(4) 形容词词组，(5)分词词组，(6)不定式词组，(7)副词词组。正是这些词组，承担着各种句法功能：谓语、主语、宾语、定语、状语……其中除谓语外，都可以作为调整的对象。

如何把这些词组正确地分析出来，是语法分析部分的一个主要任务。上述几种词组中需要专门处理的，实际上只是动词词组和名词词组。不定式词组和分词词组可以说是动词词组的一部分，可以与动词同时加工：动词前有 to，且又不属于动词词组，一般为不定式词组；-ed 词如不属于动词词组，又不是用作形容词，便是分词词组；-ing 词比较复杂，如不属于动词词组，还可能是某种动名词（如为一般的动名词，则可与其他词构成名词词组），如既不属动词词组，又不为动名词，则是分词词组。形容词词组确定起来很方便，因为可以构成形容词词组的形容词在词典中已得到“后置形容词”特征。只要这类形容词出现在“名词＋后置形容词＋介词＋名词”这样的结构中，形容词词组便可确定。介词词组更为简单，只要同其后的名词词组连结起来也就构成了。比较麻烦的是名词词组的构成，因为由连词 and 和逗号引起的一系列问题需解决。

为了简要地说明计算机如何确定词组，下面从翻译规则系统中引出部分框图作为示例。



通过这部分框图，不带连词和逗号的各种名词词组（N; N_1.....N_m; A_1.....A_m+N; A_1.....A_m+N_1.....N_m; Art+N_1.....N_m; Art+A_1.....A_m+N_1.....N_m; Adv+A_1.....A_m+N_1.....N_m; Art+Adv+A_1.....A_m+N_1.....N_m）便可构成。

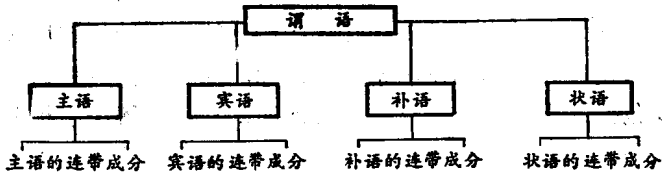
词组构成了，调整词序的对象有了，至于是否需要调整以及调整到什么位置，这要根据英汉语的对比特点而定。比如英语的后置定语，译成汉语时一般应移到被定语前。又如英语状语的排列通常是"动词＋方式状语＋地点状语＋时间状语"，而译成汉语时一般要调整为"时间状语＋地点状语＋方式状语＋动词。"

不过在调整词序之前，还有两个基本前提要解决①。

① 分清层次：调整词序必须按语言的一定层次进行，也就是说，必须按照一定的加工顺序进行。不分层次，势必调整得乱七八糟。层次如何划分？这要取决于句中各句法单位之间联系的性质。

一个简单句一般由主语和谓语这两个成分组成。这两个成分在语法上是相互联系和制约的。这个句子扩展之后，扩展成分和句中某个成分也相互联系和制约着。根据成分之间联系的特点（每个成分只能和另外一个成分发生直接联系，而一个成分可以成为若干成分联系的中心）来看，谓语是句中最大的联系中心，而跟谓语直接联系的各个成分还可能带有其他连带成分，形成各种次中心。一个简单句的层次结构可以图示如下：

① 关于这点，可参看刘涌泉、刘倬、高祖舜《机器翻译中的词序问题》，载《中国语文》1965年第3期。



从图中可以看出，一个简单句的结构可以分为三层：谓语单独构成一层，称之为谓语层；跟谓语直接联系的成分（如主语、宾语、补语、状语）构成一层，称之为直接成分层；各直接成分的连带成分，又构成一层，称之为间接成分层。处于直接成分层的成分，统称为直接成分；处于间接成分层的成分，统称为间接成分。

层次分清了，才能得到正确的加工顺序：先调整间接成分的位置，然后再调整直接成分的位置。如果相反，间接成分就要被丢到一边，找不到它的直接联系中心。

② 确定轴心：调整词序必须要有轴心。没有轴心，无从谈调整。轴心不能随便确定，轴心必定是成分的直接联系中心。

成分处在不同的结构层上，直接联系的中心不同，因而它的轴心也不同。直接成分的轴心是谓语，因为它们和谓语直接发生联系。间接成分的轴心一般是与之有直接联系的各直接成分，也可以是另一个间接成分（如果它是那个间接成分的连带成分的话）。直接成分的轴心是调整词序的主轴心（因为谓语是句中最大的联系中心），而间接成分的轴心是调整词序的次轴心。

调整词序中的最后一个问题是如何划分句子成分①。一般划分句子成分，主要是根据逻辑语义原则。这样划分出来的成分不能满足机器翻译的需要，因为它们只能表示成分的功能意义，而不能表示成分的分布关系（成分与成分之间的联系，成分在句中的位置）和两种语言的对比差异。然而后者对机器翻译是极端重要的。道理很简单：一个成分，如果本身不能表示分布关系，机器便得不到它在句中与哪个词发生联系和处于什么位置的任何信息；如果本身不能反映出两种语言的对比差异，机器便不能在综合加工时直接得到把原文改造成译文的任何根据。因此，我们根据机器翻译的特点和要求建立了一套特殊的成分体系，即所谓中介成分体系。

这套成分体系所依据的原则有三个：（1）逻辑语义原则。根据这一原则，划分出了六大类句子成分，即主语类、谓语类、宾语类、补语类、定语类、状语类。（2）结构层次原则。根据这一原则，划分出直接成分（我们以A代表它）和间接成分，而间接成分又根据它的联系词的远近分作两种：是前一词或后一词的称为近间接成分（以B代表它），不是前一词或后一词的称为远间接成分（以C代表它）。（3）对比差异原则：根据这一原则，划分出能反映该成分在两种语言中位置相同与否的前移成分、原位成分和后移成分，同时还划分出介词性成分和非介词性成分等。

根据这三项原则划分出来的成分，既能反映源语的深层结构，又能明确无误地反映被转换成的译语深层结构。具体地说，我们不仅能象通常人们所做的那样，只分析出修饰语、中心词等等，而且还分析出它们各属于哪一层。同时，我们还能通过第三原则，顺利实现源语深层结构到译语深层结构的转换。道理很简单，深层结构的成分在各语言间的差别并不大，

① 参看刘涌泉、高祖舜、刘倬《机器翻译浅说》128—134页

通过前两个原则，一般来说，既找到了源语的又找到了译语的深层结构的成分。差别主要表现在这些成分的分布排列上（当然还表现在这些成分的表层结构的形成上），不过这些成分排列上的差别，通过对比差异原则加以调整便可消除，从而转换过程自然形成。

这套成分体系包括几十个中介成分。实践证明，它不仅解决了相关分析独立综合的问题（即不管翻译什么语言，原文分析是什么样，而译文综合都使用这一套），而且为译文综合创造了极为有利的条件。例如，机器见到“宾A”，便可知道，它是同谓语发生联系的直接成分，没有移位指示，故可置之不理。又如见到“前介定B”，便可知道，它是一个近间接成分，是介词短语作定语用，有前移信号，而且是要移到前一词（成分）之前，因为B成分的联系词是前一词或后一词。

语法分析的最终目的就是要得出这类中介成分。

3. 译文综合：

原文分析是整个翻译规则系统中最复杂的一部分，相对来说，译文综合比较简单，事实上它的一部分工作（如该调整哪些成分和调整到什么地方）在上一阶段已经完成。这一阶段的任务主要是把应该移位的成分调动一下。

如何调动，即采取什么加工方法是一个非常重要的问题。根据层次结构原则，我们认为下面的方法是一种合理的加工方法（不但适合于俄汉机器翻译，同样也适用于英汉机器翻译）：首先加工间接成分，采取从后向前依次取词加工的方法，也就是从句子的最外层向内层加工的方法；其次是加工直接成分，采取依成分取词加工的方法。如果是复句，还要分别情况进行加工。一般复句，在各分句内部各种成分调整之后，各分句都作为一个相对独立的语段处理，采用从句末（即从句点）向前依次选取语段的方法加工；包孕式复句是采用先加工插入句再加工主句的方法，因为插入句如不提前加工，主句中跟它有联系的那个成分一旦移位，它就失去了自己的联系词，整个关系就要混乱。

译文综合的第二个任务是修辞加工，即根据修辞的要求增补或删掉一些词：比如可以根据英语不定冠词、数词与某类名词搭配增补汉语量词“个”、“种”、“本”、“条”、“根”等；再比如还可以根据有even（甚至）这样的词出现，给谓语前加上“也”字。又比如还可以根据主语中有every（每个）each（每个）all（所有）everybody（每个人）等词，给谓语前加上“都”字等等。

译文综合的第三个任务是查汉文词典，根据译文代码（实际是汉文词典中汉文词的顺序号）找出汉字的代码。

4. 译文输出：

通过汉字输出装置将汉字代码转换成文字，打印出译文来。

以上概括地描述了机器翻译的整个过程，同时简要地说明了为实现这一过程而采取的某些原则和方法。下面拿一句话作例子来看看这个过程是如何体现的。见《英汉机器翻译过程示例》。

五、简单的历史回顾

英美及苏联开始研究机器翻译的时间比较早，1946年计算机刚一问世，英美学者就开始讨论用它作翻译的可能性，并在1954—1956年相继进行过几次试验。除了这三个国家，日

本、法国、意大利、罗马尼亚、匈牙利、捷克、德意志民主共和国、加拿大、德意志联邦共和国、比利时、瑞士等国也陆续开展了机器翻译的研究。

我国进行机器翻译研究开始于1957年。我们主要是研究了俄汉、英汉机器翻译。俄汉方面，计算所、语言所、情报所合作研究制定了一套俄汉机器翻译规则系统，并于1959年国庆十周年前夕在我国自制的104电子计算机上成功地进行了试验（我们是世界上第五个进行机器翻译试验的国家，翻译的材料比前面一些国家的难度大，由于当时没有汉字输出装置，译出的句子是汉字代码）。之后，上述三个所，与北京俄语学院等单位协作，在过去研究的基础上又制定了一个新方案。英汉翻译方面，语言所和北京外国语学院协作，于1960年初编制了一套英汉机器翻译规则系统。由于当时人力有限，有些部分编制得比较简单。但是，从这个系统中也取得了不少经验。1975年底以来，先后开展了几个项目的英汉机器翻译研究（包括英汉题录翻译和全文翻译），并且已经取得了一定的试验成果。另外，还进行了德汉、法汉以及多种语言机器翻译的探索研究。

从国际范围来讲，机器翻译的发展经历了一条曲折的道路。大体上说，五十年代初到六十年代中叶为大发展时期，一些国家投资多，研究单位也多。但是由于当时对机器翻译的复杂性认识不足而产生了过分的乐观情绪。六十年代中叶到七十年代初由于遇到了困难而处于低潮时期，有的国家散布悲观情绪，有的国家受了影响，致使机器翻译未能大踏步前进。七十年代以来，机器翻译又进入了扎扎实实的复兴时期。越来越多的人认识到，在当今“情报资料成灾”的时代，有必要利用新技术来解决翻译问题。科学技术方面的新成就，以及机器翻译理论和实验方面已经达到的水平，目前已经可以使人们有根据提出向实际应用阶段过渡的问题了。

据报道，目前世界上已有十来个面向应用的机器翻译规则系统，其中一些是机助翻译系统，有的甚至只是让机器帮助查词典，但是据说也能把翻译效率提高50%。这些系统，可以说都还存在一些问题，有的系统，人在其中参与太多，有所谓“译前加工”、“译后加工”、“译间加工”，离真正的实际应用还有一段距离。

六、目前存在的问题

前面已经说到，机器翻译是一门综合性学科，它的实现有赖于语言学、数学、计算技术和自动化技术等科学部门的成就和共同努力。

现在的情况是，快速（每秒运算一亿五千万次）、大容量（2048兆字节）的计算机已经可以满足机器翻译的实际要求，软件技术的发展也为编制机器翻译程序提供了很大方便。另外，高速的输出装置也已齐备。从技术条件来说，目前所差的只是输入装置显得落后。由于文字识别已成为国际间重要的人工智能课题，因而可以相信，自动阅读装置在不久的将来也会完善起来，并付诸实际应用。

机器翻译研究中目前存在的最主要的问题是对语言的研究不够。为了让计算机作翻译，必须预先制定出适合于计算机使用的词典和语法。但是，做好这项工作是相当困难的。困难在于，人们还没有详尽地揭示和描述出所研究的语言的全部规律。拿英汉机器翻译来说，人们还没有在语言的一切平面上发现它们的对比差别：语法的、语义的和语用的（Pragmatic）等等。困难还在于，尽管传统语言学、结构主义语言学和转换生成语言学都从不同角度为机

英 汉 机 器 翻 译 过 程 示 例

原文输入		1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 The results of such measurements are converted into the strength index necessary for evaluating the coking property .
原文分析	查词典	冠词, 名词, 介词, 形容词, 名词, 动词, 动词, 介词, 冠词, 名词, 名词, 形容词, 介词, 动词, 冠词, 名词, 名词, 2.0* 17.4 1.0 7.2 7.2 1.5 7.3 4.18 2.0 3.1 3.3 10.1 4.4 7.2 2.0 17.2 3.1 ‹25›** ‹28› ‹15› ‹3› ‹6› ‹27› ‹10› ‹18› ‹9› ‹5› ‹22› ‹101›
	语法分析	$6_7+7=7$谓语，被动 (削‹3›) 动名Ⅱ，偏谓 $1_2+2=2$*** $4_5+5=5$ $9_{11}+10_{11}+11=11$ $15_{17}+16_{17}+17=17$ 主A 前介定B‹…110›**** 介宾A‹121› 前定B 前介状B‹116› 宾B
译文综合	调整词序	$(3_6+5‹110›)_2+2=2$ $12_{11}+11=11$ $(13‹116›+14+17)_{12}+12=12$ $8_{11}‹121›+11+11$
	查词汉文典	‹28›(这样)‹15›(测量)‹110›(的)‹25›(结果)‹6›(转变)‹121›(为)‹116›(对于)‹9›(评价)‹5›(炼焦)‹22›(性能)‹18›(必需的)‹27›(强度)‹10›(指数)‹101›(.)
译文输出		这样测量的结果转变为对于评价炼焦性能必需的强度指数。

* ×,×是根据该词的语法语义特点划分的小类，例如冠词中的2.0是指定冠词，名词中的3.1是指抽象名词。动词中的7.3是指某种及物动词。

** ‹25›为该词的译文代码，即 result 的意义（结果）。这里的数字是假定的。

*** $1_2+2=2$ 表示 1 已变为 2，用同号结成一团，移位时一起移。

**** ‹…110›：110 为“的”字的代码，…表示翻译时“的”字要放在介词词组后。