第一篇 《现代汉语语法信息词典》导引

第一章 语言信息处理与语法研究

1.1 语言信息处理的研究内容

“让计算机能用自然语言同人自由交流”一直是人们的梦想。随着社会生活的日益信息化，特别是因特网的迅速扩张，人们要实现这个梦想的愿望越来越强烈。所谓自然语言，指的是人们日常使用的语言，如汉语、英语、日语等。它是相对于人造的计算机语言而言的。计算机系统内部是通过特定形式的“语言”（二进制代码序列）传递信息、协调动作的。为了让计算机为人类服务，人同计算机也必须互相沟通。当前在绝大多数情况下，人同计算机通讯时，特别是当人告诉计算机“做什么”和“怎么做”的时候，所使用的语言仍然局限于程序设计语言、操作系统的命令语言以及窗口界面上的图标或菜单等。如果计算机能够“理解”自然语言，用户就能够通过自然语言使用数据库、专家系统、管理信息系统等各种软件，用户同因特网的沟通就更直截了当，那将一扫计算机屏幕前沉闷枯燥的气氛，计算机的环境将变得更加引人入胜，家用电器的智能化也将更上一层楼。因此，“自然语言理解”一直是计算机科学中的一个富有挑战性的课题。

从计算机科学的角度看，自然语言理解的任务是建立一种计算模型，这种计算模型能够像人那样“理解”自然语言。这就有必要给出关于“理解”的定义。然而，由于自然语言固有的复杂性，人们对自己理解语言的机制也还是不甚了了。说话人可以用不同的话表达同样的意愿，也可以用同一句话表达不同的意思。反过来，对于同一句话，不同的听话人也会有不同的反应。人与人用自然语言(包括口头的与书面的)进行交流之所以没有困难，是因为交流总是在一定的环境中进行的，交流双方的知识背景一定有共同的部分，而且交流的目的大体上也有了预设。现在的计算机智能还远远没有达到能够像人一样了解环境与理解语言的水平，即使在可预见的将来也达不到这样的水平。因此，给“自然语言理解”下一个本质性的定义是极其困难的。不过，由于语言是信息的载体，关于计算机对自然语言的理解一般可以根据实用的信息处理的观点来进行评判，实际上这些评判方法的原理同计算机科学史中著名的图灵试验（Turing test）[92]是相符合的。 如果计算机系统实现了（1）人机会话，或（2）机器翻译，或（3）自动文摘，或（4）抑扬顿挫带有感情地朗读文章等语言信息处理功能，则认为计算机具备了一定程度的理解自然语言的能力。由于这些系统，除了分析输入给计算机的文章或话语之外，还需要具备生成自然语言的语句或文章的功能，因此，在计算机科学中，除了“自然语言理解”，也常常使用“自然语言处理”或“语言信息处理”这些意义相近的术语。本书倾向于使用“语言信息处理”。按照国家标准 GB 12200.1－90 的定义，“语言信息处理”指“用计算机对自然语言的音、形、义等信息进行处理，即对字、词、句、篇章的输入、输出、识别、分析、理解、生成等的操作与加工”[1]，其内容是相当广泛的。观察计算机系统所处理的语言信息，大致上可分为两类。一类是模式信息，如声音和图像，它们是语音识别和文字识别的前期处理对象。另一类是符号信息，如书面语的文本或者作为汉语语音识别结果的音节符号，它们是代码化了的，或者更确切地说，计算机只将每个字符的编码看作处理对象。利用键盘进行人机会话，对存储于计算机系统内的文本做摘要，进行检索、校对、翻译，乃至让计算机“理解”人类的语言，所有这些工作，计算机所处理的对象都是符号信息。通常文献中所说的“语言信息处理”是指其处理对象为符号信息，本书也是在这个意义

上使用“语言信息处理”这个术语。为了实现语言信息处理的各种功能，人们开发自然语言的词法分析、句法分析、语义分析、语境分析等技术，这些技术的理论基础是计算语言学。计算语言学是植根于计算机科学、语言学、数学和认知科学等学科而成长起来的一门新兴学科。计算语言学既对语言信息处理技术的发展起指导作用，又受语言信息处理实践的推动和检验。计算语言学和语言信息处理技术的发展是相辅相成的。

自然语言信息处理经历了艰难曲折的发展过程。事实上，数字电子计算机在非数值领域的应用最早是在语言信息处理领域内开始尝试的。电子计算机问世不久，人们就开始了机器翻译试验。但无论同计算机科学技术本身的发展速度相比较，还是同计算机适用于各行各业的应用技术的发展速度相比较，语言信息处理的发展都是相当缓慢的，道路是曲折的。

本世纪 50 年代后期及 60 年代前期美国出现过机器翻译研究的第一次热潮。1966 年美国科学院语言自动处理咨询委员会发表的 ALPAC 报告[2]给机器翻译泼了一瓢冷水，语言信息处理有过一段沉寂期。自 70 年代后期以来，由于计算机技术的飞速进步和语言学理论的发展，由于一些机器翻译系统和数据库自然语言界面进入实用，更由于社会需求的推动，语言信息处理研究重新进入繁荣期，然而道路依然崎岖。原定 90 年代初完成的国际上两个大型机器翻译研究计划（欧共体的 EUROTRA 和日本与 4 个邻国的 ODA）都未能达到预期的目标。90 年代初一些学者倡导的基于语料库的统计学方法同样碰到重重障碍。国内外都有相当一部分专家对自然语言处理的现状、理论基础、技术路线在进行冷静的思考，一些学者认为至今尚未能跨越“语义障碍”，同时也在酝酿着新的突破[3]。近年来，Internet 迅速扩张，大量的信息犹如潮水般涌来，这些信息的主要载体仍然是自然语言，人们渴望发展自然语言信息处理技术以实现文本自动分类、文献检索、信息提取、自动翻译、自动文摘、自动勘校，以加速信息、知识与文化的交流，促进社会、经济、科学的进步，显然这是每一个国家都面临的挑战。语言信息处理技术的发展又有了新的强大的推动力量。人们已经了解到，语言信息处理技术有着广阔的应用领域。已有一些语言信息处理系统形成产品，进入了市场。同时，人们在开发语言信息处理系统时所创造的各种分析技术，所积累的诸如电子词典、语料库等语言数据资源也会被集成到各种信息处理系统中，从而提高信息处理系统的智能水平。语言信息处理产业崛起的前景已经呈现在人们眼前。

由于语言与思维、文化的密切关系，语言研究已成为西方现代哲学和人文科学发展的突破口。语言科学是人文科学中的领先科学，是人文科学与自然科学之间的桥梁，在整个科学体系中具有与哲学、数学相当的地位[4]。形式语言学的成果促进了计算机科学的发展已成为科学史上的佳话。由于在当代语言学研究中引进了数学方法和计算机技术，语言学本身也产生了飞跃，出现了许多新的分支交叉学科，其中计算语言学是最活跃的一个分支。

当前国外的语言学研究很多是围绕着一个中心课题展开的，这个中心课题就是同研制智能计算机有关的语言信息处理问题[5]。语言信息处理研究所取得的理论成果会对哲学与人文科学的发展产生重要的影响，其社会意义可能更在技术、经济意义之上。

智能的本质是当代科学难题之一。自然语言理解机理的探讨也是关于人类智能本质的探讨的一个重要组成部分。要实现自然语言理解，最终必须了解人是如何理解语言的以及儿童是如何学会母语的。退一步，如果能构造出人学外语的模型，对实现自然语言的计算机理解也会有重要的启示[6]。不同的语言学理论对人类的语言现象作出了不同的解释，各种争论之所以相持不下，是因为对大脑作为智能活动（包括语言活动）的物质基础的功能还未能透彻了解。经过长期的进化，人类的大脑才变成今天这样一个集成度极高的柔性信息处理系统，它不仅能够根据记忆、依据经验进行思考判断，还会在实践中学会新的知识，并能自行对学到的知识进行再组织。至今，大脑理解语言的认知过程还是一个谜。在计算机上建立一个模拟语言理解过程的认知模型（现在的自然语言处理系统是这种模型的雏形），可以为观察大脑这个黑匣子的活动提供一个“窗口”[7]。利用计算机不仅成功地模拟了逻辑思维，而且也在模拟形象思维和灵感方面进行了探索。自然语言理解的研究可以为智能科学的突破贡献力量[8]。

我国的学者研究语言信息处理自然以汉语作为主要研究对象。按照国家标准 GB 12200.1 — 90 的定义，汉语信息处理有时又称中文信息处理[1]。作者观察到当前学术界、产业界实际上是在两种意义上使用“中文信息处理”这个术语。一种意义是汉字信息处理，包括汉字字符集的确定，汉字编码，汉字的输入、输

出与远程传输，汉字字形的存储与生成以及中文编辑排版等应用系统。另一种意义才是本书所着眼的语言信息处理问题。当然，“汉字信息处理”和“汉语信息处理”的分界线也不是绝对清晰的。以中文键盘输入为例，开始以字（词）为输入单位，基本技术是编码，属于汉字信息处理范畴。当发展到以语句为单位时，要采用一些句法分析或语义分析的技术，就进入了汉语信息处理的境界。

追溯历史，我国也是世界上最早开展自然语言信息处理研究的国家之一，早在 1957 年我国就进行了机器翻译研究[9]。不过关于自然语言处理的较大规模的、比较系统的研究直到 80 年代中期才开始，是比较晚的。鉴于国情，我国的学者将主要的精力集中于实用系统的开发，机器翻译仍是最热门的课题[10]。1988 年推出的“译星”翻译软件是国内最早商品化的机器翻译系统，这几年来，版本不断更新，最新的版本“译星 99”已于 1999 年开始销售。国家 863 高技术计划支持的智能型英汉机器翻译系统 IMT/EC 装入了袖珍型电子词典，取得了很好的经济效益。桑夏公司将机器翻译技术融入因特网，建立了命名为“看世界（Readworld）”的网站。通过这个网站，不通晓英语的人也可以浏览因特网上的英语信息，可以选用英汉对照的方式，也可以只显示中文或英文。国家 863 计划支持的由中科院计算所和北京大学计算语言学研究所合作开发的汉英机器翻译系统也取得了很好的成绩，具有相当的发展潜力。其他如日汉、德汉、俄汉、汉蒙、汉藏等不同类型的机译系统也在开发中。一些典型的处理模式信息的研究课题，像语音识别、语音合成、文字识别等技术在早期均采用模式识别、神经网络等方法，很少涉及语言学的知识。随着研究的深入，研究者近来都提出了借用自然语言处理的成果改善系统性能的要求。目前，我国语言信息处理研究的发展势头是令人鼓舞的。不过，从总体上看，我国的语言信息处理研究与当前的国际水平比较，还是有一定差距的。理论研究的基础相对薄弱，理论成果较少。这种现象在科学技术的其他领域也许同样存在。对于从事语言信息处理研究的我国学者来说，需要着重探讨的是本领域的一些特殊问题。

同发达国家使用的若干种语言特别是英语和日语相比较，汉语信息处理起步虽然较晚，不过数十年来，关于其他语言的计算机理解的研究同样也未能取得突破性进展，而且正当中国人遭遇十年浩劫无法进行正常的科学研究时，国际上的机器翻译研究也处于相对沉寂的状态。直到 80 年代前期，除了中国，国际上还极少有关于汉语的自然语言处理研究的报道。到了 80 年代中期，中国的学者在比较先进的语言学理论指导下，在相当先进的计算机环境中开始了这项研究，避免了发达国家早期探索所走过的一些弯路。经过十多年的努力，汉语信息处理以及中国学者的工作已得到国际学术界和产业界的重视。人类的各种自然语言有着深层的相似性，汉语信息处理同其他语言有很多的共性，当然汉语信息处理也有自己的特性[11]。同科学技术的其他领域一样，在语言信息处理领域，中国学者也面临竞争和挑战，不过这个领域却为中国学者留下了更为广阔的空间，提供了更多的机会。汉语的“根”在中国，国内学者同汉语最亲近，最易把握汉语信息处理的特殊性。关键的问题是如何处理好既要努力同国际研究接轨又要充分把握汉语信息处理特殊性的关系。在艺术界人们承认最有民族性的艺术也最有国际性。在语言信息处理领域也应作如是观。中国学者只要充分认识到自己的优势，善于扬长补短，一定可以在语言信息处理领域为中国的发展和世界的进步做出自己独特的贡献。

1.2 语言信息处理系统的基本模型

机器翻译系统是典型的自然语言处理系统，其应用价值也最明显。当代机器翻译系统的模型可用图 1.1 表示：

图 1.1 反映的是基于规则方法的机器翻译系统的基本模型[12]。90 年代，机器翻译研究还发展了基于统计与基于实例等的各种模型[13]。不过，当前世界上实际运行的多数机器翻译系统基本上仍以基于规则的模型为基础。从这个基本模型可以了解到，机器翻译系统的基本原理乃是要素合成原理[14]。首先将原文的句子分解成基本构成要素(词，惯用语等)，这样才可以查词典，才好运用语法规则找出句子的结构，这就是句法分析(包括词法分析)，并通过语义分析及语境分析排除不适当的歧义，从而形成原文的机器内部表示。于是可在结构的层次上进行转换，得到译文句子的结构，并选择适当的译词，以后再进行词序调整、虚词增删及形态变化，最终得到译文的表层句子。

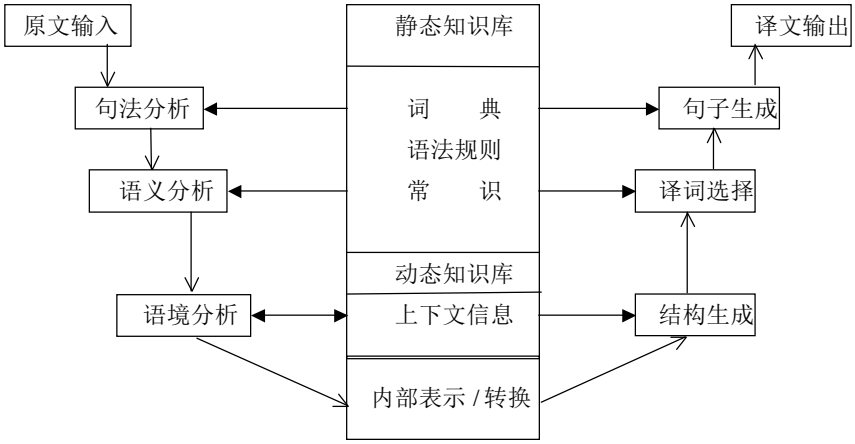

图 1.1　机器翻译系统基本模型

这样的机器翻译系统除了一般的计算机系统都有的硬件和软件(分析与生成程序)外，还有一个特别的组成部分，即语言知识库，包括静态的词典、语法规则库等，也包括动态的上下文相关信息。而且，在机器翻译系统中，语言技术(既包括存储在计算机系统内的语言知识库，也包括语言知识的归纳、表示与运用)是居于核心地位的技术。同硬件、软件相比较，目前语言技术发展尚不够成熟，成为机器翻译开发过程中必须攻克的难关。

不难理解，词典与语法规则库中需要注入大量的语言知识，上下文相关信息的提取当然也要依靠语言知识的运用。更重要的，是要在适当的语法理论指导下，找出一种合理的形式系统。这种形式系统不仅可以用来表达不同层次的语言知识，而且需要保证自然语言句子的表层线性序列与其内部表示之间以及不同语言的内部表示之间具有可计算性，也就是彼此之间能通过一系列规则、操作或过程进行转换。可以说，语言技术特别是语言知识库的质量已成为自然语言处理系统成败的关键。

除了少数实验模型，大多数自然语言处理系统都包含句法分析这个步骤[15]。狭义的语法分析就是指句法分析(包含词法分析)，广义的语法分析则包括句法分析、语义分析和语境分析。

在自然语言理解的领域中，越来越多的论著强调语义分析的重要性，这是容易理解的。在汉语中，很容易举出例子：

例 1　猴子吃香蕉

例 2　学生吃食堂

例 3　老乡吃大碗

对于这些例子，仅仅在句法的层次上分析出“吃”是动词，“香蕉、食堂、大碗”是名词，且动词与名词之间是述宾关系，并不能妥善地解决机器理解与翻译的问题，必须进行语义分析，系统要在词典与知识库中为分析程序提供语义信息，如给“吃”附以“动物的一种行为”，在句子中需有“施事”与“受事”，通常只有动物类的名词才可以充当其“施事”，只有食品类的名词才可以充当其“受事”，还要给“猴子、学生、老乡”分别附以“动物、人类”的语义标记，给“苹果、食堂、大碗”分别附以“食品、处所、器具”的语义标记[16]。

同样，也有充分的理由认为语境分析是不可缺少的，例如要将“小张打针去了”这句汉语译成英语或日语，至少要从上下文中弄清楚小张是病人还是护士。系统要有一个动态的知识库存放分析语境所得到的信息。分析程序要参照动态知识库，说明只有突破一个句子的界限，在篇章的范围内才可能正确地理解每一个句子[17]。

人们大致上采用两种方式使用自然语言进行交流[18]。一种是“意合法”，发话者的几个关键的词语，就可以让听话者捕捉到有关的信息，这当然要依赖于当时的上下文环境。另一种是“结构法”，即人们平常说的、写的句子，虽然表面上这些句子看来只是音节或词的线性序列，但实际上是有一定的结构的。在这两种方式中“结构法”是基本的。目前，计算机能处理的也就是这样的有合法结构的句子，其中心任务

就是通过句法分析、语义分析和语境分析得到句子结构的形式化的机内表示。句法分析、语义分析与语境分析，这三者之间的关系应当是以句法分析为主，词法和句法知识总是最基本的，也是研究得比较成熟的。适度的语义分析和语境分析是必要的，可以消除句法分析后残留的歧义结构。句法分析固然解决不了关于理解的全部问题，但也不宜对语义分析与语境分析期望过高。因为反映客观世界的语义系统（或者说知识系统，概念系统）即使能建立起来，也会十分庞杂，语境分析更是没有边界，很难形式化。因而，当前开发自然语言处理系统的正确策略应该是将三者有机地结合起来。这也许是最有效且最经济的原则。

即使人们一般地承认句法分析在自然语言处理中的重要作用，但是当谈到汉语时，人们的怀疑又增加了。由于汉语是我们的母语，在日常交流中运用“意合法”得心应手，而汉语的形态又不发达，对汉语语法的形式系统缺乏清晰的认识，因此，当借重句法分析开发汉语信息处理系统而得不到理想的效果时，往往对汉语句法分析的作用产生怀疑也是可以理解的。不过作者注意到各种自然语言的共性，因而难以相信在语义分析方面汉语会超越其他语言而提前到达胜利彼岸的乐观估计。相反地，作者认为当前句法分析在汉语信息处理系统中仍然处于举足轻重的地位，同句法分析做得比较成熟的英语、日语相比，仍有很多潜力可以发挥。不过这需要仔细分析汉语句法分析的特殊困难及其产生的原因，并采取恰当的对策。因此，面向计算机处理需要的汉语语法的研究以及句法分析算法的研究当前还是汉语信息处理研究的重点，至少应是重点之一。

1.3 汉语自动分析的特殊困难与对策

这个问题显然涉及汉语的特点，主要是汉语语法的特点。要确切地了解计算机分析汉语的特殊困难，就要拿汉语同其他语言比较。因此，关于其他语言的知识是必要的。限于作者的知识领域，也考虑到多数读者的实际情况，本节在论述汉语语法特点时主要是同英语、俄语与日语进行比较。按照语言类型学的观点，世界上的语言大致上可以分为“孤立语”（又叫“分析语”）、“屈折语”和“粘着语”3 个基本类型。汉语是孤立语，日语是粘着语，俄语是屈折语，英语就其主要形态特征而言也应划归屈折语。这样，本节的论述在相当大的程度上覆盖了 3 种基本类型的语言的特点。

作者所从事的或所了解的语言信息处理研究的实践表明，汉语信息处理碰到了很大的困难[97]。从外语（特别是英语）到汉语的机译系统同从汉语到外语的系统相比较，成功的例子要多得多。即使像并不需要深层句法分析的文献检索、信息提取等应用系统，也必须跨越词语切分与词性标注的障碍。作者以为，这种情况同汉语的特点特别是汉语语法的特点有密切的关系。造成汉语自动分析困难的原因可以从以下几个方面进行探讨。

1.3.1 同一词类可担任多种句法成分且无形态变化

查阅英语、俄语或日语的词典，除了能查到单词的释义、用例外，还可以查到单词所属的词类或者说单词的词性，甚至不查词典，仅根据单词形态上的特征也能判别某些词的词性。在英语中以 ation 结尾的单词一定是名词，俄语也有准确表示词性的词尾。在英语与俄语中，专有名词的第一个字母一定是大写的。在句子中英语的名词前通常有不定冠词 a（an）或定冠词 the 。在德语中，不仅专有名词，凡是名词的第一字母都要大写，句子中的名词前也要有冠词。在日语中，以する结尾的一定是动词，形容词则以い结尾。

在屈折语中，不同的词类有不同的变化。俄语和德语的名词、代名词、数词都有性、数、格的变化（俄语有 6 种格、德语有 4 种格）；形容词除在修饰名词时要与被修饰的名词的性、数、格保持一致外，还有原级、比较级与最高级的变化。英语虽属屈折语，但其名词变化较少。英语名词的数有单、复之分，名词的格只有所有格还残存形态上的标志，英语名词中只有极少数还区分性别（actor 男演员，actress 女演员。这不是语法概念中的性，但对汉英翻译有影响）。英语的人称代词仍有主格、宾格、所有格的变化（如 I，me，my）、单复数的划分（I，we；me，us；he，they）、性的区别（he，she；him，her）。英语形容词修饰名词时虽然已不受性、数、格一致性的约束，但形容词自身仍有原级、比较级与最高级的变化。屈折语

的动词变化更是复杂。英语动词作谓语时需采用限定形式，与主语的人称、数需保持一致。日语作为粘着语，其体言（名词、代名词、数词）虽然没有形态变化，但在句子中作不同的成分时必须后置不同的助词（如作主语时后置が或は，作宾语时后置を，作定语时后置の），其用言（动词、形容词、形容动词、助动词）也有复杂的形态变化。将以上种种现象加以概括，可以认为，无论是屈折语还是粘着语，在对这些语言作句法分析时，词性知识在大多数情况下是已知的或者可以根据形态变化、粘着成分等信息加以获取。语法分析就有了比较可靠的前提条件。尤其有重要意义的一个语言事实是在这些语言中词类和句法成分之间有相对简单的对应关系。在英语中“大致说来，动词跟谓语对应，名词跟主宾语对应，形容词跟定语对应，副词跟状语对应”，而“汉语词类和句法成分的关系是错综复杂的”[19]。在汉语中，名词除了主要担任主宾语外，也可以直接担任定语，在一定的条件下还可以担任谓语；作定语虽然是形容词的重要功能，但形容词与名词的主要区别在于形容词经常直接用作谓语和补语。而名词不能作补语，也很少作谓语。动词的主要功能当然是作谓语，但汉语的动词直接用作定语、动词与形容词直接用作主宾语都不是罕见的现象。概括地说，汉语的词类是多功能的。关于词类语法功能的详细讨论请参阅本书第三章。

朱德熙先生明确指出“汉语的名词、动词、形容词都是‘多功能’的，不像印欧语那样，一种词类只跟一种句法成分对应”[19]是富有真知灼见的，对于从全局上把握汉语语法的特点以及汉语同印欧语的区别是有指导意义的。不过，作者认为不能机械地绝对地理解朱先生的这个论点。实际上，朱先生在阐述这个论点时，是留有余地的，只是“大致说来”，并且进行了深入的分析。英语名词是可以直接修饰名词的，如“machine translation”、“case grammar”等，但在多数印欧语中，经常直接作定语修饰名词的是形容词，名词只有加上形容词后缀转化为形容词之后或者变成所有格形式才能修饰名词。英语的名词和形容词虽然可以出现在谓语部分，但谓语部分必须要有一个限定形式的联系动词，名词和形容词担当的只是这个联系动词的表语。英语的动词一般跟谓语对应，英语每一个句子必须有一个而且只能有一个限定形式的动词作为谓语部分的主要成分，且与主语保持人称与数的一致。不过英语的动词也是可以出现在主宾语或者定语的位置上的，只不过出现在这些位置上的英语动词不能是限定形式而只能是非限定形式（包括不定式、动名词、现在分词与过去分词）。换句话说，英语动词的一种形态只与一种句法功能相对应，如果它要担任另一种功能，则必须转化为另一种形态，尽管形态不同，也还都是动词。因为非限定形式的动词仍保留了动词的基本特征，它仍有体和语态的变化，可以受状语修饰，如果是及物动词，还可以带宾语。任何一本英语语法教科书在讲授不定式、分词与动名词时，也都是放在有关动词的章节中，因此，如果采取较为宽容的态度，可以认为英语的动词也是多功能的，但它在担任不同的功能时需使用不同的形态，其他词类（如代词、名词）也部分地具有这种特点。

再看日语。日语的动词有终止形、未然形、连体形、连用形、推量形等多种形态，作谓语时用终止形，作定语修饰名词时用连体形。日语的动词如果要作主宾语则要用连体形后接形式体言（如 こと，もの等）。日语的动词也是有多种功能的，担任不同句法成分的动词也要采用不同的形态[20]。

与英语、日语不同，汉语的动词无论担任什么句法成分，其形态都是不变化的，其他词类也是如此。这才是对汉语自动分析有着本质影响的特点。

与词类的多功能相联系，还有一个兼类词的问题。英语中有同形兼类词，如 kiss。在词典中可以查到 kiss 既是动词又是名词，两者的形态没有差别。但是在句子中，两者的差别还是会显现出来的。名词 kiss 前面常有限定词 a，the，my，her 等。动词 kiss 要发生限定形式与非限定形式的种种变化。因此在英语中 kiss 是兼类词，名词 kiss 与动词 kiss 是两个不同的词，不能将名词 kiss 的功能并入动词或者将动词 kiss 的功能并入名词。日语中サ变动词的词干与名词也是同形的，但在句子中动词与名词的区别也是显然的。汉语的情况则有明显的不同。汉语的词类虽有多功能，但又不能无限制地扩充某个词类的功能。例如，不能将作状语也扩充为名词的功能，而副词则是基本上只能作状语的一类词。汉语中有个别的词，如“重点、决心”等，它们既有名词的功能，可以作主宾语；又有副词的功能，可以作状语。需要将它们分别归入名词与副词。“重点、决心”这些词成了兼类词。汉语中也有兼类词，这是一方面。另一方面，汉语的兼类词，无论在不在句子中，形态都是一样的，如“这篇文章的重点是第 3 段”中的“重点”是名词，“这篇文章重点讲述信息抽取问题”中的“重点”是副词，两者的形态没有任何差异，这就使得在词的兼类与词类的多功能之间划一条清晰的界限变得困难了。词类的多功能与词的兼类是语法研究的一个难点。第三章

将详细讨论这个问题。句法分析要依靠系统中的词典，词典对兼类词是要明确规定的，因此汉语句法分析的困难仍在于词类具有多功能且无形态变化。

1.3.2 汉语句子的构造原则与短语的构造原则基本一致

汉语的短语（phrase，朱德熙先生的论著中叫“词组”）在汉语词组本位语法体系中有着极其重要的地位。这是因为汉语短语的构造原则与句子（这里指的是“单句”）的构造原则基本上是一致的[19]。在词组本位语法体系中，短语是一种静态的抽象的句法结构，它不与具体的话语相联系。而任何短语只要能单独站得住，带上句调后就能表示相对完整的意思，而成为汉语的句子了。在汉语中，从短语到句子是一种“实现”关系。只要把各类短语的结构和功能都描述清楚了，句子的结构实际上也就描述清楚了，因为句子不过是独立的短语而已。这里“实现”关系是相对于“组成”关系而言的。

在汉语中，由词构成短语是“组成”关系。按照组成方式的不同，可以把汉语短语的结构划分为主谓、述宾、述补、定中、状中、联合等各种类型。假设越过短语这个层次，由词直接构成句子，其组成方式或者说句子的结构基本上也就是这几种。因此可以说汉语句子的构造原则与短语的构造原则基本上是一致的。所以，朱德熙先生主张汉语语法应以词组为本。当然，在“词组为本”的汉语语法体系中，除只有一个词的句子外，通常不认为汉语的句子由词直接组成，而是由词先组成短语，再由短语实现为句子。汉语语法的这个重要特点当然也是同其他语言相比较显现出来的。英语句子的结构模式是“主语部分＋谓语部分”。汉语句子并不限于这种唯一的模式。各种类型的自由短语都可以实现为句子。每一个英语句子（包括从句）的谓语部分必须有一个用限定形式动词充当的谓语动词。担任英语句子中除谓语之外的其他成分的语言单位（词或短语，不包括从句）如果包含动词，那么这个动词必须使用非限定形式（不定式、分词或动名词）。这就是说，构造英语句子使用一套规则，构造英语短语则使用另一套规则，两者是不同的。例如：

例 1　He drives a car .

例 2　To drive a car is not difficult .

例 3　She has a friend driving a car .

例 1 中的 drives 作为句子谓语部分的主要动词，使用的是限定形式（第三人称的现在时），例 2 中的 to drive a car 是句子的主语，其中动词使用了不定式 to drive，例 3 中的 driving a car 是 friend 的定语，其中的动词用了现在分词形式。再看与这几个例子相对应的汉语。

例 4　他开车。

例 5　开车不难。

例 6　她有一个开车（的）朋友。

“开车”在例 4 中担任谓语，在例 5 中担任主语，在例 6 中不加助词“的”也可以，即“开车”担任定语。无论担任什么句法成分，“开车”的形态都是一样的，既是由同样的两个词“开”与“车”组成的，又是按照同样的结构关系即述宾关系组成的。而且，“开车”这个短语本身在一定的话语环境中就可以实现为一个句子，仍是这两个词，仍是述宾关系。在英语中，无论是“drives a car”，“to drive a car”还是“driving a car”都不是句子，只是一个短语。

汉语短语结构的另一个重要特点是各类短语的组成成分又可以是各种类型的短语。这些句法成分包括主谓短语中的主语与谓语，述宾、述补短语中的述语、宾语与补语，偏正短语中的定语、状语与中心语。这表现出了汉语句法成分特有的套叠现象[21]。当然，各种自然语言在句法结构上都具有递归性。汉语表现这种普遍的递归性的特殊之处在于短语担任不同的句法成分时形态不发生任何变化。

例 7　她完成了任务。

例 8　我知道她完成了任务。

在例 7 中，“她”是主语，述宾短语“完成了任务”是谓语，主谓短语“她完成了任务”实现了句子“她完成了任务。”即短语带上句号成了句子。在例 8 中“她完成了任务”是“知道”的宾语，“我”与述宾短语“知道她完成了任务”构成的主谓短语“我知道她完成了任务”最后才实现为一个句子。“她完成了任

务”无论作为可以实现句子的短语还是作为另一个短语的某种句法成分，其形态都是一样的。它担任“知道”的宾语时也不需要添加任何关联词。英语就不同了，请看与例 7、例 8 相对应的两句英语。

例 9　She has finished the task .

例 10　I knew that she had finished the task .

在例 9 中，主要动词 finish 用了“现在完成时”的形态；例 10 是主从复合句，主句的主要动词“knew”的宾语是例 9 中的句子，这表现了句法结构的递归性。在这个从句中，finish 仍需要使用限定形式，不过要改为“过去完成时”，宾语从句前面一般说来要加上连接成分（这里，that 是连接成分，还有很多其他的词可以作为连接成分。当使用 that 作连接成分时，有时可以省略，那是另一回事）。从汉语句法成分的套叠与英语主从复合句的构造进一步论证了汉语的短语与句子用的是同一套构造规则，而英语短语的构造规则与句子（包括从句）的构造规则是不一样的。

更深入一步考察将发现汉语中的主谓结构不仅可以作更大的短语中的宾语、主语和定语（相当于英语的宾语从句、主语从句和定语从句），而且可以作谓语。由于英语中每个句子必须有一个限定形式的谓语动词，而且也只能有一个限定形式的谓语动词（如果不考虑并列情况），因此在英语中是没有“谓语从句”这个概念的。但在汉语中，就有主谓谓语句。例如，“我肚子痛”中“肚子痛”就是谓语位置上的主谓结构。有人不同意这种观点，认为“我肚子痛”是在“我的肚子痛”中省略了“的”字。其实不然，“我肚子痛”可以扩展为“我也肚子痛”，那就不能认为“我”与“肚子”之间省略了“的”字。正因为汉语短语有上述结构及功能上的特点，在汉语语法中不需要引进英语中的“从句”（clause）概念。对照例 8 与例 10 来解释这个问题。在例 8 中，成句之前，“我知道她完成了任务”只是一个主谓结构的短语，谓语“知道她完成了任务”又是一个述宾结构的短语，其中宾语是另一个主谓结构的短语“她完成了任务”，可以如此继续。这种分析方法是按照短语构造的统一模式进行的。并无必要认为“她完成了任务”是“知道”的宾语从句。

在词组本位语法体系中，由短语实现的句子都是“单句”。不过，在结构上，汉语的短语可能相当于英语的短语，也可能相当于句子；可能相当于英语的简单句，也可能相当于主从复合句。不仅如此，比短语低一个层次的汉语合成词的组成方式也主要是这些。这正是汉语多级语法单位的构造的一致性与简明性之所在。因此，研究短语的结构对理解汉语多级语法单位的构造都是极富启发意义的。汉语短语在汉语语法中所占的重要地位现在就更加清楚了。朱德熙先生关于汉语语法应是“词组本位”的主张是非常有道理的。当然，汉语短语的概念并不能完全覆盖英语复合句的概念。在汉语中，也有“单句”与“复句”之分。复句是比单句高一个层次的语法单位[22]，复句是由分句组成的。无论是单句还是复句，又统称句子。研究汉语复句的问题将归结为研究构成复句的分句之间的逻辑关系及其表示方法，本书就不深入讨论了。

汉语的这个特点不仅同英语比较而存在，同日语比较也同样存在。日语的每个句子必须包含一个终止形的用言，而用言的终止形除了作句子的谓语之外是不能出现在其他地方的。日语的用言文节要作主宾语时需用连体形后接形式体言，要作定语也需使用连体形。日语的主谓结构虽然可以作谓语（日语句子中有大、小主语之分，这一点与汉语主语的套叠有些相似），但大、小主语要后接不同的格助词，主谓结构作定语时，其中主语后的格助词也要发生变化[20]。

朱德熙先生认为上述两个特点是关系到对汉语语法的全局认识的。造成这两个特点的根源都在于汉语词类没有屈折语那样的形态变化[19]。应当注意到，汉语与英语、日语等语言的差别对人学习的影响与对机器分析的影响是大不一样的。复杂的词形变化及多变的语句构造规则对人的记忆是个负担，中国人学外语是需要付出艰苦的努力的。而当代计算机系统具有大容量的存储与快速检索功能，对付规则的或特殊的形态变化都不会有困难，相反地，计算机程序可以从词语的形态变化中找到语法分析的根据与线索。例如英语句子中的主要谓语动词容易根据形态确定，主句与从句也比较容易区分，这些对句法分析都是至关重要的。外国人学汉语不会没有困难，但困难不在于对汉语语法基本规律的掌握。无论是中国人还是外国人都有共同的关于客观世界的知识，都有共同的用“意合法”表达知识、交流信息的能力，因而，人对汉语词类的多功能与多级语言单位的结构的一致性是容易理解、容易掌握的。相反地，计算机没有这些知识与经验，通常电子词典所包含的信息又相当贫乏，像分析英语那样，仅依靠一些语法公式来分析汉语会碰到更大的困难是不难理解的。

除了上述两个特点外，其他的一些特点也增加了计算机分析汉语的困难。

1.3.3 汉语的语序

由于汉语缺乏形态变化，粘着的助词又不是必不可少的，因此语序在组词造句中的作用就显得突出了。确定的语序曾被认为是汉语的特点之一，而且通常认为汉语的语序是 SVO[23, 24]。这个论断是否成立也要看拿汉语同哪种语言相比较。如果同俄语比较，汉语的语序确实不够自由。如果同英语或日语比较，汉语的语序未必很确定。通常认为英语是 SVO 语言，日语是 SOV 语言。请看以下 3 句汉语。

例 11　我吃了苹果。

例 12　苹果我吃了。

例 13　我苹果吃了。

这 3 种语序都是常见的。因此，朱先生认为在汉语中 SSV 与 SVO 都是常见的重要句式，很难承认汉语的语序只是 SVO。再看：

例 14　那个学生今天上午读完了这本小说。

这句话包含了 4 个较小的短语：“那个学生”，“今天上午”，“读完了”，“这本小说”。如果重新排列这 4 个短语，得到“今天上午那个学生这本小说读完了”，“这本小说那个学生今天上午读完了”……等等都能成为合法的可以理解的句子。因此，认为汉语句子的语序相当灵活也不是没有道理的。而语序的灵活不仅给句法分析过程甚至给句法规则的表述都带来了困难。

如果不从句子着眼，而从短语着眼，汉语的各种类型的短语的内部语序却是严格固定的。主谓短语是主语在前，谓语在后；偏正短语是修饰语在前，中心语在后；述宾、述补短语是述语在前，宾语、补语在后。例 14 的 4 个短语内部的语序是不能改变的。“学生那个”、“上午今天”、“完了读”、“本小说这”等都是让人莫名其妙的。

语序问题再次说明了汉语语法应以短语为本位的观点是合理的。汉语句子的分析应建立在短语分析的坚实基础上。作者曾建议语音识别与拼音汉字转换最好以短语作为一个层次的处理单位，正是基于这样的认识[25]。

1.3.4 汉语中的虚词

同样由于汉语缺乏形态变化，虚词在句子中的作用就受到了重视[24]。指称和陈述是语言表达的两种基本形态[26]。当实现指称和陈述的互相转化时，汉语虚词的作用是重要的。例如：

发工资 ———— 发工资的

我写 ———— 我写的

喝茶 ———— 喝的茶

即在适当的位置加上助词“的”表示陈述的谓词性成分转化成了表示指称的体词性成分。又如：

教授 ————（已经）教授了

春天 ————（快）春天了

即表示指称的体词要转化为表示陈述的谓词性成分时，又加了语气词“了”。不过，如果将虚词的作用局限于组词造句，虚词固然重要，但常常又是可以省略的。古诗有“两个黄鹂鸣翠柳”，现在经常说“睡弹簧床”、“出口美国”、“吃大碗”， 可以将它们解释为分别是由“两个黄鹂在翠柳上鸣叫”、“在弹簧床上睡”、“向美国出口”、“用大碗吃”变换来的，即可以认为汉语句法结构中的介词在某些情况下是可以省略的。这与日语的句子成分必须附着后置的格助词是大相径庭的，在英语中该用介词的地方通常也是不能省略的。

现代汉语中“的”字使用频度最高，主要用作助词。前面讲了助词“的”的重要作用，但很多情况是否用“的”是相当灵活的。口语中经常用“我哥哥”、“他父亲”代替“我的哥哥”、“他的父亲”。即使像

“我的眼镜”、“她的钱包”等短语中的“的”通常认为是不能省略的，但如果把它们放在更大的语境中，“的”也可以省，例如“昨天我眼镜被打碎了。”

“了”、“着”、“过”这 3 个助词和副词“将”、时间词“将来”通常被认为是汉语时态的标志。但汉语句子的时态并不是显性表示的，这些词并不是非用不可。例如：“（我）看电视。”这句话可用于回答下面的任何一句问话。

昨天晚上你干什么了？
明天下午你干什么？
刚才你在干什么来着？
你在干什么？

也就是说，“（我）看电视。”这句话表示的时态可以是过去时，也可以是将来时，还可以是“过去进行时”或“现在进行时”。如果脱离上下文，要将“（我）看电视”，这句话译成英语，并且要显性地正确地表达英语的时态，那是很困难的。现在语法分析技术通常以句子为单位，结合上下文的语境分析技术尚待发展。通俗地讲，汉语的时态是“无”，英语的时态是“有”。“无中生有”比“从有到无”要困难得多，这是不难理解的。

“被”字句与“把”字句经常被列为汉语的典型句式，实际上，“被”字与“把”字也是可以不用的。例 12 可看作是将“苹果被我吃了”中的“被”字省略了。例 13 可看作将“我把苹果吃了”中的介词“把”省略了。更简单的，还有：

例 15　　苹果吃了。

这句话表达的既可能是陈述句“苹果被吃了”，也可能是祈使句“把苹果吃了”。

总之，原本可以作为分析的线索的虚词在汉语中常被省略无疑给语法分析增加了新的困难。另外，即使虚词未被省略，虚词与同形的实词的区分又是一个难题。如“在、给、跟”等既是介词，又是动词。本来这也属于一般的同形兼类问题，只不过在句法分析时对虚词的作用所寄托的期望较高，而使得这个问题更加突出。

1.3.5　汉语的书写习惯

由于汉语的语素绝大多数是单音节的[19]，单音节的语素又用单个的汉字书写。古汉语中多音节词很少，这就形成了汉语文献都是汉字连篇书写的传统。到了现代，汉语的书写方式有了较大的变革，如分段，加标点符号等，这为阅读提供了方便，提高了阅读效率。不过现代汉语仍是按句连写的，即一句话中的汉字是一个接一个连写的。现代汉语中已涌现了大量的多音节词（书面上即是多字词），按句连写与多音节词的使用是不协调的，影响了人们的阅读效率[27]，不过由于人的智力足以克服这个障碍，人们尚未形成进一步变革汉语书写方式的强烈愿望。当用计算机处理汉语时，这个问题就严重了。从汉语句子中辨识出词或者按词切分句子是汉语分析的必由之路。解决歧义切分问题成了汉语信息处理研究中无法回避的基础课题[28]。

英语等欧洲语言基本上以词为单位书写，词与词之间留有空格。切分问题或者不存在或者相对简单。日语同汉语一样也是按句连写的，在一句话中，作为其表记符号的汉字与假名也是一个接一个连写的。因此日语分析也是先要解决词语切分或辨识问题，不过，通常的日语文章是汉字假名混用的，汉字表记概念词，假名表记功能词或用言的词尾，而且不同的格助词各司其职，这些特点可以在日语词法分析中发挥作用。汉语全用汉字书写，缺乏类似的可供利用的表层信息，汉语分析困难，切分也困难，这是可想而知的。

在研究词语切分时，人们首先注意到歧义切分问题，并投入了大量的精力。另外未定义词问题也是需要深入研究的。对于英语，未定义词是很容易发现的，只要是词典中未登录的词（在句子中有空格与其他的词隔开）就是未定义的。对于汉语，问题就复杂得多。“面的价格”（这里的“面的”指一种曾为北京市民服务过的黄颜色的面包型出租汽车）如果能切分成“面的 价格”，而在词典中查不到“面的”，当然认为“面的”是未定义词。问题是切分程序在词典中找不到“面的”，却可以查到“面”和“的”，从而把“面的价格”切分为“面 的 价格”。这里的“面”可以理解为“米面”的“面”。因而这里识别不了未定义词，

在此基础上继续进行分析和翻译，岂不是“失之毫厘，差之千里”，要闹大笑话。

前面讨论汉语短语与句子的构造一致性时也提到了汉语合成词基本上也是由两个语素按“主谓、述宾、述补、定中、状中、联合”等方式组成的。这两个语素也可能单独成词，这就造成了短语与合成词之间界限的模糊，短语的语义通常可由构成成分的意义组合而成，合成词的词义往往不等同于构成成分的意义的组合，或者发生转义，或者有引申。因此提取与辨识汉语文本中的未定义词是重要的、必要的，但这个问题同歧义切分纠缠在一起，大大增加了汉语分析的难度。

这里只讨论了由于汉语语法特点而造成的汉语分析的特殊困难。当然，任何一种自然语言，其分析和理解都是困难的。既不宜低估也不宜夸大汉语的特殊性。作者相信只要深入细致地进行思考与探索，终究会有所领悟与突破。不过作者又清醒地认识到限于当前的主、客观条件，并不奢望汉语分析能在短时期内取得突破性进展。作者认为在以下两个方面进行开拓与积累，总是会有收获的：一个方面是语言知识库，另一个方面则是受限汉语。

关于受限汉语的研究[29~32]，作者曾发表文章讨论它的必要性与意义[31，93]。自然语言处理技术经过 50 余年的探索与实践，虽然取得了相当大的进步，但在处理大规模的真实文本或随意的话语时仍然是举步维艰。实用的自然语言处理系统对自然语言总是自觉不自觉地进行了某些限制。这使人们容易认识到受限语言的研究是有价值的。作者积十余年在汉语信息处理研究中的经验，近几年来认识到在语言信息处理技术的发展进程中，受限汉语规范的制订与应用是必要的。不过，受限汉语的研究决不是消极地回避困难，它必须建立在对自然语言的复杂性和技术现状的全面了解的基础上，它可以起到里程碑的作用。受限汉语的规范成为自然语言处理系统的目标。随着技术的发展，规范可以不断地修订，逐步接近自然语言，而每一步的目标都是明确的，是可以实现的。关于受限汉语的进一步讨论，也许会离开本书的主题，这里不再展开。

关于语言知识库，与受限汉语的情况有所不同，作者不仅在刚刚进入自然语言处理领域时就已经认识到了它的重要性，而且十余年来持之以恒地进行语言知识库的建造工作[33，34]。对于计算机处理自然语言所需要的语言知识库，作者又认为它是不应该受到限制的。只要计算机的运算速度与存储容量允许，只要建造语言知识库的人力、财力允许，知识库的规模越大越好，知识越丰富越好，结构越灵活越好，使用越方便越好。这样的语言知识库当然不是一蹴而就的，作者选择了《现代汉语语法信息词典》作为第一阶段的基础工程。任何一个自然语言处理系统都需要一部机器可读的词典，这是不言而喻的。但在语言信息处理技术发展的早期，电子词典所包含的信息是很少的，系统的能力经常取决于语法规则的数量与质量。随着技术的发展，词典在自然语言处理系统中的地位越来越重要，在整个语法理论中的地位也变得越来越重要了。为了充分描述语言现象的多样性，研究者不再走单纯增加句法规则的路子，而转向将规则归纳为少量的一般化原则，同时将词典作为语法的一个重要的有机组成部分，词典中为每个词项所附加的信息同语法规则相结合，可以实现由词项驱动规则，词典在语句分析与语句生成中就可以发挥更大的作用。反映在应用领域，机器翻译也采用了词专家系统技术[35]。电子词典已成为自然语言处理实用系统开发的基础。对于汉语来说，词典的重要性就更加突出了，汉语的形态不发达，适用于汉语自动分析的形式系统也不够成熟，这种客观现实则要求研究者从机器处理的需要出发，深入地考察汉语的语言事实，系统地总结汉语语法知识，并且以既便于语言学家表述又便于机器使用的形式把这些知识表达出来。作者及其同事们正是从这种理念出发，在朱德熙先生的“词组本位”语法体系的指导下，经过十余年的努力，研制了一部《现代汉语语法信息词典》。本书的以下各章将详细介绍这部词典的内容、规模、设计思想及其应用，书后所附 10000 词的实例将把词典的全貌展现在读者面前。