第一章　引 论

§1.1 课题的提出

本课题的研究工作是尝试在句法和语义两个层级上归纳现代汉语短语结构[1]的组合规则，解决计算机分析现代汉语短语时碰到的结构歧义问题。

有别于以往主要是面向人的语法研究，本课题的研究是面向计算机的。语法研究的应用对象由过去主要是面向人发展到现在还面向计算机，而且后一个面向显得越来越迫切和重要，这是计算机科学技术飞速发展以及信息社会对信息自动化处理的要求不断提高的必然结果。而目前的信息处理技术，越来越多地需要对自然语言进行深层分析，比如机器翻译，自动文摘等，就是如此。开发这类应用系统，要求计算机掌握尽可能多的有关自然语言的知识和非语言知识，前者又包括句法知识（含跟句法现象有关的语音韵律知识[2]）、语义知识乃至语用知识等等。下面这个简图可以用来大致说明自然语言处理的知识需求情况。

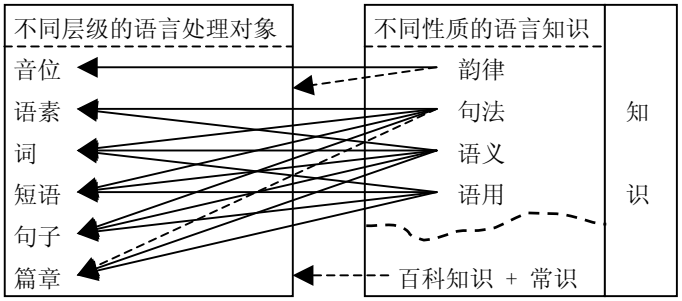


图 1-1：自然语言处理所需知识简要示意图

说明：

箭头表示“作用于”

实线箭头表示直接作用

虚线箭头表示间接作用

衡量一个自然语言处理系统的水平，可以看它处理到语言单位中的哪个层级，同时也要看它对不同性质的语言知识掌握到了一个什么样的水平。无论是比较传统的基于规则的处理策略，还是 90 年代以来方兴未艾的基于统计的方法，在对语言知识的需求这一点上实际上都是共同的。所不同者，是走规则路线的研究者一般诉诸于专家的理性知识，由人来对语言知识进行抽象（比如以带有合一条件的规则形式给出），而走统计路线的研究者一般求助于计算机对大规模语料库的统计分析，由计算机来抽象出语言知识[3]（比如以一定的数据结构记录的统计结果等）。两种路线孰优孰劣，不能笼统判断，只能跟具体的应用目标结合起来，由实践结果来评价。统计方法已经在像语音识别、自动分词和词性标注这样相对浅层的自然语言处理中有不俗表现，但在深层分析方面（比如分析句子的树结构或者句中成分间语义关系等）也还没有显出特别的优势。于是又有学者提倡把两种方法结合起来使用[4]（比如通过统计，给出带有概率值的规则）。在我们看来，无论采用哪种方法，首先都要求人自身先对自然语言有深入的了解。就规则方法来说，这一点是显然的；就统计方法来说，虽不那么明显，但道理也是一样的。现有的对自然语言深层知识的统计，一般是建立在经过标注的熟语料库基础上的。而从生语料库到熟语料库，就具体的加工方式而言，当然有人工方式，也有计算机自动加工方式或者人机互助的方式等等，但加工什么内容，标注哪些信息，仍然取决于人对自然语言的认识。

具体到中文信息处理方面。如果以处理对象的单位大小为指标，宏观地看，中文信息处理技术已经走过了字处理阶段，分词和词性标注（词处理阶段）也有了基本可以实用的成果[5]，目前可以认为是进入到句处理的前期阶段，即如何来对短语结构进行自动分析的阶段，包括划定短语边界、分析短语结构的内部句法关系、给出结构成分间的语义关系等等不同深度的分析。这样定位，并不是无视目前也有研究者在开展更大单位（譬如篇章）上的中文信

息处理研究[6]。只是就这个领域的总体发展情况来看，目前以在句子一级上开展研究工作为主。另外本着解决问题由易到难，由简单到复杂的原则，把中文信息处理目前的发展定位在重点解决短语结构的分析问题这么一个阶段，也是适宜的。而作为汉语研究者，需要考虑的就是，（1）就中文信息处理目前这样的定位而言，重点应该为计算机提供什么样的语言知识？（2）在现有的技术条件和语言学研究水平上又能够提供多少？

选择本研究课题，基本上就是带着这两个问题开始思索的结果。

对前一个问题的回答，主要是根据中文信息处理已有的研究成果和从目前的实际需求出发，初步确定了本课题研究的主要内容。从 80 年代中后期开始到现在，研究人员已经在有关汉语词语的语法功能分类和属性特征描述[7]，以及实词的语义属性描述方面开展了卓有成效的工作[8]，为计算机分析汉语的短语结构打下了一个很好的基础。而目前迫切需要进一步解决的问题，是对汉语短语的结构功能特征及组合时的条件进行全面系统地研究。在这个研究过程中得到的结果，不仅可以回过头去检验以往对词的语言知识的概括是否合适，从而进行相应的调整，而且也可以根据需要对词语的句法语义属性知识进行一定的扩充。同时，这部分知识是计算机进行短语结构分析必不可少的，只有尽可能把现代汉语短语结构的组合条件描述清楚，才能指导计算机分析出短语的正确结构，给出正确的语义解释。而从发展趋势来看，越来越多的高级自然语言处理应用系统的开发，诸如汉外、外汉机器翻译，中文信息的提取，人机对话等等，也都离不开这部分语言知识的支持。

对第二个问题的回答，则是结合我们对目前现代汉语语法理论和具体语法规律的研究水平的认识，大致确定本课题研究应该追求的合理的目标。现代汉语语法研究进入 90 年代以后发展至今，突出的特点是在以往对具体语言事实的描写基础上，大大加强了理论和方法上的探索[9]。不少研究工作尝试以新的角度来观察现代汉语的语法现象，发现以前没有注意到的语言问题，或者对原来已经提出的问题进行重新解释。在句法、语义、语用多个层面，以及从词到短语结构、到句子句式句型，乃至篇章话语等各级语言单位，形成了多层面多角度全方位的研究态势。不过，不管理论方法如何变化，目前情况下，语言研究的具体成果要能为计算机所用，都必须兼顾两点要求：一是要能够形式化，二是对处理对象要有普遍的可操作性。以这两个标准来衡量，基本就框定了在现阶段能够提供给计算机用来分析汉语短语结构的语言知识的水平。而从积极的角度说，这有助于我们从一开始就能对本课题研究工作有一个合理的目标期待。首先，对于所谓句法、语义、语用等不同层面的规则限制，目前比较适宜走以句法为主，语义为辅的路线。对语用知识，则可进行小心尝试。其次，由于所谓“大规模真实语料”的情况太复杂[10]，就目前语言学研究的水平来讲，提炼短语结构的组合规则，直接面对真实语料并不合适。而应该从相对抽象的句法结构分析入手[11]。因此，我们给出的规则也并不奢望能解决分析真实句子时碰到的由于复杂语义和篇章层面的因素造成的诸多问题。当然，中文信息处理的客观要求是必须面对实际文本，对此，目前可以考虑用适当的策略来对付真实文本中大量存在的“不听话的真实”。

在整个研究过程中，面对上述第一个问题，促使本文作者关注这项研究的实用价值，而对第二个问题的思考，则引导作者从计算机的角度来对现有的现代汉语语法理论和具体的语言研究工作进行评估，进而自觉地追求本课题研究所希望达到的，对现代汉语语法理论建设有所贡献的目标。

§1.2 面向计算机的语言学研究工作的模式

面向计算机的语言学研究工作，涉及的范围相当广泛。研究模式根据工作内容或者目标的不同也有差别。这里不作全面讨论。下面概述跟发掘语言知识相关的研究工作的一般模式，目的是为本课题的研究工作勾勒工作方式上的背景。

面向计算机开展语言研究工作的语言学家实际上可以看作是处在跟计算机以及语料形

成的一个三角关系中，见下面图 1-2。

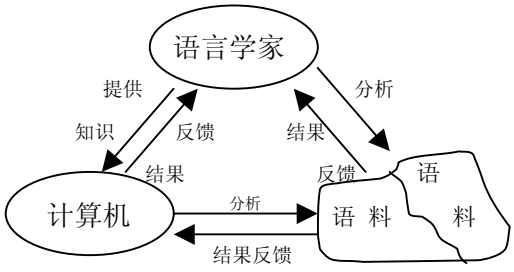


图 1-2：　语言学家、计算机、语料构成的三角关系

这是一个抽象环境尽量简化的示意。箭头表示谁“向”谁，或谁“对”谁，比如语言学家向计算机提供语言知识，计算机对语料进行分析等等。其中语料以不规则形状表示，是指其范围不确定。内部以不规则线分割，是表示语料可以按照不同标准大致区分为：包含各种复杂情况的真实语料与比较理想的“纯语料”；或者未经加工的生语料与所谓加工过的熟语料；或者合法的（或可接受的）语料与不合法的（或不可接受的）语料等等不同情况。此外，这里的“计算机”是指由硬件加软件构成的一个信息处理系统，换言之，上图已经把计算机程序设计员这个也是由人充任的角色包含在“计算机”中了。

前面提到过的两种主要的自然语言处理策略，基于规则和基于统计的方法，在上图中也都能得到反映。如果采用基于统计的方法，注重点就在上图等腰三角的底边上，即由计算机对语料（一般得是熟语料）进行统计得到语言知识（一般表现为参数），再利用得到的参数对语料进行分析，根据分析得到的反馈结果来调整已有的参数，从而提高分析能力。比如基于错误驱动的语法规则自动学习方法的研究[12]，计算机可以用错误率为指标评价当前习得的语言知识的适用水平，进而作相应的调整。也就是说，语言学家可能提供一些初始的语言知识（譬如将生语料加工成熟语料的工作），而大部分归纳和调整语言知识的工作是由计算机来完成的。如果走基于规则的研究路线，注重的就是上图等腰三角的两个腰，即由语言学家提出一套语言知识（比如可以从对有限语料的分析中初步归纳出来）给计算机用，再根据计算机反馈的结果来改进原来的语言知识，形成一个循环处理系统。显然，在这个系统中，给出和调整语言知识这两步工作都主要是由语言学家完成的。

本课题的研究工作走的是基于规则方法的路线，要求语言学家对语言知识要有全面系统的清晰认识。对此可以从语言知识的性质和形式两方面来看。从性质上讲，语言知识可以分为基于范畴（Category）的“属性：值”型知识（ATTRIBUTE:value）和基于规则（Rule）的“条件-->动作”型知识（CONDITION-->action）[13]。范畴用来刻画语言对象的一个或一组特征。规则用来表述范畴间的关系。“特征”的数量是不确定的。一个范畴可能刻画几个特征，一个特征也可能有几个范畴都能刻画它。举例来说，“名词”是一个范畴，它可以刻画一个具体的名词在几个方面的句法特征，如能受数量词修饰，能充当主宾语等等。逻辑上，所有规则都可以表示为 P->Q 这样的蕴涵式。两个命题 P 和 Q 则分别建立在已知范畴的基础上，规则因而实际上表述了命题所涉及的范畴之间的关系。比如，可以有这样的规则，如果 W 是名词（P），那么 W 能作主语（Q）。显然，这条规则在“名词”跟“主语”两个范畴间建立起了一种联系，尽管这条规则所描述的联系是粗糙的，甚至不那么正确，但是，以这样的方式建立范畴之间的联系，是分析语言的结构时必不可少的。而语言学家所要做的，正是去寻找正确的和好的联系。从形式方面看，语言学家要考虑的就是以何种形式化的方式[14]把范畴知识和规则知识组织起来，使得更有利于计算机处理。而所谓语言知识的形式化，就是以一套严格定义的符号系统来精确地表达语言知识，包括范畴的符号化和规则的公式化。这里不展开讨论有关各种形式化方法的细节。因为就勾勒研究工作的一般模式这个目的而言，在直观的层面上理解形式化就基本满足要求了，即范畴知识一般用词库（机器可读词典 MRD）来负载，规则知识则由所谓规则库（规则的集合）来承担。

有了上述对语言知识的认识，语言学家的两步工作，第一步实际上就可以更具体化为（1）确立一定的范畴体系并基于这一范畴体系对具体的语言成分进行属性赋值操作（即建立词典）；（2）给出能正确地描述范畴之间关系的规则（即建立规则库）。相应地，第二步工作也就是根据计算机的分析结果是否跟预期相符，来调整原有的范畴体系和具体语言成分的属性取值，以及相关的规则（即改进词典和规则库的内容）。而实践中，通常在开始这两步工作前，应该先规划一套初步的语言知识形式化表达体系。这样才能在一个严密的表达系统内具体展开上面两步工作。

宏观地看，投身于这一领域的语言学家的主要研究工作基本上就可以用上述模式来加以概括说明。不同研究者的工作差别也主要表现为对上述模式中各环节的侧重不同，即有的研究者可能相对更关注范畴的建立；有的研究者则选择在具体落实语言成分的属性值上花功夫；当然也有学者对发现范畴之间的联系（即寻找语言成分之间组合的制约规则）更感兴趣；还有的学者则乐于从事形式化表达体系的研究，等等。

§1.3 开展本课题研究工作的基础

开始研究工作之前，无疑应该先看看，对于解决本课题所关注的问题，前人在建立范畴（包括具体落实到语言成分的属性值上）和发现规则（特别是跟排除一些典型歧义格式有密切关系的规则）方面，以及在形式化表达体系方面，都做了哪些研究。同时，由于我们的研究主要靠计算机分析一定的语料，再根据反馈的结果来驱动作进一步的改进，因此，对开展本课题研究所依托的计算机系统环境以及选用的语料作些说明，也是十分必要的。

1.3.1 前人的研究

如果以“建立范畴，归纳规则，加以形式化表示”这样的模式来审视本课题研究相关领域现有的研究格局，大致可以概括为：范畴和规则体系已经有了一定基础。相比之下，范畴知识的具体成果相对较为丰富，规则知识的具体成果则显得不足。对于以深加工为应用目标的中文信息处理系统的要求来说，这两部分紧密相连互为依存的知识，总体上还都不能做到很令人满意的程度。形式化表达方法自乔姆斯基 50 年代创立形式语法理论以来，目前已经发展出诸多体系[15]，如 CFG（上下文无关文法）、GB（支配约束理论）、LFG（词汇功能语法）、GPSG（广义短语结构语法）、FUG（功能合一语法）、HPSG（中心词驱动的短语结构语法），以及其他从非乔姆斯基路线发展起来的形式语法描述体系，如 CG（范畴语法）、LG（链语法）、TAG（树连接语法）等等，虽然这些理论在具体表达方式以及各自确立的原则上存在差异，各有侧重，但由于这方面工作带有很强的技术性，因而构造具体的语言知识形式化表达系统，一般倾向于根据实际需要，以某种形式表达体系为主，同时也吸收其他方法的合理的表达机制。比如国内学者在中文信息处理实践过程当中提出的 MMT 模型[16]，基于依存语法的汉语分析模式[17]，基于范畴语法的汉语组合类型语法模式[18]，以及汉语完全句法树模型[19]等等。值得指出的是，不同的形式化方法在表达上的差异对一个自然语言处理系统的性能固然会有影响，但处理结果的好坏，恐怕关键还是取决于语言知识本身的多少，即发现了多少有用的范畴，给出了多少足够排除歧义的句法结构规则。

国内现代汉语语法研究在 80 年代中期以前的发展水平，主要以建立在结构主义语法理论和方法基础上的研究成果为代表。这其中又以朱德熙先生提出的词组本位语法体系[20]最为系统和精炼。该体系比较全面地建立起了汉语短语结构层面的句法范畴，同时也引入了一定的语义范畴。并从对汉语具体语法现象的研究中，形成了应在不同层级上确立语法理论范畴的观念，比如结构形式层面的句法范畴，意义层面的语义范畴，表达层面的语用范畴等等。

90 年代以来在国内汉语研究界得到广泛讨论的所谓“三个平面”的语法观[21]，实际上在朱德熙先生 80 年代初的语法论著中也已经有了明确的思想萌芽。汉语研究界自 80 年代后期直至进入 90 年代以来，一方面仍有许多学者在原来结构主义的研究路子上继续探索，另一方面也有越来越多的学者开始采用格语法、配价语法、语义指向分析、依存语法、GB 理论、话语分析（DA）等新方法，进行汉语研究新的尝试。结果是确立了不少新的对分析汉语有用的句法语义语用范畴[22]。譬如“自主动词和非自主动词[23]、一价名词和二价名词[24]、有标记成分和无标记成分[25]、主题延续性和行为延续性[26]”等等。学者们利用这些范畴，对汉语的很多语言现象做了更深入的分析。与此同时，结合中文信息处理应用系统开发和理论探索的需要，学术界也开始规划语言知识基础工程的研究和建设。这方面的工作从 80 年代中后期开始以来，发展至今，突出地表现在对汉语句法范畴和语义范畴的符号化及大规模的词典化上。具体成果以《现代汉语语法信息词典》[27]、《信息处理用汉语语义词典》、《现代汉语述语动词机器词典》[28]等机器可读词典为代表。其中《现代汉语语法信息词典》直接建立在词组本位语法体系的框架上，实现了对 5 万多汉语通用词语的句法属性特征的全面系统地描述，是第一部大规模的以形式化方式表述汉语语法研究成果的计算机可用词典。另外两部语义词典，则是学者在语义场、格语法等语义学理论的框架下对汉语的实词（特别是动词）进行具体的语义属性刻画的结果。这些词典现在都已经在一些跟中文信息处理相关的应用系统中发挥实际的效用。

跟上述范畴知识的研究工作相比，规则知识的研究相对薄弱一些。在已有的范畴研究基础上对规则知识作大规模系统地研究还比较少见。不过，这方面的探索工作实际上一直都有学者在尝试。国内直接面向中文信息处理的研究工作中，理论探讨方面如马希文（1989）[29]和冯志伟（1993）[30]等，具体规则的研究方面如范继淹（1979）[31]、曹敏（1990）[32]、马真、陆俭明（1997）[33]、孙宏林（1997）[34]和詹卫东（1997b）[35]等。此外相当多的汉语研究工作虽然不是直接面向中文信息处理，但其中有不少研究实际上可以看作是跟发现汉语的语法规则紧密相关的，只不过这些研究通常都带有比较强的描写色彩，而不是显性地以我们上文所说的规则形式呈现出来。比如李子云（1991）[36]和吴竞存、梁伯枢（1992）[37]的研究。另外还有近年来出现的一大批以自觉的理论立场为背景，形成鲜明学术风格的研究工作，如沈阳（1994）[38]基于配价理论和生成语法的空范畴理论对汉语动词句位结构以及 NP 所做的全面分析，袁毓林（1994，1995）[39]、郭锐（1995）[40]、张国宪（1995）[41]等人基于配价理论对汉语名词、动词短语、形容词等展开的配价描写，和张伯江、方梅(1996)[42]等在认知功能语法背景下对汉语主位结构和焦点结构所做的研究等，这些都可以看作是在不同层面上揭示汉语的语法组合规则。虽然这些研究工作不以中文信息处理的应用为直接目标，在规则的发现和表达方面还是以往面向人的做法，但对以中文信息处理为直接目标的规则研究，无疑也能提供有力的支持。

海外的自然语言处理研究，近年来也是范畴知识的研究进展比较显著。其中语义方面尤为突出。这既包括在理论上探讨如何构建语义知识的表示体系，如 Fillmaore（1982）[43]关于“框架语义学”的研究，也包括大规模的语义知识工程实践。如 WordNet[44]、FrameNet[45]、MindNet[46]等计算机用语义词典的开发，Chodorow 等（1985）[47]和 Ide 等（1993）[48]所做的有关从机器词典中自动抽取语义知识的研究工作，以及黄居仁等（1998）[49]从汉语名量搭配词典中自动抽取汉语名词语义分类的研究。而在结构规则的研究方面，90 年代以来国际计算语言学界统计方法风头日盛，纯粹基于规则方法进行自然语言处理研究的工作相对较少。但在面向实用开展的自然语言处理系统的研究中，用复杂特征集以及合一运算来组织语言规则知识，广泛地吸收各种形式化表示方法的特长，体现实用特色，也不乏代表性的工作，如 Karen Jenson，George E. Heidorn，Stephen D.Richardson 等人的研究[50]。语言学界的情况则是，一方面以乔姆斯基为代表的形式主义学派不断向抽象度更高的理论目标进发，关注

所谓普遍语法的原则问题。与此同时，“词汇主义”的倾向开始得到普遍认同[51]；另一方面，非乔姆斯基阵营的学派，如认知语法，功能语法等迅速发展壮大，这一阵营里学者们的研究兴趣则基本集中在从人的认知心理以及对世界的感知方式等语言外部因素出发来追求对语言现象做出解释[52]。因而总体来说，直接以在句法语义层面发现结构组合规则为目标的研究不多。当然，跟上面对国内语法研究所做概述的情况类似，在不同背景下展开的语法研究，都可能对以信息处理为目标的规则研究提供支持。

以上极为概括地描述了前人的研究工作。主要目的是廓清目前在这方面研究的总体格局，特别是对海外学者研究情况的了解，有助于为进行本课题的具体研究提供认识上更广阔的参照系。有关的具体研究工作，在下文讨论具体问题时还会提及。

1.3.2 我们的立场和主张

结合以上对相关研究的背景简评，在具体论述本课题研究工作之前，还有必要表明我们的理论立场和对研究工作的一些原则性主张。

- ◆ 就汉语短语结构的句法描写而言，词组本位语法体系所建立的理论框架和在这个框架下开展具体研究积累起来的成果，无疑可以看作是目前的“巨人之肩”。我们将以此为起点开始本课题的研究工作。
- ◆ 对于语义知识的组织方式，格语法、配价理论等尽管在理论背景和具体操作上差别显著，但在出发点和目标上其实有很强的一致性。我们将在吸收有关语义理论合理的精神内核基础上，根据实际需要加以拓展。
- ◆ 从生成的角度看，语言中一个具体的表达形式（譬如实际使用中的一个句子），是各个不同层面的语言知识（规则）乃至非语言知识叠加作用的结果[53]。从理解的角度看，对用来分析一个表达形式的语言知识加以归纳，就要特别强调知识的层次性。当然，也应该清醒地认识到，在必须用还原主义的两条腿走路的同时，最好还要有一个整体主义的大脑。即分清层次，同时也不要忘记系统。这实际上也是一张纸的两面，本来就是一体的。
- ◆ 追求理论系统的严密并不就一定意味着要以丧失灵活性和牺牲真实性为代价。我们不主张以所谓“句法自主”来保证句法系统的纯洁性，但把结构形式和语义、语用层面的语言现象笼而统之地一块儿处理，给出一些模糊的解释，也不是科学研究的态度。合理的假设应该是，各子系统相对独立，同时又以一套确定的机制相互影响。
- ◆ 在本体论的意义上探讨到底哪个子系统起决定作用，哪个子系统是受别的子系统支配制约的，当然有不朽的价值。但在目前，我们更愿意在方法论的层次上处理各个子系统的关系，尽量淡化各子系统之间地位上的主次之分。如果一定要有所偏重，那么，本着易于形式化，概念易于把握的子系统优先的原则，我们主张以句法层面的知识为主，语义层面的知识为辅，来组织短语结构的规则。
- ◆ 任何“好的语言理论”都有其适用范围。这个适用范围可以指归纳的语言知识所在的不同层次（或语言系统的不同子系统），也可以指面对的对象是语言中不同的层级单位（比如词、短语结构、句子、篇章等）。我们以为，对超出其适用范围的“过分”的要求，一个理论完全可以说“不”。
- ◆ 正如对人进行语言能力考核要分级别一样，对机器理解语言的能力做评判，也要有正确的层级观念。不同的自然语言处理应用系统，理解语言到多“深”、多“广”，根据需要不同可以制订不同级别的目标。虽然通常呈现出来的结果是一个整体面貌，但科学的评价仍然应对系统在不同层次上的表现进行分项打分。这种观念的一个自然同时也是合理的推论就是，不要奢求一个只具有“短语结构”分析能力的系统，能把“句子”分析得有多么好。

◆ 本课题研究工作的直接对象是“短语结构”，而不是“句子”。但在研究过程中，由于我们强调研究是面向实用的，因此或者从策略上，或者在理论探索上，我们都将尽量追求对短语结构规则的研究能有效地向分析句子的层次靠拢。

1.3.3 开展本课题研究的计算机环境和语料

本课题研究工作基本是在一个汉英机器翻译系统的开发调试环境下进行的。这个汉英机器翻译系统的语言内部知识表示包括线图（Chart）、树结构（tree）和特征网络（feature network）三种形式。汉语的结构分析采用的是改进的线图分析（Chart Parsing）算法。系统的目标是能处理开放的汉语语料，处理的语言单位定在句子一级。目前这个系统语言知识库的规模是，汉英词典收汉语词近 5 万条。规则库通用规则 300 条左右[54]。已经调试过 4 千个多汉语句子。详细情况可参见刘群等（1997）[55]。

这个机译系统主要的处理环节可以概括为以下四步；

（1）对输入的汉语句子进行分词和词性标注；

（2）进行句法分析产生汉语的句法结构树（带有复杂属性特征描述，其中包含句法结构信息和语义搭配信息）；

（3）将汉语句法树转换成英语句法树；

（4）经过必要的结构变换和词形变换，由英语句法树生成英语句子。

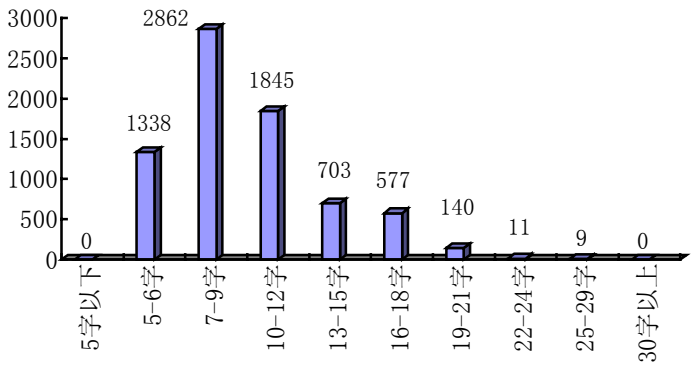
我们的研究工作假定上面第一个环节的分析结果都是正确的[56]。在这个前提下把目光对准第二个环节的分析任务，并且是集中在对短语句法结构的分析上，而不是将整个“句子”尽收眼底，尽管在实际应用中，我们必须面对整个句子而不仅仅是“短语结构”。上面第三和第四个环节的工作，不在本文研究范围内。

本课题研究中用来调试规则所选取的语料，主要来自北大计算语言学研究所主持汉英机器翻译评测工作所用的题库，并根据需要作了适当的调整。按句中标点符号的分布来划分，语料的总体情况如下表所示[57]（语料样例可参见附录四）：

序号	特征	数量		序号	特征	数量
1	不含标点的（词组）	61		7	出现省略号的句子	0
2	出现引号的句子	52		8	出现感叹号（感叹|祈使句）	132
3	出现书名号的句子	6		9	出现问号的句子（疑问句）	591
4	出现括号的句子	5		10	出现逗号的句子（复句）	1764
5	出现破折号的句子	14		11	不出现逗号的句子（单句）	7485
6	出现冒号的句子	3		共 10113 句	总字数:107463	用字总数: 2610

（表 1-1： 测试语料句子类型 ）

其中不出现逗号的句子，即是一般所说的单句[58]，占全部语料的 74%。单句中按字数不同来统计句子数的结果可表示为下面的直方图（图中柱体上方数字表示句子数）：



（图 1-3：调试语料中“单句”按字数不同进行统计的分布图）

§1.4 本文的结构安排

本文主要分为两大部分来展开论述。

（1）现代汉语短语句法语义范畴的确立及主要短语结构的分析规则（第 2、3 章）

（2）现代汉语短语结构歧义格式分析及排歧研究（第 4、5 章）

本文第 2 章在已有的关于词的句法功能分类及属性描述的理论研究和工程实践基础上，向短语的句法功能分类和属性描述进行了拓展。并在已有的配价理论、格理论等语义理论基础上，提出了表示汉语实词和短语语义信息的“广义配价模式”，确立了描述短语句法功能特征和语义信息的基本范畴。第 3 章则以此为基础，尝试尽可能全面而细致地给出现代汉语短语结构的组合规则。

现代汉语短语结构歧义是进行汉语句法分析的一大主要障碍。本文第 4 章对此进行了全面调查，对短语歧义格式的不同类型做了区分，并对各种类型的短语歧义格式的分布情况进行了统计。针对一些典型的歧义格式，第 5 章在排歧策略方面进行了探讨。

在上面两部分主要内容之外，本文还单辟一章（第 6 章）对本课题研究工作的实验结果加以了说明。最后，在结束语（第 7 章）中对本课题研究工作进行了总结，简要概括了本课题研究取得的主要成绩，讨论了本课题研究工作对中文信息处理研究的意义，并提出了进一步研究的计划和目标。

附 注：

[1] 本文的研究对象是现代汉语的短语结构，但限于目前的研究水平，对什么是短语这个问题还难以用概括的表述来准确的定义。众所周知，汉语研究中有关什么是词、什么是句子之类的基础性问题目前都还没有得到圆满的解决。当然，问题的困难性并不是我们不给出短语的定义的借口。之所以没有在正文中给短语下一个完整定义，主要是因为就本文的研究目的而言，有一个严格的定义固然好，没有定义也不怎么妨碍开展具体的研究工作。这里我们可以给出关于短语的一些宏观说法，期望有助于读者对短语概念的理解。(1) 从形式语法的角度看，短语是产生式规则生成的所有终结符（参见§3.1）；(2) 从本文主张的语言理论立场来看，短语是个纯粹抽象的句法功能单位概念，它可以通过我们假设的句法结构概念来定义，即可以出现在任意一个句法结构（本文所说的句法结构，仅指下文中确定的“主谓、述宾、述补、……”等 9 种结构范畴，参见§2.1.2）中的结构位置上的语言成分。句子是不能出现在这些结构位置上的。不难看出，这只是把短语的定义问题转化为句法结构的定义问题了，而结构的定义问题我们则处理为先验基础概念不加定义（参见§2.1.1）。

[2] 参见吴为善（1994）《句法结构和音节组合》，载 邵敬敏 主编《九十年代的语法思考》，P236-247，北京语言学院出版社 1994 年版；《刘丹青（1994）《汉语形态的节律制约——汉语语法的“语音平面”丛论之一》，载邵敬敏主编《语言研究与语言应用》，P144-155，北京语言学院出版社 1994 年版。

[3] 参见周强（1995）《基于语料库和面向统计学的自然语言处理技术介绍》，载《计算机科学》1995 年第 4 期。Su,Keh-Yih, Tung-hui Chiang and Jing-Shin Chang. 1996. An Overview of corpus-based statistics-oriented (CBSO) techniques for natural language processing, Computational Linguistics and Chinese Language Processing, vol.1, no.1, August 1996, 101-157. Taiwan.

[4] 参见 Klavans,Judith L. and Philip Resnik, eds. 1996. The Balancing Act : Combining symbolic and statistical approaches to language, the MIT Press；周明 等（1994）《统计与规则并举的汉语句法分析模型》，载《计算机研究与发展》1994 年第 2 期。

[5] 参见冯志伟（1997）《自然语言的计算机处理》，上海外语教育出版社 1997 年版；冯志伟（1992）《中文信息处理与汉语研究》，商务印书馆 1992 年版。

[6] 参见吴立德等（1997）《大规模中文文本处理》，复旦大学出版社 1997 年版。

[7] 参见俞士汶 等（1998）《现代汉语语法信息词典详解》，清华大学出版社 1998 年版。

[8] 参见陈力为、袁琦（1995）主编 《中文信息处理应用平台工程》有关文章介绍，电子工业出版社。以及詹卫东、常宝宝、俞士汶（1998）《基于词组本位语法的语义模型》，载新加坡《中文与东方语言信息处理学会学报》(Communications of COLIPS) Vol.8,No.1,1998。

[9] 参见陆俭明、郭锐（1998）《汉语语法研究所面临的挑战》，载《世界汉语教学》1998 年第 4 期。

[10] 这实际上就是促使我们思考现代汉语研究的对象问题。可参见朱德熙（1988）《现代汉语语法研究的对象是什么》，载《语法研究和探索》（四），北京大学出版社 1988 年版。
[11] 马希文（1989）《以计算语言学为背景看语法问题》（载《国外语言学》1989 年第 3 期）中谈到，“在我们研究语言时，为了使表层的材料整齐划一，也有必要设想这些材料的深层结构。设想多深都没有关系，只要能保证生成的表层结构是正确的就行。”我们的想法跟马先生这个思想是一致的。
[12] E.Brill (1995) Transformation-based error-driven learning and natural language processing : a case study in part-of-speech tagging, Computational Linguistics, Vol.21, Number 4,1995.
[13] 参见白硕（1995）《语言学知识的计算机辅助发现》，P5-7，科学出版社；冯志伟（1992），P116-123。
[14] 参见徐烈炯（1988）《生成语法理论》，上海外语教育出版社。P74-76。
[15] 参见翁富良、王野翊（1998）《计算语言学导论》，中国社会科学出版社 1998 年版。
[16] 参见冯志伟（1992）《中文信息处理与汉语研究》，商务印书馆 1992 年版，P71。
[17] 参见周明、黄昌宁（1994）《面向语料库标注的汉语依存体系的探讨》，载《中文信息学报》1994 年第 3 期。
[18] 参见翟成祥 等（1991）《汉语组合类型语法》，载《中文信息学报》1991 年第 3 期。
[19] 参见吴蔚天、罗建林（1994）《汉语计算语言学 —汉语形式语法和形式分析》，电子工业出版社 1994 年版。
[20] 参见朱德熙（1982）《语法讲义》，商务印书馆 1982 年版；朱德熙（1985）《语法答问》，商务印书馆年版。
[21] 参见范晓（1996）《三个平面的语法观》，北京语言学院出版社 1996 年版。
[22] 参见鲁川（1988）《汉语句子的语义成分与语用成分》，载《语法研究和探索》（四），北京大学出版社 1988 年版。
[23] 参见马庆株（1988）《自主动词和非自主动词》，载《中国语言学报》第 3 期，商务印书馆。
[24] 袁毓林（1994）《一价名词的认知研究》，载《中国语文》1994 年第 4 期。
袁毓林（1995）《现代汉语二价名词研究》，载沈阳，郑定欧主编《现代汉语配价语法研究》，北京大学出版社 1995 年版。
[25] 张国宪（1998）《语言单位的有标记与无标记现象》，载邵敬敏主编《句法结构中的语义研究》，北京语言文化大学出版社 1998 年版。
[26] 参见徐赳赳（1990）《叙述文中“他”的话语分析》，载《中国语文》1990 年第 5 期。
[27] 同注 6。
[28] 参见陈小荷（1998）《一个面向工程的语义分类体系》，载《语言文字应用》，1998 年第 2 期；林杏光等主编（1994）《现代汉语述语动词机器词典》，北京语言学院出版社 1994 年版。
[29] 参见上面注 10。
[30] 冯志伟（1992）《计算语言学对理论语言学的挑战》，载《语言文字应用》1992 年第 1 期。
[31] 范继淹（1979）《“的”字短语代替名词的语义规则》，载《中国语文通讯》1979 年第 3 期。
[32] 曹敏（1990）《计算机自动分析量词短语的方法及规则》，载《中文信息学报》1990 年第 1 期。
[33] 马真、陆俭明（1996）《“名词”+“动词”词语串浅析》，载《中国语文》1996 年第 3 期。
[34] 孙宏林（1997）《从标注语料库中归纳语法规则：“V+N”序列实验分析》，载《语言工程》，清华大学出版社 1997 年版。（全国第四届计算语言学联合学术会议论文集）
[35] 詹卫东（1997b）《PP<被>+VP1+VP2 格式歧义的自动消解》，载《中国语文》1997 年第 6 期。
[36] 李子云（1991）《汉语句法规则》，安徽教育出版社 1991 年版。
[37] 吴竞存、梁伯枢（1992）《现代汉语句法结构与分析》，语文出版社 1992 年版。
[38] 沈阳（1994）《现代汉语空语类研究》，山东教育出版社 1994 年版。
[39] 同注 18。
[40] 郭锐（1995）《述结式的配价结构与成分的整合》，载沈阳、郑定欧主编《现代汉语配价语法研究》，北京大学出版社。
[41] 张国宪（1995）《论双价形容词》，出处同上。
[42] 张伯江、方梅（1996）《汉语功能语法研究》，江西教育出版社 1996 年版。
[43] Fillmore,C.J. 1982. Frame semantics, In Linguistics in the morning calm, The Linguistic Society of Korea ed. Hanshin Publishing Co. Seoul, 111-137.
[44] Miller,G., et al. 1990. Introduction to WordNet: an on-line lexical database. In International Joural of Lexicography 3, No.4, 235-244.
[45] Bake, C.F., C.J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In Proceedings of COLING'98 , 86-90.
[46] Richardson,S.D.,William B.Dolan, and Lucy Vanderwende. 1998. MindNet: acquring and structuring semantic information from text. In Proceedings of COLING'98, 1098-1102.
[47] Chodorow, M., R.Byrd, and G.Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. In Proceedings of the 23rd Annual Meeting of the ACL, 299-304.
[48] Ide, N., and J.Veronis. 1993. Extracting knowledge bases from machine-readable dictionaries: Have we wasted our time? In proceedings of KB&KS'93 (Tokyo), 257-266.

[49] Huang, Chu-Ren, Chen Keh-Jiann, and Gao Zhao-Ming. 1998. Noun class extraction from a corpus-based collocation dictionary: an integration of computational and qualitative approaches, Quantitative and Computational Studies on the Chinese Language, Benjamin K. T'sou et al. eds., 339-352.

[50] Karen Jenson, George E.Heidorn, and Stephen D.Richardson, eds. 1993. Natural language processing: the PLNLP approach, Kluwer Academic Publishers

[51] 黄昌宁、陆镜光（1998）《现代语言学给我们的启迪》，载《世界汉语教学》1998 年第 4 期；黄昌宁（1993）《关于处理大规模真实文本的谈话》，载《语言文字应用》1993 年第 2 期。

[52] 张敏（1998）《认知语言学与汉语名词短语》，中国社会科学出版社 1998 年版；戴浩一、薛凤生（1994）《功能主义与汉语语法》，北京语言学院出版社 1994 年版。

[53] 对此不妨举一个简单的例子略作说明。比如语言成分间的搭配关系。动词“吃”的后面可以加名词构成述宾结构。这是一条句法规则，极为概括，覆盖面也最广，但同时搭配限制也最弱。也就是说，概括性强，但精确度差。如果引入语义搭配规则，就可以使概括度降低，但精确度提高。比如可以限制动词“吃”后面带的名词宾语，语义上应具有[+可食性]这样的特征。这样，就可以把“桌椅门窗……”等等相当多的名词排除出能跟“吃”搭配的名词之外。在这个限制层面上，还可以将[+可食性]进一步细分为[+液体]和[+固体]两种情况，因为普通话中，只有[+固体]的食物才用动词“吃”，而[+液体]食物，用动词“喝”。此外，如果考虑“吃”的前面名词，即“施事”（Agent）搭配成分，也会有语义要求，并不是所有的名词都能发出“吃”这个动作，即能成为“吃”的施事的名词，语义上应该具有[+有生命]这样的特征。甚至这样的限制，在有些语言中仍嫌概括度过高，还要进一步限制，以提高精确度。比如在德语中，对应汉语“吃”的动词是“essen”，但不象汉语中“吃”的施事可以是“狗、猫、猪”等等动物，德语中“essen”的施事只能是“人”。参见韩万衡（1997）《德国配价论主要学派在基本问题上的观点和分歧》（载《国外语言学》1997 年第 3 期）。搭配限制到语义层面还没有完，还可以继续到语用层面进行限制。比如在回民社区中，“吃”的对象跟汉族人就不完全一样，因为回民有宗教禁忌，不能“吃猪肉”。这就是由于社会文化传统等方面的语用因素造成的搭配限制。从以上对语言成分间搭配限制的简单分析中不难看出，一个实际的语言表达式，正是这些不同层面的因素共同起作用造成的结果。当我们反过来对一个已有的语言表达式进行剖析时，同样应该持这种层次观点，对复杂的语言现象，力求在多个层面上进行分析。极端而言，关于语言成分间搭配性质的分析，几乎可以认为是语言分析的全部内容。而正是在研究搭配的过程中，我们注意到，语言成分之间搭配的合法与否，不是由一个或几个简单因素决定的，而是由诸多的在不同层面上具有不同性质的因素决定的。

[54] 词典中对每个词（包括语素和成语、习用语等等语言成分），都描述了比较丰富的语法功能信息，这部分信息主要来自北京大学计算语言学研究所开发的《现代汉语语法信息词典》；此外，对实词还描述了一定的语义信息，参见詹卫东（1998），出处见上面注 7。有关词语语法语义属性信息的说明，在下文论述具体规则时将会提及。机器翻译系统中的现有规则是在调试了 4000 多例句基础上形成的，参见詹卫东（1997a）《现代汉语本位语法体系在机器翻译中的应用及其问题》载吴泉源、钱跃良主编《智能计算机接口与应用进展》，电子工业出版社 1997 年版（第三届中国计算机智能接口与智能应用学术会议论文集）。需要特别说明的是，机译系统中目前的规则不是专门针对汉语短语结构总结出来的，而是根据调试实际语料的情况归纳的。本文的研究工作则是在已有规则的基础上，集中讨论现代汉语短语结构的规则，希望能够融入更多的语言学研究成果，把有关短语结构规则的语言知识挖掘得更深入一些。

[55] 刘群等（1997）《一个汉英机器翻译系统的计算模型与语言模型》，出处同上。也可参见詹卫东（1997a）。

[56] 事实上这仅仅是为了使我们的研究工作可以集中在短语结构层面所做的一个假设。尽管目前见诸报道的汉语分词以及词性标注软件的正确率大都声称在 95%以上，但实际分词中存在的问题仍然很多。汉语文本中的人名、地名、机构名称、缩略语、未登录词辨识等，目前都还有不少问题尚未解决。对此，本文一概都不加以讨论。

[57] 统计是按照确定的顺序进行的，表中序号大的情况不包含序号小的情况。另外需要说明的是，这些语料只是供测试用，并不是本文研究所要分析的全部对象。其中包含冒号、引号（即含引语）、省略号、分号、书名号、括号等等的句子，以及含“吧、呢、啊、吗、呀……”等等语气词，含“哦、哎、……”等等叹词，含“喀嚓、扑通、……”等等象声词的句子，目前我们的分析系统都还没有全面考虑，也都不在本文的研究范围之内。

[58] 这只是简单的说明，不是关于“单句”的定义。同样地，所谓“出现逗号的句子”，也不是在语言学意义上对“复句”的定义。但值得庆幸的是，这样“简单蛮横”的说明（或者看法）基本不妨碍我们下面进行具体的研究工作。本文研究工作中调试的对象主要就在这七千多“单句”范围内。对其他句子的分析虽然也根据需要有所兼顾，但目前相对比较少。需要说明的是，计算机分析系统对一个具体句子的处理，最终效果的好坏，牵涉到程序环境，语法规则，词典信息等诸多因素，因此调试工作是相当花时间和精力的。作为学位论文的研究，而不是应用工程项目，我们更关心的是短语结果规则整体框架方面的问题，因而对调试句子的具体数目没有严格要求。