80年代以来汉语信息处理研究述评[①]

——作为现代汉语语法研究的应用背景之一

詹 卫 东

提 要 本文对80年代以来汉语信息处理领域的研究工作进行了简要的总体概括。目的是探讨如何面向信息处理开展现代汉语的语法研究。通过评述80年代以来信息处理领域主要的三大块研究工作，包括对国外理论方法的引介及结合汉语处理进行的宏观层面的一些思考、相关应用系统的研制开发、汉语知识库的建设和语法规则的发掘等，我们得到的一个初步认识是，面向信息处理的现代汉语语法研究，应在背景清晰定位明确的前提下，大力加强对用于计算机的汉语短语结构规则的研究。

关键词：信息处理 语法研究 语言知识 范畴 规则

一

面向信息处理的语言研究带有学科交叉的边缘性质，在拓宽研究领域、扩大研究视野的同时，新开辟的研究空间也难免随带着一定程度上的背景模糊和层次不清。本文希望通过对80 年代以来中国大陆计算语言学界、汉语研究界分别从不同角度展开的有关汉语的信息处理研究做一番梳理工作，从而能对汇聚于“汉语信息处理”这一面大旗下的各路研究队伍的研究面貌及研究成果有一个清醒的全局认识，并在此基础上，探讨服务于信息处理的现代汉语语法研究的发展方向。

类似于通常的大多数总结性文章，本文的目的也在于能对将来进一步的研究工作，在内容选题和方法上提供一定参考。特别是对语言学专业背景的研究人员，在投身这一研究领域时的研究取向及策略能有所启示。就这样的目的而言，选择80年代以来这个时间段作为我们所讨论问题的一个外在坐标，仅仅起到让我们的讨论相对集中一些的作用。限于资料和篇幅，本文基本没有涉及海外学者的相关研究工作。副标题中不说是“现代汉语研究”，而是谈“现代汉语语法研究”，主要是因为本文不谈语音处理方面的研究工作。另外，本文所指的语法研究，从宽理解也包括语义内容。

下面第二节我们简要说明目前语言信息处理研究的宏观模式及格局。第三节我们将 80年代以来国内在汉语信息处理领域展开的研究工作大致划分为三部分，并分别对这三部分内容展开述评。第四节在述评的基础上澄清我们对一些论争和理论问题的思考，进一步明确汉语语法研究人员以信息处理为背景开展相关研究时，应关注什么样的问题，遵循何种原则和标准，并说明我们对面向信息处理的现代汉语语法研究的发展方向的看法。

二

自然语言的信息处理是跟计算机的诞生几乎一同开始的一个多学科交叉研究领域。来自计算机科学、语言学、数学等不同学科的研究人员构成了目前这一领域的主要研究力量。随着计算机应用的日益普及，其功能也从主要是数值计算发展到以非数值信息处理为主。不管是数值还是非数值信息，计算机处理信息的一般模式都可以归结为以下三部分来看。

（1）处理对象（输入）：有限种符号的有限长序列($M=a_1\ a_2\ \ldots\ a_n$)

（2）处理过程（运算）：用事先编制好的程序对其进行有穷次的变换

（3）处理结果（输出）：产生新的符号表达式 (M')

把自然语言作为输入在计算机中进行处理时，上面模式中的（2）在实现策略上可以有不同的选择。譬如早期人机对话系统采用的简单模式匹配方法；后来发展起来至今仍在广泛使用的基于规则的处理方法；以及近年来日趋流行的语料库统计方法等等。大而言之，规则

[①] 本文在跟中科院计算所二室刘群副研究员的多次讨论中获益良多，在此深表感谢。

方法和统计方法的并存，互相之间既补充又竞争的关系，形成了当前自然语言处理领域理论和技术策略取向的基本格局。

在我们看来，无论是哪种方法，都可以抽象为两部分来看，一是关于自然语言的知识，一是表述知识的机制。我们假定有关自然语言的知识是客观的，那么知识本身对规则方法和统计方法应该是共同的，没有差异。这样，比较规则方法和统计方法的差异，很显然就可以归结为表述知识的机制的不同。一般而言，规则方法最常见的是以一定的形式文法系统来表述自然语言中大小成分间的组合规则；统计方法则以各种统计数据来显示语言成分间的组合可能性。不少论文在篇首例行的谈及这两种方法的优劣比较时，通常会归结为，在实际操作上前者的知识来自专家的内省，后者则是由计算机从真实语料中统计得来；在效果上则是前者的知识颗粒度大，后者的知识颗粒度小；以及在面对处理对象时前者的鲁棒性（robustness）差，后者的鲁棒性强等等。我们认为，这样的比较似乎显得很直观，但却是一些似是而非的粗糙意见，并没有深入到两种方法的实质。实际上，在考虑规则方法跟统计方法的异同时，真正应该回答的问题是，两种方法在组织语言知识时各自的困难和负担在哪里，对语言知识的控制方式如何，系统的总体效率和代价怎样，等等。而在说哪一种方法对自然语言处理更有用时，也不应该是笼统地下结论，而应该是对不同层次和级别的自然语言处理问题分开来讨论，譬如，统计方法用于自动分词和词性标注以及语音识别等领域，取得了比较好的效果，用在句法结构和语义的分析上会怎样呢？

本文不展开讨论上述格局中规则方法和统计方法各自具体的优劣。我们倾向于这样来看待目前的格局，即无论是哪一种方法，最终都需要依赖可靠的语言知识驱动计算机正确地处理自然语言。就目前对自然语言知识掌握的水平而言，两种方法都还有许多研究工作要做，尚不到争一日之短长的时候。此外，把两种方法对立起来看仅仅是一种视角，自觉地审视二者的共性并互相补充，对研究工作可能更有启示。事实上，已有不少研究人员用统计的方法发现规则，再用得到的规则进行分析处理；或者利用统计方法在传统的上下文无关文法的规则中加入概率权值得到概率上下文无关文法的产生式规则，都显示出将二者结合起来的某种趋势。

统计方法涉及到较多的数学公式，考虑到本文主要意图在于能对文科背景的研究人员进入汉语信息处理这一研究领域开展研究工作提供一定参考，因此下面的讨论基本集中在汉语信息处理中跟规则方法相关的研究工作上。

就规则方法而言，人要做的工作主要包括：

① 人以自身的理性思维从自然语言中抽取可被形式化的语言知识。

② 以一定的形式化方法表述这些语言知识。

③ 将这些语言知识算法化后编制成程序输入计算机。

上述工作一般说来应该是由语言学工作者和计算机科学工作者共同来完成的。而就最典型的情况来说，语言学工作者在从纷繁复杂的语言现象中抽取可形式化的语言知识方面担负的任务多一些；计算机科学工作者则在以一定的形式模型来表述语言知识以及如何将语言知识算法化编制成程序实现方面更能贡献力量。

在基于规则方法的框架中，语言知识最概括地讲可以分为范畴和规则两部分内容。而所谓从自然语言中抽取语言知识，也就是由人来为自然语言建立有限的范畴，并以有限的规则来表述这些有限范畴之间的有限关系。80 年代以来国外语言学流派纷呈，理论迭出，无非是在语言知识的抽取中，对确立哪些范畴以及采取何种表达方式来组织规则系统有不同选择的结果。国内计算语言学界、汉语研究界在面向信息处理开展的语法研究方面，也同样如此。下面我们基于这种认识，展开具体评述。

三

80 年代以来国内在汉语信息处理领域的主要研究大致可以分为三大块：

（1） 引介国外理论方法并结合汉语特点探讨汉语的计算机处理理论问题。

（2） 各种跟汉语的信息处理相关的实验和应用系统的研制开发。

（3） 汉语知识库的建设及汉语语法规则的发掘。

如果我们以对汉语语言知识的抽取水平作为一个标尺，来衡量在这个不算太短的时期内，面向信息处理的现代汉语语法研究的发展程度及状况，可以得到的一个大致清晰的初步

结论是，在确立现代汉语语法语义的范畴方面取得了较多成绩，在规则方面虽有一定探索，但跟范畴方面的进展相比则显得不足。综括而言，就是对汉语知识规律的发掘，整体水平仍难以满足计算机处理的需要。

尽管这样的观点多少会被斥为保守或至少是态度不甚乐观等等，但下文对 80 年代以来汉语信息处理领域三大块主要研究工作的评述，仍然可能是支持上述认识的。不管怎样，清醒地审视过去和现状，才能有效地去设计未来。

需要说明的是，这里罗列三部分研究工作的内容，只是就最主要的方面勾勒而已。分成三块很大程度上也是为了叙述的方便起见，并不见得实际的研究就只有这些并一定以这样的面貌呈现。由于本文主要是在考察 80 年代以来汉语信息处理研究基本状况的基础上，探讨以信息处理为应用背景进行现代汉语语法研究的问题，因此我们对上述（1）、（2）两块研究内容的评述相对简略。目的着重在把背景性的内容拉开拉宽。对（3）这一块的研究工作，其中包括已经建成相当规模的知识库，以及虽然规模不大但堪称积极探索的对汉语语法规则的相关研究，评述则相对前两部分要详细一些。

下面先来谈（1）这一块的研究工作。

这一块的工作以引介国外计算语言学领域的理论方法为主。而学者们在介绍国外较之国内先行了许多步的理论和方法的同时，也有不少人结合汉语自身的特点，对这些理论和方法做了深入一步的探索。

从抽取语言知识的角度看，众多的语法理论大致可以分为两类：一类侧重从语言事实中发现范畴，建立规则（相当于我们在第二节中提到的步骤①的工作）。像美国的描写语言学理论（结构主义语法）、法国特尼埃尔的依存语法（配价语法）、菲尔摩的格语法、韩礼德代表的系统功能语法、Langacker 倡导的认知语法等都属于这一类；另一类侧重如何将已发现的语言知识用一定的形式化方式加以描述（相当于我们在第二节中提到的步骤②的工作）。自乔姆斯基 50 年代末提出转换生成语法以来，到 80 年代蓬勃发展、蔚为大观的一系列跟形式语法密切相关的语法理论，诸如扩充转移网络（ATN）、支配约束理论（GB）、功能合一语法（FUG）、词汇功能语法（LFG）、定子句语法（DCG）、中心词驱动的短语结构语法（HPSG）、广义短语结构语法（GPSG）、范畴语法（CG）、链接语法（LG）等等，就都属于这一类。计算语言学的迅速发展为上述两类语法的各种理论提供了一个良好的演练舞台。这些各具特色的语法理论在计算语言学的各个应用部门，诸如机器翻译、人机对话等领域，几乎都已得到广泛实验。当然，这些研究主要的自然语言背景也差不多都是英语。

国内 50 年代末就已开始的机器翻译星星之火，到了 80 年代初期再度燃起，成为当时自然语言处理研究领域的主要风景。代表性的研究工作基本收录在陆续出版的 3 卷《语言和计算机》论文集中。而对国外相关领域的介绍，理论内容相对较少，主要偏重在各种上机实现的系统方面（所谓的第一代、第二代乃至第三代人机对话系统等等）。范继淹、徐志敏、李家治、陈永明、冯志伟等人的介绍及其所研制的实验系统的报告，是这方面的代表。在那个重整旧山河、开创新局面的短暂过渡期内，当务之急很明显是百废待兴中先把框架搭起来。如果对当时艰苦的硬件环境稍加留意，就不难了解研究人员在表现出急迫的使命感的同时投入了多么大的工作热情。

将国外的语法理论方法全面系统的汉化是自 80 年代中后期开始的。随着《中文信息学报》在 86 年底的创刊，国内这一领域的研究者有了一块稳固的研讨阵地。介绍国外各种语法理论的文章成为国内研究汉语信息处理的重要参考。而语言学界《国外语言学》和《语言文字应用》两份杂志也一直扮演着这方面文献热情的传播者的角色。此外，自 1991 开始的两年一度的全国计算语言学联合学术会议，为研究者提供了宝贵的交流和学习机会。从每次会议的论文集大致可以看出一个阶段内中国计算语言学理论和实践的发展水平。

90 年代初国内有三本引论性质的介绍计算语言学的专著问世。钱锋、陆致极、刘开瑛等学者将有关自然语言处理这一领域研究的基本理论方法加以总结形成系统，基本上反映了国外一个时期内的基础研究面貌。这三本引论性质的著作对研究人员形成一个关于计算语言学的整体印象助益良多。此后到 90 年代中后期，国内又陆续有这类系统地研究介绍计算语言学的专著出版，其中冯志伟、姚天顺等学者的著作堪称代表。跟早一批的著述相比，这时的著作一方面补充了有关国外更新的研究状况的内容，另一方面对国内学者所做的系统研究和理论探索也多有涉及。

相对于具体语法理论的介绍旺势而言，对自然语言理解做深层次的带有哲学色彩的思考，在国内就显得非常冷清。80 年代中期宁春岩在《语文研究》上发表的一篇《自然语言理解中的几个根本问题》，以及他译介的美国哲学家休伯特.德雷福斯（Hubert L.Dreyfus）的专著《计算机不能做什么－－人工智能的极限》，是做此努力的少数工作之一。此后语言学界有袁毓林在 1993 年发表的《自然语言理解的语言学假设》一文，对一些自然语言理解的底层问题做过分析。此外似乎在语言学界的刊物上就罕见这方面的文章了。

值得一提的是，国外在对多种形式语法理论进行广泛实验之后，效果似乎并没有理论提出之初那般的轰动。自然语言处理的困境仍未得到实质上的改善。这些理论多数属于我们上文所说的第二类，是侧重对语言知识加以形式化描述的。而对自然语言知识本身的发掘则很有限。尽管表达方法先进了，所表达的内容却并不一定就会跟着有实质上的提高。这也就是这些语法理论难以从根本上解决自然语言处理问题的症结所在。

总括而言，(1)这一块的研究工作对确立我国计算语言学领域的宏观格局起到了决定性的作用。对这一领域许多现象的观察，以及对研究课题的把握，都离不开这个大背景。

下面再简单谈谈（2）这一块的研究。

这一块的研究工作主要涉及实践应用，带有浓厚的工程技术色彩。结合汉语的实际情况，用计算机对汉语进行信息处理，首先就碰到汉字的输入输出问题。跟汉字的输入密切相关的研究是汉字编码。国内在经历了所谓万“码”奔腾的汉字编码战国时代之后，这方面的问题基本解决。目前汉字输入已不构成汉语信息处理的障碍。不仅如此，从键盘到 OCR 到手写识别到语音输入，汉字的输入方式已经是多种多样，能够满足多种需要了。跟汉字的输出密切相关的是汉字字库的信息压缩技术。享有“当代毕升”美誉的北京大学教授王选与其同事一道研制成功的汉字折线段压缩技术，很好地解决了这个难题。从而划时代地使汉字文献的印刷出版告别铅与火，进入电子时代。

跨过了字处理难关的科研人员在继续迈步前进时，又迎面遭遇到汉语特有的自动分词困难。由于汉语书面语分句按词连写的习惯，词与词之间没有像拼音文字那样的天然空格隔开，这样，计算机面对的汉语整句输入，实际上就是单个方块汉字线性排列的字符串。要像人一样对句子进行处理，就必需把这一串字符切成合乎人的语感的一串词。这几乎是我们进行其它所有跟自然语言处理相关的应用开发，诸如机器翻译、人机对话等的前提。然而不像在解决字处理难题时那么走运，在分词问题上，尽管我们的许多计算机自动分词应用系统都宣称达到 90%以上的正确率，但由于一方面在理论上没有最终解决汉语词这个语言单位的性质问题，另一方面也是更重要的方面是汉语词本身的特点造成困难，国家虽然已经出台了分词规范（“信息处理用现代汉语分词规范[1]”，中国国家标准 GB13715），但在实践中仍有相当多的分词歧义问题、未定义词问题等困扰着研究人员。

不管怎样，汉语切词软件目前也算基本上达到实用要求了。分词结果作为后续处理的输入基本能够满足要求。相比之下，比分词更进一步的，同时也是自然语言处理核心部分的句法分析，情况则是更加不尽人意。就汉语的特点而言，句法分析的很大一部分工作实际上可以看作就是短语（词组）结构分析。当汉语信息处理迈入句处理（短语结构分析）阶段时，碰到的是比字处理和词处理阶段更多而且更大的困难。以语言信息处理中最引人注目的机器翻译领域为例，吴蔚天等在设计汉外机器翻译系统 Sino Trans 时提出的汉语完全语法树模型，即是在尝试建立适合计算机使用的汉语句法分析模型方面所做的努力。虽在一定程度上推动了计算机分析汉语句法结构水平的发展，但在直接发掘汉语语言知识方面，即揭示汉语的语言成分组合规律方面并没有太多进展。

不过，值得指出的是，除了机器翻译之外，还有如汉语生成、篇章理解、信息检索、自动文摘、自动校对等等应用系统的开发，都对汉语语言知识提出了迫切的需要。这些需要从一开始就是今后也还将是汉语研究的主要驱动力。在这一点上，计算机科学工作者跟语言学研究人员应该结成攻关同盟军，充分发挥各自的知识优势来排疑解惑。如果过分强调实用或受到应用系统开发人力财力及时间上的限制，计算机科学工作者就难以做到真正消化吸收语言学界已有的理论和具体语言知识的成果。语言学研究者也会因为唯实用与否论成败而进入思考语言问题的误区，难以真正推进汉语语法研究的发展。

---

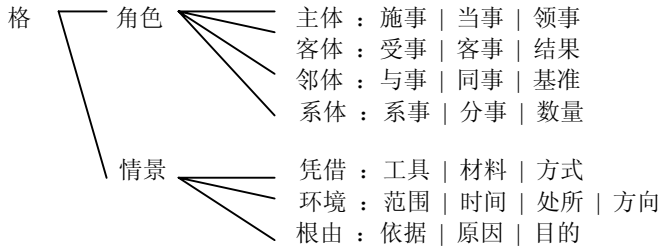
[1] “信息处理用现代汉语分词规范”1992 年被批准为国家标准，1993 年 5 月 1 日在全国正式实行。

下面我们要评述的第（3）块的研究工作，是国内研究人员在直接发掘汉语语言知识方面所做的尝试和努力。其中就包含了集中计算机科学和语言学两方面的研究人员来共同开展攻关工作的范例。

我们先来看（3）这一块研究中两个规模比较大的构建汉语知识库的工作。

一个是鲁川、张普、林杏光等倡导的基于格语法的汉语语义格关系研究。另一个就是朱德熙、陆俭明、俞士汶等倡导的基于词组本位语法的现代汉语语法信息词典的研究。从抽取汉语语言知识的角度看，这两项工作在建立汉语语义知识和语法知识的基本范畴体系方面分别都做了相当深入的探索工作。

《动词大词典》在将格语法理论落实到汉语一个个具体的动词上去的口号下，首先提出了一个由 22 个格组成的汉语格系统。这 22 个格以分成两大类、7 小类的格局组织成汉语动词的格关系系统。可图示如下：



而对动词本身，《动词大词典》“根据动词所表明的动作或状态相关的主体客体间的语义关系”，将动词分为 6 个次类。分别是：

他动词：主体是动作的发出者，动作涉及客体。	如：吃、重视、研究 等
自动词：主体是动作的发出者，动作不涉及客体。	如：走、跑、毕业 等
外动词：主体不是动作的发出者，动作涉及客体。	如：碰见、知道、懂 等
内动词：主体不是动作的发出者，动作不涉及客体。	如：病、死 输 等
领属动词：表示领属关系的动词。	如：有、拥有、具有 等
系属动词：表示系属关系的动词。	如：是、等于、属于 等

基于上面的总体框架，《动词大词典》对 1000 多个现代汉语常用动词按义项进行了动词语义格关系的描写，并对每个动词的各种格框架情况给出相应的例句来说明。比如：

爱护 <他动> 爱惜并保护。
【 基本式 】[施事{军人、猫、图书馆} + 爱护 +
受事{儿童、士兵、眼睛、身体}] 读者要爱护图书
【 扩展式 】[系事]〈作为读者〉每个人都应该爱护图书。| 你〈作为公司的职员〉应爱护本公司的名誉。 ……

从这里所举的简单样例不难看出，《动词大词典》实际上是对汉语动词跟名词性成分的语义搭配进行了概括描写。在学术界对汉语语义研究中应设立多少格尚无结论的情况下，这部词典的编者出于语言工程的考虑，花大力气从具体描写每个动词的格框架做起，应该说是很有些魄力的。但语义描写从理论框架的确立到具体落实中每个词项的把握，没有哪一步是易走的坦途。正如《动词大词典》的编者在序言中指出的那样，深化现代汉语格关系的研究是没有止境的。从动词格框架系统中各个格的确立是否恰当（包括名目和数目），到具体每个动词在刻画其格框架时如何取舍，从仅仅是举例性地罗列可跟动词发生格关系的名词，到在动词的语义组合框架中研究名词的语义特征和类别从而整理出名词的语义分类体系，真正能让计算机用起来能发挥效力的汉语语义模型的建立还有很长的路要走。不管怎样，先有一个一定规模的探索性成品出来作为后继者的参考，是经验也好是教训也好都会是有启发意义的。

《现代汉语语法信息词典》以朱德熙先生提出的词组本位语法体系作为设置各项语法范畴的理论基础。首先是选取一些具体的功能标准确定了汉语的词语分类系统，并对照一个词

语的句法功能表现按义项把它归入某个词类；然后是以功能理念指导词语语法属性项目的设置，并根据一个词语的实际用法情况标记它的属性值。词典中的属性项目设置得相当多。例如作为研究重点的动词在词典的总库和分库中共设立了 100 多项属性，来标记一个动词能否重叠、能否直接受名词修饰、能否作“有”的宾语、是带体词宾语还是带谓词宾语等等。根据每个词在这些属性项目上的取值，我们就大致上可以确定一个词在实际的汉语话语中出现时的分布状况。我们简单举下面三个动词为例说明。

词语	同形	义项	助动	外内	体谓准	双宾	着了过	重叠	VVO	离合	单作谓语	单作补语	兼类	……
保管	1	保存			体		着了过				可			
保管	2	担保			谓									
帮		帮助			体	双	着了过	VV			可		q	

“保管”有两个义项，作“保存（物品）”理解时，它可以带上“着、了、过”等助词；还可以单独作谓语（如“我保管”）；带体词性宾语等。而当它作“担保”理解时，它就不具备这些功能性质了。从分布上看，“保管”一词的两个不同义项，在句法上有大致的互补分布关系。“帮”除了作动词外，还兼属量词类（标记为 q ）。这样描述记录一个词的功能用法特征的方式很容易让人联想到所谓复杂特征集（complex feature set）之类的形式化手段。实际上，“现代汉语语法信息词典”可以说正是复杂特征集的形式化描述方法在汉语词语的语法知识形式化方面的一次大规模实践。只不过在离散式的复杂特征集外衣下，关于 5 万多汉语词语的语法知识仍是靠着词组本位语法理论统一起来的有机整体。像上面例中“体谓准”这样的属性项目名称，就直接取自朱德熙先生关于汉语动词宾语性质的区分。

虽然跟语义知识的复杂性相比，语法知识相对容易把握一些。但也并不是没有问题存在。“现代汉语语法信息词典”为各类词设置的属性是否包含冗余信息，具体词语在相关属性之间取值上的一致性是否得到保证，词语的语法特征信息作为静态的孤立的标记，在它参与组合时如何变化，也即语法信息词典对一个词的用法所作的记录跟它在实际语言中出现时的丰富性相比还有多大差距等等，都是有待进一步研究的问题。这直接影响到词语的语法属性在用于计算机对汉语句子进行句法分析时能发挥多大作用，以及如何能更有效地发挥作用。这些都是急迫需要攻关的课题。

这里介绍的两项研究工作并不直接属于信息处理领域的某个具体的应用系统，而是出于通用的考虑，可以看作是在面向汉语的信息处理这个大目标下展开的构筑基础平台的工作。规则方法也好，统计方法也好，在具体操作中其实都离不开这样的语言基础知识库的支持。不过，如果仍然回到我们一贯的范畴加规则的语言知识结构视角上来。不难发现，上述汉语知识库的建立，主要也还是句法语义范畴的设置，并没有涉及范畴之间关系的研究，即缺乏直接服务于信息处理的汉语句法规则的总结。而下面将要提及的一些面向汉语信息处理对汉语语法研究展开的宏观探讨，以及一些针对汉语特定的具体问题的研究，应该可以看作是涉足汉语句法规则研究的努力。其中宏观研究方面有代表性的包括马希文的《从计算语言学角度看语法研究》，冯志伟的《计算语言学对理论语言学的挑战》和他提出的潜在歧义结构论，白硕的《语言学知识的计算机辅助发现》，罗振声等对汉语句型的自动分析和分布统计的研究，以及一些学者提出的受限汉语（Restricted Chinese）的研究等等。这些研究工作在把汉语研究置于计算机信息处理这个广阔的应用背景这一点上都给人以启发。白硕的研究工作提出了一套利用计算机来辅助发现汉语语法知识的体系，并从数学上给出了证明，同时以动词的小类划分为实例进行了小规模的实践检验，很具参考价值。谈到这里，有一点需要顺便指出，白硕在他的论文最后，单辟一章“结果的语言学解释”，对由计算机计算抽取的汉语知识给出语言学上的直观解释。这是非常可取的做法。而时下不少研究工作，尤其是一些用统计方法来处理语料的文章，疏漏了这方面的必要说明，罗列了一大堆统计公式再配合一两个简单的例子，让读者特别是文科背景的读者丈二和尚摸不着头脑。这对研究工作的进展实际上起不到多大的推动作用。当然这只是我们的一家之言。下面回到正题。除了上面举出的研究工作之外，面向信息处理展开汉语具体问题的研究则可以举下面这些例子，包括马真、陆俭明对汉语“名词”+“动词”词串组合歧义的研究，孙宏林从标注语料库中归纳汉语“V+N”序列的语法规则的实验分析，以及詹卫东对汉语“P<被>+VP_1+VP_2”歧义格式自动排歧的探索等等。不同于上文介绍的较大规模的语义和语法知识库的建设重在范畴的设立，这里有关

汉语语法知识的概括，重在规则的发现。具体地说，这些研究着重于在已知词项自身的范畴属性值基础上，来发现多个词项在组合时的相互制约条件，或者是根据语符串的上下文环境来判断一个语符串的内部层次和关系。譬如同样是语料中出现的“V+N”序列，在“维护大局积极进取”中，“维护”跟“大局”形成述宾式 VP；而在“维护大局的稳定”中，“维护”跟“大局”不发生直接的结构关系。同样是“P<被>+VP_1+VP_2”排列格式，“被警察抓住 vp_1 罚了款 vp_2”，介词“被”的辖域（scope）一直到第 2 个 VP“罚了款”；而“被老师批评 vp_1 写了检查 vp_2”，“被”只管到第 1 个 VP“批评”，“写了检查”则不是“被动”的。不难看出，要让计算机能正确判断这些结构的组合格局和内部关系，必须建立在对汉语的短语组合规则有充分详细的描写的基础上。我们已经有了比较系统的汉语词语语法语义范畴体系（当然这些范畴仍有进一步调整改进的必要），接下来的工作重心就应该是在短语的范畴研究和规则研究方面加大力度了。

四

上面把 80 年代以来汉语信息处理领域研究工作的基本状况做了简要评述。而对本文开头提出的关于如何面向信息处理展开汉语语法研究的问题，我们形成的初步认识是，在已有研究的基础上，应充分重视给计算机用的汉语短语结构规则的系统研究。这样不仅可以检验现有的范畴体系，更能为发展出一套完整的用于信息处理的汉语语法打开局面。这一节我们再在述评的基础上，对面向信息处理进行汉语语法研究的两个相对宏观的问题发表一点看法，希望能对投身这一领域的语言学研究人员提供参考。

一是如何看待 80 年代以来五花八门的各种语法理论以及语料库方法的崛起对汉语语法研究的意义。

对于各种语法理论的认识，我们仍然发挥在第三节开头部分已经阐明的观点，即从抽取语言知识的角度看，80 年代以来发展起来的诸如 GPSG、LFG、HPSG、DCG 等等，以及乔姆斯基生成语法学派不断翻新的理论体系（包括 X-bar、GB、θ — theory 等等组成部分），大致上都可以看作是将语言知识加以形式化的方法。以这些理论为背景来审视汉语的语法研究，最大的意义是促使研究者把问题看得更深入更清晰。譬如在全面描写汉语词语的语法功能属性时采用复杂特征集的方式，对组织汉语词语的语法功能项目，以及如何确定某个特定词语的功能值，提供了比较好的描述手段。同时也使得我们对词的用法性质的认识，类似于语音学中对语音的认识从元音、辅音到音位，再到区别性特征那样不断深入了。换言之，这些理论本身并不一定就能给汉语语法研究带来更丰富的有关汉语结构规律的知识，但因其视角的新颖和描述手段倾向于严密的形式化表述，可以促进我们认识上的清晰化和对汉语语法研究的规范化。以上是从这些理论的共性来看。如果我们强调这些理论的个性，带来的问题就是，这些理论如果分别用于汉语语法研究，是效果一样呢，还是各有高下？对这个问题，笔者本人没有太多的实践经验可资比较，国内也很少看到真正用某个理论一以贯之地组织出一套较为完整的汉语语法的研究成果，因此目前还难以回答。但如果允许根据有限的经验做一些尝试性的猜想的话，我们的看法是，这些理论用于组织汉语语法知识的大效果应该是差不太多的。差别只在技术细节和具体的计算机实现上。

这里值得一提的是，国内学者黄曾阳积多年研究心得，提出面向整个自然语言理解的理论框架——概念层次网络理论（HNC），对传统的基于句法知识的语言表述及处理模式提出挑战，代之以语义表达为基础来对汉语进行理解。就其理论模式的思想原则而言，我们认为，HNC 仍然是试图以有限的形式符号去控制无限的自然语言意义。在 HNC 理论的宣言中，有限的形式符号的组织方式是以所谓有限的句类、有层次的网络概念体系的面貌出现的，并断言这建立在对人类大脑语言感知过程的模拟基础上，可以完备地表述自然语言任何语句的语义结构。事实上，按照我们对自然语言知识的分析，从句法入手还是从语义入手，并不是本质上的差别。问题的关键在于，从句法入手，就要分出大大小小的句法类，并以这些类为基础去进一步表述各类之间可能的关系，给出同类关系和异类关系的判别依据，在需要的情况下引入语义知识来共同帮助判断；从语义入手也是采用这个基本模式，同样要分出大大小小的语义类，给出把实际的语言成分归入这些类的操作标准，并在对语言成分进行正确组合或分解时给出基于语义类的判别依据。这样看来，HNC 理论虽然作为一条新路，探索精神值得我们钦佩，探索的方向也反映了目前自然语言理解领域研究的趋势（即对自然语言语义知

识的迫切需要），但实际落实起来，要做的工作仍然是困难重重（无法回避给意义分类的问题）。乐观地展望这一新思路的未来前景固然可以鼓舞人心，务实地来思考理论内在的问题以及实践起来可能碰到的障碍应该是更能真正推动研究工作的进展的。我们期望沿着这条道路探索下去，会有更多的实际的关于自然语言的规律性知识被挖掘出来，能够有效地服务于自然语言理解的各项应用。

再谈语料库方法的崛起。这主要表现为近年来大规模语料库的积极建设以及语料库自动检索、查询、标注等自动加工工具的不断提高和完善。基于经验主义的思路来抽取汉语的语言知识从根本上讲离不开人的语言知识的有力支持。譬如建设汉语的树库，选用什么语法体系确立标注集，人工标注到什么深度，都得事先由人来决定。再比如所谓的从语料库中发现汉语的分析规则，同样也得事先由人来设计好一定的知识模板，有目的性的去发现。当然同时我们又必需看到，随着计算机硬件软件技术的飞速发展，对语料的处理能力已经今非昔比。如果有好的统计模型和质量可靠的标注语料库支持，利用计算机发现汉语的结构规律知识是非常值得探索的道路。总之一句话，高性能的计算机无疑是语言研究的有效工具，但光有简单的语料库和计算机程序，统计出来的语言知识必然很有限（譬如词频或简单的互信息等等）。无论多好的工具，还得人来驱动。

二是在信息处理的背景下如何看待汉语研究界长期以来积累的一些争议问题，以及近年来提出的各种语法体系对汉语信息处理中语法研究的意义。

众所周知，汉语研究过程中遗留了一大堆争议问题，诸如“词类问题”、“汉语中词这一单位的有无问题”、“汉语的句子是主谓模式还是话题陈述模式”等等。汉语信息处理研究这一应用背景为我们认识这些问题提供了一个非常有利的视角。拿“词类问题”来讲，争议最大的是“词无定类”还是“词有定类但类无定职”。在我们看来，一个词的词类属性，仅仅是该词一个比较重要的功能值而已。给词定一个词性，不是自然语言处理的最终目的。词性只是分析的手段之一。我们说“劳动”这个词只属一个类（譬如动词），或者说“劳动”兼属两个类（譬如动词和名词），都并不是问题的实质。实质问题是，我们拿“劳动”的词类属性来干什么？以及为了这个目的，是给“劳动”定一个词性好，还是让它兼两个类好。根据笔者本人以形式化规则的方式来组织汉语知识的经验，坚持“词无定类”原则带来的后果是词典中的描述负担重；坚持“词有定类但类无定职”原则的后果是规则的负担加重，而词典的负担轻。权衡之后是“词有定类”更加可取一些。这里我们不展开来谈这个问题。我们认为，如果研究者更多的从这样的角度来看待词类问题，会对研究工作有更好的推动。再譬如“汉语中有没有词”这个问题，也是如此。简单地拿印欧语来对比一下，说汉语的“字”如何如何独特，跟“word”如何如何不同，实际上解决不了任何问题。对计算机处理来说，与其问“汉语中有没有词这个单位”，不如问“汉语中的字在组合时有什么规则”。因为，无论设不设立“词”这个单位，纯粹的语法研究也好，为信息处理服务也好，都得回答一个共同的问题，即“小的单位是怎样组合成大的单位的”。同样，汉语的句子构造是“主谓模式”还是“话题陈述模式”，也不是一个真正的问题。真正的问题是，如果坚持主谓模式，就得回答汉语中哪些成分在什么条件下可以做“主”，哪些成分在什么条件下可以当“谓”。而坚持话题陈述模式，就得说清楚“话题”和“陈述”分别是由哪些成分充任的；有无形式标记；二者的结构规则是什么，等等。回答不了这些问题，主谓和话题陈述的争论就没有多大意义。

联系上述争议问题，近年来出现了一些语法体系，如“小句中枢说”、“字本位语法”等等，和更早的“句本位语法体系、词组本位语法体系”等并立当今汉语研究界。如何看待这些在不同的历史时期提出的汉语语法体系？我们还用上面的眼光和思考方式，联系信息处理对汉语语言知识提出的要求来看。无论哪一种体系，都应该以对汉语语法知识的发掘程度及表述上的简洁程度为衡量标准。从这点上说，上述体系在发掘汉语语法知识方面各有着重，也各有可取之处。譬如按句本位的研究路子，对汉语的句型知识可能就会有较深入的总结。按小句中枢，对汉语的复句系统，应该更能产生研究成果。按词组本位，对汉语的短语结构规则就容易深入探求其规律。按字本位的眼光，对汉字组词的规则就更加重视。这些理论体系之间，不是简单的一个否定另一个，而是对汉语语法知识的各个层次、各个方面有不同侧重，从不同途径去挖掘汉语语法知识。谈到表述的简洁程度，因为汉语在构词造句的各个层次上存在着大致的同构倾向，而词组本位体系牢牢扣住这一点来组织汉语的语法知识，总体上来看，比句本位体系表述汉语语法知识要精练一些。至于字本位理论，因提出时间不长，尚未见到关于汉语具体语言规律的系统的知识表述问世；而“小句中枢说”虽有邢福义先生

的《汉语语法学》做了比较系统的阐发，但就具体语言知识的挖掘而言，跟词组本位语法体系并没有什么本质的不同，只是在关注的重点方面有所偏重，目前还都不能做深入比较，给出有启发意义的评价。

综括起来，我们倾向于在清晰的应用背景下，以理论的目的性为驱动来思考上述争议问题和不同的理论体系。有时候，表面的争端并没有像强调的那么尖锐。似是而非的问题下面才是需要真正去认真思考求索的真知。这里我们不妨打一个比方。几个瓶子里分别装有不同的液体，如果一个没有嗅觉和色觉但会认字的人来用它们做菜。他就只能根据瓶子外面贴的标签来决定如何使用它们。如果贴有“酱油”标签，就用它来增加咸味；如果贴有“黄酒”标签，就用它来祛除荤菜的腥味，等等。计算机处理自然语言，大致上也就是这样。计算机系统就像那个没有嗅觉和色觉但能认字的人一样。它只能根据事先贴好的标签来决定怎么做。譬如我们给“劳动”和“光荣”分别贴上“动词”和“形容词”的标签，同时又让计算机知道“动词”后面加一个“形容词”可以构成合法的主谓结构，表达一个完整的意思（相当于那个人知道“酱油”可以增加咸味），计算机就能造出“劳动光荣”这样动听的汉语句子来。但问题在于，我们在给“劳动”贴上“动词”标签的同时，还给“企图”也贴上了同样的“动词”标签。计算机当然就根据上述知识也“理直气壮”地造出让人觉得别扭的“企图光荣”来了。很显然，现在人看了这个糟糕的结果后有两个地方需要检查检查。一是标签是否贴得合适；二是如果标签贴得合适，是不是该告诉计算机，有的“动词”跟“形容词”不能随便组合成“主谓结构”来表达意思（相当于告诉那个人不是所有的“酱油”都能给菜带来可口的咸味，对“酱油”可能还要进一步细分来加深认识）。

我们以这个也许算不上很恰当的比喻来结束本文。只是想强调，对面向信息处理从事汉语语法研究的人来说，真正的问题始终应该包括：（1）、我们该为汉语准备多少标签？（2）、给汉语中的任意一个成分（无论大小包括语素、词、短语、句子等等）贴上某一个标签，意味着什么？（3）、关于这些标签之间的相互搭配使用关系，我们掌握了多少？从本文对 80 年代以来汉语信息处理研究基本状况的总结来看，接下来的研究工作重点首当其冲应该是回答第三个问题，即汉语一类词跟一类词之间的组合规则，一类短语跟一类短语之间的组合规则到底是怎样的。而在回答这个问题时，应该时刻想到答案是给计算机用的。对计算机，不妨尽可能地把它想象得傻一些，所以给出的答案务必清晰明确。

附记： 本文对 80 年代以来汉语信息处理领域的有关研究工作的评述，是极为粗线条的。一方面受篇幅限制，一方面也因为作者目力所及十分有限，加上难以算入眼光敏锐者行列，有的重要研究工作文中没有提及势必难免，而提到的研究工作定位评述不当也是很有可能，因此，疏漏有误之处恳请专家学者指正。另外，本文参考文献较多，文后仅按出版时间顺序列出，不再与文中对应一一注释。

参考文献 （按出版时间排序，同年之内单篇论文在书之后）

[1] 冯志伟《国外机器翻译的新进展》，载《国外语言学》1980 年第 1 期。
[2] 董振东《逻辑语义及其在机译中的应用》，载《情报科学》1980 年第 2 期。
[3] 朱德熙《语法讲义》，商务印书馆 1982 年版。
[4]《语言和计算机》（1）中国社会科学出版社 1982 年版
[5] 李家治、陈永明《机器理解汉语——实验 I》，载《心理学报》1982 年第 1 期
[6] 范继淹、徐志敏《RJD-80 型汉语人机对话系统的语法分析》，载《中国语文》1982 年第 3 期
[7] 冯志伟《国外自然语言理解系统简介》，载《计算机科学》1984 年第 2 期。
[8]《语言和计算机》（2）中国社会科学出版社 1985 年版
[9] 宁春岩《自然语言理解中的几个根本问题》，载《语言研究》1985 年第 2 期。
[10]《语言和计算机》（3）中国社会科学出版社 1986 年版
[11] 休伯特.德雷福斯（Hubert L.Dreyfus）著《计算机不能做什么－－人工智能的极限》宁春岩译 马希文校 三联书店 1986 年版
[12] 陆致极《关于广义短语结构语法》，载《国外语言学》1986 年第 4 期。
[13] 鲁川 梁镇韩《信息处理用规则汉语》，载《中文信息学报》1987 年第 4 期。
[14] 张潮生《格语法与自然语言处理》，载《中文信息学报》1988 年第 4 期。
[15] 冯志伟《中文科技术语的结构描述与潜在歧义》，载《中文信息学报》1989 年第 2 期。
[16] 冯志伟《中文科技术语的歧义结构及其判定方法》，载《中文信息学报》1989 年第 3 期。
[17] 马希文《从计算语言学角度看语法研究》，载《国外语言学》1989 年第 3 期。
[18] 孙茂松 黄昌宁《汉语中的兼类词、同形词类组及其处理策略》，载《中文信息学报》1989 年第 4 期。
[19] 钱锋 《计算语言学引论》 学林出版社 1990 年版

[20] 陆致极 《计算语言学导论》 上海教育出版社 1990 年版
[21] 冯志伟《汉语句子描述中的复杂特征》，载《中文信息学报》1990 年第 3 期。
[22] 刘开瑛 郭炳炎《自然语言处理》科学出版社 1991 年版
[23] 张潮生《语义表达的一些性质》，载《中文信息学报》1991 年第 1 期。
[24] 翟成祥 王岩冰 张家重 徐家福 《汉语组合类型语法》，载《中文信息学报》1991 年第 3 期。
[25] 张普《信息处理用现代汉语语义分析的理论与方法》，载《中文信息学报》1991 年第 3 期。
[26] 陈肇雄 张玉洁 张祥《SC 文法功能体系》，载《全国人工智能和智能计算机学术会议论文集》，电子工业出版社 1991 年版。
[27] 冯志伟 《中文信息处理与汉语研究》商务印书馆 1992 年版
[28] 黄昌宁 苑春法《国外语料库述评》，载《机器翻译研究进展》，电子工业出版社 1992 年版
[29] 冯志伟《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。
[30] 黄昌宁等《语料库、知识获取和句法分析》，载《中文信息学报》1992 年第 3 期
[31] 陆俭明《八十年代中国语法研究》，商务印书馆 1993 年版。
[32] 袁毓林《自然语言理解的语言学假设》，载《中国社会科学》1993 年第 1 期。
[33] 黄昌宁《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
[34] 吴蔚天 罗建林《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。
[35] 林杏光 王玲玲 孙德金 主编《现代汉语动词大词典》，北京语言学院出版社 1994 年版。
[36] 林杏光审定、鲁川主编《动词大词典》，中国物资出版社 1994 年版。
[37] 孙宏林《信息处理用汉语语义词典的描述方法》，载《现代语言学·第三届全国语言学会议论文集》语文出版社 1994 年版。
[38] 沈家煊《R.W.Langacker 的“认知语法”》，载《国外语言学》1994 年第 1 期。
[39] 罗振声 郑碧霞《汉语句型自动分析和分布统计算法与策略的研究》，载《中文信息学报》1994 年第 2 期。
[40] 徐通锵《“字”和汉语研究的方法论——兼评汉语研究中的“印欧语的眼光”》，载《世界汉语教学》1994 年第 3 期。
[41] 冯志伟《自然语言机器翻译新论》，语文出版社 1995 年版。
[42] 姚天顺 等《自然语言理解 —— 一种让机器懂得人类语言的研究》，清华大学出版社&广西科学技术出版社 1995 年版。
[43] 白硕《语言学知识的计算机辅助发现》科学出版社 1995 年版。
[44] 俞士汶《关于受限的规则汉语的设想》，载《语言现代化论从》第二辑，山东教育出版社 1995 年版。
[45] 周强《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 2 期。
[46] 冯志伟《论歧义结构的潜在性》，载《中文信息学报》1995 年第 4 期。
[47] 冯志伟 《自然语言的计算机处理》 上海外语教育出版社 1996 年版。
[48] 黄昌宁 夏莹《语言信息处理专论》 清华大学出版社、广西科学技术出版社 1996 年版。
[49] 邢福义《汉语语法学》东北师范大学出版社，1996 年版。
[50] 刘伟权 王明会 钟义信《建立现代汉语依存关系的层次体系》，载《中文信息学报》1996 年第 2 期。
[51] 俞士汶 朱学锋 王惠 张芸芸 《现代汉语语法词典规格说明书》，载《中文信息学报》1996 年第 2 期。
[52] 姬东鸿 黄昌宁《汉语形容词和名词的语义组合模型》，载《中文与东方语言信息处理学会通讯》（Communications of COLIPS）Vol.6 No.1 1996 年 6 月
[53] 马真、陆俭明《“名词”+“动词”词语串浅析》，载《中国语文》1996 年第 3 期。
[54] 周强 俞士汶《汉语短语标注标记集的确定》，载《中文信息学报》1996 年第 4 期。
[55] 陆俭明《关于语义指向分析》，载《当代中国语言学》总第一期，1996 年。
[56] 孙茂松 黄昌宁 方捷《汉语搭配定量分析初探》，载《中国语文》1997 年第 1 期。
[57] 孙宏林《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）
[58] 孙健 张尧 王启祥《汉语受限语言的设计与应用》，载《中文信息学报》1997 年第 3 期。
[59] 周强 张伟 俞士汶《汉语树库的构建》，载《中文信息学报》1997 年第 4 期。
[60] 黄曾阳《HNC 理论概要》，载《中文信息学报》1997 年第 4 期。
[61] 詹卫东《PP<被>+VP1+VP2 格式歧义的自动消解》，载《中国语文》1997 年第 6 期。
[62] 朱靖波 张玥杰 姚天顺《面向数据的句法分析技术》，载《中文信息学报》1998 年第 1 期。
[63] 张树武 黄泰翼《汉语统计语言模型的 N 值分析》，载《中文信息学报》1998 年第 1 期。
[64] 刘志文 等《自然语言语句的 HNC 表示》，载《语言文字应用》1998 年第 2 期。