关于汉语信息处理的认识及其研究方略

俞士汶　　朱学锋

（北京大学计算语言学研究所 北京 100871）

[**摘要**]在总结长期实践经验的基础上，笔者分析了为什么自然语言处理是一个相当困难的研究领域，而汉语信息处理是更加困难的研究领域。面对日益强烈的社会需求，汉语信息处理的研究方兴未艾。笔者探讨了开展这项研究的技术路线，特别强调了语言知识库建设的重要性。

[**关键词**]自然语言处理，汉语信息处理，语言知识库

[**中图分类号**]TP391

Chinese Information Processing and its Methodology

Yu Shiwen　　Zhu Xuefeng

Abstract: Based on the experiences of long-term research practices, the authors analyze why the Natural Language Processing is a very difficult research domain, especially the study of Chinese Information Processing is more difficult. To meet the demands of informative society the study of Chinese Information Processing is well under way without signs of slackening. The authors discuss the methodology of Chinese Information Processing and emphasize the importance of Language Knowledge Base.

Keywords: Natural Language Processing, Chinese Information Processing, Language Knowledge Base

1. 引言

2001 年 5 月国家语委在无锡召开了语言文字应用研究“十五”科研规划论证会议。笔者对中文信息处理在整个会议进程中得到的重视有相当强烈的感受。无论是领导干部的讲话[1]还是国家语委提出的《语言文字应用研究“十五”项目指南（征求意见稿）》，以及专家们的发言与论证，都充分表述了中文信息处理技术对我国社会的信息化进程和信息产业发展的战略意义。作为一名长期从事语言信息处理技术研究的专业人员当然深受鼓舞。现在的条件同 10 多年前草创时期相比，实在是好得太多了。笔者愿在本文中阐述对自然语言处理特别是汉语信息处理的一些基本认识，也阐述了汉语信息处理技术研究应该采取的技术路线，强调了语言知识库建设的重要性。这些认识主要是自己长期研究实践经验的总结，理论水平可能不高。期望能达到向先进学习、与同行交流的目的。

投稿日期：2002 年 2 月 10 日

基金支持：国家自然基金 69483003、973 项目 G1998030507-4、863 项目 2001AA114040、北大 985

作者信息：俞士汶，男，1938 年 12 月生，北京大学计算机科学技术系教授；**朱学锋**，女，1937 年 12 月生，北京大学计算语言学研究所副教授；两人的研究方向皆为计算语言学。

2. 自然语言处理——难

语言研究确实很难。道理并不复杂。首先，人们研究任何事物和学问总是要依靠思维。研究语言同样离不开思维。可是思维（至少逻辑思维）又要用语言来表达。也就是说，语言既是研究的对象，又是研究的工具。当其他领域的学者将自然现象、社会现象、生理现象等作为对象加以研究时，就没有这样的尴尬。第二，语言现象是无限的，而从事语言研究的人所能利用的资源总是有限的。只用有限的资源去解决无限的问题，实在太困难了。如果语言信息处理的研究者不预先明确研究的范围和目标，甚至给人以任何问题都能解决的假象或模糊认识，结果往往是从期望的高峰跌入失望的低谷。第三，从事语言信息处理研究，最得力的工具自然是计算机。可是，当前可以利用的通用计算机不论功能多么强大，仍然被约束在冯•诺依曼的体系结构内。它的本质功能只不过是对一种表现形式的符号串实施一连串的但总是有限步的变换，而得到另一种表现形式的符号串。这个过程同人的思维过程、认知过程是大相径庭的。如果没有跳出这个窠臼，却声称可以在这样的计算机上再现人脑的“理解”机制，即使充分肯定研究者的宏图大志，也要冷静地指出这是对自然语言理解的困难估计不足。

下面的例子也许可以把这个问题说得更明白一些。笔者偶然读到《今日民航》2001年9月号上的一篇关于“沙漠化”的文章，这是一篇新闻报道，应该是写给普通人看的。笔者读到其中的这样一段文字：

> 几年前由于种植籽瓜有利可图，使大批的种植者就到过渡带来开垦，
> ……。
> 在这样的绿洲和沙漠过渡带开垦，极易造成风蚀。

却遇到了困难。对于删节号前的那句话，每一个字都认识，也没有专有名称，可是试读了两遍，就是读不通。因为运用自己的语言知识和常识，对后半句进行切分，只能得到“就”、“到”、“就到”、“过”、“到过”、“过渡”、“带”、“来”“带来”、“开垦”这样一些词语，组织不成可以理解的句子。直到读到删节号后面的那句话，才“顿悟”到一个并不深奥的专业知识：在绿洲和沙漠之间存在着“过渡带”。再返回到前面那句话，这时自己的脑海（知识库）中已经有了关于“过渡带”的知识，因而可以实现正确的切分：

> 使/ 大批/ 的/ 种植者/ 就/ 到/ 过渡带/ 来/ 开垦/

理解它也就不存在困难了。其实，机器处理这段文字的困难还不仅限于此。像“籽瓜”这两个字连在一起也是少见的。笔者只是猜想大概是指一种专门用来取籽食用的瓜。在这样的知识或“预设”的指导下，才可能辨识出“籽瓜”这个词，才能正确切分前半句话。在汉语自动分析技术中，通常把切分作为处理的第一步，正确的切分是理解的基础。这个例子又反过来说明，只有理解了，才能正确切分。对于这段文字，人能理解的关键是“过渡带”和“籽瓜”这两个概念。笔者的亲身经验说明，人即使事先并没有学习过这些知识，但是通过下文可以“领悟”这两个概念。实际上人的理解能力还不限于此。由于当代人有了“环境保护”和“防止沙漠化”的观念，就依据这里所引用的两句不连续的话还可以做出文章的摘要：“为

了防止沙漠化，要停止在绿洲和沙漠之间的过渡带发展种植业”。读者不难想象，当前机器的智能同人的智能相比，该有多大的距离！要害在于人脑的这种“领悟”和“推理”的机制是难以形式化的，至少目前还没有这种形式化的成果。因此，计算机也就无法自动填补知识的空缺。目前，人脑的认知机制还是一个谜，这是实现“自然语言理解”的真正障碍。

科学不会停止探索。脑科学、认知科学、语言科学、计算技术的发展和融合也许终将突破这个障碍，不过，这大概不会是 10 年或 20 年之内可以办到的事。

3．汉语信息处理——更难

已经相当普及的术语“中文信息处理”所指的研究工作实际上可划分为两个层次：一个是“汉字信息处理”，“中文”指“文字”；另一个是“汉语信息处理”，“中文”指“文章”。当然这两个层次也有联系。成千上万的汉字与几十个西文字母的区别已在中国造就了汉字信息处理产业。至于汉语信息处理，除个别例外，还没有成熟到这个程度。汉语信息处理属于自然语言处理的范畴，其最高境界当然也是汉语理解，也就是实现计算机对汉语的领会与运用。为了逐步解决这个难题，中国学者既要认识到汉语和其他语言的共性，也要认识到汉语区别于其他语言的特性。汉字信息处理中一些最基础的课题，像汉字大字符集、汉字编码等在英语世界是不存在的。即使在混合排版、混合字符识别等少数任务中，汉字同其他文字的关系也只是“物理”的或“混合”的关系。进入语言（语句、话语或篇章）层次，情况就不同了，各种语言面临的共同问题多了，无论是英语还是汉语，都会碰到像词语多义、语言单位定界、句法结构、语义表示、指代与省略、隐喻等这样一些基本的问题，如果进行机器翻译或语言比较研究，相关语言之间的关系则是相互依存的。从发展水平看，计算机还不能像人一样理解汉语，同样也不能理解英语。但也不好说汉语信息处理同英语信息处理现在都处于同一条起跑线上。只不过汉语信息处理不像汉字信息处理起步时同英语世界的差距那么大。同时，也应当注意到汉语有自身的特点，汉语信息处理有特殊的难题。认识到共性，有利于学习先进的理论、技术与经验，有利于汉语信息处理研究同国际接轨，有利于成果的传播。认识到特性，可以为汉语信息处理研究争取更广阔的发展空间。

本节从计算机处理的角度讨论现代书面汉语的特点。这个问题很多学者和笔者都曾探讨过。希望这里能谈得更深入一些。

1．语言单位。关于作为研究对象的语言单位，学者们有很多论述。笔者认为，以多大的语言单位作为信息处理的对象至少要顾及 3 个因素：①应用目标，②技术与理论的发展水平，③语言类型。表达完整知识或信息的语言单位应该是一篇文章或一本书，尽管通常也认为句子是表达相对完整的意义的语言单位。香港城市大学郑锦全教授曾作过一个有趣的实验[]：看《明报》的一则新闻的最后一句，看不懂，倒着往回多看一句，还是不懂，再往回多看一句，如此继续，直到可以理解为止。实验说明，由于汉语文本中有大量省略、指代的句子，计算机孤立地处理一个句子，或者难以理解，或者产生歧义，是不奇怪的。但目前的技术又还不容易驾驭篇章这么大的单位。甚至连处理有显式标记的段落也还困难。当前绝大多数语言信息处理系统（如机器翻译）是以句子作为基本处理单位的。Chomsky 形式语法的产生式规则的起始符就是句子 S。实际应用基于统计的 n 元语法时，n 一般不大，实际上也是约束在一个句子的范围内。朱德熙先生也认为最大的语法单位是句子。有些应用研究，如信息提取和自动文摘，固然要以篇章为对象，但也要以句子处理为基础。

2．句子界定。对于英语和日语，从篇章中分割出句子是很简单的事，而且句子还是很清晰的语法单位。英语句子一定包含一个由限定形式的动词担任的谓语，日语句子一定以终止形的动词结束。可是汉语中句子同句子之间的界限并不清晰。古汉语不使用标点符号，断句是

大学问。现代书面汉语虽然使用标点符号，但标点符号并没有承担界定句法单位的功能。若以句号作为句子的结束标志，句子可能很长。句号之前的内容可能是一个句群或段落。若认为逗号可以作为句子的结束标志，则很多句子又是不完整的，有缺省的句子机器是很难分析的。随手抄录 2002 年 1 月 26 日《参考消息》中一句话：

车臣武装分子和世界其他地区的恐怖分子是一丘之貉，应该合力打击他们。

可以把这句话看作是一个由两个分句构成的复句，问题是机器如何判断第 2 个分句有省略，省略的是什么，“他们”又指代谁。又如朱德熙先生举的一个例句[2]：

你得藏在一个你看得见他，可是他看不见你的地方。

这里，逗号的左右两部分又不是分句。句号结束的只是一个单句，但它内部却包含了一个复句结构形式：“你看得见他，可是他看不见你”。再从《科技术语研究》2001 年 4 期 14 页摘录一段文字如下：

新一届测绘学名词审定委员会的主要特点是年青化，吸收了一些工作在教学、科研前沿的青年专家学者，充分发挥他们接触新知识多，对名词工作热情高、活力大的特长，同中老年专家共同做好新一届委员会的名词审定工作。

这段文字共有 99 个字，一逗到底。人读起来，通顺易懂。可是计算机处理可就难了。其中第 3 个逗号的作用同其他 3 个逗号的不一样。试用几个机器翻译系统进行了翻译，没有一个系统能给出可以使用的译文。

自 80 年代中期以来，汉语信息处理学界就将句子的词语切分作为一个重要的攻关方向，开发了很多软件，发表了很多文章，还制订了国家标准。但却很少有人直接提及现代书面汉语的断句问题。也许这个问题对人来说是不成问题的，但对计算机仍然是个难题。

3．词语切分。需要把一个句子进一步划分更小的语言单位并不是汉语的特殊课题。在英语中，尽管 word 同 word 之间留有空格，也是有切分问题的。只不过由于英语的 word 在句子中有形态标志，虚词同实词的词形不同，名词前常有冠词或介词，专有名词的第一个字母要大写，这些因素使得英语的切分相对容易些。日语的书写方式同汉语相同，有同汉语相似的切分问题，但日语多种文字（汉字、平假名、片假名）混合使用以及助词（Postpositional）的不可缺省，使得日语切分也相对容易些。在汉语中，标点符号之间只是连续的汉字序列。由一个单音节的语素构成的单纯词同不成词的语素之间的界限不清晰，按同样的方式（如定中、状中、述宾、述补、主谓等）构成的复合词和短语的界限也不清晰，因此，词语切分始终是一个难题。

4. 词性标注。由于属于同一词类的词呈现诸多相同的语法属性，因此词性（Part Of Speech）对于语言信息处理是最便于应用的。Chomsky 形式语法的产生式规则的终极符就是词性符号。克服数据稀疏、使 n 元语法实用化的解决方案之一是用相应的词性序列替换词本身的序列。为了得到 n 元语法的各种参数，需要将一定规模的语料进行切分并标注上词性。因此，词性标注又成为语法分析和大规模语料库深加工的必要步骤。当有了一部划分了词类的机器词典，词性标注的主要工作就是消解兼类词在实际文本中的歧义。任何语言中起语法作用的

虚词的数量都不多，但使用频率很很高。英语中像 and, by, in, of, the 等虚词同实词是不同形的，现代日语中的助词“に、て、は、を”没有汉字表记形式，语法功能比较明确。英语和日语的虚词的这些特性使得词性标注便于找到参照点，英语和日语的实词在句子中有形态变化，因此英语和日语的词性标注也就相对容易。可是汉语的情况却不一样，首先，虚词同实词在词形上没有区别。“把”既是介词，也是动词和量词；“和”既是连词，也是动词；“在” 既是介词，也是动词和副词。因此，汉语的词性标注缺乏最便于把握的线索。实词中的兼类词（如名词的“锁”与动词的“锁”，区别词的“共同”和副词的“共同”）在使用时也没有形态上的区分（如“门锁”与“锁门”，“共同利益”与“共同奋斗”）。汉语的实词在使用时既没有形态变化，又表现出多功能性（如动词呈优势分布的主要功能虽然是担任主谓结构中的谓语和述宾、述补结构中的述语，但也可以担任主语和宾语，而且形态没有任何变化）。专有名词同普通词在形态上也没有任何差别，这些特点给词性标注带来本质性的困难。

汉语的词语切分与词性标注是两件事，又常常结合起来同步进行，在自然语言处理流程中相当于其他语言的词法分析，它是后续的句法分析、语义分析和语境分析的基础。汉语的词语切分与词性标注也有独立的应用领域（如面向 Web 的海量信息管理）。因此，高性能的词语切分与词性标注软件的价值是不容低估的。

5. 句子的语序。英语是 SVO 型语言，日语是 SOV 语言。很难按照类似原则把汉语归类。汉语句子中词语的排列顺序是相当自由的。例如，

那个学生昨天上午看完了这本小说。

这句话的语序也可以改变为

这本小说那个学生昨天上午看完了。

这本小说昨天上午那个学生看完了。

昨天上午这本小说那个学生看完了。

意思没有变化。因此词语在句子中的位置信息对它的句法功能的提示甚少，这当然增加了句法分析的难度。

分析句子时，把组成这个例句的几个短语“那个学生”、“昨天上午”、“看完了”、“这本小说”各自看作一个整体是方便的。当然这些构成成分本身也是有结构的，可以注意到这些短语的结构是稳定的，内部顺序是不能改变的。

不过，也不能认为汉语的语序无关紧要。像汉语的时间状语“昨天上午”通常就不能放在谓语动词“看完了”的后面。又如，“她是昨天上午回的家。”这句话，其构成成分的位置就很难移动。无论对于分析还是生成，汉语语序规律的深入探讨都是必要的。

6．汉语的句法结构。如果排除由一个自由的语素或词实现的句子，汉语的单句都是由短语实现的，因为任何类型（主谓、述宾、述补、定中、状中等等）的自由短语加上句调都可以实现为句子[2]。任何一种汉语短语结构的构成成分的顺序是固定不变的，或者说短语的结构稳定。因此，将短语结构的研究作为汉语语法研究的中心课题是合理的。其合理性也为大规模汉语语言工程实践所证实。如果以数理语言学或计算语言学中广为流传的理论体系作参照，将汉语短语结构研究看作汉语语法研究的重心也是有理据的。上下文无关语法产生式规则中的所有标识符除唯一的开始符和少量的终极符（词性）外，大量的非终极符代表的都是短语。当代颇有影响的 GPSG 和 HPSG 更直接把短语结构（phrase structure, PS）放在语法理论的名称中。

汉语的短语和复合词是按照同样的结构规则构造的，可以将尚未实现成句子的短语或复

合词统一叫做“句法结构”。同时由于单音节的成词语素同不成词语素的界限也是模糊的，因而短语和复合词的界限是模糊的。

汉语的所有句法结构规则都是允许嵌套的。其他语言的嵌套结构也是常见的，像英语的从句也可以包含另一个从句。不过，汉语句法结构的嵌套有两个特点：其一是句法结构的嵌套不需要加其他连接词或关连词，其二是主谓结构的谓语还可以是另一个主谓结构[3]。这两个特点使得除词语切分的困难之外又多了一个短语定界的困难，汉语自动分析真可谓雪上加霜。

7. 虚词的省略。前面已提到汉语的虚词同实词同形所造成的麻烦，虚词的省略也造成很大的困扰。同孤立的一句“鸡不吃了。”不一样，“苹果吃了。”在语义上不会有歧义。“苹果”只可能是“吃”的受事。但在句法上又有歧义：可能是被动陈述句“苹果被吃了。”省略了“被”，也可能是祈使句“把苹果吃了。”省略了“把”。

8. 汉语的时态、语态和语气。由于缺乏严格的形式标记，在一句话的范围内辨别时态（过去、现在、将来、进行、完成等）、语态（主动态与被动态）、语气（真实语气与虚拟语气）也是不可能的。“看电视”可以用于回答以下任何一个问题：

“你昨天晚上干什么了？”

“你明天晚上干什么？”

“现在你在干什么？”

又如“你什么时候回家？”这句话用于询问未发生的事，“你什么时候回家的？”用于询问已发生的事。两句只差一个“的”。现代汉语中使用频度最高、已经不堪重负的助词“的”又挑起了一个表示时态的重担。如果认为整个句子末尾的“的”（假定能鉴定出该“的”不是先同最近的一个单词或短语组合）都表示过去时态，那也不对。“我会永远爱你的”表述的却是对未来的承诺。

自然语言信息处理（更具体地说，就是机器翻译）是当代电子计算机在非数值领域的最早应用，已经有 50 多年的历史了。然而，无论同计算机科学技术本身一日千里的发展相比较，还是同计算机在各个领域的成功应用相比较，自然语言处理技术的发展都是相当缓慢的，历经坎坷，至今未能取得重大突破。综合上面的分析，概括地说，汉语的形态不发达，适用于自动分析的形式标记相对贫乏，自动分析的难度绝不会比其他语言更低。笔者也认为中国学者研究汉语信息处理具有天然的潜在优势，特殊的困难也许为中国学者留下了更有广阔的发展空间，汉语理解的研究也许能为解开“人类智能本质”这个世界性难题做出贡献。

4．语言知识库的重要地位

认识到汉语理解研究的困难，就需要把汉语信息处理研究看作是一项长期的任务，不宜期望一蹴而就。为了实现良性循环与可持续发展，在今后的五到十年内，比较现实的技术路线是将自然语言处理研究作为语言工程来实施，必须面向应用，争取尽快为社会做出贡献，从而得到回报，继续为汉语理解研究提供支持，向最高境界前进。

既是工程项目，就需要有规模的控制，“受限汉语”[4]还是一个值得探讨的题目；也要有质量的指标，要有检验措施。笔者曾组织过 863 项目的评测，现在接受 973 项目组织的评测(姚天顺教授主持)。应该说评测促进了研究。当然，评测的规模有待扩大，其规范性、权威性、公开性也有待进一步提高。

尽管是工程项目，在实施过程中也要促进理论研究。对理论研究成果也要有检验手段。只是自我欣赏或相互恭维不会给理论研究的进步带来任何好处。

在语言工程实施过程中，要重视人才培养。知识经济的竞争归根到底是人才的竞争。

任何豪华、壮观的建筑都是建立在地下的坚实的基础上的。语言工程也有应用与基础之划分。由于语言工程的价值体现在它的应用上，人们重视应用研究是理所当然的。不过，在20 年的发展过程中，汉语信息处理基础研究的薄弱制约了应用研究的发展这一事实已经为越来越多的学者所认识。

基础研究要做的事情很多。在语言工程中，笔者认为最重要的基础研究是语言知识库的建设。为了提高语言信息处理系统的智能水平，最容易想到的就是给计算机装备足够庞大的知识库，知识库包括各种形式和内容的机器词典、规则库、语料库等。假设，机器词典中收录了“籽瓜”和“过渡带”这两个词，对第 2 节中的例句，至少机器实现正确切分的可能性就存在了。不过，最容易想到不等于最容易做到。因此，自然语言处理系统的研究者应当把语言知识库作为自然语言处理系统的基础设施，下大力气建设好，因为语言知识库是自然语言处理系统的必要组成部分，语言知识库的规模与质量是自然语言处理系统成败的关键。

笔者自 1986 年北大计算语言学研究所成立以来，就与全所同仁一道为建设语言信息处理综合知识库而努力。十六年来，积累了一些成果和经验。

《现代汉语语法信息词典》是北大计算语言所综合语言知识库的第一块基石[5]。这部电子词典的研制历史已有 16 年。收录词语超过 7.3 万。依据语法功能分布，建立了词语分类体系，完成了这 7.3 万个词语的归类。在分类的基础上，更进一步按类详细描述每个词语的多种语法属性。朱德熙先生的词组本位语法体系对本词典的研制起了指导作用。因为词典中描述的语法属性基本上就是词语之间的组合关系以及词语担当句法结构中的成分的能力。笔者之所以首先研制这部语法知识占主体的电子词典，是由应用系统开发的需求驱动的。足够大的规模、合理的结构、丰富的信息、准确的描述、广泛的适用性都是这部词典影响日益扩大的内在因素。

在《现代汉语语法信息词典》的基础上，北大计算语言所又着手大规模标注语料库的建设[6]。到 2002 年 2 月底完成 2700 多万字的语料的切分和标注，其中包括 1998 年全年《人民日报》。标注集除了《现代汉语语法信息词典》中的 26 个词性代码外，还包含人名、地名、团体机构名称等专有名词标记；对语素 g 划分了子类，如 Ng, Vg, Ag；对动词和形容词，标示了他们的名词用法和副词用法；总共约有 40 个标记。除了在词语的层次上进行标注外，还对短语型的地名、团体机构名称也加注了特别的符号。这个语料库是一个现代汉语语言知识的宝库。

《现代汉语语法信息词典》与大规模标注语料库相结合，又得到新的有价值的资源。如带词性的词频统计可以填补汉语学界的空白。进而可以将词语的各种语法属性值从定性的“可否”型改造为定量的概率型[7]。

当要求提高对语言信息处理的智能水平时，必须将词语层次的直接匹配与变换提升到概念的层次。基于概念的文献检索与信息提取就需要一部反映同义关系、反义关系、上下位关系、部分-整体关系、成员-群体关系等内容的中文概念词典（Chinese Concept Dictionary, CCD）。国际上已经有了这种架构的在线词典 Wordnet。开发 CCD 应当保持同 Wordnet 兼容，这样既可以参照已有成果，避免重复，还可以为跨语言的信息处理架设桥梁。北大计算语言所正在开发这样一部 Wordnet-like Chinese Concept Dictionary[8]，现在已取得阶段性成果。

除了上述成果外，北大计算语言所的语言知识库目前还包括面向汉英机器翻译的语义词典、汉语短语结构知识库、不同级别对齐的英汉双语语料库等资源，也包括为构造知识库所开发的工具软件，如词语切分与词性标注软件、语料库精加工软件、自动注音软件、CCD 可视化辅助开发软件。很多工具软件都有独立的应用价值。

北大计算语言所在从事包括语言知识库建设在内的基础研究时，始终既注意把握基础研究的内在发展规律，又注意顺应科学技术的历史潮流，满足客观需要。以《现代汉语语法信息词典》为主体的系列成果已转让到世界各地：美、德、法、日、韩、瑞典、新加坡、香港、台湾以及境内。包括 Microsoft, IBM, Intel, Xerox, Fujitsu, Toshiba, Matsushita, NTT, Canon, Sail-Labs（德国）, Enpia（韩国）,联想，青鸟等 IT 界著名企业在内的约 50 多所大学、研究所和公司在使用北大的这些成果。

作为综合型语言知识库，当然还有很多工作要做。北大计算语言所作如下规划：

（1）语言单位的多样化。以词为基础，向短语与语素两个方向扩展。句子与篇章的知识也是需要关注的。

（2）语言知识的多样化，由句法知识向语义知识、概念、语用知识和构词知识等多方向辐射。

（3）语种的多样化，由单语言（汉语）向多语言发展。除英语外，考虑到为“数字奥运”服务，也要适当扩充其他语种。

（4）领域的扩充。除包括常识等通用语言知识外，还要增加专业领域知识。专业术语库是必要的补充。目前首先考虑信息科学技术领域。

（5）开发方法的多元化。专家知识是必须依赖的。大规模深加工的语料库将成为获取语言知识的新源泉。建立单个知识库之间的连接，形成一体化的知识库。

（6）支援应用系统的开发以检验知识库的适用性和质量。目前注目于机器翻译系统和信息提取系统。开发应用程序接口，优化查询、统计等应用界面，让更多的人使用并检验这个知识库。

（7）以知识库为基础，探索、创建新的语言模型。可以在可靠的语言资源的基础上验证结合统计方法与规则方法的概率语法模型。

5．结语

我国西部大开发的战略已经开始实施。2001 年，我国成为国际贸易组织 WTO 的正式成员，我国申办 2008 年奥运会也获得成功。中国社会与国际接轨的信息化进程明显加快，对语言信息处理技术提出了强烈的需求。我国著名语言文字专家、97 岁高龄的周有光先生对汉语、汉字在新世纪的发展寄予热望。他在 20001 年 9 月份的《中国语文现代化学会通讯》上发表文章，认为“21 世纪，华语（笔者注：周先生指的就是‘汉语’）将在全世界华人中普遍推广”。同月在南京召开的“首届华文传媒论坛”也认为，“中文极有希望成为世界上第二大媒体语言”。在这样的形势下，献身汉语信息处理事业已不再局限于中华儿女的民族自豪感和拳拳爱国心，潜在的巨大经济利益已经驱使 IT 行业的众多跨国公司跻身这个新的热门研究领域。

由于语言信息处理技术需要语言学、数学、认知科学、计算机科学等多学科的相互融合，现在最缺乏的是文理兼通的人才。北京大学计算语言学研究所除了从事语言信息处理的基础研究和应用研究[9]，也在人才培养方面作了一些工作。但在汉语信息处理的宏伟事业中，北大计算语言学研究所的工作只是沧海一粟。国家“十五”计划执行伊始，教育部和国家语委组织制订并论证今后五到十年语言文字应用研究的规划，确实是及时的。北大计算语言所非常高兴能使自己的局部研究融入总体规划，并在主管部门的指导下为规划的实施奉献绵薄的力量。

参考文献

[1] 袁贵仁，以规范标准建设为核心，开创语言文字应用研究新局面，《语言文字应用》，2001

年，3 期，3-8
[2] 朱德熙，汉语语法讲义，北京：商务印书馆，1982 年
[3] 陆俭明，汉语句法成分特有的套叠现象，《中国语文》，1990 年第 2 期
[4] 俞士汶、朱学锋，受限汉语研究的必要性， 王均主编《语文现代化论丛（第三辑）》，150-160
[5]、王惠，《现代汉语语法信息词典》的新进展，《中文信息学报》，2001 年，第 1 期，59-65
[6] 俞士汶、段慧明、朱学锋等，大规模标注汉语语料库开发的基本经验，ICCC2001 主题报告（新加坡），Proceedings of ICCC2001, 56-60
[7] 俞士汶、段慧明、朱学锋, 汉语词的概率语法属性描述，《语言文字应用》，2001 年，3 期，21-26
[8] Yu Jiangsheng, Yu Shiwen, Liu yang, Zhang Huarui, Introduction to CCD, Proceedings of ICCC2001, 361-366
[9] 俞士汶, 计算语言学的应用研究与基础研究，见《中国中文信息学会二十周年学术会议》，北京：清华大学出版社，2001 年 11 月，54-65

2002 年 2 月 10 日第一次寄稿
2002 年 2 月 19 日第二次寄稿