计算语言学的应用研究与基础研究*

俞士汶

北京大学计算机科学技术系
北京大学计算语言学研究所
北京，100871
Email: yusw@pku.edu.cn

摘要：本文回顾北大计算语言学研究所的发展轨迹，探讨计算语言学研究中应用研究与基础研究的关系。本文也概要介绍了北大计算语言学研究所在中文信息处理领域所取得的一些基础研究成果以及正在做的和准备做的一些研究工作。

关键词：计算语言学，应用研究，基础研究，现代汉语语法信息词典，汉语标注语料库，中文概念辞书

Applied Research and Foundational Research in Computational Linguistics

Yu Shiwen

The Institute of Computational Linguistics,
Peking University, Beijing, 100871
Email: yusw@pku.edu.cn

Abstract: Reviewing the developing history of the researches conducted in the Institute of Computational Linguistics of Peking University (ICL/PKU), this paper discusses the relation between applied research and foundational research in Computational Linguistics. This paper briefly introduces not only the achievements in foundational research of Chinese information processing obtained by ICL/PKU, but also the projects which ICL/PKU is doing and plans to do.

Keyword: Computational Linguistics, applied research, foundational research, Grammatical Knowledge-base of Contemporary Chinese, Chinese Tagged Corpus, Chinese Concept Dictionary

---

* 有关研究得到 973（G1998030507-4）、国家自然科学基金（69483003）、985 等的支持

作者信息：俞士汶，男，1938 年 12 月生，教授，主要研究方向：计算语言学

About Author: Yu Shiwen, Male, born in Dec. 1938, Professor, interesting in Computational Linguistics.

一.引言

今年是中国中文信息学会成立二十周年，同时也是其二级学会计算语言学专业委员会成立十五周年，对于全国从事语言信息处理研究与开发的学者，自然是欢庆丰收的节日。北京大学有个小小的计算语言学研究所，今年恰好也满十五周岁。笔者自 1978 年参加计算机-激光汉字照排系统(748 工程)原型机研制起，20 多年来一直在中文信息处理这块园地中劳作。北大计算语言学研究所在这个领域的应用研究和基础研究方面也取得了一些成果。本文回顾北大计算语言学研究所的发展轨迹，探讨计算语言学研究中应用研究与基础研究的关系。期望本文能起到交流心得的作用，也算是献给中国中文信息学会成立二十周年的一份薄礼。

为了简洁，本文有时用“ 计算语言所 ”替代“ 北京大学计算语言学研究所 ”。

二.应用研究与基础研究的关系

任何一门学科都有基础研究，也有应用研究，也都需要处理好基础研究和应用研究的关系。作为一个普遍性课题，很多人都研究过它，真知灼见自然不少。关于这个问题的一些思想、见解和主张会对国家、一个部门或一个单位的宏观决策产生影响，自然也受到人们的重视。不过，笔者只是基层单位的一名普通教学研究人员，只有相对的自由去选择做什么，不做什么。因而，笔者以为，自己的经验如果对情况相似的学者、年轻人有一些参考价值，也就心满意足了。

2.1 应用研究带动基础研究

笔者的早期学术经历主要是研制国产计算机的系统软件，特别是操作系统[1]。1986 年自主选择转向计算语言学研究时，曾对自己的优势与劣势作过分析。数学功底好，熟悉计算机语言，了解形式语言理论与编译原理，喜欢探究语言现象和语言学问题，这些是优势。但当时，除一两名合作伙伴外，硬、软资源一无所有。研究工作从何做起，是一个十分现实、十分严峻的问题。

进入“ 计算语言学 ”这个新领域，要了解其基础理论、技术发展脉络和研究热点，自然有大量的专著和文献需要阅读。1979 年自己正忙于开发汉字照排系统中的分时操作系统，抽空去听了并无直接关系的“ 自然语言理解 ”讲座，当时决没有想到竟为 7 年后选择新的研究方向作了铺垫。真可谓“ 学海无涯 ”，学问和知识永远是有用的。不过，1986 的情况就不同了。需要尽快找到切入点。实际上，笔者及其同仁独立从事自然语言处理研究是从开发应用软件开始的。1986-1987 年间北大同日本松下公司合作开发中文文字处理机（Word Processor）。北大的主要任务是开发一个固化在文字处理机里的中文输入软件。当时中文输入技术的主流是以字为单位，汉字-代码对照表中包含一些词只能算是点缀或词处理的“ 萌芽 ”。正是由于笔者当时开始了计算语言学研究，在松下公司硬件优势的基础上，自然地想到将语言分析技术引入中文输入软件。结果，计算语言所开发了一个以词为基础、以句子为变换单位的中文输入软件[2]，一次变换正确率就达到 88%以上。无论设计思想，还是技术实现，当时都是很先进的，并对以后的智能输入软件的开发产生了广泛的影响。

1986-1990“ 七五 ”期间，笔者又主持了“ 七五 ”攻关项目“ 机器翻译自动评估软件 ”

的研制，计算语言所还参与了“ 日汉机器翻译系统 ”等的开发。

上述这些应用研究都取得了评价甚高的成果，发展的潜力也是有的。不过，笔者及其同仁在其后的相当长的一段时间内并没有去积极发展这些应用软件，而把主要精力转向计算语言学的基础研究。笔者之所以会作出这样的决策，是因为通过开发应用系统更深刻地认识到基础研究的重要性与迫切性。

在中文输入软件[2]中，将包含大量汉语语法知识的词典和语法规则库引入代码-汉字变换过程是该软件的创新点，可是计算语言所当时做这件事却是打了一场“ 遭遇战 ”，完成的是“ 急就章 ”。事后痛感要想在计算语言学领域有所作为，必须系统、深入地学习语法知识并将人能理解的知识转化为计算机可操作的数据格式，有关机器翻译的研究当然更加深了这种认识。应用研究的实践所提出的需求正是研制《现代汉语语法信息词典》[3,4]的原动力。

北大计算语言学研究所的其他基础研究也都是在应用研究的带动下进行的。

2.2 基础研究的内在发展规律

应用研究驱动下的基础研究的目的明确，成果适用性强。不过，既是基础研究，就一定要考虑成果的广泛适用性，不宜迁就个别应用系统的特殊需求。基础研究需要坚持相对的独立性。

撰写《现代汉语语法信息词典》的规格说明书时，把它的广泛适用性放在首要考虑的地位，对自然语言处理有用的语法知识都要尽可能充分地发掘出来，把它纳入其中。拼音-汉字转换程序需要词语的拼音，通常开发面向文本的机器翻译系统时，并不需要，不过将来集成了语音识别和语音合成的口语翻译又缺不了拼音。动词和形容词的重叠信息，像AABB（如：慌慌张张），A 里 AB（如：慌里慌张）等，拼音-汉字转换程序和汉外机器翻译系统的汉语自动分析程序都是需要的，但外汉机器翻译系统中的汉语自动生成程序就不一定需要。当形式划一的语法规则还没有能力描述远距离依赖关系时，介词库中的“ 后照应词 ”与“ 后照应类 ” 这两个属性字段也许发挥不了作用，但是针对性强的介词短语识别专用程序就可以实现它们的价值。《现代汉语语法信息词典》中有那么多词语语法属性信息（仅动词就有 100 多项语法属性），并非每个应用系统全都需要，但应用系统都可以从中裁剪出所需要的信息。《现代汉语语法信息词典》的成功也正在于此。词典总体结构的设计和数据格式的选择当然也考虑了词典的通用性。在大致分类的基础上分类描述每个词语的详细的语法属性信息的总体构思适应了语言信息处理技术的发展，因而使《现代汉语语法信息词典》保持了长远的生命周期。《现代汉语语法信息词典》是计算语言所奉献给语言信息处理大厦的第一块基石。

基础研究也要上规模。像电子词典、语料库这样的基础设施建设，必须有足够大的规模才有意义。计算语言所于 1986 年开发第一部语法电子词典时，就将词语总数定在 4 万[2]。《现代汉语语法信息词典》由 5 万多词语发展到 7.3 万词语，也是实际应用单位提出的需求。

基础研究有其内在的发展规律。当《现代汉语语法信息词典》基本成形并产生效益之后，研究工作该向何处发展？考虑到《现代汉语语法信息词典》中的语法属性字段的值的类型多数为“ 可否型 ”或“ 是非型 ”，在此基础上建立语法规则，仍然难免“ 说一不二 ”或“ 非此即彼 ”， 缺乏柔性。实际取值也主要依赖专家的知识，真实语料的例证相对地少了

一些。1990 年代，语料库语言学迅速成长，如果能以大规模真实语料的统计数据为依据，用概率值重新描述词的语法属性，就可以为语言信息处理建立新的语言模型（如规则与统计相结合的概率语法模型）。因此，从 1998 年起计算语言所开始探索词的概率语法属性描述,并承担了国家社科基金“ 九五 ”语言学科重大课题之子课题“ 词的语法属性描述研究（97@yy001-6）”的研究。这项研究必须以大规模深加工的语料库为基础，因此，将大规模语料库的加工提上议事日程。当然，大规模标注语料库的用途不限于此。当Internet迅速扩张、网上信息像潮水般涌来时，原来基于字符串匹配的全文检索技术已不适应新型搜索引擎更关注查准率的需要，以对中文文本进行词语切分与词性标注为前提开发搜索引擎成为人们新的追求。实现这个技术的一个途径是准备好大规模的已经切分好与标注上词性的语料作为样本供机器学习。总之，无论对计算语言学还是对传统语言学的研究，大规模标注语料库都是极为宝贵的资源。

在有了具有相当规模的标注语料库之后，对词的概率语法属性描述这个课题进行了实验性研究，已取得一些有意义的成果[5]。

到 1990 年代后期，计算语言所开发大规模标注语料库已有了很好的基础，一是有了《现代汉语语法信息词典》，二是有了一个性能优越的词语切分与词性标注软件。

从 1999 年 4 月起，计算语言所与FUJITSU合作，开始对 1998 年全年《人民日报》的语料进行词语切分和词性标注的加工（计 2600 万字，到 2001 年 10 月底完成 10 个月的、约 2100 万字的工作量）。这件事本身也是一项重要的基础研究。这项成果是计算语言所奉献给语言信息处理大厦的第二块基石[6,7]。

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辞书（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。但是，这种架构的词典所反映的知识体系一定存在与民族、社会、文化、语言相关的部分，因此，中文概念辞书CCD也不可能是Wordnet的直接翻译或汉化。况且，计算语言所的《现代汉语语法信息词典》、《现代汉语语义词典》、大规模标注语料库也为开发CCD提供了丰富的资源[14]。

当然，基础研究的内在发展规律也需要顺应科学技术的历史潮流，既要善于利用已经具备的客观条件，又要能促进应用系统的发展。否则，基础研究不可能有生命力，也不可能进入良性循环。

2.3 基础研究课题的选择

自然语言处理的最高境界是让计算机“ 理解 ”人的语言，实际上这是在探索人类本身的智能的奥秘，其难度是可想而知的。计算语言学是自然语言处理的指导理论，基础研究课题很多，如语言模型、语法体系、分析算法、语言习得与认知模型、知识表示与知识获取等等。无论个人还是一个小的研究集体，都不可能全面出击。笔者 15 年来选择基础研究课题，主要考虑以下因素。

（1）以汉语为主体。这似乎是不言而喻的，因为自己是中国人。正因为是中国人，往往

过高估计自己对汉语的认识与掌握。国外语言学理论与语言信息处理技术发展十分迅速，要学的东西太多了，把每种新的理论或技术拿过来，将对象换成汉语试一试，都会有心得，都有文章可做。但这样即使“接轨”，也只是“跟踪”，很难有真正的“创新”。站在计算语言学理论与方法的高度，对汉语语言事实进行周密的调查、总结，这件事迟早是要做的。在文化艺术领域，人们承认最有民族性的作品也最有国际性。笔者相信，在语言信息处理领域也应作如是观。事实证明，坚持以汉语为主体，这条路是走对了。在实践中，笔者对汉语分析的特点有了更深切的把握[6]，基础研究才会有的放矢。

（2）发挥交叉学科的优势。北大具有文理结合的综合优势，适合开展计算语言学的基础研究。朱德熙、陆俭明两位先生有敏锐的洞察力，率先与计算机学科结合，更是计算语言所难得的机遇。文理结合的优势，很多人都是认识到了的。但不同学科的交叉必须是深入的。来自不同学科的专家只要多进行一些交流，就有可能撞击出新思想的火花。但要将思想、理念变成物化的成果，则一定需要有相当数量的某个学科的专家深入了解并掌握自己原来不懂的另一学科的知识、视点、方法而形成新的特长与优势。笔者的学术背景是数学与软件。偏离信息技术的主流，长期地、深入地学习相对冷僻的语言学科的知识，对于原本属于计算机学科的人来说，也是一项相当理性的选择。计算语言所选择的研究课题不仅能够发挥文理两方面人才的优势，更要着眼培养文理兼通的两栖人才，因而研究队伍不断壮大，人才素质不断提高。也就是既出成果，又出人才。

（3）选题要有前瞻性。从事基础研究的学术带头人，特别是在信息技术领域，在中国当前环境中，不宜追逐已经热起来的潮流，而要看得更远些，更深些。进行基础研究要甘于坐冷板凳，要坚持厚积薄发，不宜急功近利。计算语言所的很多工作都是 10 多年前规划并开始研究的，在相当长的时间内并没有见到效益，但全所同仁坚持 10 多年，锲而不舍，终于迎来了受学术界和产业界认可的大好局面。

（4）处理好同应用研究的关系。尽管坚持基础研究的相对独立性，但笔者和计算语言所的同仁从未脱离过对应用系统的探索。博士生、硕士生作学位论文，他们在研究中也构造了多个不同规模的应用系统，计算语言所还同科学院计算所（刘群副研究员）合作，共同开发汉英机器翻译系统，已有８年之久；这两年又联合清华大学（孙茂松教授）共同承担了国家 973 项目中的汉英机器翻译研究任务。近几年还开展了信息提取（Information Extraction）研究，这项研究得到国家自然科学基金和北大-IBM 创新研究院的支持。正是在应用系统开发的实践中，可以发现应用系统所需要的共同的基础。这样的基础研究成果才是适合客观需求的。

（5）要善于扬长避短。在众多的基础研究课题中，既有轻重缓急的区别，也有是否适合自己做的问题。计算语言所 10 多年前选择《现代汉语语法信息词典》，正是充分估计了语法研究在语言信息处理领域中的全局地位以及朱德熙先生的词组本位语法体系的成熟程度和广泛影响。10 多年前，已有人强调语义研究在汉语信息处理中的重要性，但笔者认为当时条件尚不具备，计算语言所仍选择系统的语法研究作为突破口。

三.北大已取得的若干基础研究成果

经过 15 年的积累，计算语言所在语言信息处理领域取得了如下成果：

（1）现代汉语语法信息词典

（2）包含 600 多条语法规则的现代汉语短语结构知识库

（3）100 万字的切分、标注、注音语料库

（4）包含数万个句对的英汉对照双语语料库

（5）词语切分与词性标注软件

（6）文本注音软件

（7）中国古诗词计算机辅助研究系统

（8）包含 5000 多条目的“英日德汉对照计算语言学术语库”

以上成果的知识产权都属于北京大学。另外，与中科院计算所共享知识产权的成果有：

（9）包含 4.9 万词语的现代汉语语义词典

（10）在 1998 年 863 组织的评测中译文质量名列前茅的汉英机器翻译系统

与 Fujitsu、《人民日报》社共享知识产权的有

（11）现代汉语标注语料库

正在研制的还有“中文信息提取系统”、“中文概念辞书”等。

现在，这些成果已传播到世界各地：美国、日本、德国、法国、韩国、新加坡、瑞典以及中国的香港、台湾和境内。包括 Microsoft、Xerox、Intel、IBM、日本 Fujitsu、Matsushita、NTT、Toshiba、Canon、德国 SailLabs、韩国 Enpia、国内联想、青鸟等著名企业在内的 50 多个大学、研究所和公司都采用了北大的成果。应用系统涉及的语言除汉语外，还包括英语、日语、德语、法语、韩国语、蒙古语和藏语。这些成果在语言信息处理领域产生了广泛的影响，取得了明显的社会效益和经济效益。限于篇幅，这里不能一一介绍。只介绍如下成果的概要，贯穿其中的是关于基础研究的一些认识。

① 现代汉语语法信息词典[3,4]

这部电子词典的开发历史已超过 15 年。收录的词语（entry）超过 7.3 万。如果将 1 个词语的一个语法属性信息定义为一个信息单位，词典的总信息量约为 350 万个单位，可见规模之庞大。本词典的研制虽然是一项语言工程，但也涉及汉语语法的一系列基本理论问题：词语观、词类观、兼类处理策略[15]、语法功能界定与词语的语法属性确定等等。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。《现代汉语语法信息词典》的研制者将词组本位语法关于汉语语言单位的划分、词类划分的本质依据、复合词与短语的构造规则、词类与句法成分的关系、句法结构与其构成成分之间的关系、短语与句子的关系等一系列经典论述[16]所阐述的原则全面贯彻到词典研制的全过程。由于主要研制者对语言信息处理技术的需求又有清晰的理解，并有开发应用系统的实际经验，因而能依据这些原则于 1990 年制订出词典的详细规格说明书，这个规格说明书在此后的十余年间基本没有变化，成为所有参与词典研制的研究人员共同遵守的规范。这个成功的环节也应归因于主要研制者长期开发系统软件的学术背景，充分认识到软件工程中的前期工程

——需求分析的重要性。在词典研制过程中，坚持了如下原则：通用与专用相结合，以通用为主；词语分类与词语属性描述相结合，以属性描述为主；以句法知识为主，适当结合语义知识；依靠专家知识与利用计算机辅助获取语言知识相结合，主要依靠专家知识；在软件技术方面采用通用的成熟的关系数据库管理软件作为支撑软件。正因为有坚实的理论基础，又坚持了这些合理的原则，尽管词典规模庞大，开发周期长达 10 余年，人员也有变动，但整个研制过程却是有序的，基本上没有出现反复。不仅《现代汉语语法信息词典》这一成果本身成为计算语言所开展语言信息处理研究的基础，而且在研制过程中所积累的经验也在计算语言所此后的研究项目中发挥了重要的借鉴作用。成功固然带来欢乐和满足，过程也经常给人以喜悦和充实。科学研究的过程是长期的，只有进入“过程重于成果”的境界，生活才永远充满朝气和喜兴。

关于这部词典在语言教学中的应用也在探讨中[8,9]。

② 现代汉语词法分析器

这里介绍的“现代汉语词法分析器”包括“词语切分与词性标注软件”[18]和“精加工软件”两个分立的软件。

纵观汉语信息处理的全过程，除了那些立足于字符串匹配的技术外，凡涉及语言的深层分析，通常都包含词语切分与词性标注这个步骤。即使不与涉及句法、语义、语境的分析器相衔接，分立的“词语切分与词性标注”软件也有广泛的应用领域。正因为看到这个前景，从 1992 年起笔者就安排研究生开发“词语切分与词性标注”软件。当时，多数研究是把“词语切分”和“词性标注”分成两个过程。笔者考虑到计算语言所已有数万词的语法信息词典可以利用，便提出了“词语切分”和“词性标注”一体化的要求。因而，从 1993 年起计算语言所就有了这样一个基本的工具软件，经过若干届博士研究生的改进（其中孙斌的贡献是关键的），并与计算语言所相关的研究工作同步进化，该软件逐渐成熟。现在该软件不仅增加了人名、地名等专有名词识别和短语型机构名称识别等功能，还加上了注音功能，有兴趣的读者可以访问本所的主页，实际测试该软件。该软件不仅在本所的研究工作中发挥了重要作用，而且也同《现代汉语语法信息词典》一样传播到了世界各地。

笔者通常回避直接回答关于这个软件的精度的提问。同是中文，有些文言文，笔者也不容易读懂。读不懂与不会正确切分有很大关系；碰到某些翻译文章，可能要几个来回才能正确切分，也才能理解。当然这些是极端的例子。不过，这些事实说明了精度与对象（实际文本）是密切相关的。另一方面，说这个软件在完成全年《人民日报》语料标注这个大规模的语言工程中承担了 96%以上的工作量，则一点也不夸张。这也是计算语言所勇于承担这项工程的基本条件之一。

由于语料库加工的工程浩大（2600 万汉字），即使不到 4%的错误若全部依靠人工校对，工作量也是难以承受的。继续提高通用切分标注软件的精度不仅受到限制，也要降低速度等其他性能指标。而速度快正是计算语言所切分标注软件的优良性能之一，有了这个优点，该软件被很多检索引擎的开发者看中。

正像没有包治百病的药方一样，也不应期望一个软件解决所有的问题。语料库标注工程的主要技术负责人段慧明及其合作伙伴在长期的实践中，摸清了自动处理产生的切分标注错误的分布规律，理解并利用《现代汉语语法信息词典》中所包含的知识，开发了专门

订正错误的“精加工软件”，使得自动加工的结果可以达到请人“欣赏”的境界，从而大幅度减轻了人工校对的工作量和劳动强度，保证了工程的进度和质量。这个软件的成功开发给人以重要的启示：对科学研究最投入的人是最有收获的人，也是成长得最快的人。

③大规模现代汉语标注语料库

计算语言学的一个重要分支——语料库语言学在 1990 年代得到迅速的发展。语料库语言学的要旨是让计算机从急速膨胀了的大规模真实文本语料中直接学习到自动处理语言信息的知识。例如，对大量语料进行简单的统计，就能得到“汉字使用频度”这样的知识。不过，原始语料的利用价值或者说无指导的机器学习的潜力是有局限性的。如果在原始语料中预先注入一些语言学知识，譬如，对一部分语料进行词语切分和词性标注的处理，然后再将加工好的语料作为训练样本提供给计算机，计算机就会学得更多更好，也能掌握自动切分与自动标注的本领。另一方面，利用标注好的语料就可以统计带词性的词频、同形异类词的分布等一系列过去不可能得到的、应用价值更高的数据。因此，深加工的语料库成为计算语言学的基础资源，而语料库的深加工技术也成为计算语言学的研究热点之一。

计算语言所与Fujitsu合作研制的《人民日报》标注语料库是当今世界上规模最大的中文标注语料库[6,7]。如果不作加工，现在建几十亿字甚至更大的原始语料库并不困难。如果只是全自动加工，对加工的精度不作苛求（不要求严格的校对，实际上就是不要求注入专家的知识），建立不同层次的深加工语料库也不困难。因此，加工的精度才是衡量语料库的价值的最重要的指标。

目前，《人民日报》标注语料库的加工只包括词语切分和词性标注，词性标注也只使用 40 多个标记，标记集不算大。不过，该标记集是在《现代汉语语法信息词典》这个基础成果的基础上制订的[7]，《人民日报》标注语料库和《现代汉语语法信息词典》可以建立直接的联系，由此提供的信息远比分立的语料库或词典都要丰富得多。另外，这个语料库还标注了专有名词，包括短语型专有名称，为人名、地名、机构名称这一类未定义词的自动识别提供了丰富的知识（构造知识及语境知识）。总的来说，现在的加工层次还不算深，但研究工作应当深入。1995 年，计算语言所曾同新加坡国立大学配合，开发了一个小型的汉语树库[19]。正因为有了这样的前瞻性研究，在对大规模语言工程进行决策时，才能做到胸有成竹，制订切分标注规范时，才能做到取舍得当。

《人民日报》标注语料库最显著的特点还是其卓越的质量。计算语言所每完成《人民日报》一个月语料（约 200 多万字）的加工任务，就将结果交付 Fujitsu 验收，若合格，Fujitsu 出具验收报告给计算语言所。前 8 个月的验收报告表明，每个月的错误率（指不符合《规范》的比率）不超过千分之二。在前 6 月的全部结果出来之后，另外又花了大量时间进行一致化处理，进一步减少了错误。现在公开的《人民日报》1998 年前半年 1300 多万字的标注语料库就是这样一个高质量的语言知识库。

为了保证加工质量，首先制订了《现代汉语语料库加工——词语切分与词性标注规范》[7]。本《规范》具体、明确，易操作。本《规范》是在词组本位语法理论体系指导下、吸收语料库加工的长期实践经验[18,19]、集思广益、经反复推敲而制订的。《规范》既是发展“词语切分与词性标注软件”的需求说明，又是人工校对的准则。整个工程历时 3 年，人员变动甚多，但由于《规范》稳定，工程质量就有了基本保证。质量保证的第二个要素是

人员的素质。计算语言所的骨干已有 10 余载从事语言信息处理研究的功底。校对人员至少有语言学研究生的水平，并经过严格的培训。质量保证的第三个要素是自动化程度的不断提高。研制《现代汉语语法信息词典》，计算机起了辅助作用；而标注语料库的制作，计算机则要起关键的作用。一方面，不进行人工校对，即不投入专家知识，要建成高质量的标注语料库会陷入空谈，另一方面，如果过分依赖人工校对，则不仅投入过大，不堪重负，而且大量的手工操作必然有随机性错误，反而难以消除。因此，又要努力使人工干预减到最少。②中介绍的“精加工软件”的开发正是向这个方向迈出的塌实的一步。

④古代诗计算机辅助研究系统

现代汉语与古代汉语有很大的不同，这是显而易见的，但汉语自古至今一脉相传，又是不争的事实。如何对现代汉语和古代汉语作定量的对比研究显然是一个饶有兴趣、且有深远意义的课题。计算语言所主要从事现代汉语信息处理研究，从语言学功底考虑，研究涉及历史语言学的问题确实没有现实的优势。不过，笔者笃信学海无涯，研究无限。况且，计算语言所掌握用计算机处理语言学问题的比较深入的理论和技术，这个优势同古代汉语研究相结合，应该能够迸发出新的火花。古代汉语研究与现代汉语语法研究的结合也会推动现代汉语语法研究的深入和语言信息处理技术的发展。考虑到，同古文相比，古代诗歌更接近当时的口语，因此，计算语言所首先选择唐宋诗作为研究对象。古诗研究也给工程技术主导的环境增添了诗情画意，对培养文理兼备的人才会有潜移默化的影响。

从 1993 年起，计算语言所一直有学生在这个全新的领域进行探索[20,21]。在近 8 年积累的基础上，胡俊峰博士建立了一个规模适中、包含诸多创新成果的“古代诗词计算机辅助研究系统”[11,22]。目前纳入该系统的唐宋诗语料达 640 万字。该系统的主要特点是实现了基于“词”的诗句或诗篇的处理（这里的“词”不是“宋词”的“词”，而是指汉语语法研究中的语言单位“复合词”中的“词”。过去写文章，常用英文“word”注释这个“词”，本文放弃这么做。原因是现在认为英文的“word”与中文的“字”有着更多的相似点。这个问题超过了本文的范围，暂且放下）。现有的中文古籍计算机处理系统，无论它界面上有多少功能（如：检索、统计、索引、缉佚、分类等），其内核程序采用的都是汉字串匹配技术。而计算语言所的系统却从按句连写的诗篇中分离出“词”。然后，以“词”为着眼点考察诗句、诗篇、诗人乃至朝代的特征。

首先遇到的问题是如何界定“词”？古诗中单纯词比比皆是和现代汉语中复合词大量出现都是平凡的事实。问题在于古诗中有没有复合词？在这个基本理论问题的引导下，胡俊峰的研究取得了如下成果：（1）提出用于界定复合词的“相对共现度”、“插入率”等概念，并建立了相应的计算公式，使词汇自动提取研究突破了仅依靠“频度”、“互信息”等常规统计数据的局限，为后续研究奠定了坚实的基础；（2）建立了唐宋诗的词汇知识库，除了词频、词性、注音等常规信息外，还包括词汇的对仗信息、共现信息、词汇-作者分布信息，词汇-时代分布信息等，这是中国古诗研究中的第一个这样的知识库；（3）在词汇知识库基础上建立的唐宋诗辅助研究系统所包含的像诗句相似性检索等功能是其他古文献系统所没有的；（4）还进一步探索了汉语构词规则的自动提取、词汇语义关系的自动发现、词汇的意象分类和词汇联想语义网络的构建等更深入的问题，在词汇语义统计分析的方向上迈出了第一步。有些结果是颇有趣味的。

可以看到，唐宋诗辅助研究系统所采用的若干关键技术对现代汉语也是适用的。例如，复合词自动提取技术可以借鉴于现代汉语半固定短语的识别，因而对专业术语的发现和专业术语库的建立有很大的参考价值。这正好回归到笔者与同仁决定开展古汉语研究的初衷。

四.北大正在做和计划做的工作

当然，对已有的成果还要继续发展。像《现代汉语语法信息词典》，还要不断提高质量。借清华大学出版社预约出版《现代汉语语法信息词典详解（第二版）》的良机，不仅全面修订了该书的正文，而且对书中所附的 1 万词语样例的全部语法属性逐一进行了复查，除订正讹误外，还补充了实例和说明，反映了计算语言所在词的语法功能研究领域取得的最新成果。这些成果无疑会进一步提高词典数据库的质量，还准备发展词典的数据结构，以便对每个词语的每个语法属性，都可以提供从《人民日报》标注语料库中选择的适当例句。到 2002 年 4 月，1998 年全年《人民日报》2600 万字语料的加工任务将按计划完成，中文信息处理学界将有一个更大规模的汉语标注语料库可以利用。完成这个语料库同《现代汉语语法信息词典》的连接，并利用这个语料库完善词的概率语法属性研究[5]。

由于计算语言所已建好了语法知识库的基础，今后会向语义倾斜。计算语言所将继续倾注热情与力量于中文概念辞书 CCD 的研制。笔者认识到凡涉及语义、概念，难度一定很大，必须谨慎从事。计算语言所从 2000 年初开始筹划这个项目，2000 年 9 月启动预备性工程，到 2001 年 4 月编制了中文概念辞书的规格说明书和 1600 个概念的样例，并在“第二届汉语词汇语义学研讨会（2001 年 5 月，北京）”上向海内外专家广泛征求了意见。后来又研制了辞书可视化辅助编制软件 VACOL(Visualized Auxiliary Construction of Lexicon) 。2001 年 7 月才正式进入规模开发阶段。预计到 2002 年 4 月会得到可供研究使用的成果。

以往研究关注的语言单位主要是语句。更大的单位，如篇章，还没怎么涉及。像“信息提取”这样的研究已对篇章信息提出了强烈的需求。王厚峰博士在指代消解方面的探索[10]在篇章处理方面开了一个头。除了代词有指代问题，实际上名词也有指代问题，还有省略的补足问题，总之，这方面需要投入力量继续进行探索。

正像一个有用的人才除了有语言能力外，还要有专业知识。计算机语言信息处理系统也需要通用领域与专业领域的互相配合。计算语言所的词语切分与词性标注软件在结构上配备了多部用户词典，但这些用户词典目前都是空的，需要用户自己用专业词典去填补。计算语言所的人员大部分是计算机专业的，应该有基础开发信息科学与技术领域的术语库，计算语言所即将开展术语自动提取研究，并准备按照 CCD 的框架组织计算机专业术语库。

显然，汉语和外语的对比研究对机器翻译和跨语言信息处理具有重要意义。计算语言所已经有了数万句对齐了的汉英双语语料库，正在建造数十万量级的汉英对应的短语库。

计算语言所将在自己开发的语言资源的基础上继续“汉英机器翻译”与“信息提取”等应用研究，并将应用研究提出的问题和需求不断反馈到基础研究中去。

计算语言所还将扩充、完善古代诗词辅助研究系统，继续注意将现代汉语的广度研究与贯通古今的纵深研究相结合。

最重要的，计算语言所将集成这些单项研究的成果，构造综合型语言知识库[12,13]。期望这个综合型语言知识库在语言信息处理的研究和汉语语言学的研究中发挥更大的作用。

五．结语

步入新世纪，北京大学计算语言学研究所迎来了自 1986 年成立以来的最好的发展时期。“ 一叶知秋 ”，正是社会信息化的浪潮与知识经济的躁动为计算语言学和自然语言处理技术开拓了广阔的发展空间。而坚持十余载的基础研究为现在的发展奠定了基础，创造了条件。不过，我们设想，在 15 年前，当“ 计算语言学 ”在中国还是一个相对生僻的技术术语的时候，如果没有朱德熙先生在北京大学创立了计算语言学研究所，再设想，15 年来，如果没有计算语言所这样相对宽松、优越的学术环境，计算语言所也许坚持不了如此长期、相对冷漠的基础研究。当然，国家有关科研项目和基金的支持，也为这些基本研究提供了最基本的保障。笔者藉此机会向所有支持北大计算语言学研究所的部门、单位、领导、师长和朋友表示诚挚的谢意。也愿与风雨同舟十余载的全所同仁在共享辉煌的同时继续互相鞭策，永葆学术青春。

北京大学计算语言学研究所毕竟只是一个小小的基层研究单位，力量、资源有限，中文信息处理的更多工作需要更多的专家，特别是青年学者的投入。笔者期望北京大学计算语言学研究所奉献的小小的浪花能汇入中文信息处理发展的长河。

参考文献

[1]杨芙清、俞士汶，操作系统结构分析，北京：北京大学出版社，1985 年

[2]俞士汶, 中文输入中语法分析技术的应用, 《中文信息学报》, 1988，2 卷 3 期, 20-26

[3]俞士汶、朱学锋、王惠，《现代汉语语法信息词典》的新进展，《中文信息学报》, 2001，15 卷 1 期, 59-65（封三）

[4]冯志伟、曹右琦，评《现代汉语语法信息词典详解》，《中文信息学报》，1999 年 1 期，封三

[5]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

[6]段慧明、松井久仁於、徐国伟、胡国昕、俞士汶，大规模汉语标注语料库的制作与使用，《语言文字应用》，2000 年 2 期，72-77

[7]俞士汶、朱学锋、段慧明，大规模现代汉语标注语料库的加工规范，《中文信息学报》，2000 年第 6 期，58-64

[8]亢世勇、朱学锋、俞士汶，《现代汉语语法信息词典》在计算机辅助语言教学中的应用，第二届中文电化教学国际研讨会论文集，P250-255，2000 年：广西桂林

[9]刘云、俞士汶、朱学锋，《现代汉语合成词语数据库的开发及应用》，第二届中文电化教学国际研讨会论文集，P273-278，2000 年：广西桂林

[10]王厚峰、何婷婷，汉语中人称代词的消解研究，《计算机学报》，Vol.24, No.2, 136-143,2001/7/28

[11]胡俊峰、俞士汶，唐宋诗之计算机辅助深层研究，《北京大学学报》（已录用，将载于 2001 年第 5 期）

[12]朱学锋、俞士汶，自然语言处理与语言知识库，见罗振声、袁毓林主编，《计算机时代的汉语汉字研究》，清华大学出版社, 1996 年, P107-118

[13]俞士汶、段慧明、朱学锋，综合型汉语知识库及其在汉语教学中的应用, 第四届全球华人教育资讯科

技大会主题报告，《Proceedings of GCCCE2000》, P12-19, 2000 年 5 月：新加坡
[14]Yu Jiangsheng, Yu Shiwen, Liu Yang, Zhang Huarui, Introduction to Chinese Concept Dictionary, Accepted by ICCC2001(Singapore)
[15]俞士汶、段慧明、朱学锋，语言工程中同形及兼类词语的处理策略，见黄昌宁、张普主编《自然语言理解与机器翻译》，北京：清华大学出版社，2001 年，P211-218
[16]朱德熙，《语法讲义》，北京：商务印书馆，1982
[17]周强、段慧明，汉语语料库加工中的切词与词性标注处理，《中国计算机报》1994 年第 21 期 85-87
[18]周强、俞士汶，一个人机互助的汉语语料库多级加工处理系统 CCMP，见陈力伟、袁琦主编《计算语言学进展与应用》，北京：清华大学出版社，1995 年，50-55
[19]周强、张伟，一个改进的汉语短语自动界定模型，中文电脑国际会议 ICCC'96（新加坡）论文集，75-81
[20]刘岩斌、俞士汶、孙钦善，古诗研究的计算机支持环境的实现，《中文信息学报》，1997 年第 1 期，27-36
[21]穗志方、俞士汶、罗凤珠，宋代名家诗自动注音研究及系统实现，《中文信息学报》，1998，第 2 期，44-53
[22]俞士汶、胡俊峰，唐宋诗之词汇自动分析及应用，已在台湾中研院第 3 届汉学会议上报告，2000 年 6 月，台北市，论文集在编辑中