北京大学语言知识库概况

俞士汶
北京大学计算语言学研究所，北京 100871，中国
yusw@pku.edu.cn

北京大学计算语言学研究所自 1986 年成立以来，在面向信息处理的语言知识库建设方面取得了如下成果：

（1）现代汉语语法信息词典
（2）大规模现代汉语基本标注语料库
（3）现代汉语语义词典
（4）中文概念词典
（5）英汉对照双语语料库
（6）信息科学技术领域术语库
（7）现代汉语短语结构知识库
（8）中国古代诗词语料库
（9）服务于语言知识库建设的各种工具软件

这些知识库都具有相当大的规模。像（1）《现代汉语语法信息词典》收词超过 7.3 万，在依据语法功能分布对 7.3 万词语进行分类的基础上，又逐类描述每个词语的详细的语法属性。像（2）“现代汉语基本标注语料库”的规模已超过 3000 多万汉字。这些成果也比较成熟，像《现代汉语语法信息词典》已经有了 17 年的历史。很多成果已经传播到世界各地。根据协议得到北大成果许可使用权的国内外单位约有 60 家。有兴趣者可在网站（www.icl.pku.edu.cn ）上自由下载相当一部分成果。

这些知识库之间有内在的联系和协调的分工。

《现代汉语语法信息词典》是北大语言知识库的第一块基石。“大规模现代汉语基本标注语料库”就是在它的基础上开发的。《现代汉语语法信息词典》和“基本标注语料库”主要描述汉语词语的语法知识。

面向机器翻译的（3）“现代汉语语义词典”和面向文本内容处理的（4）“中文概念词典”从不同侧面描述了汉语词汇的语义知识，这两部词典目前收入的词条（或概念）均超过 6 万。

跨语言处理是语言信息处理的重要方向之一。“现代汉语语义词典”和“中文概念词典”的每一个词条都有对译的英语词。（5）“英汉对照双语语料库”以更大的对译单位（文章、段落、句子、短语）覆盖两种语言。目前对齐了的英汉对译的句对已经超过 10 万。这些双语资源可以支持机器翻译和跨语言文献检索、信息提取等研究。

从成果（1）到成果（5）汇集的都是日常生活语言的知识。成果（6）信息科学技术

领域术语库则为计算机提供专业知识。

从成果（1）到成果（6）聚焦于词汇知识。成果（7）“现代汉语短语结构知识库”描述的则是句法结构知识，含 600 多条扩充的上下文无关句法规则。

从成果（1）到成果（7）都是关于现代汉语的。建设（8）“中国古代诗词语料库”的目的是开展古诗词计算机辅助深层研究，进行古汉语和现代汉语的对比研究。

北大计算语言所充分认识到建设语言知识库必须仰仗专家知识的注入。在接受中文系、英语系教授指导的同时，北大计算语言所也努力培养本所的文理兼通的人才，鼓励他们潜心投入语言知识库的建设。当然，辅助知识库建设的一系列工具软件的开发和运用同样是不可缺少的。成果（9）包含的“现代汉语词语切分与词性标注软件”、“现代汉语文本注音软件”、“双语语料库构建工具集”具有通用性。面向中文概念词典 CCD 研制过程的可视化词典辅助构造软件 VACOL 对研制 CCD 起了关键作用。

这些知识库所汇集的语言知识及其表述形式独立于特定的语言信息处理系统、语言理论和实现算法，使得这些知识库得以广泛传播。从方法论角度考虑，这些知识库的建设既采用并服务于基于规则的方法，也采用并服务于基于统计的方法。“大规模基本标注语料库”及其开发工具“词语切分与词性标注软件”是最典型的例证。

得到《汉语语言与计算学报》主编赖金锭博士的大力支持，这个专辑发表北大语言知识库的如下 5 篇规格说明书：“北大语料库加工规范：切分·词性标注·注音”，“现代汉语语义词典规范”，“中文概念词典规格说明”，“北京大学汉英双语语料库标记规范”，“现代汉语短语结构知识库规格说明书”。此前，《现代汉语语法信息词典》的规格说明书早在 1996 年第 2 期《中文信息学报》上发表。1998 年和 2003 年北京清华大学出版社还出版了专著《现代汉语语法信息词典详解》的第一版和第二版。《北京大学现代汉语语料库基本加工规范》在《中文信息学报》2002 年第 5，6 期连载。

衷心期望得到专家和朋友的指教，使北大语言知识库建设尽可能少走弯路。

谨向赖金锭博士和本专辑的客座编辑王惠博士致以诚挚的谢意。

Outline of the Language Knowledge-bases at Peking University

Shiwen Yu

Institute of Computational Linguistics, Peking University, Beijing, 100871, China

Abstract: *This paper provides a general picture of five large-scale language Knowledge-bases of Peking University, and essentializes a distinct but complementary relationship among them, as well as the technical principle for developing the Language Knowledge-base.*