现代汉语多级加工语料库简介

俞士汶　段慧明　吴云芳

北京大学计算语言学研究所

北京大学计算语言学研究所研制的现代汉语多级加工语料库包括5200万字的基本加工（词语切分、词性标注、命名实体标注、注音）语料库、2800万字的同形标注语料库，它们都是综合型语言知识库的组成部分。此外，还有56万字语料标注了并列结构。

语料加工必须遵循规范。在实施大规模加工之前，必须制定好规范。北京大学计算语言学研究所制定规范遵循两个原则：（1）吸收语言学研究的成果；（2）适应自然语言处理的需求。

《现代汉语语料库基本加工规范2001年版》连载于《中文信息学报》2002年第5，6期。此文获评中国科学技术协会优秀论文（参见中国科学技术协会奖状）。

《现代汉语语料库基本加工规范2003年版》发表于新加坡《汉语语言与计算学报》2003年第2期。2003年版与2001年版相比较，规范的基本内容未变，依据的语言学基础没有变化，只是细化了一些标记，并增加了注音。

这里发布的1998年1月份《人民日报》基本标注语料库（约200万字）遵循的是2003年版规范。

同形标注语料库就是在基本标注语料库的基础上增添“同形”信息，即词语的粗粒度义项的标注。这里发布的1998年1月份《人民日报》同形标注语料库中的“同形”信息源自《现代汉语语法信息词典》“同形”字段，用以区分同类词语中的不同读音、不同词项以及不同义项。

这里还发布了56万字的并列结构标注语料库及其《说明》，该语料库及其《说明》曾在北京大学计算语言学研究所的网站上发布过。

北京大学计算语言学研究所关于多级加工语料库的研制得到多个单位、多项基金的支持，并得到学术界多位同仁的协助。过往发表的论著、报告等资料反映了有关方面的贡献，并表达了诚挚的谢意，恕不在此重复。

附记：北京大学计算语言学研究所还有700万字的（细粒度）义项标注语料库（“义项”信息源自《现代汉语语义词典》以及树库等语料库资源，待发布。