现代汉语系统语言资源开发及相关研究

亢世勇
鲁东大学文学院
山东省语言资源开发与应用重点实验室

主要内容

- 现代汉语系统语言资源
- 1、《汉字义类信息库》（电子版）
- 2、《汉语语义构词信息库》
- 3、词网—《新编同义词词林》
- 4、《现代汉语新词语信息电子词典》
- 5、标注句子语义成分语料库
- 6、文本蕴涵信息库
- 7、汉语拼音词汇数据库

1、《汉字义类信息库》（电子版）

- 该信息库以国标6763个字形为基础，按照“一字一条原则、一义一条原则、语法意义原则、补充原则”将6763个字形分化为17430个字位，利用数据库软件详细描述了每一个字位的义项、读音、同形、同音、语义类、词性、成词与否等属性信息，标注信息采用简单明了的汉字、字母、数字，以数据库文件格式保存。该信息库是进行字义标注、字音标注、构词语料库标注的基本资源，也可应用于字义、字音、字形以及字与词之间关系的研究，还可以应用于汉字教学。

2、《汉语语义构词信息库》

- 以《同义词词林》为基础，结合《现代汉语词典》《新词语大词典》选取了52366个双音合成词，然后将《汉字义类信息库》中的信息用计算机给这些合成词中的每个字标注义类标记和简单释义，经过人工校对，建成大型的《汉语语义构词数据库》。

3、词网—《新编同义词词林》

- 面向信息处理使用，根据冯志伟教授的知识本体的思想及其构拟的ONTOL-MT语义分类体系为指导。
- 以普遍性原则、完备性原则、明晰性原则、多角度原则语义分类的原则构拟了新的语义分类体系，共分出的15个大类、203个中类、1477个小类。
- 以结构匀称性、突出现代性、语词义位性为收词原则共收录13万多词语进行归类标注。
- 该资源可应用于自然语言理解、机器翻译、信息检索、语言教学及写作当中。

4、《现代汉语新词语信息电子词典》

- 收录了1978年以来出现的新词语3万8千多个，分总库、名词库、动词库、形容词库、新造词库、外来词库、旧词新用库、方言词库等，详细描述了这些词语的语法属性、部分语义属性和语用属性、来源信息、构词法信息以及丰富的例句。
- 以数据库文件格式存储，是进行现代汉语新词语研究、词汇历时研究、汉语教学、汉语信息处理的重要资源。

5、标注句子语义成分语料库

- 国家社科规划项目任务
- 人民日报、奥运会新闻、中小学语文课文等语料。
- 标注信息：分词、词性、句法成分、句法语义成分、词语语义类。
- 可用于自然语言处理理解、分析、生成，语言学研究等。

6、文本蕴涵信息库

- 国家863项目“基于人类认知的语义知识融合、学习与计算技术”中的任务
- 文本蕴涵关系：一个连贯的文本T和一个被看作是假设H的文本之间的一种关系。在T的语境下，如果H的意义可以从T的意义中推断出来，那么就说T蕴涵H。 语言中的同义形式。
- 标注中小学语文课本 82万多字， 以XML格式存储 。
- 可用于自然语言理解、信息检索、机器翻译以及对外汉语教学。

7、汉语拼音词汇数据库

- 在新修订汉语正词法的指导下，收录现代汉语词汇12万左右，进行语音标注。
- 可以作为词典注音、正词法的基本依据。

8、方言有声数据库建设计划

- 中国语言资源有声数据库，是国家语委正在开展的一项语言工程，旨在用现代信息技术、遵循统一的工作规范和技术规范、将中国各县域的语言实态记录下来，归档建库，永久保存。
- 希望作为承接山东各地语言资源有声数据库建设的实施单位 。
- 参照中国语言资源有声数据库建设领导小组办公室《中国语言资源有声数据库调查手册 汉语方言》以及李宇明《论中国语言资源有声数据库的建设》，制定《中国语言资源有声数据库烟威地区建设实施方案》，并做好了各种硬件准备。
- 列入山东省语言资源开发与应用重点实验室建设计划当中。

9、中小学生语言偏误语料库的开发

- 按照大规模原则、全面系统原则、平衡性原则、代表性原则、真实性原则、开放性原则收录语料建立大规模中小学生语言偏误语料库，开发中小学生语言学习的宝贵资源。
- 目前已经录入偏误语料近10万个记录，300多万字。

感谢

感谢各位专家学者！

欢迎批评！