《人民日报》语料中并列结构的标注

吴云芳

2003-11-7

0 引言

并列结构是语言信息处理中的难点，无论对英语、日语还是汉语，莫不如此。为了对汉语并列结构进行系统的、定量定性的研究，笔者手工标注了《人民日报》1998 年 1 月 1－10 日语料中的所有短语层面的有标记并列结构，语料共计约 56 万字，标注出并列结构 7215 个。基于这样的语料，笔者完成了自己的博士学位论文《面向中文信息处理的现代汉语并列结构研究》（2003 年北京大学中文系，导师陆俭明、俞士汶）。现把笔者标注的语料公布于众，与各位专家学者共享，希望能在此基础上作进一步的研究，共同努力，揭开现代汉语并列结构的神秘面纱。

1 标注符号

语料中用“{ }”标注出并列结构。

例如： 人类/n 的/u {生存/vn 与/c 发展/vn} 还/d 面临/v 种种/q {威胁/vn 和/c 挑战/vn} 。/w

当出现嵌套并列结构时，用嵌套的“{ }”来标注。

例如： 江/nr 泽民/nr 指出/v ，/w 我们/r 将/d 继续/v 坚持/v {{和平/ad 统一/v 、/w 一国两制/l} 的/u 基本/a 方针/n 和/c {发展/v 两岸/n 关系/n 、/w 推进/v 祖国/n 和平/ad 统一/v 进程/n} 的/u 八/m 项/q 主张/n} ，/w

2 标注范围

我们并没有标注出语料中所有的并列结构，而只是标注的短语层面的有标记并列结构。具体而言，我们没有标注下面几类并列结构：

（1） 并列复句。

例如：尊老爱幼/i 既/c 是/v 中华民族/n 的/u 传统/a 美德/n ，/w 又/c 具有/v 重大/a 的/u 现实/n 意义/n 。

（2） 跨越“，”的并列结构。

例如：制止/v 追求/v 表面文章/i ，/w 搞/v 花架子/n 等/u 形式主义/n ，

（3） 无标记的并列结构。

例如：宾主/n 进行/v 了/u 亲切/a 友好/a 的/u 交谈/vn 。

3 后记

虽然笔者经过了多次修改校正，但标注错误还是在所难免。敬请各位专家学者使用此语料的过程中，给予我们批评指正。标注过程中，笔者也切切实实体会到了并列结构处理的困难。有时并列结构的边界问题，即使是人来判断也要颇费思量。而且，不同的人来判断，或

者是同一个人在不同的时间来判断，结果也会存在差异。

现代汉语并列结构有许许多多未解的难题期待着我们去回答。笔者的博士学位论文《面向中文信息处理的现代汉语并列结构研究》只是并列结构研究的一个小小部分。若有专家学者对拙作稍感兴趣，请发 E_mail 至：wuyf@pku.edu.cn。