基于藏语语料库的词类分类体系研究

作者：才让加 扎洛

发布时间：2008-10-13 浏览数：24

对于英语、法语和德语等西方语言而言，词与词之间一般采用自然的空格作为切分标记，但在汉语、藏语和日语等语言的实际切分中没有自然的空格作为标记，这就需要通过人工或机器对其进行词与词的切分和标注。汉语和藏语在形式上有较大的差异，但在功能上却有着深层的相似性，由于汉语在语料库切分标注规范的研究方面已有很多成果，而这些成果为我们研究和确定藏语语料库词语分类体系很好的参照，因此在对藏语文本进行切分标注时首先解决的是藏语词类的分类体系和标记集的制定，其次是根据藏语词类的分类体系和标记集建立藏语语料库切词词典，最后对原始文本语料进行切分和标注。在这里词是指事物间区别意义的语法单位。词语的标准可划分为三种：形态标准、意义标准和语法功能分布标准。藏语的黏着性特点，单纯地使用任何一种方法来进行藏语词语的分类很难得到最佳的分类体系，由于藏语具有格词表征词与词之间的明显的形式特征，动词又保留着形态变化的基本特征，而格词、形容词等具有广义上的形态变化特征，因此利用语法功能与形态特征进行藏语词语的分类应该是一种既符合藏文自身特征又可以被计算机自动处理的方法。因此，我们在藏语词语分类体系的构建上，采用先分虚实，再确定大类，在大类的基础上分出小类，然后根据需要分出不同深度的子类。为了便于计算机自动分析和处理，我们重点考虑以下四点：一是划分出来的兼类词尽可能的少，二是有利于句法分析，三是便于操作，四是与藏语传统语法保持最大的一致，根据这四点考虑，结合藏语信息处理的实际需要，从总体上采用了以语法功能分布为主，形态变化为辅的分类标准，依据这一标准对藏语词类进行了归类，并在藏语自动分词系统和汉藏机器翻译系统中得到了部分应用。在大类的划分上一是与传统藏语语法分类体系保持衔接，二是在大类的基础上依据分类标准建立划分基本类的可操作的判别准则。为了使藏语语料库词语分类体系具有规范性、稳定性、针对性、实用性和继承性，青海师范大学藏文信息处理与机器翻译省级重点实验室从2002年开始进行了藏语语料库的多级加工研究，对1000万字的藏语原始语料进行切分和标注实验（这些原始语料包括藏文典籍、中小学教材、藏文期刊和报纸、藏文网站界面文字、现代藏文文选和藏文版政论著作、藏文版法律法规等），通过藏语语料的切分标注实验，2005年12月提出了《藏语语料库分类体系及标记集》(讨论稿)，2007年6月提出了《藏语语料库词语分类体系及标记集（V1.0）》。

责任编辑：宗哲

文章出处：中国藏学网

本文注释信息：

标签：才让加　扎洛

无法找到该页

您正在搜索的页面可能已经删除、更名或暂时不可用

请尝试以下操作：