汉语语料库的文本描述

傅爱平*　　张弘
中国社会科学院语言研究所，北京

提要
本文用通用可扩充置标语言 XML 定义了一个汉语语料文本描述模式 XML Schema for Corpora（简称 XSC），作为汉语书面语语料的通用描述规则，描述各种原始的和带标的语料文档。希望能够容纳不同的标记集，兼顾各种不同类型的标注需要。用 XSC 描述汉语语料文本，有助于保持语料的原貌、表现语料样本的篇章组织形式、反映语料中蕴涵的各种语言信息、记录语料的说明性信息。在 XSC 的约束下标注、并通过了格式验证的语料文档，已经完成了从非结构化数据到 XML 结构数据的转换，可以直接装入 XML 数据库进行管理和应用。本文介绍了 XSC 的设计思路、基本框架和主要内容，并通过基于 XSC 开发的语料库实例说明了 XSC 对语料文本的描述功能。

关键词
汉语语料库　**语**料文档的描述　**基**于XML的文本数据结构
语料库标注　**XML** Schema

A DESCRIPTIVE TEXT DATA FORMAT FOR CHINESE LANGUAGE CORPORA

Aiping Fu **Hong Zhang**
Chinese Academy of Social Sciences, Beijing

ABSTRACT

The paper proposes a general-purpose text data format for documents in Chinese language corpora. The format describes the archival structure and other attributes of the documents by a set of markup elements built using XML Schema. So it is called the XML Schema for Corpora, XSC for short. The XSC is intended 1) to carry the basic textual structural information of the documents in both raw and annotated corpora, 2) to describe the linguistic features in annotated corpora based on the different annotations, 3) to be open-ended in the sense that document-specific element types can be used, by user's customization within the hierarchical and nestable framework of the XSC, 4) to allow the documents to be converted into an XML data file and processed using automatic tools such as XML database management system, indexing software, and other transformations. In this paper the framework and the applications of the XSC are presented, with some instances taken from the XSC-based Chinese language corpus built by the authors.

KEYWORDS

Chinese language corpora **D**escription of the corpus documents
XML-based text data structure **C**orpus annotation **X**ML Schema