# 人机交互式的汉语辞书编纂系统

傅爱平　吴　杰　张　弘　李　芸

**摘　要**　人机交互式的汉语辞书编纂系统是一个计算机网络应用系统，在互联网上为汉语语文辞书的编纂提供协同工作环境。这个系统以语言数据资源的建设和应用为基础，借鉴语言信息处理的相关研究成果，融入汉语语文辞书编纂的知识和经验，用人机交互方式管理词典编纂的工作流程，包括新编词典立项、词条结构设置、选词立目、词条编写修改和审定、词条编审历程的保留与追溯、辅助词典成书等，同时也集成了多个语料库、已有辞书和词表，提供给词典编者作为参考。文章简要论述这个系统的设计思想、主要功能和关键技术。

**关键词**　计算机辅助辞书编纂　语言数据建模　汉语辞书数据库　基于 XML 的语言数据处理

## 一、概　　述

人机交互式的汉语辞书编纂系统（以下简称“编纂系统”）是中国社会科学院语言研究所研制的一个计算机网络应用系统，用于汉语语文辞书的编纂。这个系统涵盖了汉语语文辞书编纂的完整流程，包括新编词典立项，词条结构设置，选词立目，词条编写、修改和审定，词条编审历程的保留与追溯，辅助词典成书等，同时也集成了各种语料库、已有辞书和词表，供词典编者参考使用。

多年来，传统的汉语辞书编纂工作一直是手工操作，费时费力，效率低，很不适应当前科学技术迅速发展、信息数量与日俱增的形势。近些年，国内先是语言信息处理领域的学者提出了辞书编纂自动化的必要性，中国大百科全书出版社和商务印书馆先后尝试用计算机辅助编纂词典，北京大学计算语言学研究所和教育部语言文字应用研究所也研制了各自的词典编纂系统。中国社会科学院语言研究所有着数十年汉语辞书编纂研究和实践的历史，积累了丰富的专业知识和经验。我们研制这个辞书编纂系统，是希望以这些知识和经验为基础，应用计算机软件工程和网络技术，在大规模语言数据资源和互联网上人机交互机制的支持下，把语言数据建模、语言信息处理与辞书编纂过程结合起来，改变以往辞书编纂和修订全部由手工操作的工作方式，提高辞书编纂的质量、效率和科学性，同时为辞书研究和汉语词汇研究提供数字化的语言资源。

人机交互式汉语辞书编纂系统建立在 TOMCAT 网络发布系统上，在 WINDOWS SERVER下采用浏览器/服务器方式运行。系统程序用 JSP、JAVA、HTML 等语言编制，主要

用 Berkeley DB XML(以下简称 BDB XML)数据库系统在后台管理各种语言数据。

## 二、编纂系统的结构设计

编纂系统的总体设计思路是:以语言数据资源的开发和管理为基础,借鉴语言信息处理的相关研究成果,融入汉语语文辞书编纂的知识和经验,用人机交互方式管理编写词典的工作流程,提供编者需要的各种资料和信息。整个编纂系统由人机交互式工作流程控制平台、在编词典数据库和语言数据资源库三个子系统组成。图 1 是编纂系统的组织结构图,其中:

(1) 人机交互式工作流程控制平台:按照词典编纂的流程,根据词典编者的不同权限,提供从词条结构定制、选词立目、词条编写、修改和审定、编者信息交流,到词条过录、排序、生成检字表、输出词典的各种处理功能和操作界面。在处理过程中编者可以随时调用语言资源库的内容,进行语料检索或统计,查询各种参考词典,也可以回溯在编词典库中保存的词条修改和编审的历史记录。

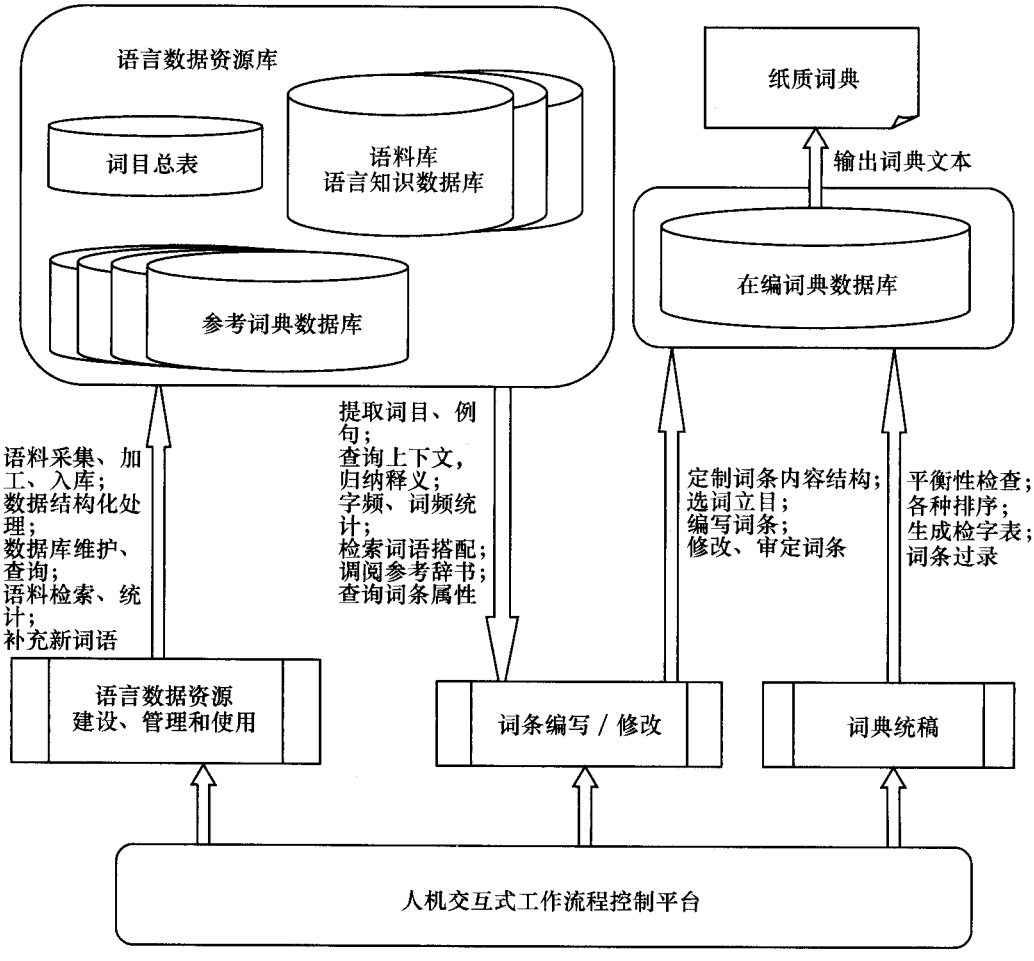


**图 1 编纂系统的组织结构**

(2) 在编词典数据库:在编词典是编纂系统人机交互式工作流程的主要操作对象和产出目标。在编词典的每个词条以义项为单位存储,每个义项的内容由各种属性或特征组成。在编词典数据库建立之初,要先由主编根据编纂系统提供的词条结构模型来确定词条的内容结构,再由编纂系统生成词条编写界面。编者就在这个界面上编写词条。编好的词条可以再修改或提交审定,在编词典数据库会保留修改和审定的记录(包括:修改/审定者、改动内容、修改/审定时间等),以供日后查询,也能为每个编者保存个人编写日志。在编词典一旦编写完成,编纂系统会自动把它的副本转为参考词典。

(3) 语言数据资源库:由词目总表、参考词典数据库、语料库及其检索统计模块组成。词目总表用开放的方式尽量多地收录现代汉语的词语,记录每个词语的各种属性/特征,主要为选词立目提供素材,也可以在编写词条时供编者参考。参考词典数据库收集各种已有词典的各个版本,供用户在编写词条时随时调阅参考,也可以用于词典查考和词汇研究。语料库里集成了编纂词典需要的各种语料。检索统计模块在编写词条时随时调用,对集成在系统里的语料库和数据库进行检索和统计。

在整个编纂系统的设计中,自然语言数据资源的形式化描述和结构化处理是基础性的工作,有两个主要内容:一是用数据建模的方法研究汉语语文辞书的内容结构和汉语语料库的文本结构,建立辞书内容结构模型和语料库文本描述模型;二是研制词典内容结构化处理和语料文本描述的软件工具,建立基于 XML 的词典数据库和语料库。这些语言数据资源支撑着整个编纂系统的构造和运行。

## 三、编纂系统的主要功能

### 1. 辞书编纂业务和系统管理

编纂系统以人工编写词典的知识和经验为基础,用归纳与分析结合的方法,对编写汉语语文辞书的全过程进行需求分析,提出要解决的问题,建立需求模型,描述整个系统的任务流程,确定系统的总体结构和设计方案。在编纂系统中,面向用户的全部应用功能集成在人机交互式工作流程控制平台上,位于系统的前台。

这些应用功能可以分为三类:编写业务流程功能、编写业务辅助功能和系统管理功能。前两项包括词典编写过程中的各项操作,有词条内容定制、编写任务分派、个人任务、待编词目、词目确认、词条编写、词条初审、词条终审、编写进程处理、词条统计、词条提取和词条删除,还有缺字处理、规范用词的补充等。第三项系统管理功能包括辞书浏览、语料检索、项目管理、词表管理、人员管理、消息管理、数据管理等。详情可参见编纂系统操作指南(编纂系统课题组 2011)。

进入编纂系统的每个用户都有自己的权限:主编、组长、编者、访客。权限不同,能够使用的功能也不同:

访客:辞书浏览、语料检索。

编者:除访客的全部权限以外,还有:待编词目、词条编写、词条初审、词条统计和词条提取。

组长:除编者的全部权限以外,还有:人员管理、任务分配、进程管理、词条终审。

主编:除组长的全部权限以外,还有:词条内容定制、词目确认。

此外还设有系统管理员。他除了有上述所有权限以外,还有系统后台的全部管理权限,负责系统数据安全、数据备份、用户管理、项目管理等工作。

图 2 是人机交互式工作流程控制平台的主要功能。

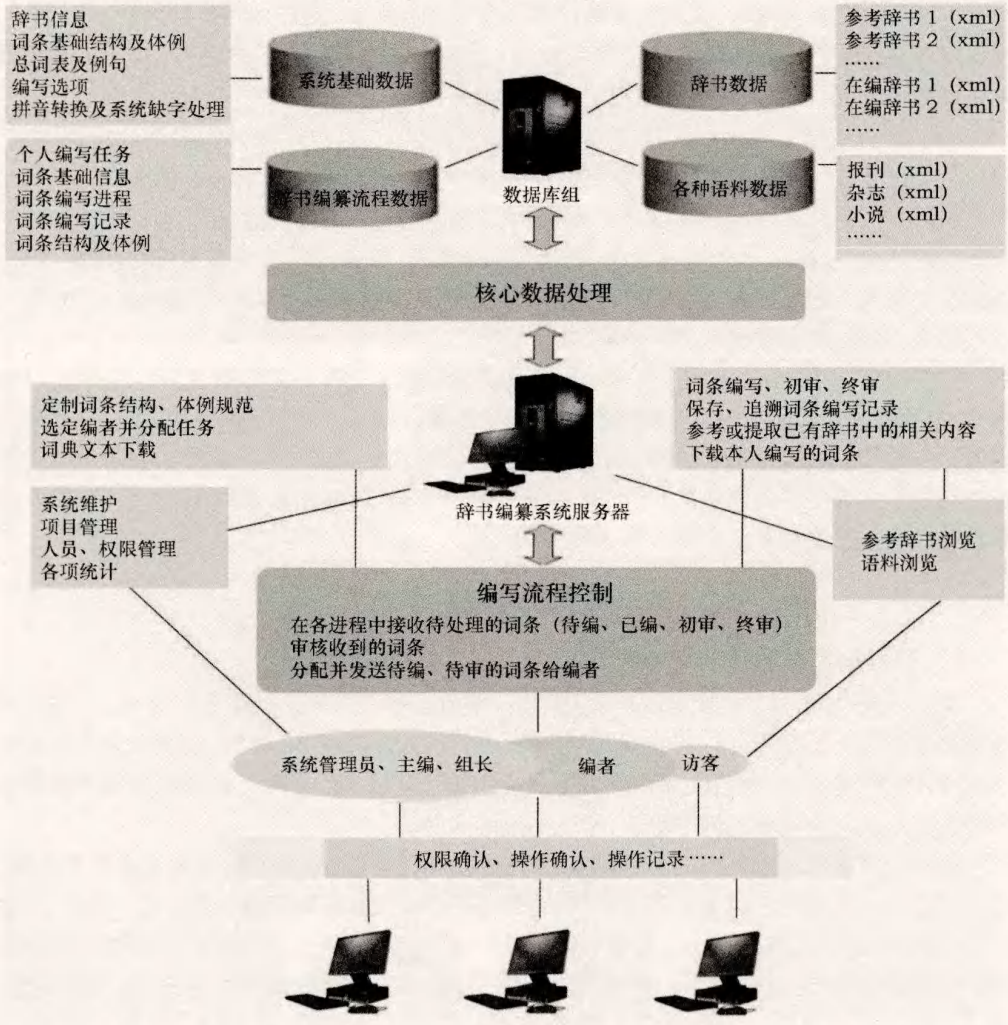


**图 2 工作流程控制平台的主要功能**

## 2. 基于 XML 的词典数据资源管理

编纂系统里的词典数据库有两种:在编词典和参考词典。前者是系统的主要操作对象和产出结果(可以同时编写多部在编词典),后者供词典编者参考(目前系统收录了六部参考词典)。词典数据资源管理系统在编纂系统的后台运行,负责所有词典的内容管理,主要是词典内容描述、词条信息标注、建立词典数据库、词典数据库管理和词条内容查询。

从语言信息处理和辞书数字化的角度来看,汉语语文辞书里蕴含着大量系统的汉语文

字、语音、词汇、句法、语义、修辞、语用等信息。要把词典作为一种语言数据资源用计算机来处理,首先需要把文本形式的词典转换成结构化的词典数据库。这就需要对词典的内容进行结构化的描述和组织:描述词条内容的表现形式和其中蕴含的语言知识;并把这些信息组织成合理有效的数据结构。

在编纂系统中为了给汉语语文辞书建立数据结构,我们提出了用 XML Schema 表示的辞书内容结构模型 XML Schema for Dictionary(以下简称 XSD)。在这个模型支持下对每一部词典做结构化处理,建立辞书 XML 数据库,用原生 XML 数据库系统 BDB XML 管理和访问[1],形成了基于 XML 的词典数据资源管理系统,它可以创建、管理和访问编纂系统里的全部词典数据库,主要功能是:

(1) 用 XSD 描述词条内容,建立词典的数据结构;

(2) 根据 XSD 用自动标注程序对词典文本做 XML 标注,描述词条中包含的各种属性或特征,再通过人机交互方式校对,得到词典的 XML 文档;

(3) 根据 XSD 用通用 XML 软件工具对标注好的词典 XML 文档进行良构性(well-formedness)和有效性(validation)检验,保证数据的有效性和一致性;

(4) 把通过检验的词典 XML 文档以节点方式批量导入 BDB XML 数据库系统的容器中,并根据访问方式设定多线程(进程)锁策略,自动生成词典数据库。不同的词典在数据库系统的容器里用命名空间加以区别;

(5) 词典数据库建立以后,根据编纂系统前台可能提出的各种检索需求,针对 XML 文档元素设置相应的索引类型,编制适当的索引策略,以达到快速读写的要求。此外,通过事务子系统、锁子系统和日志子系统来处理系统的并发策略;

(6) 根据编纂系统前台客户端生成的 XQuery 查询语境,对 XQuery 导航函数解释执行,转换和返回 XQuery 的查询结果,实现多层次元素及属性查询、多元素复合查询、多容器查询等检索要求。

除了在编词典数据库以外,目前词典数据资源管理系统里可用的参考词典数据库有《现代汉语词典》第 4 版、第 5 版等六部,共 307670 个词条。

### 3. 语料资源的建设、管理和使用

编纂系统集成了多个语料库供编写词条时查询,语料资源管理系统在后台运行,目前有内置语料库六个,外部语料库一个。用于词典编纂的语料库可以有多种来源、多种类型、多种篇章形式、多种标注方法。为了能够在编纂系统里用统一的方法使用不同的语料资源,我们提出了汉语书面语语料的文档描述模型 XML Schema for Corpora(以下简称 XSC),用以描述各种汉语语料的文本组织形式、表现语料库中标记的语言知识信息、记录语料库的说明性信息。

XSC 规定了语料 XML 文档的语法格式,在它的约束下,经过标注的各种语料库可以自动生成 XML 结构的文档,然后在通用的 XML 开发应用环境下做各种加工处理。例如用 XML 解析器对语料文档进行良构性和有效性检验,保证数据结构和内容都符合 XSC 的规定。经过验证的 XML 语料文档,已经完成了从非结构化文本到 XML 数据结构的转换,可以直接导入 XML 数据库系统进行管理、提供访问,或者用通用编程接口来实现各种应用。

这样就能够在同一个数据库平台上,用同一种方法和技术管理和访问多个不同类型的语料库。

之所以选择通用可扩充置标语言XML来描述语料库,是因为它除了能兼容多种标注需求以外,还是正式发布的国际标准,在规范性和通用性方面有优点,便于有效地规范语料文档的数据结构,有助于与国际语料库编码标准接轨,也有助于语料库的数据交换和资源共享。

语料资源管理系统使用原生XML数据库系统BDB XML,以XSC为基础,建立了一个多种语料库文本标注、文档管理和数据处理的集成环境。利用BDB XML的XML文档分析器、XQuery查询引擎和独特的索引系统,优化语料内容索引策略,建立基于成本的查询方案,实现对XML文档节点、元素、属性以及元数据的灵活索引,在多层次元素及属性查询、多元素复合查询、多容器查询等检索环境中,使复杂的XQuery语句快速命中目标,提供检索结果。目前在编纂系统里内置的六个语料库全部采用这种基于XML的方法和技术。

此外,编纂系统还有外部的动态语料库。这是一个原始语料库,用动态跟踪的方式采集了24种报纸语料,单份报纸的时间跨度为2—13年,共采集了142年次、约40亿字,用基于Apache Solr的分布式语料全文检索系统提供查询。

表1是编纂系统里可用的语料库:

**表1**

| 名　　称 | 字　　数 |
|---|---|
| 报摘语料库 | 840万 |
| 广电语料库[2] | 1.5316亿 |
| 平衡语料库[3] | 4560万 |
| 语文教材语料库 | 7万 |
| 文学作品语料库 | 707万 |
| 近代汉语专书语料库 | 98万 |
| 动态报纸语料库(外部) | 40亿 |
| 合计 | 42.1528亿 |

## 四、编纂系统研制中的关键技术

### 1. 基于XML的语言资源处理方法

编纂系统用基于XML的语言数据资源处理方法设计数据结构,这包括以下几方面的探索性工作:词典的内容描述和数据组织,语料库的文本描述和数据组织,以及应用原生XML数据库系统处理语言数据资源,研究和开发实用的软件技术和应用系统。

(1) 基于XML的词典内容描述和数据组织

文本形式的词典可以认为是一种用非结构化形式表现的、具有半结构化特征的语言数据。我们用辞书内容结构模型XSD为词典做数据建模,用XML Schema定义词条的内容和词典的组织结构,提取词条中蕴含的各种语言学信息,把文本形式的词典转换成词典数据

库。在 XSD 里,词典以词条为基本单位,由众多词条组成,每个词条含有形、音、义、用法等各种属性。一部词典的全部内容表现为树形结构,树的第一层节点是词条,每个词条的各层下位节点是这个词条的各个属性。全部词条的属性和属性之间的关系构成了一部词典的内容结构。在 XML Schema 框架下,所有代表词条属性的节点都表现为元素或子元素,对这些元素进行定义和约束,就可以得到各个属性节点的确切定义。通过 XSD 对一部词典进行结构化标注,再把文本形式的词典转换成 XML 原生数据库,就能够系统地组织并完整地描述词典内容的表现形式和其中蕴含的语言知识。这种词典数据库也是一种词语知识库,它不仅可以用于词典的编纂、查考和典藏,也可以为语言研究、词汇研究和语言工程提供数据资源。

以往的辞书数字化工作大多是把词典做成二维表,再用关系型数据库来处理。我们选择 XML Schema 代替关系型数据模式作为词典数据建模的方法,是因为 XML 的数据结构适合描述语文词典的结构形式,XML Schema 的树形数据模式正好完全体现了词条结构的层次关系和管辖关系。用 XML Schema 可以方便地描述不定长内容的词条属性(例如词条的释义部分);描述不定量重复出现的词条属性(例如词条释义中的例句);描述词条中的嵌套关系(例如多层级义项的嵌套)。更重要的是,可以根据词条描述的需要,动态地为 XML Schema 补充子树或元素、变更对已有元素的约束,只要不改变原有的树形架构,就不会影响它的兼容性。这些都是关系型数据模式不容易做到的。(傅爱平等 2009:28)

辞书内容结构模型 XSD 也有一种通用性:它定义的是汉语语文辞书中每一个词条所有可能的属性,以及每一个属性所有可能的取值(属性值)。这样就可以涵盖多部词典的内容和组织结构。也就是说,同一个 XSD 可以描述多部语文词典。在编纂系统里,每一个在编词典数据库和六部参考词典数据库都是用这个 XSD 定义的,它们都在 BDB XML 数据库系统上用统一的方法建立和管理,用统一的技术提供查询,获得了理想的使用效果。

除了通用性以外,XSD 还有某种抽象性:它描述的词典内容模式是一种底层数据结构,与词条及其属性或特征在具体词典中的表现形式没有关系。比如对异形词的处理,有的词典用"同××"表示,有的词典用"也作××"表示,还有的词典两者都用或者更随意。无论在具体词典中表现如何,在 XSD 中都定义成一个可选的属性"异形",其属性值为"是"或"否"。这样就把词典数据的内容和形式分离开了。词典编者只需集中精力琢磨词条内容,无须考虑词条体例的表现形式,有关体例样式的工作都由编纂系统通过"词条定制"的功能用人机交互的方式来完成。

(2) 基于 XML 的语料文本描述和数据组织

用于词典编纂的语料库有多种类型,收录了各种各样的篇章样本。这些语料样本或者表现为原始文本的形式(可带有原生标记[4]),或者是带有附加标记的形式(带有非原生标记[5])。目前国内语料库研究和开发的情况是,不论带标语料库还是原始语料库,只要研究或应用目的不同,就会有不同的标记集和标注规范,也就有各自的语料库管理和检索系统。在分析了各种类型的语料库及其加工现状之后,我们用 XML Schema 构建了一个语料文本描述模型 XSC,定义语料标注的描述规则,描述语料的各种原生标记和非原生标记。目的是客观地表现语料文本的原貌,兼顾各种不同类型的标注需要,尽可能容纳不同的标记集,

描述各种原始的和带标的语料库。与此同时以 XSC 为基础，建立了一个语料库文本标注、文档管理和数据处理的语料资源管理系统。

XSC 面向多种类型的汉语书面语语料。傅爱平等(2011)认为，语料库中不管是原生态的标注，还是非原生的标注，通常都主要描述三类信息：一是篇章组织和文本结构信息，即组成语料文本的篇章、段落、句子、词语等语言结构成分，语言成分在文本中是以文字符号等实体形式表现的；二是语言知识信息，是语料在词汇、语音、语法、语义、语用等各个层面的属性或特征，它们附着于各个语言成分之上；三是功能性或说明性信息，有两种：第一种是主题、语体、作者、出版者、版本、承载媒体、出版时间等，一般附着于语料的单位样本之上；第二种是校注、言者角色、言语伴随行为、言语环境等关于文本正文的说明，一般情况下，它们附着于各个语言成分之上。语料文本描述模型 XSC 的主要任务就是描述这三类信息，即描述各种汉语语料的文本组织形式、表现语料中标记的语言知识信息、记录语料库的说明性信息。XSC 定义的是语料库的描述规则，不管语料库是原始的还是带标的，不管标注的是哪些信息，XSC 都应该能用 XML 把这些语料文档表现出来。

一般来说，语料标注的主要对象一是语言成分，二是语言知识信息。后者是语言成分的属性、特征或语言成分之间的关系。在语料文本中，语言成分是文字或符号的实体形式，关系附着于成分之上。不管出于什么需要、用哪个语法体系去研究语言，语言成分及其关系都是基本的研究对象。语料标注不论采用多少种标记集，也无非是从不同的角度去描述各个语言成分及其关系。所以在 XSC 中，“成分”和“关系”是最基本的描述对象。

在 XSC 中语料库表现为树形结构，含有元素和属性两类结点，用元素来定义语言成分，用元素的属性来定义关系。在语料文本中，语言成分有其客观实体表现，描述语言成分的主要原则是客观，要尽可能反映它们的原貌。XSC 能够根据不同的需要，描述各种语言成分及其在语料中的各种出现方式。语言成分的属性或特征，以及语言成分之间的关系是带有主观性的信息，描述它们的原则是兼容。XSC 可以提供一种自选参数的兼容机制，描述每个语言成分的多种关系，尽可能表现各种语法体系和分析方法所需要的标记信息。

在编纂系统中，我们应用 XSC 描述语篇的组织和语言成分的各种组合，能够兼顾不同类型的语料文本和语言分析方法，生成 XML 结构文档并自动导入 XML 数据库，使用通用的软件工具管理和访问语料库。编纂系统中六个内置的语料库(约 2.15 亿字)都是用这个 XSC 定义的。它们在 BDB XML 数据库系统上用同一种方法建立和管理，在统一的语料资源管理平台上提供查询，获得了预期的使用效果。

**2. 词典编纂流程的功能集成设计**

编纂系统的前台是人机交互式工作流程控制平台，它除了为用户提供参考词典和语料库以外，主要任务是词典编写全过程的业务流程控制。这个平台的设计既体现了人工编写词典的知识和经验，也集成了一些应用计算机和网络技术提高词典编纂效率和科学性的功能。下面是几个例子：

（1）可视化的词条编写界面

编纂系统给编者提供了可视化的操作界面（见图3）：

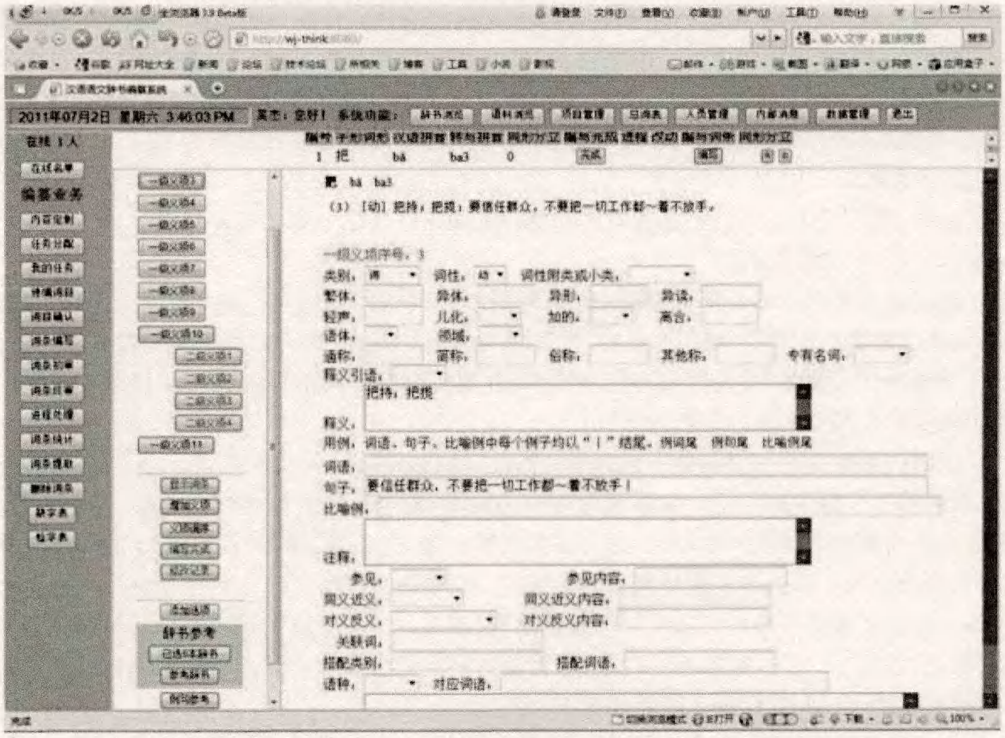

**图3 编纂系统的操作界面**

这个界面的主要功能是人机交互编写词条，同时也提供编纂系统的管理和语言数据资源的使用。

词条编写是整个编纂业务的核心部分，主要有词条内容定制、任务分配、编者任务、待编词目、词目确认、词条编写、词条初审、词条终审、进程处理、词条统计、词条提取、删除词条、缺字处理等多项功能，逐一列在编写界面左端。编写界面的主要部分用来表现词条的内容和结构。

在词条界面上，内容按义项显示，左侧给出了整个词条的结构。语文词典的一个词条下面可以有若干个义项，义项下面还可以有子义项。在辞书内容结构模型 XSD 里，我们用树形结构定义这种义项之间的嵌套关系，体现在词条编写界面上，是主义项、一级义项、二级义项等的层级结构视图。一个词条可以只有一个主义项，也可以有若干个一级义项或二级义项，通过这个结构视图，一个词条的内容结构可以一目了然。编者可以根据需要打开各个义项查看或填写内容，也可以增加或删除各级义项，还可以对义项重新排序。

词条编写的操作以义项为单位，编写界面的中心视图是每个义项的内容，用列表框给出每一个属性，编者只需要按照视图的提示在属性框中填入内容。有的属性值编者可以自主填写，比如词义和例句。有的属性值规定了取值范围，比如词性，只能在给定的参数里选

择,这样有助于表达形式的规范和平衡。词典编纂是多人参与的项目,各位编者在专业水平、编写经验和表达习惯上都会有所不同,比如词条属性值或标记符号的使用就可能因人而异。编纂系统在词条编写界面提供了选项和赋值两种内容填写方式。前一种只能在词条属性值的取值范围内选择,后一种也会对编者自主填写的内容做一些检查校核。目前采用选项操作的属性有:词目类别、词性、词性附类或小类、儿化、语体说明、领域说明、释义引语、搭配类别、语种、参见、同义近义、对义反义等。各个选项的参数根据需求可以由主编随时增加或修改。

此外,编写界面还提供了参考词典、语料库、词条修改记录等供编者调用。

(2) 词条内容的定制

在编纂系统里新编一部词典时,先要定制新词典的词条内容。词条内容定制的意思是:以辞书内容结构模型XSD为基础,根据新编词典的需要,确定词条里要包含哪些属性、属性之间有什么关系、对属性值有哪些约束,并规定词典输出的体例或版面格式。词条内容定制实际上是对词典内容的设计。

汉语语文辞书的一个词条下包含形、音、义、用法等各种属性,比如字形、拼音、词性、释义、例句等。根据各自不同的编纂理念和应用需求,各个词典对属性选择不同、多少不等。主编给新词典做了词条内容定制以后,编纂系统会根据定制的结果,按照XSD的树形结构规则,自动生成一部新的在编词典的内容结构(是XSD的一棵子树),再根据这个内容结构在后台自动生成新在编词典数据库的数据结构,在前台自动生成供编者使用的可视化操作界面以及数据显示格式和数据保存格式,为新词典的编写做好准备。利用词条内容定制功能,编纂系统可以同时创建几部新的在编词典。

在编纂系统中,利用词条内容定制还可以控制在文本形式下词条输出的体例或版面格式。控制词条输出体例的意思是,指定词条中部分属性名和属性值的标志符或缩略符,用于文本形式的词条数据输出。例如在有的词典文本里,词条的属性"词性"用外加□表示;属性"例词""例句""比喻例"用"◇""|"等符号表示。通过词条结构定制可以给属性值指定表达符号和位置信息(分为属性值前附加、属性值间附加或属性值后附加等几种位置)。控制文本形式下词条输出的版面格式,是为了把词条从编纂系统的词典数据库里取出来,按照印刷文本的形式呈现给词典编者。版面格式的控制主要包含词条中各个属性排列的顺序、各属性值的显示格式(例如空格、折行、缩进等)。处理得当的版面格式能够在词典的编写过程和排版过程之间起到沟通的作用,编者可以比较直观地看到词条的基本排版样例。

词条内容定制也可以在词典修订时用来变更原有的内容格局,还可以在已有词典的基础上减去一些属性项,不需改动内容,直接自动生成原词典的属性缩减本。

在编纂系统的XSD里目前一共有39个属性,供定制词条内容结构时选择。不够的话,还可以扩充XSD,添加新的属性或属性值。对于不同词典的内容需求,XSD中包含的属性就像是个最小公倍数,能够兼容各种属性。这得益于辞书内容结构模型XSD的通用性和抽象性,它们是词条内容定制功能得以实现的基础,也为XSD提供了更多的应用空间。

（3）词典数据的内容与形式相互分离

前面说过，辞书内容结构模型 XSD 定义的是抽象的词典数据，它描述词条所有可能的属性，也定义每个属性所有可能的属性值，跟词条及其属性在具体词典中的表现形式没有关系，这样就把词典数据的内容和形式分离开了。

这种分离的作用是，利用词条内容定制功能可以控制文本形式下词条输出的体例或版面格式，还可以规范词条输出格式、标点符号、特殊标记等。目前编纂系统已经对非正体、词类、语体说明、语用说明、注释、用例、外来语等词条属性的表达方式做了一致性处理，以避免输出时的随意性。

词典数据的内容与形式分离，还使我们能够在不同的设备上用不同的格式表现同一部词典的内容，输出便于人们查阅的各种文本形式。排版印刷格式是其中之一，还可以是网页格式、在移动设备上表现的格式（比如手机上显示），等等。另外还有词典的排序，可以根据拼音字母做正序排列，也可以做逆序排列。想用什么符号表示每一个属性，也可以自行设定。

（4）词条编写记录的保存和回溯

一个词条在编写、审校过程中，编写人员和审校人员可能做多次修改。保留词条编写和修改的记录并根据需要回溯词条的编写过程，对于语文词典的编纂来说十分重要。以前人工编写的时候，编者大都在卡片上用不同颜色的笔来做历次记录。通过这些记录可以追溯词条编写的过程，反映语言和词汇的变化，回顾历任词条编者的工作思路。

编纂系统提供了“保存修改记录”和“回看修改记录”的功能，根据编者的要求，把每一次编写和审校的信息记录和保存下来，供日后回溯。这些信息包括：修改者、修改前后的内容、修改时所处的进程、修改提交日期，还可以留下修改备注（包括修改原因、参考资料、遗留问题等）。回看修改记录时，会突出显示修改前后不同的内容。另外系统还有“撤消修改”的功能，在编写过程中可以根据需要随时恢复某次修改前的词条内容。这些功能有助于追溯词典的编审历程，不仅对编写词条有用，对词典修订和词典研究也有用处。

（5）编纂进程的动态处理

进程处理是控制编纂业务流程的功能，由主编或组长操作，分成编写、初审、终审、定稿等几个进程。在不同的进程中，词条根据需要在编者、组长、主编之间往返传递。进程处理的作用是帮助主编和组长了解每个编者当前的工作进度，掌握每个词条当前所在的进程和处理状态，在此基础上设定权限、分配任务、了解编写中的问题、组织协同作业，借助网络平台处理词条编写的各个环节，科学合理地管理词典编纂的全过程。

在词条编写过程中，每个编者也都可以在自己的词条编写界面中看到当前词条的操作进程、修改状态、当前进程是否完成等信息，明确自己的任务和工作进度，还能通过进程处理功能与其他编者交流信息、配合工作。

## 五、下一步工作

人机交互式汉语辞书编纂系统是面向应用的计算机网络服务系统，能否在语言技术、计算机网络技术和词典编者的智慧之间实现最佳的结合，还需要在实际运行中验证和完

善。编纂系统和用户之间应该经历一段较长时间的磨合,根据用户的体验和意见不断改进系统的设计和功能。

当前信息技术的发展日新月异,新的产品和网络应用模式不断涌现,互联网上开放式的网络词典编纂和服务已经成为汉语辞书编纂的一个新的应用方向。我们为这个编纂系统研制开发的关键技术和核心模块,有些已经具有开放应用模式的功能,可以作为下一步工作的基础。这种开放模式通过互联网(包括移动互联网平台),一方面为公众提供词语查询服务,另一方面开放词典编写平台,让所有对词典有兴趣、有建设性见解、愿意有所贡献的各界人士都能参与词典编写:提出新的条目,修改已有的词条,给出形、音、义、用法、来源等新的信息。读者和编者之间形成互动关系,利用先进的信息技术和海量的互联网资源,实现庞大的用户群和众多领域专家之间的分工协作,使词典的编纂和研究更好地适应语言生活的实际。

**附 注**

[1] Berkeley DB XML是由美国Sleepycat Software公司开发的开源嵌入式数据库管理系统。

[2] 广电语料库采用了中国传媒大学传媒语言语料库的一部分原始语料,谨向中国传媒大学国家语言资源监测与研究中心有声媒体语言分中心诚致谢意。

[3] 平衡语料库采用了国家语委语料库的一部分原始语料,谨向教育部语言文字应用研究所诚致谢意。

[4] 原生标记描述语料文本的篇章组织形式(如册、篇、卷、章、回、节等)、对正文的说明(如校注等)等信息。

[5] 非原生标记描述语料在词汇、语音、语法、语义、语用等方面的属性或特征(如词性、短语结构、语法功能、语义关系等)。

**参考文献**

1. 常宝宝. 基于语料库的双语词典编纂平台的构建. 辞书研究,2006(3):122—133.

2. 傅爱平,吴杰,李芸. 汉语语文词典的词条结构模型. 辞书研究,2009(2):28—37.

3. 傅爱平,张弘. 汉语语料库的文本描述. //汉语语料库及语料库语言学圆桌会议论文,香港,2011.

4. 刘辉. 词典微观数据结构标准化和关系数据库设计. //罗益民,文旭主编. 中国辞书学会双语词典专业委员会第七届年会论文集. 成都:四川人民出版社,2007.

5. 陆汝占. 汉语词典编纂一体化环境(上、下). 辞书研究,2000(2):37—48,2000(3):35—43.

6. 张弘,傅爱平. Berkeley DB XML在语料库管理中的应用. //第八届两岸三院信息技术交流与资源共享研讨会论文集. 台北,2010:355—365 .

7. 章宜华,刘辉. 基于微观数据结构的双语词典生成系统初探. 外语与外语教学,2007(8):61—64.

(中国社会科学院语言研究所 北京 100732)

(责任编辑 李潇潇)