陕西省旅游景区公示语翻译语料库系统的设计与实现*

董李鹏[1]，高东怀[2]，朱益平[3]，张知元[1]
（1.西北大学 现代教育技术中心，陕西 西安 710069；2.第四军医大学 网络中心，陕西 西安 710032；
3.西北大学 外国语学院，陕西 西安 710069）

摘 要：根据语料库语言学的基本观点，结合陕西省旅游景区的文化特色及多样性特点，对陕西省旅游景区公示语翻译语料库的主题栏目进行了规划，开发了一套基于 PHP 的陕西省旅游景区公示语翻译语料库系统。该系统的设计与实现对国内其他省市公示语翻译语料库的建设和应用研究具有一定的参考价值和启示作用。

关键词：语料库；公示语；旅游景区；翻译；查询检索
中图分类号：TP393 文献标志码：B 文章编号：1673-8454(2012)23-0054-04

一、引言

语料库（Corpus）通常是包含数以万计字的机器可读的语言材料集，它不同于档案，通常是被挑选出来并经过处理的文本，可用来代表特定的语言变体或流派，因此可作为一个标准的参考[1]。人们通过语料库观察和把握语言事实，分析和研究语言系统的规律[2]。公示语是在公共场合所展示的文字，具有特殊的交际功能以及提供信息和完成指令的作用。

目前，常用的语料库可分为三大类，分别为：译文语料库（translation corpus）、类比语料库（comparable corpus）、对应语料库（parallel corpus）[3]。译文语料库以收录译文为主，其宗旨在挖掘翻译语言本身的特征；类比语料库收录同一种语言的原生文本和翻译文本，它们之间无翻译对应关系，但在时代、体裁、主题等方面具有可比性，可用来研究翻译语言的特点；对应语料库收录原文与译文双语平行对照文本，为了能方便、精准的检索到所需要的语料，开发者通常会按事先设定好的标准对语料进行句或段的对齐[4]。陕西省旅游景区公示语英汉/汉英翻译语料库则属于对应语料库，同样它也属于双语“专用性”语料库。本文构建的语料库系统作为旅游解说系统中重要的软件组成部分，它将大量实际应用中的高质量的公示语汉英翻译语料经过处理后整合起来提供给所需的用户进行检索、研究，该系统的实现对国内其他省市公示语翻译语料库系统的设计与开发具有一定的借鉴意义。

二、公示语翻译语料库主题栏目划分

陕西省是旅游文化大省，共有人文景观 800 多处，是文物古迹荟萃之地，是重要的国际文化旅游胜地之一。首先，课题组成员分头实地调研，广泛采集资料，然后，对所收集的资料和实地采集的译例进行分析、归类，并对陕西省旅游景区的文化特色和多样性特点进行系统的分析，初步规划了语料库的主题栏目，共分为 10 大类和 48 个小类，如表 1 所示。语料库资料的收集整理和栏目的划分为系统开发奠定了基础。

表 1 陕西省旅游景区公示语翻译语料库主题栏目

一级分类	二级分类	一级分类	二级分类
地文景观类	华山	古建筑类	大雁塔
	翠华山		小雁塔
	……		……
水域风光类	壶口瀑布	民俗文化景区类	党家村
	洽川湿地		北广济街回坊
	红碱淖风景区		高家大院
生物景观类	朱雀国家森林公园	旅游度假区类	渭水园
	太平国家森林公园		新桃花源山庄
遗址遗迹类	大明宫国家遗址公园	古寺庙类	大慈恩寺
	大唐芙蓉园		荐福寺
	……		……
帝王陵墓类	黄帝陵	博物馆类	陕西历史博物馆
	秦始皇陵		秦始皇兵马俑博物馆
	……		……

注：“……”：表示省略的语料栏目。

* 基金项目：本文系陕西省社会科学基金项目“陕西省旅游景区公示语语用翻译研究”（项目编号：09K008）的阶段性研究成果之一。

三、公示语翻译语料库系统总体设计

1.设计原则

参照国内外的相关研究课题，确立了本系统的设计原则，主要是：

（1）实用性和易用性：系统在设计初期考虑到管理员和普通用户需求，并吸纳其他知名系统的设计理念，尽力达到功能完善、简单易用的用户体验。

（2）易管理和可维护性：管理员可利用浏览器登录到此系统进行管理与维护，以保障其有效的运行。

（3）可扩展性：系统在开发时，出于长久的考虑，预留了接口，作为未来新功能的扩展。

（4）安全性：实现简单的用户权限（超级管理员、普通管理员、普通用户）分配，保障了系统可被安全的访问。

笔者所建立的公示语翻译语料库系统主要有两大功能。管理语料数据库，利用该系统提供的插入、删除、更新等功能来完成语料数据库的管理，比较重要的是语料加工功能，利用特殊的方式将语料篇章分割成若干句子以记录的形式存入数据库中；语料检索功能，可利用词组或句子作为检索关键词进行分类检索和全库检索。设计公示语翻译语料库系统应紧紧围绕这两大功能。

2.系统框架设计

从用户角色的角度出发，系统的功能结构如图 1 所示。

图 1　旅游景区公示语翻译语料库系统功能结构

普通用户有发表评论和查询检索语料的功能，他们可以通过发表评论的功能对语料库系统进行评价并给出宝贵的意见和建议，以便进一步完善该系统；也可利用查询检索的功能对语料库中的资源进行查询和研究。

管理员（超级管理员和普通管理员）具有分类管理、用户管理、友情链接管理、语料管理、评论信息管理、系统配置、查询搜索 7 大类功能。

（1）分类管理：可对语料库资料来源的一级分类和二级分类进行管理，包括添加、删除、编辑等功能。

（2）用户管理：可对已注册的合法用户（管理员和普通用户）进行管理，实现用户的删除、编辑、审核，查看总用户数等功能。

（3）友情链接管理：管理员可根据友情链接的分类（单语语料、双语语料、论坛沙龙、软件下载、旅游资讯）添加或删除相关友情链接。

（4）语料管理：管理员可对语料库中所有的语料进行修改、增加和删除操作。在添加语料的同时，系统可利用正则表达式强大的查询匹配功能，将语料篇章分割为若干条中英文相对应的平行语料，以记录的形式存入语料库之中。

（5）评论信息管理：主要实现评论信息的查看和删除，评论版块提供了用户与管理员交流的平台，用户可以对语料库系统提出宝贵的意见和建议。

（6）系统配置：主要提供查看管理员手册、修改管理员手册、修改用户使用帮助、修改联系方式等功能。

（7）查询检索：系统不仅提供全库检索和分类检索功能，还加入了精确和模糊检索功能，关键词（Keywords）以中文或英文的任意字符串为主，之间以空格分隔，系统加入自动识别和词组过滤机制，上述功能使用户能更准确的搜索到所期待的结果。语料检索页面如图 2。

图 2　语料库系统检索页

3.技术路线

陕西省旅游景区公示语翻译语料库系统是由西北大学外国语学院和现代教育技术中心联合研发。主要采用目前 Web 开发的绝佳组合 LAMP（Linux+Apache+MySQL+PHP）。后台采用服务器端脚本语言 PHP，前端采用 Web 标准 DIV+CSS 实现网页布局，并融入了最关键的技术正则表达式（Regular Expression），它自身具有一套非常完整的、可以编写模式的语法体系，提供了一种灵活且直观的字符串处理方法[5,6]，利用它使得在语料添加、模糊查询、精确查询等功能的实现中发挥了巨大的作用。

4.系统架构

本系统采用 B/S 体系架构(Browser/Server,浏览器/服务器模式)[7],从逻辑上可分为表现层(Presentation layer)、业务逻辑层 (Business logical layer) 和数据访问层(Data access layer)。浏览器、Web 服务器、数据库服务器分别和表现层、业务逻辑层、数据访问层相关联。如图 3 所示,当用户或管理员通过终端浏览器向 Apache 服务器发出查询语料等请求时，服务器解析请求并与数据库服务器进行交互完成业务逻辑处理，随后将查询到的信息以 HTML 的形式返回到客户端浏览器中。

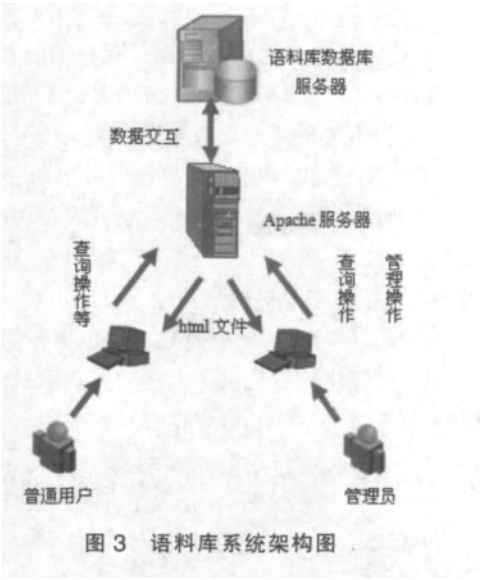

图 3 语料库系统架构图

5.系统数据库设计

本系统采用 MySQl5.1.41 数据库，根据系统的功能需求和各模块设计,需建立 7 个数据表,分别为:语料文章信息表、语料句词信息表、语料分类信息表、用户信息表、友情链接信息表、评论信息表、系统配置信息表。

用建模工具 PowerDesigner 对系统数据库图进行规划,如图 4 所示。语料文章信息表(articles)包括文章编号、文章英文标题、文章英文内容、文章中文标题、文章中文内容、文章特色图片、用户编号、语料分类号字段,其中用户编号与用户信息表中的用户编号相关联，语料分类号与语料分类信息表中的分类编号相关联；语料句词信息表(sentences)包括句词编号、句词英文内容、句词中文内容、文章编号,用户编号、语料分类号字段,其中文章编号与语料文章信息表中的文章编号相关联，用户编号与用户信息表中的用户编号字相关联，语料分类号与语料分类信息表中的分类编号相关联；语料分类信息表(sorts)包括分类编号、父分类编号、分类名称和分类描述字段;用户信息表(users)包括用户编号、用户名称、用户密码、用户角色等字段;友情链接表(links)包括链接编号、链接名称、链接所属分类、链接具体地址字段;评论信息表(messages)包括评论信息编号、评论信息内容、用户编号字段，其中用户编号与用户信息表中的用户编号相关联;系统配置信息表(options)包括配置信息编号、配置信息名称、配置信息分类等字段。

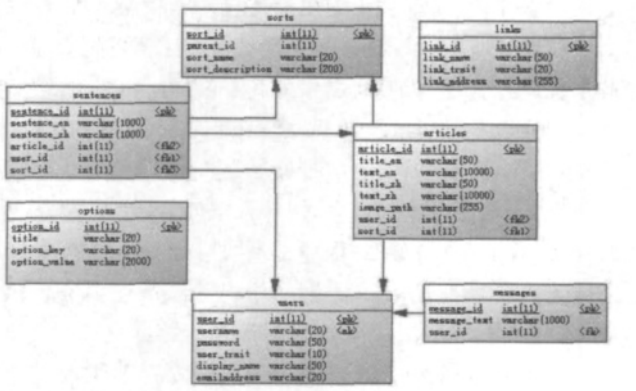

图 4 系统数据库图

四、系统实现

下面从系统的逻辑框架出发分别阐述数据访问层、业务逻辑层、表现层的实现过程。

1.数据访问层设计

数据访问层即持久层，它是一组封装了对数据库进行 CURD(创建、更新、读取、删除)操作的类。系统中数据访问层设计了 8 个类,如表 2 所示。

表 2 系统中定义的类说明

类名	说明
db.class	数据库操作类
sorts.class	语料分类操作类
users.class	用户信息操作类
articles.class	语料文章操作类
sentences.class	语料句词操作类
messages.class	评论信息操作类
Links.class	友情链接操作类
option.class	系统配置操作类

(1)sentences 类的设计

以 sentences 类为例说明类中定义的方法。sentences 类定义了对语料句词信息表的各种操作方法，此类中的方法定义如表 3 所示。

(2)方法的实现

下面以 sentences 类中 getSentences()方法为例,阐述方法的实现，当管理员想要编辑或查询某一篇语料文章

表 3　sentences 类中定义的方法说明

方法	说明
function searchSentences	检索并返回符合条件语料句词
function searchSentencesTotal	检索统计并返回符合条件语料句词的条数
function getSentences	检索并返回某一篇文章的语料句词
function getSentencesTotal	检索统计某一篇文章语料句词的条数
function getSentence	得到某一条句词的具体信息
function addSentence	增加一条语料句词
function updateSentence	更新某一条语料句词
function deleteSentence	删除某一条语料句词

所对应的平行语料时，点击相应按钮，这时程序会调用getSentences()方法，参数为此文章的编号、起始序号、每页显示的条数，随后程序会组合生成 select 语句在 sentences 表中进行查询，将查询到的每一条记录的每个字段分别存入 sentence 对象的成员中，然后将每个对象存入数组 sentence_array 中保存起来，即为所查询的结果。

2.业务逻辑层设计

业务逻辑层处在三层架构中最关键的位置，起到了承上启下的作用，他主要负责从表现层获得用户输入的数据，并调用数据访问层提供的相应方法完成和业务需求有关的功能。在此，介绍一下用户检索语料的策略在该层中的实现。

语料库系统的查询检索功能分为全库检索和联动分类检索，为了实现查询检索的精确度，系统还加入了精确和模糊两种检索模式，另外还加入了本系统的一大创新点“词组过滤机制”。当用户在搜索框中输入以空格分割的中英文关键词，选择相应的搜索模式，点击搜索时，系统会调用封装在 sentences 类中的 searchSentence ()方法，参数为关键词、栏目一级分类、栏目二级分类、起止序号、每页显示的条数、搜索模式，然后系统会根据传入的参数判断是英文还是中文关键词，并经过一系列复杂的程序组合生成 SQL 语句，如果为精确搜索，系统会利用 MySQL 提供的扩展正则表达式中的 regexp 操作符来精确匹配关键字，如果为模糊搜索则不使用正则表达式，随后将查询到的每一条结果再利用正则表达式提供的搜索替换操作符 preg_replace 在关键词出现的地方加上特殊标记，以备输出时做高亮显示，最后循环以对象的形式存入数组中。

“词组过滤机制”是本系统不同于其他语料库系统的一个地方，例如：用户检索 temple 但不想让结果出现 museum，可以在搜索框中输入“temple –museum”。这一机制的实现过程为，当用户输入关键词点击搜索时，系统会对关键词进行分析，如果发现某个关键词前面有“–”号标记，程序将会在组合 SQL 语句中加入 “and no regexp 关键词”，这样在执行 SQL 语句时，有此关键词的语料将会被排除在外。

3.表现层设计

表现层是人机交互的接口，主要是处理用户的输入和回显系统处理后的数据。当用户发出检索请求时，服务器端经过处理将所查询到的平行语料存入对象数组中并返回，然后程序再循环从数组中取出相应的平行语料，如果语料条数过多，则可调用分页操作类中的分页方法进行处理并分页显示，显示结果时，关键字红色高亮显示，以便快速定位关键词在语料中的位置。

五、总结

本文从语料库语言学的观点出发，规划了陕西省旅游景区公示语翻译语料库的主题栏目，设计和实现了本语料库系统。作为陕西省的首个旅游景区公示语翻译语料库系统，它的成功构建与开发为陕西省旅游景区公示语的搜集整理、理论探究、创新发展、综合利用注入新的活力，而且有助于传播旅游文化，增进中外人民的友谊，让世界进一步了解陕西，让陕西走向世界，为提升陕西作为文化旅游大省的国际竞争力和文化软实力贡献一份力量。本系统目前为止基本实现了语料库系统所需的功能，但仍然有一些问题没有考虑周全且有很多高级的功能未能实现，例如：字词频统计、词性还原、中文分词等。因此，接下来的工作是加固系统的基本功能并尽可能地实现系统的高级功能。◉

参考文献：

[1]Taner Sezer.Corpus linguistics theory and design and application of a Turkish corpus[D].Turkey:Mersin University PhD thesis,2005.

[2]魏顺平,何克抗.小学语文教学语料库的设计与开发[J].中国电化教育,2007(245):66–69.

[3][4]李德超,王克非.新型双语旅游语料库的研制与应用[J].现代外语,2010(1):46–54.

[5]高洛峰.细说 PHP[M].北京:电子工业出版社,2009.

[6]Ben Forta(著),刘晓霞等(译).MySQL 必知必会[M].北京:人民邮电出版社,2011.

[7]百度百科.B/S 架构[EB/OL].http://baike.baidu.com/view/1477348.htm.

（编辑：王晓明）