

面向语料库机助辞书编纂系统的设计与实现^{*}

张永伟 顾曰国 胡钦谕 曹新龙

摘 要 “面向语料库机助辞书编纂系统”由中国社会科学院语言所和中国多语言多模态语料库暨大数据研究中心研发,是一个可以充分地利用各种语料库、辞书库和其他相关资源库辅助汉语辞书编纂的平台。文章介绍了该系统的研发背景、目标、系统设计与实现。此外,对系统的技术选型也进行了简单介绍,为辞书编纂或相关系统的研制提供借鉴。

关键词 辞书编纂系统 语料库 设计与实现

一、引 言

随着信息技术的不断发展,计算机辅助辞书编纂已经逐渐成为趋势。辞书编纂系统可以提升辞书编纂效率,缩短辞书编纂周期,节约辞书编纂成本,提高辞书质量。(张亚斌,赵胜男等 2017)近些年来,许多机构都研制了功能丰富的辞书编纂系统,但是在有关汉字的处理上仍存在不少问题,不完全适应汉语辞书的编纂需求。如华烨、李亮(2012)对 IDM DPS、ABBYY Lingvo Content、Tlex 等几款在国际上有影响的辞书辅助编纂系统进行了对比,指出这些系统在汉字处理、汉字样式设定、汉语界面支持、汉语有关自动化支持以及汉语语料库支持等方面尚存在许多不足。张永伟(2020)也从汉字录入、存储与显示,汉字样式设置,汉语条目辅助注音,汉语条目自动排序,汉语条目内容自动检查等方面对比了 Tlex、FLEx、Termbases、DEBWrite、Léacslann 等国外辞书编纂系统,指出这些系统在汉字处理上存在不足。

为了辅助《现代汉语词典》《新华字典》等系列语文辞书的编纂,中国社会科学院语言研究所研制了“人机交互式的汉语辞书编纂系统”(傅爱平,吴杰等 2013),但是该系统研发时间较早,在兼容性、功能性、安全性等方面都需要进一步改进。为此,我们又重新研制了“面向语料库机助辞书编纂系统”(Corpus-Oriented Computer-Assisted Lexicography,简称 COCAL),期望 COCAL 能够充分地利用各种语料库、辞书库和其他相关资源库,辅助进行各种汉语辞书的编纂。

^{*} 本文系国家语委“十三五”科研规划 2020 年度一般(委托)项目“辅助语文辞书编纂的人工智能关键技术研究”(项目编号 WT135-69)、国家社科基金项目“基于《现代汉语词典》的词汇计量研究”(项目编号 20BYY170)、中国社会科学院辞书编纂研究中心项目“面向语料库机助辞书编纂系统:V1.0 版本升级 V2.0”阶段性成果。承蒙谢冰春及本刊匿名审稿专家审阅初稿并提出宝贵意见,谨此致谢。

二、整体功能架构设计

COCAL 主要由后台管理、前台编纂以及公共应用三个系统组成。其中,后台管理系统包括辞书管理、资源管理、用户管理、系统管理四个模块,主要面向辞书管理人员(比如主编)、资源管理人员和系统管理人员。前台编纂系统包括辞书编纂、语料库检索、辞书检索、词表检索和在线资源检索五个模块,主要面向辞书编纂和审稿人员。公共应用系统包括输入助手、修改密码、登录/登出三个模块,面向所有用户。COCAL 整体功能架构设计框图如图 1 所示:

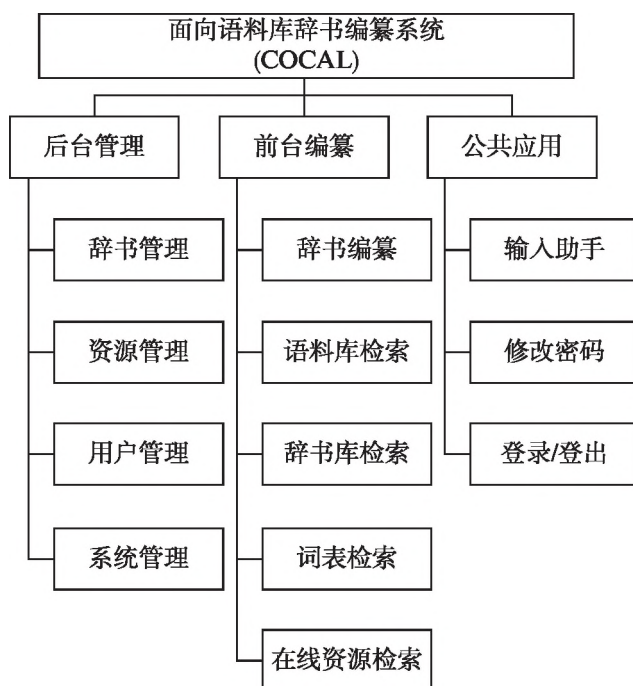


图 1 系统整体功能架构设计框图

公共应用系统中修改密码、登录/登出等是常见的功能,本文不详细介绍。输入助手是本系统实现的关键技术之一,本文将在第六部分介绍。为了更直接地展示系统实现结果,我们对重点功能给出了相应的图示。

三、技术选型与数据库实现

(一) 技术选型

COCAL 使用了主流 MVC^[1] 框架开发,采用 B/S(Browser-Server,浏览器-服务器)体系架构。用户无需安装任何软件,仅通过浏览器就可以使用系统的全部功能。所有数据均保存在中心服务器中。

COCAL 使用 Java、FreeMarker、JavaScript、HTML 等作为程序主要开发语言,使用 CSS、Bootstrap、JQuery、Ajax 等前端网页样式显示和控制技术。COCAL 使用的第三方开发框架主要有 SpringMVC 模型-视图-控制框架、Hibernate 数据库持久化框架、Apache Lucene 全文索引工具包、Apache Shiro 权限管理框架、FreeMarker 页面模板等。根据保存数据类型不同,COCAL 使用了两种数据库,分别是 MySQL 关系型数据库和 MongoDB 文档数据库。

（二）数据库实现

COCAL 管理的数据包括三种类型,分别存储在不同数据库中。词表的详细信息、条目编纂历史、操作日志等数据量大、动态增长速度快、结构复杂多变的数据保存在 MongoDB 文档数据库中;需要全文检索高级查询的数据经过分词等处理后保存在 Lucene 的索引中;其他数据保存在 MySQL 关系型数据库中。MongoDB 具有更好的伸缩性和灵活性,适用于海量数据的管理;Lucene 适用于全文检索。

COCAL 各类数据在不同数据库中的存储的情况如图 2 所示:

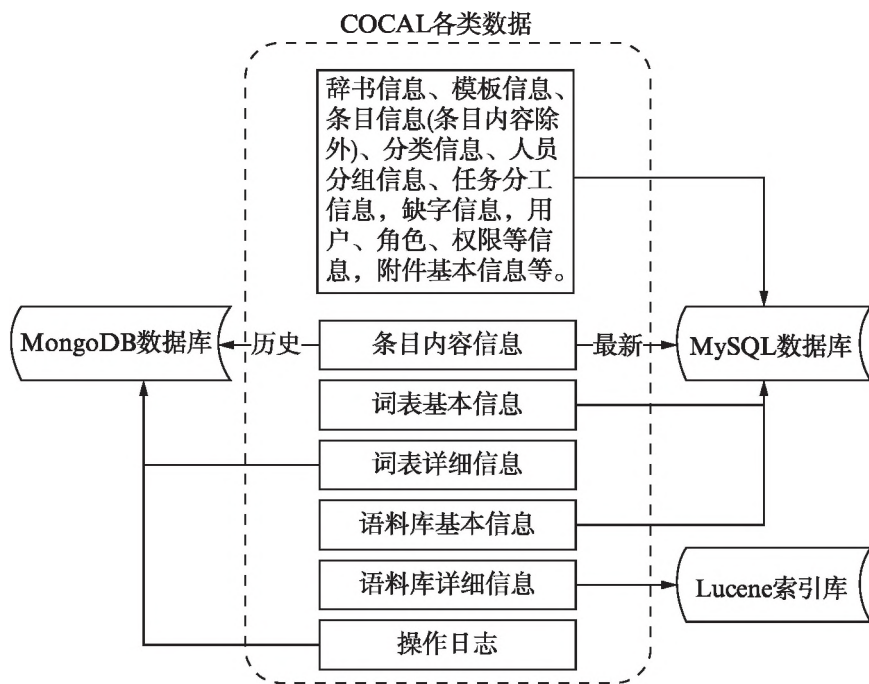


图 2 系统各类数据存储示意图

四、后台管理系统的设计与实现

（一）辞书管理

辞书管理模块是后台管理系统的重点,涉及体例、条目与人员分组等管理要素,主要由辞书管理人员操作,具体包括以下 6 个子模块:

1. 模板管理

模板管理子模块管理辞书体例。它管理编写辞书条目时需要填写哪些内容(例如词目、拼音、释义、配例等),这些内容使用什么编辑器编辑,编纂人员编纂条目内容时需要受到怎样的约束限制以及条目内容如何排版显示等。每部辞书的体例不同,条目内容和排版显示均可能存在差异。条目编纂模板和内容显示模板均可以自定义,使得系统能够用于编纂各种辞书,并且还能支持这些辞书的个性化内容显示。模板管理界面如图 3 所示。

模板管理界面包括条目编辑器、条目编纂模板、内容显示预览、内容显示模板等区域。其中条目编纂模板使用 Json 语言定义了条目编辑器包括哪些表单元素、表单元素如何排列、编纂人员编纂条目内容时受到哪些约束限制等。COCAL 自动解析条目编纂模板的内



图3 辞书管理-模板管理界面

容,生成条目编辑器。内容显示模板使用 JavaScript 语言定义了条目内容应如何排版显示,JavaScript 语言的灵活性决定了条目内容可以被灵活地排版显示。条目编辑器的输出是 Json 格式(由 Json 语言定义)的条目,COCAL 自动执行内容显示模板定义的 JavaScript 语句,接受 Json 格式的条目作为输入,输出 HTML 格式或者纯文本格式的条目内容,最终显示在内容显示预览区。

不难看出,辞书管理人员需要根据辞书体例为新建的辞书项目设计条目编纂模板和内容显示模板,这虽然对辞书管理人员的操作提出了较高的要求,但也确保了 COCAL 具备足够的灵活性。为了降低辞书管理人员设计模板时的难度,我们提供了丰富的参考示例。此外,辞书的条目编纂模板和内容显示模板一旦设计完毕,后期将无需再频繁维护,避免了辞书管理人员的持续投入。基于这样的设计,COCAL 不仅可以用于编纂汉语辞书,也可以用于编纂各种外语辞书,甚至多语辞书,具备极大的灵活性。

2. 分类管理

分类管理子模块管理辞书条目的类别信息,为任务分工做准备。类别是条目的静态属性,与辞书参编人员的分组一一对应。辞书管理人员需事先定义辞书条目的类别。类别名称和数量没有限制,编纂过程中随时可以进行增删和修改。比如条目可以分为“A—C 母”“D—F 母”等类别,也可以分为“语文”“哲社”“科技”等类别。在分类管理界面可以新增、查看、编辑、删除、查询条目类别信息。

3. 参编人员管理

参编人员管理子模块管理辞书的编纂人员和审稿人员。辞书管理人员从系统注册用户列表中选取参与辞书编纂的工作人员。在参编人员管理界面可以添加、移除参编人员。

4. 收词管理

收词管理子模块管理辞书收录哪些条目,以及各属于哪些类别。COCAL 的条目来源有三种:完全新增、从系统其他辞书条目中复制和从系统外导入。辞书管理人员通过收词管理子模块确定辞书收词规模和范围,为条目指定类别(分配分组)。收词管理界面如图 4 所示:



图 4 辞书管理-收词管理界面

在收词管理界面可以新增、编辑、复制、删除、恢复、导入、导出(Word 和 XML 两种格式)、查询条目,对条目进行分类。为防止误删除,我们将删除分为临时删除和永久删除两种,临时删除的条目可以恢复,永久删除的条目不可以。临时删除的条目依然对编纂人员可见,以文字增加删除线的形式显示。收词管理界面新增、编辑的信息只包含词目、拼音、分类等基本信息,条目详细的释义需要在前端编纂界面进行编纂。

5. 分工管理

分工管理子模块由人员分组和任务分工两个更小的子模块组成。辞书管理人员创建

条目类别后,系统自动依据条目类别为参编人员创建同名分组,人员分组名称和条目类别名称一一对应。在人员分组子模块中,辞书管理人员可以添加组员,任命或解任组长。组长可以查看、编辑、审核、管理组员承担的条目,同一个参编人员可以分属不同分组,也可以同时作为多个分组的组长。人员分组界面如图5所示:



图5 辞书管理-分工管理-人员分组界面

在人员分组界面可以添加、移除每个分组的组员,指定、取消组长。任务分工子模块主要用于为条目分配编纂人员。辞书管理人员可以为每个条目分配具体的编纂人员(我们称之为分工),取消指定的编纂人员等。任务分工界面如图6所示:



图6 辞书管理-分工管理-任务分工界面

在任务分工界面选择待分工条目后,点击“分工”按钮,在条目所属分组的人员列表中选择具体的参编人员即可实现条目的分工。条目分工后,辞书编纂人员可以在编纂时看到所分配的条目。

6. 统计

统计子模块统计辞书各分组、参编人员、编纂进度、编纂时长的详细情况,以图表形式

直观显示,便于辞书管理人员随时掌握辞书收词的整体状况、编纂进度,各参编人员的工作状况等信息。统计界面如图 7 所示:

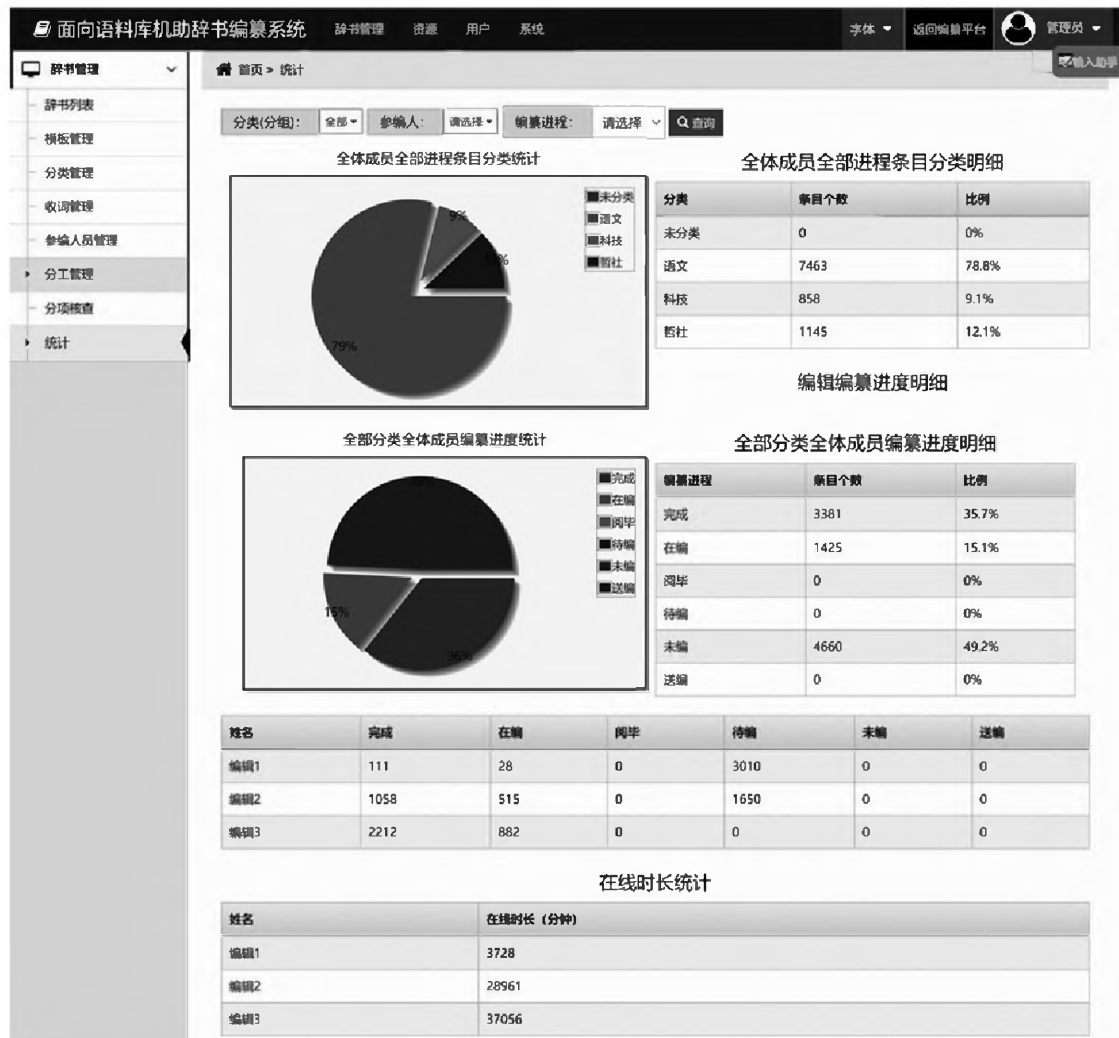


图 7 辞书管理-统计界面

在统计界面,人员分组(条目分类)、参编人员、编纂进度等作为过滤条件可以自由组合。比如,可以统计整部辞书各分组条目比例,统计某个编纂人员分配的不同条目类别比例,统计某个分组下成员们各自的编纂进度,等等。

(二) 资源管理

COCAL 的特色之一是集成了丰富的语言资源。资源管理模块主要由资源管理人员操作,具体包括以下 3 个子模块:

1. 语料库管理

辞书编纂越来越离不开语料库。资源管理人员可以通过语料库管理子模块管理辞书编纂时需要参考的文本语料库,为编纂条目提供参考例句。语料库管理界面如图 8 所示:

在语料库管理界面可以新增、编辑、删除语料库基本信息(包括语料库的名称、版本、来源等),用户上传语料后,单击“重建索引”或“删除索引”按钮,为语料库建立或者删除索引。



图 8 资源管理-语料库管理界面

2. 词表管理

这里的词表并非指词目列表,而是指供辞书编纂参考的包括词目、拼音、释文等内容在内的弱结构化数据。之所以称为词表,是为了和利用 COCAL 系统编纂的辞书相区分。资源管理人员可以将需要参考的纸本辞书数字化,分离出词目、拼音、释文等内容后,使用条目管理子模块导入系统,供编纂条目时参考使用。资源管理人员创建一个词表后,可以进入该词表的条目管理界面,管理该词表内的条目。某词表的条目管理界面如图 9 所示:



图 9 词表管理-条目管理界面

在条目管理界面,可以新增、编辑、删除和批量导入词表条目。

3. 在线资源管理

辞书编纂需要参考许多互联网资源,辞书编纂人员编纂条目时需要在不同的互联网资源中反复切换和查询,耗时耗力。资源管理人员可以在在线资源管理子模块中添加在线资源,便于编纂人员在编纂条目时参考。在线资源管理界面如图 10 所示:



图 10 在线资源管理界面

在在线资源管理界面,可以添加、编辑、删除在线资源。

(三) 账号管理

账号管理模块包括权限管理、角色管理、用户管理三个子模块,主要由系统管理人员为不同角色的用户分配权限。同一角色具有相同权限,每个角色可以拥有多个权限,每个用户可以分属不同角色。辞书管理人员、资源管理人员、系统管理人员、辞书编纂人员和审稿人员等都是系统内置的角色,已经为他们分配了相应的权限。COCAL 也支持添加新的角色,以及为角色指定不同的权限。权限、角色、用户分离的设计既保证了系统功能分配的灵活性,又保证了系统的安全性。

(四) 系统管理

系统管理模块包括日志管理、上传文件管理等子模块。其中,日志是用户在系统中操作的记录。系统管理员可以通过日志管理子模块查询自动保存的用户操作记录,通过上传文件管理子模块查看、删除或者下载文件。由于系统管理模块的各项功能较为常见,本文不展开介绍。

五、前台编纂系统的设计与实现

(一) 资源检索

资源检索模块包括语料库检索、辞书库检索、词表检索和在线资源检索四个子模块。其中,语料库检索子模块用于检索文本语料库中的例句,辞书库检索子模块用于检索系统

内在编或已经编纂完成的辞书,词表检索子模块用于检索词表中的条目,在线资源检索子模块用于检索常用的互联网资源。资源检索模块主要为辞书编纂人员编纂条目时提供参考,提高编纂效率。

(二) 辞书编纂

辞书编纂模块是前台编纂系统的重点,主要包括条目列表和条目编纂两个子模块。用户登录前台编纂系统后,首先看到的是待编辞书列表,列表中列出了所有分配了编纂任务的辞书,在辞书列表中选择某一部辞书后,即可进入条目列表界面。

1. 条目列表

条目列表子模块显示了所有分配的条目,编纂人员可以新增、删除、恢复、编纂条目,优化条目排序,统计个人编纂情况,批量保存、提交、审阅、退回条目等。其中,组长可以查看或编纂所有组员的条目。条目列表界面如图 11 所示:



图 11 条目列表界面

编纂人员可以在条目列表界面看到条目内容、分类(分组)、进程、版本等信息。单击条目后面的“编纂”按钮就可以进入条目编纂界面。

2. 条目编纂

条目编纂子模块是辞书编纂人员使用最多的功能,辞书编纂的主要工作就在条目编纂界面中完成。条目编纂界面分为条目列表区、预览区、编纂区以及资料区四个区域。辞书编纂人员可以在条目列表区查看相关条目,在预览区查看条目排版后的文本,在编纂区编纂条目内容,在资料区查询和浏览参考资料。

条目编纂进度分为未编、待编、在编、完成、阅毕五种。未分配任务时,条目编纂进度为未编;分配任务后,编纂进度改为待编。在条目编纂界面,编纂人员可以保存、提交编纂结果,组长可以保存、提交、退回或阅毕编纂结果。编纂人员、组长均可编纂待编条目。条目编纂完成后保存,编纂进度改为在编;编纂完成后提交,编纂进度改为完成。组长审阅编纂进度为完成

自动生成。资料区除了语料库检索、辞书库检索、词表检索、在线资源检索以外,还包括当前条目的修改记录。用户编纂某个条目时,系统自动将词目作为检索词在各个资源库中搜索,并在资料区中显示检索结果,辞书编纂人员可以直接点击查看。

此外,在条目编纂界面,COCAL 还提供了许多便捷功能,比如可以快速进入上一组或下一组条目(比如相同字头条目或者同一个语义类的条目等)的编纂界面,对比条目内容、恢复条目历史版本、划词查询当前辞书是否收录该词目等,这些都是 COCAL 在使用过程中由编纂人员建议增加的实用功能。

六、公共应用功能的设计与实现

公共应用功能包括输入助手、修改密码、登录/登出等功能。修改密码、登录/登出是常见功能,本文不展开介绍。输入助手用于查询、输入缺字,是 COCAL 的特色功能,也是汉语辞书编纂系统的一个难点。

缺字(又称外字)包括 Unicode 标准已编码但尚无字体字模支持的汉字以及 Unicode 标准尚未编码的汉字。国外辞书编纂系统和一些国内辞书编纂系统对缺字的支持不够友好(张永伟 2020),COCAL 解决了这个问题,能够支持全部汉字的处理。

缺字的基本信息包括编码、拼音、笔顺、笔画数、字形图片等。这些信息是普通语文辞书条目排序需要的基本信息,也是查找缺字所需要的基础信息。COCAL 支持缺字的新增、编辑、删除、查询,还会将字形图片自动转换为 SVG 和 WOFF 字体格式^[2]。此外,COCAL 还研制了输入助手,用于查询和输入缺字。输入助手输入、显示缺字的原理是:(1) 自动记录光标在系统表单中的位置,在输入助手中选择某个缺字后,将缺字信息(“« + 缺字编码 + »”形式)填入相应位置;(2) 识别所有页面中“« + 缺字编码 + »”形式的符号串,根据不同页面的需要,自动将其替换为名为@ font-face 的 CSS@ 规则^[3]或者缺字字形图片。

COCAL 界面右侧始终显示“输入助手”按钮,单击后打开或关闭输入助手。录入缺字信息后,所有用户均可以随时通过输入助手查询、输入缺字。目前,输入助手支持笔画数查询和拼音查询两种查询方式。输入助手的拼音查询界面如图 13 所示。

七、结 语

COCAL 提供丰富的语料库资源,支持协同编纂各种汉语辞书,提升了辞书编纂效率,提高了辞书质量,初步达到了研发目的。但是 COCAL 仍有许多地方需要改进。

首先,COCAL 主要是为编纂原创辞书研发,注重“编”而未强调“查”,COCAL 对有相同或类似特征的条目进行有针对性的专项核查尚存在改进空间。其次,COCAL 只支持文本语料库查询,提供例句,但未利用自然语言处理技术充分挖掘语料,提供诸如词频表、搭配列表、语块列表等功能。最后,人工智能技术已经同许多领域结合,产生了革命性的智能产品和服务,但在包括 COCAL 在内的辞书编纂系统中的应用相对滞后。利用人工智能技术可以进行义项的自动划分,“旧词新义”的自动发现,例句的辅助生成,条目内容自动检查等。在辞书编纂时集成这些人工智能技术将是 COCAL 下一阶段的目标。



图 13 “输入助手”拼音查询界面

附 注

- [1] MVC 是 Model View Controller 的缩写,是一种经典且实用的软件开发框架模式。
- [2] 之所以选择 SVG 和 WOFF 这两种字体格式,是因为它们可以被绝大多数浏览器支持,具有更好的兼容性。
- [3] CSS 的 @font-face 规则允许网页开发者为其网页指定自定义的在线字体。

参考文献

1. 傅爱平,吴杰,张弘,等. 人机交互式的汉语辞书编纂系统. 辞书研究,2013(6).
2. 华焯,李亮. 国际计算机辅助词典编纂系统管窥. 辞书研究,2012(5).
3. 陆汝占. 汉语词典编纂一体化环境(上). 辞书研究,2000a(2).
4. 陆汝占. 汉语词典编纂一体化环境(下). 辞书研究,2000b(3).
5. 张亚斌,赵胜男,何朝辉,等. 数字化辞书协同编纂系统的设计. 辞书研究,2017(6).
6. 张永伟. 辞书编纂系统的汉字处理: 挑战与解决方案. 辞书研究,2020(1).

(张永伟 顾曰国 胡钦谔 中国社会科学院语言研究所/
辞书编纂研究中心 北京 100732)

(顾曰国 曹新龙 中国多语言多模态语料库暨
大数据研究中心 北京 100089)

(责任编辑 马 沙)

“being bought” by the seller who divides, weighs, or grabs the goods in the trading process. Finally, this article suggests that dictionaries should establish new entries or add new senses to old entries where ambiguity arises.

Keywords: purchase, metonymy, cut meat (buy meat), tear cloth (buy cloth), entry inclusion

The Design and Implementation of a Corpus-Oriented Computer-Assisted Dictionary Compilation System

Zhang Yongwei Gu Yueguo Hu Qin'an Cao Xinlong

Abstract: The corpus-oriented computer-assisted dictionary compilation system was developed by the Institute of Linguistics of the Chinese Academy of Social Sciences and the China Multilingual and Multimodal Corpus and Big Data Research Center. It is a platform for Chinese dictionary compilation by making full use of various corpora, dictionaries, and other related resources. This article introduces the research and development background, goals, system design and implementation of the system. In addition, we give a brief introduction to the technical solutions selected during system development, which can provide reference for dictionary compilation or the construction of related systems.

Keywords: dictionary compilation system, corpus, design and implementation

On-*men*'s Function Mechanism from Its Collocational Restrictions

Zhang Xinhua

Abstract: When studying the Chinese suffix *-men*, we must be clear about the word that collocates with it, i. e., whether it is a pronoun or a noun. This article focuses on the latter alone. Clarifying the categorical nature of the NP that collocates with *-men* is the precondition for the study of *-men*'s function. However, this is still controversial in the academic world. This article first demonstrates that the N in N-*men* denotes kind. The category of kind is not a particular sub-class of nouns, but a higher category supervening all nouns. This function covers all particular objects (extension) based on some common properties (intension), and the properties are more prominent and the objects are suppressed. The function of *-men* is to reverse that relation. Scholars often hold that the N-*men* has a definite, but we disagree with that, and maintain that the definiteness is only a peripheral feature, and is not an solid one, for the N-*men* may indicate generic reference. Secondly, this article investigates the collocational restrictions of *-men* that concurs with attribute nouns, collective nouns, and relational nouns. We take the attribute nouns as a counterexample. Since those nouns make properties prominent and ignore objects, they reject *-men*. On the other hand, many attribute nouns can enhance their thingness and accept *-men*. In fact, the collective nouns are not a perfect noun sub-class, and have heterogeneous members. The typical collective nouns denote objects directly and reject *-men*. But such collective nouns are rare. Most of them denote kind and accept *-men*. Differently from the predominant viewpoint, this article demonstrates that the relational nouns all denote kind and accept *-men*.

Keywords: kind, plural, definite, extensive, collocational restrictions