文章编号：2095-2163(2019)04-0100-04 中图分类号：TP319 文献标志码：A

面向语言分析的语料库技术平台建设

马创新[1]，梁社会[2]
（1 江苏师范大学 语言科学与艺术学院，江苏 徐州 221009；2 南京师范大学 国际文化教育学院，南京 210097）

摘 要：为了提高语言研究者的工作效率，开发了语料库技术平台 Inspire1。本系统兼具通用性、全面性、一体化和易用性等特点，集成了语料采集、加工、统计、检索和分析等 5 个模块。利用本系统，能够使研究者直观地观察到语料库中蕴含的语言规律，从语料库中发现新的知识。

关键词：语料库；语言研究；软件

Construction of corpus technology platform for language analysis

MA Chuangxin[1], LIANG Shehui[2]
(1 School of Linguistic Sciences and Arts, Jiangsu Normal University, Xuzhou Jiangsu 221009, China;
2 International College for Chinese Studies, Nanjing Normal University, Nanjing 210097, China)

【Abstract】 In order to improve the working efficiency of language researchers, the paper designs and implements a common corpus technology platform Inspire1. The system integrates five modules. Using this system, the linguists can intuitively observe the linguistic rules, and discover new knowledge from the corpus.

【Key words】 corpus; language research; software

0 引 言

在当今大数据时代，人们可以利用的数据量每年都以指数倍增长，所以在语言学研究中，原始语料的获取已经不再是难题，而如何利用先进的智能技术高效地采集语料、加工语料和分析语料，已成为当今语料库语言学界亟需解决的重要问题。

语料库建设和应用技术能够减轻研究者的工作负担，提高语言研究的效率。因此，构建一个语料库技术平台，对于语言教学和研究有着较大的实际意义[1]。语料库技术平台建设是一项多学科交叉的复杂工作，研究者不仅要掌握先进的计算机技术和知识组织方法，还要具备深厚的语言学功底。

1 语料库处理软件概述

1.1 当前常用的语料库软件

许家金和贾云龙[2]参照 McEnery & Hardie[3]对语料库软件的分类方式，提出按照语料库软件的运行环境可以把语料库工具分为 3 类，一是运行在 DOS 环境下的工具，如：CLOC、XANADU、TACT、MiniConcordancer、MicroConcord 等；二是运行在 Windows 或其它图形操作系统中的工具，如：Wordsmith Tools、AntConc、MonoConc Pro 等；三是基于互联网的语料库网络应用工具，如：CQPweb、BYU corpora、SketchEngine 等。

李亮[4]按照语料库软件开发者的国籍来划分，当前常用的语料库软件和其来源国分别是：美国有 Conc、Paraconc、Monoconc；英国有 MicroConcord、Wordsmith Tools、Longman MiniConcordancer、Free TextBrowser、Concordance；德国有 LEXA、TextSTAT；加拿大有 Concorder；日本有 CorpusWizard；中国香港有 Concapp。从语料库软件的数量和品质两方面来看，英国在该领域占据领先地位，其次是美国和德国。

1.2 普遍存在的问题

分析众多语料库处理软件，笔者发现国内开发的语料库软件数量少、使用率低[5]。此外，这些语料库软件还普遍存在以下几方面的问题：

(1)用于分析和处理汉语语料的软件较少。汉语具有与英文不同的特点，比如在计算机字符集中，一个汉字与一个英文字母所占用的存储单元是不同的。再如汉语还存在分词连写的问题，不像英文每个单词之间都有间隔。

(2)有些语言处理软件的功能单一，并且只能完成浅层任务。仅能用于某一项具体的语言处理工

基金项目：江苏省社科基金(15YYC001)；国家社科基金(15BYY096)。

作者简介：马创新(1980-)，男，博士，讲师，主要研究方向：计算语言学；梁社会(1979-)，男，博士，副教授，主要研究方向：计算语言学。

收稿日期：2019-04-17

作，在实际的语料处理中，需要使用多个软件才能完成一项任务。

（3）有些语言处理软件易用性较差。主要表现在设计不合理、界面不友好、操作复杂、没有做到简单易用、难以在语言学领域推广使用。

为了能够切实解决语言研究中的困难，提高工作效率，针对当前语料处理软件所存在的问题，笔者提出设计语料处理软件的 4 条原则[6]：

（1）通用性原则。全世界现有语言大约在 5 000～7 000 种之间，使用人口超过 100 万的语言约有 140 多种，有文字的语言在 930 种左右。开发的软件应该具备广泛的通用性，能够处理汉语、英语、法语、俄语等使用人口较多的语言文字。

（2）全面性原则。应该开发功能集成化的“分析型深层工具”，所设计的语言处理软件不仅能够发现表层语言现象，而且能够挖掘出深层语言规律。

（3）一体化原则。软件的各项功能要按照语料处理时的先后顺序进行组合，而不是简单叠加在一起。语料采集、加工、统计、检索、分析等各项功能及其子功能之间要具有一定的逻辑关系，形成统一的功能整体。

（4）易用原则。软件设计应遵循用户至上原则，采用访谈法和问卷调查法充分了解语言研究者的需求状况。在人机接口的设计方面，做到简易直观，让用户通过很少的学习和训练，就能够使用软件[7]。

2　系统模块与功能设计

本系统使用的编程语言是 C++，编程工具是 Microsoft Visual Studio Community 2015，使用了 MFC 类库[8]。其主要功能模块如图 1 所示，分为 6 个子模块：公用模块、分析、检索、统计、加工和采集模块。公用模块的功能是选取、显示和输出语料文件的，其它 5 个子模块都要用到公用模块来选择和浏览待处理语料文件、以及显示与输出处理后的结果文件。

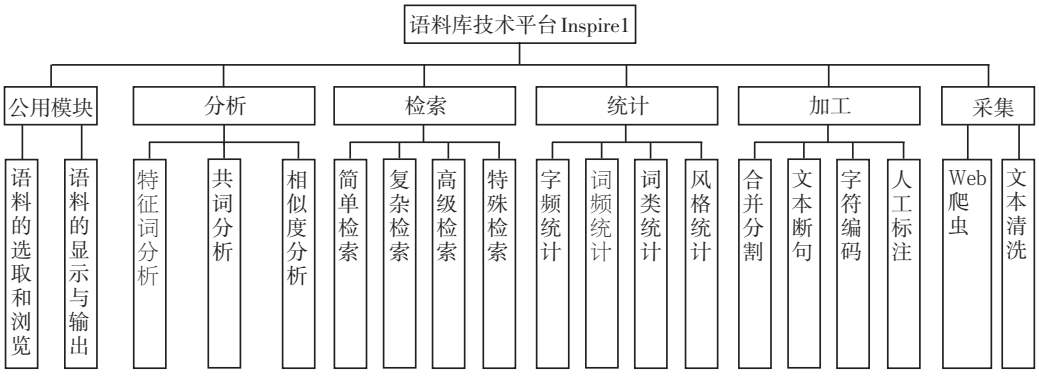

图 1　Inspire1 的主要功能模块

Fig. 1　The main functional modules of Inspire1

语料库技术平台 Inspire1 主要包括 5 大功能，对此可做阐释分述如下。

（1）语料采集功能。包括 2 项子功能：

①WEB 爬虫。用以获取指定网页中的所有链接并且保存所有链接网页到本地文件夹中。

②文本清洗。由于网络上采集下来的 WEB 资源中掺杂着大量的杂质信息，如字体信息、格式信息、广告、超链接等，需要对网页内容进行数据清洗，以去除其中的杂质。

（2）语料加工功能。语料采集过后，需要再做加工，使得语料在形式上保持一致，以便于检索和统计。这项功能包括 4 项子功能：

①语料的分割与合并。用以调整语料文件的大小。

②按照断句标记对文本做断句处理。断句标记是由使用者定制的，以此来调整语料中每个片段单位的长短。

③字符编码的转换功能。可使文本文件的字符编码在 Unicode、Big5、UTF8、GBK 等编码之间实现自由转换。

④人工标注辅助系统。在语言研究中，经常要对语料进行人工分词和标注，该系统能起到辅助作用。还能根据预定规则对标注后的语料进行检验，发现违反规则的情况就会给予提示[9]。

（3）统计功能。语言研究中经常要统计语料中的字频、词频和词类频率，这项功能包括 4 项子功能：

①字频统计。统计出语料中出现的字型数、字型出现的频次和频率。能够统计单字频率、“邻近双字”的同现频率、“邻近三字”的同现频率、以及“邻近四字”的同现频率。

②词频统计。统计出语料中出现的词型数，每

个词型出现的频次和频率。能够统计单词频率、“邻近双词”的同现频率、“邻近三词”的同现频率、以及“邻近四词”的同现频率。

③词类统计。统计出语料中出现的词类数,每种词类的出现频次和频率。能够统计单个类别的频次和频率、“邻近双类”的同现频率、“邻近 3 类”的同现频率、以及“邻近四类”的同现频率。

④风格统计。统计出语料中的词型数、词例数、词型与词例之比、平均句长、句长标准差、段落数、平均段落长、以及段落长标准差。

(4) 检索功能。可分为 4 项子功能,分别提供 4 种类型的检索。分析后,可得研究概述如下。

①简单检索。用户输入一个关键词,系统能够从语料库中查找出所有该词的用例,并用红色字体把用例中的关键词标示出来。同时,还能把包含这个关键词的文本片断全部抽取出来,存在一个新的文件中。文本片断可以是以小句为单位,也可以是以整句或段落为单位,用户能够自己定义。

②复杂检索。用户可以输入多个关键词,系统能够查找出语料库中所有这些词的用例,并用红色字体把用例中的关键词标示出来。同时,能把包含这些关键词的文本片断全部抽取出来,存在一个新的文件中。关键词之间的出现关系是“并且”还是“或者”,能够由用户来设定。

③高级检索。系统能够按照用户输入的正则表达式检索语料,并且用户可以自主设定所抽取的语料片段的形式,编辑断句标记。

④特殊检索。用于处理分词之后的文本,用户输入一个关键词,并且指定在关键词之前的词语个数、以及在关键词之后的词语个数,系统能够查找出“前词+关键词+后词”这种形式词串的所有用例,并用鲜红和深红 2 种颜色字体分别把前后词和关键词标示出来。系统还能够统计出这种形式词串的出现频率[10]。

(5)分析功能。可分为 3 项子功能,分别提供 3 种类型的分析模式。这里,可给出内容表述如下。

①特征词分析。系统能够按照预设算法提取各个语料文本的特征词,进而为文本分类,信息抽取提供技术支持。

②共词分析。系统能够对一组词两两统计其在同一篇文献中出现的频次,以此为基础对这些词进行聚类分析,分析结果能够反映出这些词之间的亲疏关系,有效地展示这些词之间的关联,进而可以分析这些词所代表主题的结构变化。

③相似度分析。系统能够通过计算文献之间在词型等级方面的相关系数,来获取量化的语言风格相似度。

3 软件系统应用流程

(1)首先利用“采集模块”的子模块“WEB 爬虫”从互联网上抓取含有语料文件的网页集合,再利用“文本清洗”模块对含有 HTML 标签和广告等杂质的网页集合进行数据清洗,得到“原始语料”。

(2)利用“加工模块”的子模块“合并分割”对文献资料作合并或分割处理;“文本断句”模块作断句处理;“字符编码”模块转换语料文件的字符编码;“人工标注”模块对语料进行分词、标注词性、标注语义角色等处理;经过此阶段的处理得到“精加工语料”。

(3)利用“统计模块”中的“字频统计”子模块统计出语料文件的字频信息;“词频统计”模块统计出语料文件的词频信息;“词类统计”模块统计出语料文件的词类信息;经过此阶段的处理得到“统计报告”。

(4)利用“检索模块”的各项检索功能,根据研究的需要,对语料文件进行检索和信息抽取,得到“检索报告”。

(5)利用“分析模块”的各项分析功能,分析特征词、共词和文本的相似度,得到“分析报告”。

本系统的应用流程如图 2 所示。

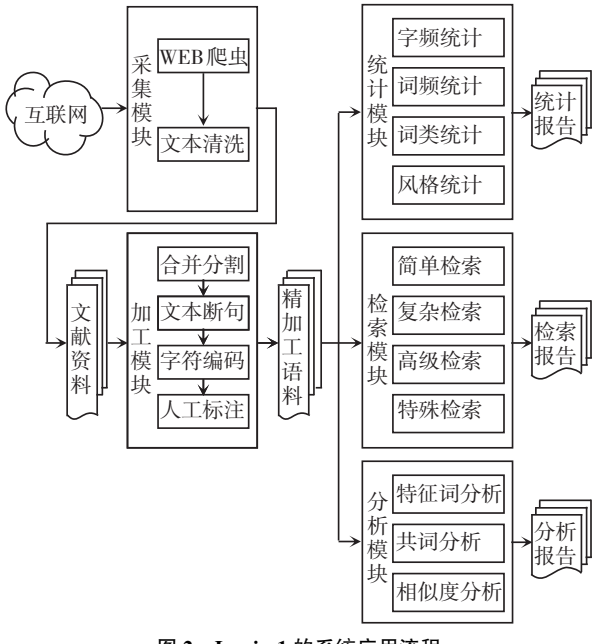

图 2 Inspire1 的系统应用流程

Fig. 2 The system application process of Inspire1

本系统初始界面的上方是一个标签视图控件,

该控件中还包含多个标签视图控件和表单视图控件，下方并排安置 2 个浏览器视图控件，其中左边控件主要用于显示输入文件的内容，右边控件主要用于显示处理结果[11]。以“简单检索”界面为例，如图 3 所示，界面的上方是提供给用户交互的界面，左下方控件中显示的是待处理的文件内容，右下方控件中显示的是以“曰/v”作为关键词的查找结果，所有符合查找条件的语句片段都显示这里。

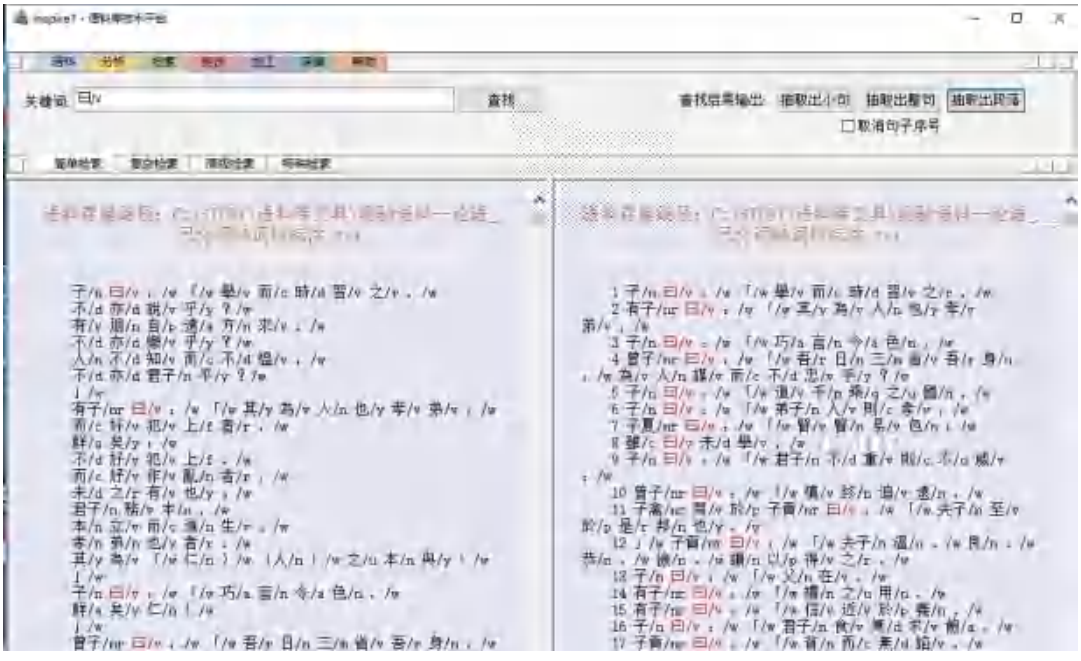

图 3　Inspire1 中简单检索的使用界面

Fig. 3　The use of simple retrieval in Inspire1

4　结束语

为了提升语言分析的效果，使研究者直观地观察到语料库中蕴含的语言规律，从语料库中发现新的知识，设计并实现了语料库技术平台 Inspire1。本软件采用面向对象的思想编程，各部分功能相互独立，具有较强的可扩展性，并且是无需安装的绿色软件，占用很少的存储空间，能够满足语料库建设和应用中的大部分技术需求。

参考文献

[1] 马创新. 语料库技术平台使用指南(语料处理软件)[2019-04-09]. http://blog.sina.com.cn/s/blog_740006d40102x448.html.

[2] 许家金，贾云龙. 基于 R-gram 的语料库分析软件 PowerConc 的设计与开发[J]. 外语电化教学，2013(1)：57-62.

[3] MCENERY T, HARDIE A. Corpus linguistics: Method, theory and practice[M]. Cambridge: Cambridge University Press, 2012.

[4] 李亮. 英语语料库检索工具的设计理念及其深层化[J]. 外语电化教学，2007(6)：16-20.

[5] 王立非，梁茂成. WordSmith 方法在外语教学研究中的应用[J]. 外语电化教学，2007(3)：3-7，12.

[6] 周晓云. 手段与效果的正比论—语言教学的现代化手段[J]. 电化教育研究，2001(12)：34-35.

[7] 马创新，陈小荷. 文献中的词型分区规律与高频特征词的发现[J]. 语言文字应用，2018(3)：124-133.

[8] MALIK D S. C++编程—数据结构与程序设计方法[M]. 晏海华，等译. 北京：电子工业出版社，2003.

[9] 马创新，陈小荷，曲维光，等.《论语》与其注疏文献对齐语料库的构建[J]. 现代教育技术，2012，22(7)：109-113.

[10] 马创新，陈小荷. 文献中的词语分布、词型等级和风格计算[J]. 中文信息学报. 2017，31 (4)：20-27.

[11] 姜秋霞. 信息技术辅助语言教育的研究范式[J]. 电化教育研究，2010(6)：107-108.