

英语教材编写过程特定内容的高亮显示

熊文新 陈国华 许家金

(北京外国语大学 中国外语教育研究中心, 北京 100089)

【摘要】针对大学英语教材编写中的词语/词组控制问题, 采用Perl和VB/VBA在Word环境下设计了一个特定词语、词组高亮凸显的计算机辅助标注系统。系统分为离线生词抽取和在线实际标注两个步骤, 界面友好, 稍加扩展适合各类语言类教材的编写。

【关键词】凸显; 教材编写; 词语控制; Perl; Visual Basic/VBA

【中图分类号】G623.31

【文献标识码】B

【论文编号】1009—8097(2008)02—0097—04

一 引言*

英语教材编写突出的问题就是对教材词汇的控制。词语复现和词语难度控制是衡量语言教材质量的一个重要标准。然而对教材的评估在实践中由于缺乏有效工具和方法, 很难得到准确的数据和量化分析^[1]。有学者专门研究过教材词汇的选择标准, 如可教性、类似性、实用性、覆盖面及定义能力等^[2]。在教材编写过程中, 哪些词语在使用该教材前就该掌握, 补充新词语应该安排在第几课, 重要词语如何复现等问题经常困扰教材编写者。当前词汇遴选和控制主要由教材的编者和编审人员凭个人经验手工实现。

在当前教材编写过程中, 虽然计算机操作, 特别是字处理软件的应用已经取代手工劳动, 但这仅停留在工作方式的改变。一些方便快捷的机器辅助处理仍比较缺乏。譬如, 对教材词汇量的控制软件就没有。

本文用Perl^[3]和VB/VBA^[4]作为开发语言, 在Word环境开发了一个特定词语/词组的凸显工具, 方便英语教材编者进行教材编写质量的控制。

二 系统设计

1 用户分析

目标用户是在Windows平台利用Word进行英语教材编写的编辑人员。他们具备基本计算机操作技能, 但并非计算机专家。他们希望在其熟悉界面完成工作, 因此系统应满足方便易用原则。

2 设计要求

教材编写和初步编校在Word下完成, 系统应能在编辑具体某篇课文时, 在屏幕上以不同颜色高亮凸显特定词语/词组

内容, 即(1)凸显该套教材中在该课文首次出现的单词(即生词); (2)以不同颜色凸显大学英语大纲中不同等级的词语(如初、中、高级); (3)凸显特定专名信息; (4)凸显某一英语学习阶段必须掌握的词组。

假定当前编辑第4课, 有静态词表中不存在的词语unlock, 并且在前三课没有出现该词, 那么就应凸显该词语。

本系统通过屏幕直观显示, 方便编辑和审校掌握本套教材中的词语使用情况, 有针对性地调整教材文本的选篇和文章排序。

3 性能描述

系统设计尽可能全地标识出需要特别显示的词语/词组。对召回率要求较高, 而准确率相对没有太高要求。因为教材编写是一个脑力劳动过程, 校对、审核都需要编者的交互。凸显部分能够缩小用户检查范围。

三 数据库资源准备

系统数据库有中学英语课程标准词汇表、大学英语教学大纲分级词汇手册及词组表、专名词表和用于解决实际文本词语和词典中记录的词目词不一致的词形还原表。

根据目标需求, 数据库设置6张表。表basicLex是中学大纲词汇, 只有词形字段。highLex和phrase分别是大学英语教学大纲的参考词表和词组表, 前者有词形及不同等级字段, 等级采用初级(B)、中级(I)和高级(H)三个级别; 后者只有词形式字段。lemma表有两个字段, 其一为文中实际可能出现的词形, 其二为相应的词语原形。专名表Proper只有词形。上述各表是各类大学英语教材编写普遍适用的资源。动态词表tempWord则针对特定教材编纂, 根据不同课文中的实际词语动态生成, 字段设计为词形和该词在实际编写教材出现的

*基金项目: 中国博士后科学基金资助项目(20070410044), 教育部人文社会科学重点基地重大项目(02JAZJD74005)、2006-2007年度全国基础教育外语教学研究资助金项目(JJWYYB2006018)

收稿日期: 2007年5月13日

先后次序。

表 1 示例表明数据库实际存储的实例。带有括号者表示词形和括号内字段形式共同构成某一记录的实际表示，如在 HighLex 中 abandon 属于 B 类；lemma 中 children 一词与 child 具有同一性；tempWord 中 unlock 在当前教材编写中首次出现在第 4 课。

大学英语教学参考大纲词表与中学英语课程标准词汇表有不少重复的地方，尤其是大学初级词汇（B 级）。设计数据库时，应该剔除中学词汇，我们设计了一个小工具，进行词表整理，防止在词语凸显时把应该掌握的中学词语当作作为大学初级词汇高亮显示。

表 1 数据库结构设计 with 实例

数据库名	表名	字段名 (属性)	示例
Dict	basicLex	word (char)	ability, join, grasp
	highLex	rank (char)	abandon(B), notorious(I), savor(H)
		word (char)	
	lemma	org (char)	children(child),was(be)
		word (char)	
	phrase	phrase (char)	a good deal, take up
	proper	word(char)	Elizabeth, John
	tempWord	word (char)	unlock(4), bedroom(5)
rank (int)			

四 程序实现方法

整个标注系统的实现分为两个阶段：脱线的数据准备和在线实际文本标注。标注时有些资源具有普适性，可以在系统运行前准备好，如各类大纲词表和 lemma 转换底表等。围绕特定教材编写的生词表创建则需要根据实际文本动态生成，需要特别处理。

1 构建服务于特定教材的生词库

(1) 采用 Perl 实现抽词模块

考虑到系统实现过程的快捷开发，我们采用 Perl 作为生词抽取的主要工具。因为 Perl 语言不仅具有强大的文本字符串处理能力，比如正则表达式和内置的哈希变量等；又由于解释执行，方便调试；同时 Windows 平台下的 ActiveState Perl 还提供 PDK 开发包，可将脚本转化为动态链接库（DLL）和可执行程序（EXE）形式，并能脱离 Perl 解释器，这就使用户无需安装 Perl 环境，也能享受 Perl 开发带来的便利。

生词抽取程序假定教材每课课文自成一个单独文本文件。本步骤接受的输入是一个记录课文文件名列表的文件；输出是一个记录生词及首次出现位置的文件。

输入文件格式为每行一个课文文件名，排列顺序按其课文目录的先后，譬如第 1 册有 10 篇课文，按照教材目录顺序排列；若是多册课文，则每下册课文紧随上册课文后排列，如第 2 册第 1 课紧随第 1 册第 10 课文件名之后。

由于生词抽取程序抽取的是超出各类大纲之外的新词语，因此，程序所需资源是前文所指的各类大纲词表。在系

统初始化阶段，首先读入这些词表，通过 Perl 内置的哈希变量记录各词语信息，作为下一步分析课文中出现的词语是否为生词的依据。

系统处理从读取用户输入的课文文件名列表入手，置计数器为 0，按行读入每一个课文文件名，并累计计数，以该实际文件名和计数值作为参数，调用具体处理模块。

在被调用程序中，系统首先读入调用程序传递过来的文件名所指向的文件内容，通过正则表达式过滤标点等无关信息内容；采用 Perl 系统函数 Split 以空格为分隔标识，切分课文为单词串；再根据系统初始化阶段构造的词语哈希表信息，清除必须掌握的各类大纲词语；针对上述处理剩余下来的词语，构造一个新的用来存储生词及其首次出现位置的哈希表。如果是首次出现，则将该单词及其传递过来的计数值输出到一个特定文件中。这样便能保存生词在教材中首次出现的课文位置。

Perl 脚本通常在命令行下由 Perl 解释器逐步执行。我们利用 PDK 工具编译生成一个能在 Windows 下独立运行的可执行程序。

(2) 采用 VB 实现图形界面

由于教材编校人员不是计算机专业人员，命令行输入对其计算机水平要求太高。因此一个方便易用的界面至关重要。

我们采用 VB 作为前端，设计了一个图形化的交互环境，将后台实际执行处理运算的 Perl 程序隐藏起来。用户只需通过鼠标操作就能完成从一套教材所有课文抽取生词并导入数据库的工作。设计界面如图 1 所示。



图1 生词及其首次出现的抽取界面

工作步骤处理流程如下：第一步通过对话框控件得到用户指定的文件名列表，即获取要处理的整套教材中所有课文；第二步以该文件名作为输入参数，采用Shell方式调用Perl程序执行生词及其首次出现位置的抽取；第三步利用DAO对象库将第二步Perl处理结果数据导入今后实际标注阶段要使用的Access数据库中；第四步结束整个处理过程。

VB和Perl程序之间的Shell调用采用异步方式实现，这使得在系统执行过程中，有可能存在Perl负责的文本处理过程尚未结束，而VB主控程序就已经跳入数据库导入阶段，造成系统出错的状态。因此我们引入IsRunning函数判断Perl调用处理是否完成，只有当Perl进程结束之后，才能进入下一流程的处理。

生词抽取处理需要一定时间，为防止用户面对系统不知所措，我们采用分步骤处理的办法，利用颜色改变作为上一阶段处理是否结束的标志。譬如，系统在执行第一步时，其他步骤的信息和按钮都是灰色的，处于非激活状态，这样可规范并控制用户的执行顺序。只有当前步骤运行完成后，紧随其后某一步骤的颜色变成蓝色，用户才可以点击对应的本步骤按钮，执行相应操作。

2 教材特定词语/词组的标注

上一小节采用VB+Perl实现了围绕特定一套教材的生词的动态生成。结合原有词表等数据信息，在本节中我们将完成在Word环境下对课文特定词语的突出显示。一个工作样例见图2所示。

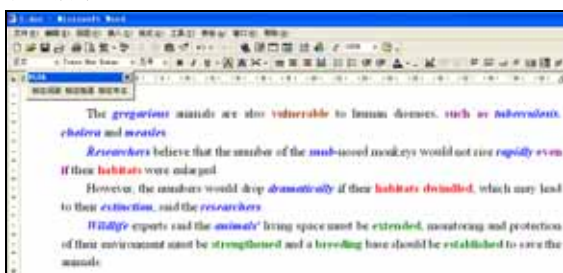


图2 Word环境下特定词语高亮显示效果

由于Word是教材编写用户的主要工作环境，又由于Office软件强大的二次开发功能，我们采用Word VBA开发一个特定词语/词组凸显模板，可无缝加载到Word的菜单栏和工具栏

中。

(1) 词语的凸显

词语凸显包括大学英语大纲不同等级词语、当前课文新出现的生词及专名等词语，以不同颜色区分。这些词语基本都是单个词语，在文本中是以空格分开的字符串。

目前设计是大学词语B类为绿色、I类为黄色、H类为红色；新词语用蓝色特别标出；专名采用暗黄色显示。

Word VBA提供一个Words集合对象，能够直接获得当前编辑文档中所有词语构成的词集。Words集合的每一项都代表一个单词的Range对象，可以修改其格式属性，无需计算单词颜色改变的起始和终止位置，而直接对单词进行凸显操作。故此，我们对词语的高亮标注采用Words对象。

并非Words集合的每个元素都是一个合适的词语，如纯数字串、标点、控制字符等都可能囊括其中。在实际处理过程中，对一些不合法的误输入也应有识别过滤机制。为此，我们编写专门的checkString函数，检查Words集合中每个元素aWord的合法性。正常情况下，函数返回值是可用来进行数据库查询的词语candidate。这样可防止系统误将一个不合法词当作新出现的生词。

课文词语在上下文语境中可能有不同的曲折形式，如动词的人称及时态变化、名词的单复数变化等，所有这些造成文中词语与词典记载的词目形式不尽相同，为此需要引入词形还原。实际处理时，我们根据一个大规模lemma底表，引入VBScript Regular Expression组件调用，创建Dictionary对象，构造文本词语与词典词形的对应关系。

标注过程是遍历当前编辑文档中Words集合的每个合法词语完成的。遍历过程中，分别用当前词语实际词形和词典还原形式，到各类词表数据库中进行匹配检索，如果在某个特定数据表的词形字段能够检出相应的词语，即Found函数的返回值为True，则根据预先定义的显示格式规范，对该词语改变其字体颜色属性，实现计算机屏幕的词语凸显。该部分伪代码形式如下：

```

For Each aWord In ActiveDocument.words
  candidate = CheckString(LTrim(Trim(aWord)))
  IF (Found(candidate)) Then
    Call Highlight(aWord)
  End IF
  IF (Found(lemmatize(candidate))) Then
    Call Highlight(aWord)
  End IF
Next

```

(2) 词组的凸显

词组凸显可分为两类：一类是由连续词串构成的词组，一类是由隔离词串构成的词组。前者如go through，可当作整体来看；后者如neither...nor...，中间可插入数量不等的其他词语。两者识别策略不尽相同。

词组一般都在某一句子内，基本不跨句出现。因此对词组的查找限定在句子范围，具体到编程，采用Word VBA的

Sentences集合对象，它能提取出选定部分、区域或文档中的所有句子。

① 连续词串的词组凸显

通过读入词组表，构造词组的dictionary对象，再对文档Sentences集合对象的每一项，逐次检查是否包含有某一特定词组，并根据该词组在文本的起止范围改变特定的字体颜色。该部分伪代码如下：

```
For sentenceNum = 1 To ActiveDocument.Sentences.Count
  For phraseNum = 0 To isPhrase.Count - 1
    If (Instr(neuSent, isPhrase(phraseNum)) > 0) Then
      tagRange = checkPos(neuSent, isPhrase(phraseNum))
      Call Highlight(tagRange)
    End If
  Next phraseNum
Next sentenceNum
```

处理对象是每一个句子，需要改变颜色的只是句中构成词组的某些词语。因此应记录变色词语在句中的偏移位置，同时还应记录该句在整篇文档的偏移位置，这样才能准确获得构成词组的词语变色范围在全文的绝对位置。

实际处理时，引入Lemma操作，匹配时注重的是词语原形的比较，但凸显时应还原为文本实际形式。这种变换往往改变词语长度，引起颜色标注错位。我们发现根据词语在句中的相对位置，比绝对位置要相对简单并且准确。处理方法如下：

将原始句切分为词向量，以数组体现；对lemma后的变换句同样操作。采用词组底表中的词组在变换句中匹配，记录变色词语在变换句中相对首词出现的偏移位置，再到相应的原始句中寻找对应词语所处的绝对位置，可以准确定位颜色变更显示的区域。

② 非连续词串的词组凸显

非连续词串的词组可以分成若干部分。单独每个部分是连续词串。如词组as far as...be concerned, 由...分隔为两部分，需要分别检索。

标注时，按句读入文本内容后，对句中可能要标注的词组每个部分，直接采用连续词串的词组检索方法。在确认某一词组的全部构件都已匹配后，才能确定对其进行高亮显示。有关该部分的词语凸显伪代码如下：

```
For phraseNum = 0 To phraseSplited.Count - 1
  offsetDoc = 0
  For sentenceNum = 1 To ActiveDocument.Sentences.Count
    If (meetCondition(phraseSplited1, phraseSplitedN)) Then
      tagRange = Range(phraseSplited1, phraseSplitedN)
      With ActiveDocument.Sentences(sentenceNum)
        Call Highlight(tagRange)
      End With
    End If
  Next sentenceNum
  offsetDoc = offsetDoc + Len(matchedString)
Next phraseNum
```

五 结语

本文就大学英语教材编写过程中的词语控制问题，采用Perl和VB/VBA设计了一个计算机辅助特定内容高亮显示的工具，既具有良好用户界面，又使开发人员利用便捷开发工具，缩短工程周期。经过初步试用，系统较好实现了原有设计思想。

本文所述高亮显示方法具有通用性。静态词表可以方便扩充完善，而生词抽取程序也适用不同教材编写工作，譬如引入各类考试词表等，可以完成考试辅导类教程的编写。由于本系统的宿主对象Word的Words对象内部具备汉语分词功能，稍加改变，也可以用于汉语教材的编写。

应该指出的是，本系统只是一个工程实现。要达到更好处理效果，还需要引入更高层的句法处理。当前还只是直接采用Lemma底表对照的词语还原技术，某些以-ed或-ing等形式结尾的词语，有可能因为被Lemma底表收录而造成错误还原，造成误标。因此一个好的Lemma底表或更好的句法分析技术将有助于此类问题的解决。

参考文献

- [1]赵勇,郑树棠. 几个国外英语教材评估体系的理论分析[J]. 外语教学, 2006, 27(3).
- [2]Richards,J.C. Curriculum Development in Language Teaching [M]. Cambridge: Cambridge University Press, 2001.
- [3]Brown, M著,顾凯等译.Pearl参考大全[M].北京:人民邮电出版社,2002.
- [4]Sanna, P.等著, 沈刚,刘景华等译. Visual Basic for Applications 5 开发使用手册[M].北京: 机械工业出版社,1997.

Highlighting Specific Contents in the Compilation of English Textbooks

XIONG Wen-xin CHEN Guo-hua XU Jia-jin

(National Research Center for Foreign Language Education, Beijing Foreign Studies University, Beijing ,100089,China)

Abstract: To help control vocabulary use in English textbooks, a computer-aided highlighting system is authored for marking up specific words and/or phrases. Perl and VB/VBA are combined to realize the highlighting function to be implemented in Microsoft Word. In this paper, the system architecture and technical implementation are detailed in terms of off-line new word extraction and online highlighting.

Keywords:Text highlighting; Textbook compilation; Vocabulary control; Perl; VB/VBA