

文章编号: 1003 - 0077 (2006) 03 - 0036 - 07

《元朝秘史》电子文本检索系统的研制*

江 荻¹, 严海林², 孙伯君¹, 斯钦朝克图¹, 孟达来^{1,3}

(1. 中国社会科学院 民族学与人类学研究所 语音学与计算语言学重点实验室, 北京 100081;

2. 北京理工大学 自动控制系, 北京 100081; 3. 早稻田大学 人文学部, 日本 东京)

摘要: 本文概要地介绍了 13 世纪《元朝秘史》的文献背景及原文所独有的复杂文本形式, 通过对文本的内涵分析和版面分析, 设计了关于《元朝秘史》电子检索系统的研制方案。其中主要解决了原文三行一体显示格式的还原问题, 而且系统可以分别对原文汉字音写、汉语译文、汉字旁译、语音语法标注等不同部分进行检索和统计。检索输出结果包括研究者最重视的传统学术章节号、卷页码、在电子文本出现的具体位置。另外, 系统对检索词采用了上下文检索技术, 输出文本包括检索词的部分上下文内容。本系统基本满足历史、文学和语言研究的应用需求。

关键词: 计算机应用; 中文信息处理; 元朝秘史; 复杂文本; 电子检索系统

中图分类号: TP391

文献标识码: A

An Introduction to the Retrieval System for "the Secret History of the Mongols"

JANG Di¹, YAN Hai-lin², SUN Bo-jun¹, Siqin Chaoketu¹, MENG Da-lai^{1,3}

(1. Lab. of Phonetics and Computational Linguistics, Institute of Ethnology & Anthropology, CASS, Beijing 100081, China;

2. Automation Department of Beijing Institute of Technology, Beijing 100081, China;

3. DEPT. of Letters, Arts and Sciences, Waseda University, Tokyo, Japan)

Abstract: This paper firstly gives a brief introduction to the background of the Secret History of the Mongols, the great book published in 13 century in Yuan Dynasty, and its special complicated original typeface in form. After an analysis to its content and page form, a scheme of electronic retrieval system has been then designed for it, which resolves the problem of returning to the original shape of the archaic writing form with three lines representing one content. Furthermore, the retrieval system also provide the functions of retrieving and counting each contents of the origins, including Chinese transliterate, Chinese translation, and phonological and grammatical markers. The retrieval result includes numbers of traditional academic chapters and sections which are very important for the users, numbers of original volumes and pages, and retrieval objects' positions in the electronic text. In addition, the system makes full use of a concordance technology, which can present retrieval units with their contexts. Generally speaking, this retrieval system can basically satisfy the needs of studying history, literature, and language from the important historical document.

Key words: computer application; Chinese information processing; the secret history of the Mongols; complicated text; electronic retrieval system

* 收稿日期: 2005 - 01 - 28 定稿日期: 2005 - 05 - 08

基金项目: 中国社会科学院重点实验室资助项目 (MZ: 101)。

作者简介: 江荻 (1954—), 男, 博士, 研究员, 主要研究方向为中国民族语言自然语言处理。

1 电子文本制作的背景及缘起

《元朝秘史》又称《蒙古秘史》。据考证,原文为畏兀尔体蒙古文,成书年代大约在13世纪中叶,作者佚名。^{*}该书原本早已散失,我们现在看到的《元朝秘史》是明代四夷馆纂辑的汉字音写本^[1]。明朝翰林译员为了教学蒙古语,训练通事和译字生,用汉字音写原《蒙古秘史》中的蒙古语,并在音写汉字旁边逐词加注汉语译文,包括标注人称变位、时态、数、格等语法成分。全书分成282节,每节之后有汉文总译,汉文题名为《元朝秘史》。

《元朝秘史》记载了蒙古氏族和部落的起源,描述了成吉思汗建立蒙古国的过程,包含了12、13世纪蒙古社会丰富的历史资料,是一部重要的蒙古史典籍。另一方面,该书特定的汉字音写注译本真实地记录了13世纪的古蒙古语,保存了大量的古蒙古语的语音和语法现象,其音写规则极其规范和严密,具有很高的历史语言学价值;《元朝秘史》同时还是一部经典的蒙古文学作品,书中用诗化的语言,塑造了鲜明的具有草原民族特色的人物形象,被誉为蒙古民族的史诗。

清代以来,《元朝秘史》成为学界关注和研究的对象,研究涉及版本源流、史地探求、语法词汇考证、蒙古文复原、汉字的拉丁转写等^[2]。据统计,有关《元朝秘史》的研究成果在世界各国数百种之多^[3~6],《元朝秘史》还有英、俄、德、法、匈、日、汉文等多个译本^[7]。联合国教科文组织在巴黎召开的纪念《元朝秘史》成书750周年会议决议认为,《元朝秘史》在人类文化发展史中留下印迹,并在世界文化史中享有崇高的地位,它的“独特的艺术、美学和文学传统及天才的语言,使它不仅成为蒙古文学中独一无二的著作,而且也使它理所当然地进入世界经典文学的宝库”(1989年6月)。中国社会科学院文学所杨义教授在“《蒙古秘史》七百六十年祭”一文中称赞说:“《蒙古秘史》是这个民族精力最旺盛、元气最充沛的时代,把内蕴的精力和元气转化为文字的伟大的产物,是蒙古族创世纪式的回忆、想象和纪录”;该书“因记载蒙古族勃兴初期史料和洋溢着浩瀚博大的狩猎游牧文化精神而驰名。它吸收远古以降蒙古民间文化精粹,开蒙古书面文化先河,乃是研究蒙古史、元史、世界中世纪史的经典文献,充满大气磅礴的史诗气息”^[8]。

为了进一步推动《元朝秘史》的深入研究,我们设计了《元朝秘史》的电子文本检索系统,为全世界热心研究《元朝秘史》的学者提供强有力的研究工具。本文的目的是针对《元朝秘史》独特的复杂文本形式,从计算机处理技术角度讨论电子文本系统的构造和检索技术以及该系统所具有的研究功能。

2 复杂文本格式的挑战

《元朝秘史》有多个版本,明初刻本分正集十卷和续集二卷,即现行十二卷本,后《永乐大典》收录本则分十五卷。本项目选用十二卷本作为电子版本研制对象,全书近30万汉字。

正如上文所说,由于《元朝秘史》是成书于13世纪的汉字音写注译本文献,因此原文面貌非常独特。在书写格式上,文献采用竖排的方式,正文是汉字音写的蒙古语词,音写汉字采用特别的方法仔细区别与汉语不同的发音,如:在汉语来母字左上角加个小“舌”字来表示汉语没有而蒙古语所具有的舌尖颤音,在汉语晓母字左上角加个小“中”字表示汉语没有而蒙古语所具有的小舌塞音,在正文同行汉字右下方加小字“惕、勒、木、黑、克、卜”等分别表示蒙古语

* 原书蒙文名《忙豁论·纽察·脱卜察安》(Monqol-un Nihuca Tobciyan,即《蒙古秘史》),蒙古文史书《黄金史》(A Itan Tobci)中还保存了《蒙古秘史》三分之二左右的佚文,但已改写,不是原文原貌。关于《蒙古秘史》成书年代学界尚有争议,根据书后“鼠儿年七月”、“写毕”等字样,学界分别推测其成书于1228年戊子、1240年庚子、1252年壬子、1264年甲子和1276年丙子等年份。概言之,《蒙古秘史》的成书时间在13世纪中叶。

闭音节尾音 * t(d)、l m、q、k、b等;正文右边以词为单位旁注汉语意译及语法形式。

除了以上复杂的书写格式外,全书章、节、段及译文编排也具有一定的复杂性。原书按照明四夷馆的分段节译编排方法,共分二百八十二节,每节正文之后有一段汉语总译,全书贯彻到底。而十二分卷中,每卷又有自己的卷标和页码,如“元秘史一·二”表示第一卷第二页,“元秘史八·十二”表示第八卷第十二页。下面影印第一卷第一页以窥全貌。

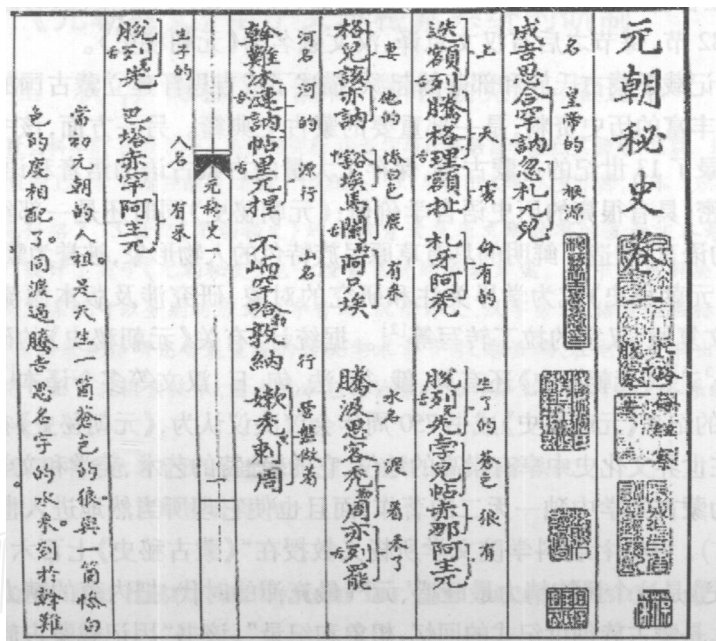


图 1 《元朝秘史》第一卷第一页影印件

《元朝秘史》文本是较典型的复杂文本格式,其特征包括排列的非线性关系以及非同列词语的对齐关系。就软件设计而言,一方面要满足原文本格式排列的显现要求,另一方面也要符合电子格式自身的存储、检索和输出制约。解决这两方面问题即本项目的主要任务,只有这样才能顺利开展对该文献所需的检索、统计、输出和编辑功能的开发。可以说,《元朝秘史》独特的复杂文本格式对该文献的电子化处理提出了设计上的挑战。

3 文本分析及文本块定义

诚如上文所述,《元朝秘史》是一种复杂文本格式的历史文献。因此,在采用现代技术进行处理之前,需要更具体剖析构成这种复杂文本的构件和细节,并且,只有通过具体的分析才能建立起计算机处理所需的模块与流程。

根据文本的构成,我们知道,该文献原文是一种三行一体格局的文本,每行的字串虽分写却不分离。因此我们的设计必须具有整体性观念,为此,我们首先定义了五类文本块类型,通过建立文本块之间的关系来实现必要的技术处理。各文本块的相关定义如下:

正文块:指汉字音写字串,该文本块是本文的主体部分;

汉译块:指每节正文之后的汉语翻译部分,这部分字串与正文不同之处在于无需右边的汉译旁注;由于这部分在处理上与正文块基本一致,因此不单独处理,下文正文块处理包括了这一部分。

旁译块:指正文右边与正文词语对齐的汉语意译字串及语法标注;

声类块:指正文左边与正文汉字对齐的表蒙古语声类(辅音)的字串;

韵类块:指原文夹在正文字串中的小写汉字字符,其作用是表示蒙古语闭音节尾音 * t (d)、l m、q k、b等的读音。

另外,表示传统分节的数字可看作分节块,表示卷码和页码的数字可看成卷页块。

需要说明的是,《元朝秘史》原文献是竖列书写,列序是从右至左,每列字序从上至下。而现代汉字文本较普遍的阅读习惯是横行排列,字序则从左至右(当然不同民族文字有不同阅读传统,这里仅就当代汉字阅读而言)。加之计算机对横行文本处理较为方便,因此本项目电子文本将原文改成横行排列处理与显示。

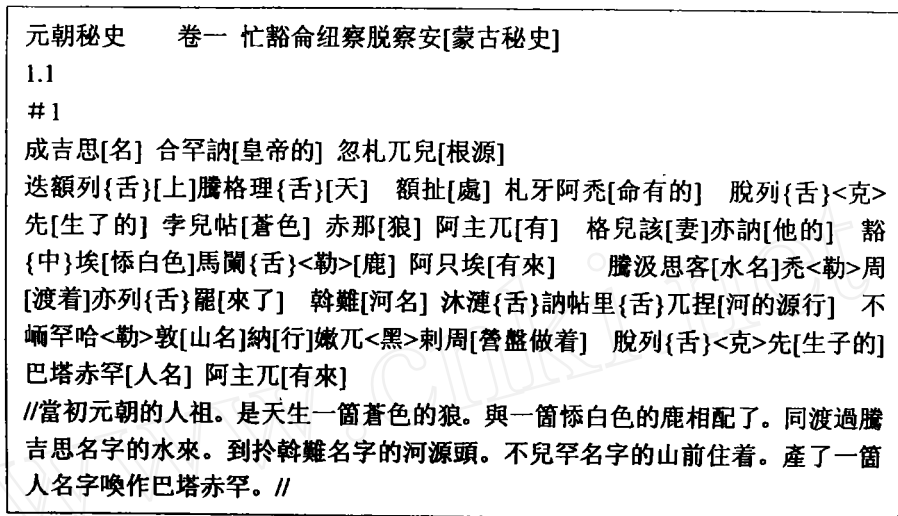


图 2 《元朝秘史》电子文本格式片段

确定了文本块,我们就可以按照线性顺序来安排文本字符串的排列,具体规则是:正文块音写字串顺序(输入)排列,包括文本中的空格、句号和换行均予以保留;汉译块连续录入,其间的空格和换行均保留,但段前与段后要添加标记“//...//”;旁译块紧接在所对齐正文字串之后,前后分别用标记“[和“]”标示;声类块紧接在所对齐正文字串后,前后分别加上标记“{和“}”;韵类块仍然保持原文本中的位置,但前后分别添加标记“和“”。

由于以上带标记字串的意译块([])、声类块({ })与韵类块()在文本中的分布经常可能连续出现,为此,可再规定三者出现顺序为先声类块({ }),次韵类块(),后旁译块([])。其中各种标记既能表明各块之间的关系,也标明了字串之间的对齐位置。图 2 是从文本中抽取的一个计算机排列片断(#1 表示传统分节,1.1 表示卷码和页码)。

在以上约定规则基础上,我们进一步讨论各模块的设计。

4 系统功能及模块设计

根据普遍的研究需要,《元朝秘史》数据系统应包括词语检索、词语统计、以及统计、检索结果输出等功能。为此,整个系统可以分为如下基本模块:显示模块、编辑模块、统计模块和检索模块,如图 3 所示。

4.1 显示模块

在输出模块方面,本文以显示输出加以论述。由于用户设备的差异,文本的行输出长度不尽一致,因此,显示模块分为不换行显示和换行显示两类。不换行显示指正文或汉译字符串宽未超出视窗边界的显示;换行显示指正文或汉译字符串宽超出视窗边界,则需要设计自动判断

超宽情况并自动换行。

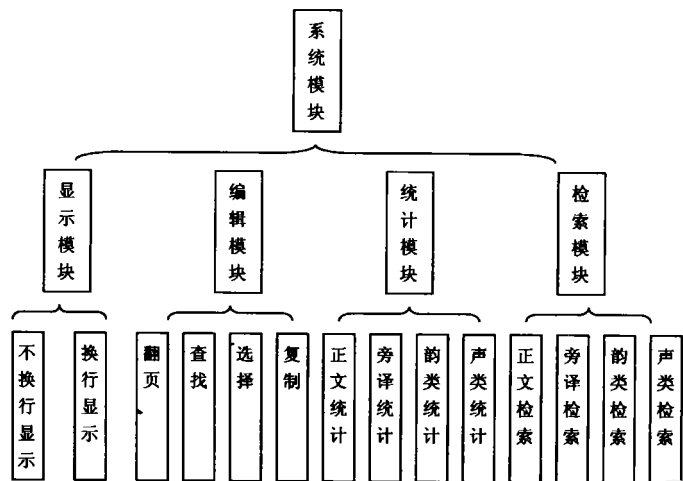


图 3 《元朝秘史》数据系统的功能模块

显示输出最重要的步骤是将线性排列的字符串还原为原文本的三行一体显示形式,这个过程主要通过对标注文本块进行编码和解码完成。编码的过程是顺序按整段读取文本,然后按照文本块进行结构类型定义:

连续的正文块(包含汉译块,以及传统章节号、卷码和页码)

```
public struct ZwStruct
{
    private int z; // 记录当前文本的传统节号
    private int y; // 记录当前文本的卷码
    private int d; // 记录当前文本的段号
    private int f; // 标记,记录此段为章节号、正文块还是汉译块
    private int l; // 记录当前文本的长度
    private string s; // 记录当前文本的内容
}
```

类似地可以定义其他文本块如下:

- 正文块上(右)边的旁译块 ZwPyStruct
- 正文块下(左)边的声类块 ZwBzStruct
- 正文块中小写形式的韵类块 ZwWzStruct

定义完这 4 种结构后,需要按照复杂文本的格式来显示,因此通过解码来进行显示控制,其步骤如下:

从 ZwStruct 中读取正文块信息,并将其按照屏宽显示,如果此段属于卷页或章节号跳转到;如果此段属于连续的汉译块,跳转到;否则执行下一步。

从 ZwPyStruct 中读取旁译块(标记“[”与“]”中的内容),并根据其位置信息定位显示在所对应正文的上方。

从 ZwBzStruct 中读取声类块(标记“{”与“}”中的内容),并根据其位置信息定位显示在所对应正文的下方。

从 ZwW zStruct中读取韵类块 (标记“ ”与“ ”中的内容),并以斜体或高亮显示。
所有信息读取完毕后输出 (显示),否则跳转到 。

图 4是本项目的显示样例。

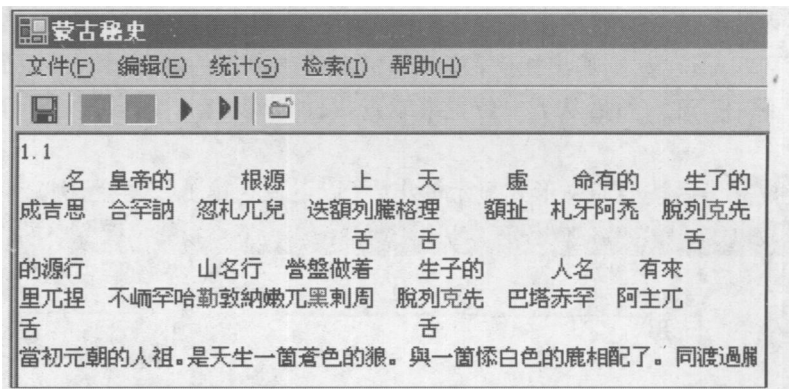


图 4 连续文本还原为三行一体显示案例

4.2 编辑模块

编辑模块包括翻页、查找、选择和复制模块。

翻页模块:原《元朝秘史》分为 12 卷,每卷连续显示,查阅时需要进行翻页。

查找模块:对连续文本的查找,查找的结果在显示页面中高亮显示。

选择和复制模块:实现一般的文本选择和复制功能。

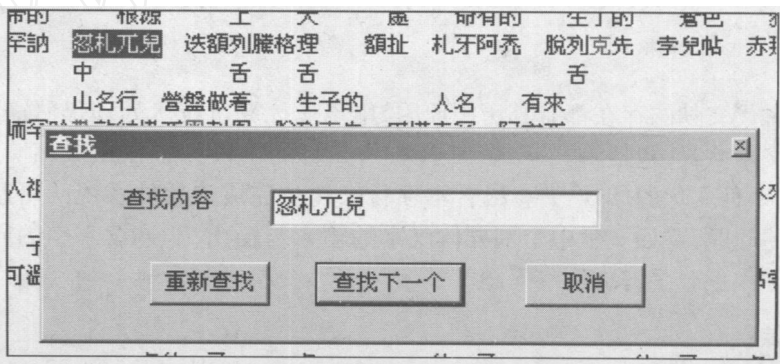


图 5 编辑模块示例

4.3 统计模块

统计模块包括分别统计正文块、汉译块、旁译块、声类块、韵类块。其基本算法如下:

从原文本中依次读取一个字符。

如果这个字符在字库中存在,则记录其个数增 1;如果这个字符在字库中不存在,则按顺序插入到字库中,并将其个数置为 1。

如果原文本读取结束,则算法结束;否则跳转到 。

统计-正文									
文件(F)									
字符	频率	字符	频率	字符	频率	字符	频率	字符	频率
成	4	吉	73	合	553	罕	120	訥	259

图 6 数据统计举例

4.4 检索模块

检索模块包括正文检索、汉译检索、旁译检索、声类检索和韵类检索。其算法如下:

将原文读入内存。

将待检索的字符或字符串从原文开始位置依次匹配。

如果找到匹配位置,则输出相应信息(包括字符所出现的卷页、章节、页码段等信息,还包括一定的上下文信息),然后从下一位置继续查找;如果没有找到匹配位置则直接跳到下一位置继续查找。

原文查找完毕后按照原文形式显示输出。(以下仅为第 1 - 2 卷示例)

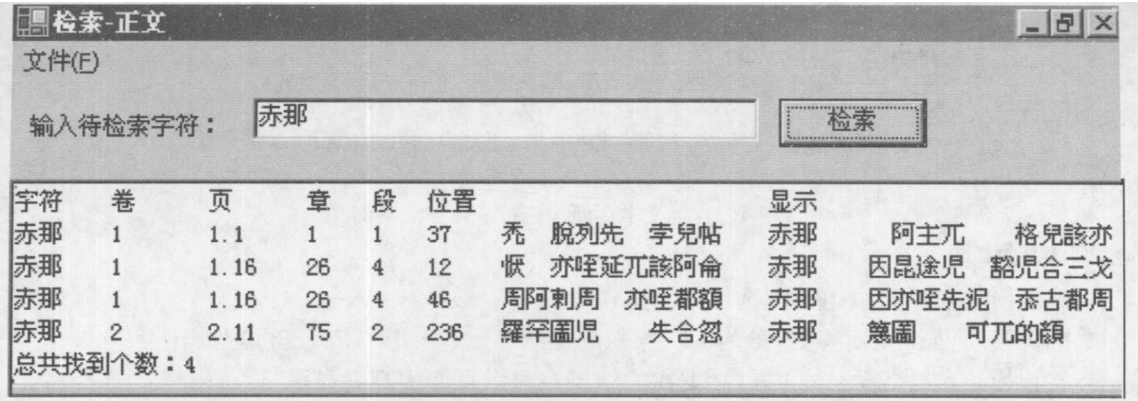


图 7 带章节、卷码和上下文信息的检索

5 结束语

复杂文本古代文献的存在都有其历史原因,用现代计算机技术处理仍然需要一定程度上保存这种复杂文本格式,包括古文字编码问题^[9],其目的不仅为了在相似环境中准确地开展文献研究,也是一种文化的体现^[10]。尽管本项目已初步完成该软件系统的研制,但仍然存在一些需要解决的问题,譬如文献中个别异体汉字的输入与输出,非连续字串的不同步检索等等,我们的下一步工作正着手解决这些问题。

参 考 文 献:

[1] 佚名. 元朝秘史 [M]. 明代四夷馆纂辑十二卷刻本影印本, 13 世纪.

[2] 栗林均, 确精扎布. 元朝秘史 蒙古语全单词、语尾索引 [M]. 仙台: 东北大学东北亚研究中心, 2001 年.

[3] 陈垣. 元朝秘史译音用字考 [A]. 陈垣学术论文集 (第二集) [C]. 北京: 中华书局, 1982 年.

[4] 小泽重男. 元朝秘史 全释 (上、中、下) [M]. 东京: 风间书房, 1984 - 1986 年.

[5] Rachewiltz, Igor de. Index To The Secret History of the Mongols [M]. B bomington, 1972

[6] 陈中永, 等. 蒙古秘史 多视角研究 [M]. 呼和浩特: 内蒙古教育出版社, 2001 年.

[7] F. W. Cleaves. The Secret History of the Mongols [M]. London: Cambridge Massachusetts, 1982

[8] 杨义. 蒙古秘史 七百六十年祭 [N]. 中华读书报, 2000 年 12 月 13 日.

[9] 张再兴. 古文字字库建设的几个问题 [J]. 中文信息学报, 2003, 17 (6): 60 - 65.

[10] 李宇明. 搭建中华字符集大平台 [J]. 中文信息学报, 2003, 17 (2): 1 - 6