

中国古代小说版本数字化和 计算机自动比对

周文业

【内容提要】 本文介绍了古代小说版本数字化目前的进展和将来的发展。古代小说版本数字化和计算机自动比对可以用来仔细比对古代小说的各种版本,对于研究版本演化有很大的帮助。小说版本数字化包括两种文字版(简体修订版、繁体原貌版)和图像版。小说版本数字化分两阶段进行,第一阶段集中于简体字版,目前已经完成《三国演义》、《水浒传》、《西游记》、《金瓶梅》、《红楼梦》等名著主要版本的简体字版。第二阶段集中于繁体字版,将完成以上名著所有影印本的繁体字版。利用计算机可完成小说各种版本的自动比对,比对结果有两种显示方式,即逐行显示和分窗口显示,两种方式各有特点。利用计算机还可实现小说的同词脱文的统计等特殊要求。小说的图像版、图文对照和图文检索清楚、明了,为小说版本研究提供了更快速、直观的工具。古代小说版本数字化工程庞大,将逐步推进。对于古代小说版本数字化应该实事求是地评价,古代小说版本数字化和计算机自动比对为古代小说研究提供了一个更加快捷、方便的工具和方法,古代小说版本数字化必将推动古代小说版本的研究,但真正的分析还必须由人工进行,计算机不可能代替人工的分析研究。

【关键词】 古代小说 数字化 版本 比对

一、古代小说版本数字化和计算机自动比对

1.1 古代小说数字化的三个发展方向

目前古代小说数字化正在向“广度”、“深度”、“真度”三个方向发展(表1):

	种类数字化	版本数字化	版式数字化
特点	广	深	真
小说数量	数量多	数量少	数量少
版本数量	一般只收入一种版本	多种版本	一般少数版本
小说种类	任何小说	主要针对多版本名著	名著
功能	一般浏览、检索	版本比对	高级浏览、检索
目的	一般研究	版本演化等深入研究	恢复原版、原貌
规模	549 种小说,八千万字	5 种小说、50 个版本、约 2500 万字	几种

1. **种类数字化**的目标是追求“广度”,收入小说的种类尽可能多,而每种小说只收入一种版本。目的主要是供学者作一般的浏览、检索、引用,不分版本,不做深入研究。如国学公司将推出的《国学宝典》系列中的《中国古代小说典》(尹小林策划、沈伯俊主编),收入古代小说八百多种,其中文言小说五百多种,白话小说二百多种,共计一万多卷,八千万字,将是目前收入小说最多的古代小说资源库。

2. **版本数字化**的目标是追求“深度”,不追求收入小说的数量,而追求一部小说版本的数量,主要针对多种版本的名著。目前已经完成《三国演义》、《水浒传》、《西游记》、《金瓶梅》、《红楼梦》五部小说 21 种版本的简体字版,正在开发繁体字版。版本数字化最终目的是为版本的深入研究服务,通过计算机版本自动比对,为版本演化等问题作深入研究提供一个快速、简捷的工具。

3. **版式数字化**的目标是追求“版式还原”,以数字化方式还原古代小说的版式。版式还原一般可以同时提供两种版式,一种是图像版,保留原版原貌;另一种是文字版,其中的文字可以进行检索、查询。两种版式完全重迭,可以自由转化。这种版式即还原了原版、原貌,又发挥了数字化的检索、查询功能,是比较理想的版式,但制作工作量很大,要对每个字做处理,成本非常高。

1.2 古代小说版本数字化和计算机自动比对

许多中国古代小说都有一个很复杂的创作、出版过程,其版本演化过程往往也非常复杂。许多中国古代小说的发展过程中都产生了多种版本,在《三国演义》、《水浒传》、《西游记》、《金瓶梅》和《红楼梦》等名著中更为突出。现在中国古代小说四大名著(《三国演义》、《水浒传》、《西游记》、《红楼梦》)都已完成数字化,出版光盘并上网,但都是为一般读者阅读,并没有把所有的版本都数字化。一般读者阅读选择最流行的版本就可以了,但对于古代小说研究来讲,只研究最流行的版本是绝对不够的。要利用计算机对古代小说版本进行真正深入、细致的学术研究工作,应该将古代小说尽可能多的版本数字化,输入计算机。

古代小说版本数字化是古代小说数字化研究中的一项基础工程,是利用计算机开展古代小说版本研究的前提和基础。只有在数字化地基础上,才能利用计算机真正开展古代小说版本的各种分析研究工作,如版本比对、分类检索,以及建立古代小说版本数据库等。古代小说版本数字化的目的,是为古代小说研究人员提供多种研究用的版本,而不是为一般大众阅读古代小说提供普及读物,这是整理校订工作所依从的原则。

古代小说版本数字化工程包括四项内容,即版本数字化、版本比对、比对结果和版本图文对照。版本数字化是基础,版本比对和图文对照是在版本数字化基础上扩充的功能,比对结果是利用版本比对软件比对的结果(图1)。



图1 版本数字化框图

1. 版本数字化:

- 版本数字化包括文字版和图像版两种形式,电子文本用于阅读、检索,而图像可以保存原版原貌。
- 文字版包括简体字版和繁体字版两种,各有不同的整理原则和用途。
- 简体字版根据排印本整理,加标点,主要用于一般阅读、检索,因此对原版文字可能作适当修订。数字化先从简化字版开始,目前已经完成《三国演义》、《水浒传》、《西游记》、《金瓶梅》和《红楼梦》等五部小说的21种简体字版。
- 繁体字版是根据原版影印本整理,主要用于版本研究,因此对文字不做任何修订,并保持原版行款。目前繁体字版刚开始整理,计划完成《三国演义》、

《水浒传》、《西游记》、《金瓶梅》和《红楼梦》等所有影印的版本全部繁体字版。

- 图像版是扫描原版而生成,目前完成了少数图像版。计划在完成《三国演义》、《水浒传》、《西游记》、《金瓶梅》和《红楼梦》等所有繁体字版的同时,完成对应的图像版。

2. 版本比对:

- 版本比对是数字化的主要功能,要对几十万字的小说进行版本比对,用人工比对是非常困难的,版本数字化后,利用计算机进行自动比对是很容易的。
- 版本比对按照输出形式,分为逐行比对和分窗口比对两种,各有优缺点。

3. 比对结果:

- 版本比对结果是利用版本比对软件,为学者直接提供若干主要版本比对后的结果。

4. 图文对照:

- 图文对照是开两个窗口,分别显示图像版和文字版,并作对比,即可看到电子文本,又可以同时看到原版原貌。

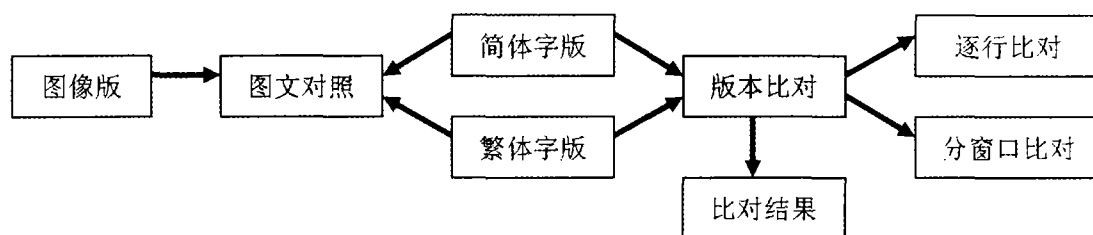


图2 版本数字化功能

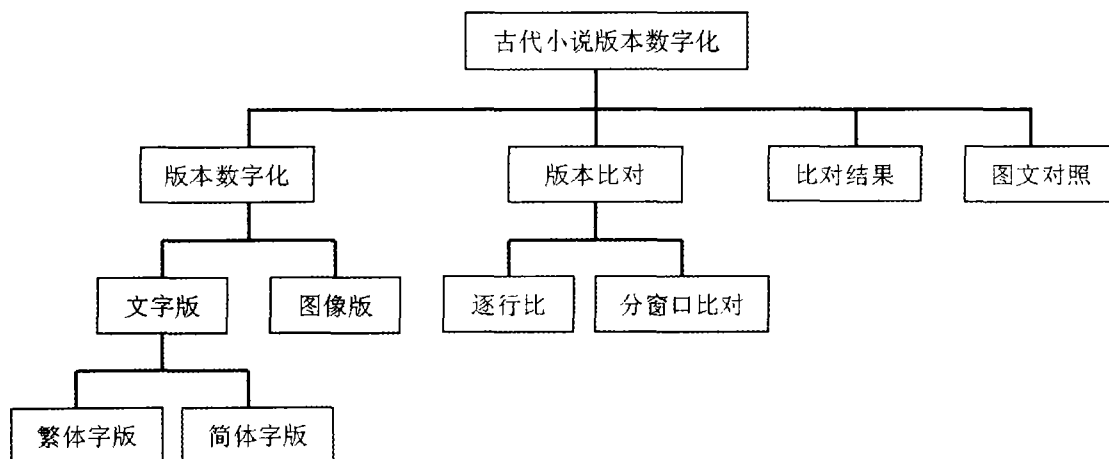


图3 版本数字化组成

二、古代小说版本

2.1《三国演义》版本

古代小说版本数字化是从《三国演义》开始的。《三国演义》版本是古代小说中版本数量最大、最复杂的,按照版本的演化发展分为“演义”和“志传”两个系列四类,版本总数有几十种。

“演义”系列又可分为无评语的早期本和有评语的后期本,早期“演义”系列版本没有评语,其中嘉靖元年本是目前已知刻印年代最早(1522年)的版本,版本价值非常高。周曰校本是“演义”系列中较早的一个版本。“演义”系列中各种评本(李卓吾本、钟伯敬本、李渔本、毛宗岗本等)是后期重要版本,有大量评语。毛宗岗本是目前最普及的版本,包括醉耕堂、芥子园、大魁堂、贯华堂、文英堂、三槐堂等几十种版本。

上图下文的《三国志传》系统,又有繁本和简本之分。叶逢春本属于“繁本”类,是除嘉靖元年本以外,刻印年代最早的版本(1548年),其版本价值非常高,可惜由于国内没有正式出版,研究还很不深入。“简本”系列版本一般认为是“繁本”的删节本,种类非常多。

《三国演义》的许多版本在国内、台湾和日本都出版了影印本,约有20种,是古代小说中影印版本比较多的小说。计划将尽可能将所有影印的版本全部制作成繁体字版和图像版。

《三国演义》在国内已有多种排印本出版,除大量的毛评本外,还有“通俗演义”系列的嘉靖本、周曰校本、李卓吾本、李渔本、钟伯敬本五种,“志传”系列“简本”的黄正甫本。日本出版了“繁本”的叶逢春本,一共有八种排印本。目前还有部分“通俗演义”系列版本没有排印出版,而各种“志传本”系列只出版了两种,绝大部分还没有排印出版。

《三国演义》版本数字化的简体字版已经完成了八个版本,其中有六种“演义”系列版本,即嘉靖元年本、周曰校本、李卓吾本、钟伯敬本、李渔本、毛宗岗本,“志传”系列版本有两种,即繁本系列的叶逢春本和简本系列的黄正甫本。

2.2《水浒传》版本

《水浒传》的版本也非常复杂,一般分为四类,即文简事繁的“简本”、文繁事简的“繁本”、繁简综合的“全传”本和腰斩删改的金圣叹本。“简本”有102回、

110回、115回、124回几种,“繁本”为100回,“全传”本为120回,金圣叹本为70(71)回。

文简事繁的“简本”文字简陋,有平田虎和平王庆故事,现存较完整的“简本”是明万历二十二年双峰堂刊刻的《水浒传传评林》,保存在日本轮王寺日光慈眼堂。

文繁事简的“繁本”相对“简本”文字描述详尽,在梁山聚义后,只有平辽和平方腊,而没有平田虎和平王庆。“繁本”中较早、较重要的是明万历三十八年容与堂刊刻的《李卓吾先生批评水浒传》,简称“容与堂本”。繁简综合的“全传”本综合了“繁本”和“简本”,在100回“繁本”基础上,增加了“简本”中的平田虎和平王庆,对文字也略作修改。现存的“全传”本最重要的是万历四十二年袁无涯刊刻、有李卓吾序、杨定见小引,也称“袁无涯本(袁刻本)”或“杨定见本(杨序本)”。

腰斩删改的金圣叹本将120回“全传”腰斩为70回,将第一回改为楔子,删去梁山聚义后的故事,以卢俊义一梦为结束,题名《第五才子书施耐庵水浒传》,有明崇祯年间的贯华堂刻本。

《水浒传》版本非常多,目前仅从上述四类版本中各选一种排印本制作了数字化的简化字版。以上四类版本都有影印本,可以由此制作繁体字版。

2.3《西游记》版本

《西游记》版本比《三国演义》和《水浒传》简单。《西游记》版本和《三国演义》、《水浒传》一样,也分为“繁本”和“简本”。目前保存的最早版本是明万历二十年世德堂本,属于“繁本”,也是最重要的版本,目前出版的各种《西游记》多数是以世德堂本为基础修订的。

除“繁本”外,《西游记》还有许多“简本”,这些“简本”多刻印在清代,与世德堂本相比,除故事叙述简略外,关于唐僧的出身也不同。世德堂本有关唐僧的出身叙述很简单,而各种“简本”中有关唐僧的出身叙述很详细。

《西游记》版本简体字版除完成世德堂本外,已经完成数字化的简本有三种,分别是名为《唐三藏西游释厄传》的朱鼎臣本,名为《西游记传》的杨致和本,和嘉庆年间刻印的《西游记原旨》。

《西游记》版本的影印本很多,主要版本都有影印本,除上述版本外,还有杨闽斋刻本、《西游记证道书》、《西游真诠》等,今后将陆续完成繁体字版。

2.4《金瓶梅》版本

《金瓶梅》版本比《三国演义》、《水浒传》和《西游记》简单得多,只有三种,即“崇祯本”、“词话本”和张竹坡评本。张竹坡评本以崇祯本为基础加评语,实际属于崇祯本系统,因此《金瓶梅》也可以只分为“词话本”和“崇祯本”两个系统。

“词话本”即《新刻金瓶梅词话》100回十卷,为明万历年间刻本,因为书名中有“词话”字样,因此一般称为“词话本”。

“崇祯本”即《新刻绣像批评金瓶梅》100回二十卷,为明崇祯年间刻本,因此一般称为“崇祯本”,有插图二百幅。张评本即《张竹坡批评第一奇书金瓶梅》,刊刻于清康熙三十四年,以崇祯本为底本,文字略有修改,主要增加了近十万言的总评、回评和读法等评论。

考虑到张竹坡本的价值在评语,在文字上与崇祯本基本相同,而目前版本数字化还未开展评语数字化,因此《金瓶梅》简体字版目前只完成“词话本”和“崇祯本”。而繁体字版将根据《金瓶梅》的各种影印本制作。

2.5《红楼梦》版本

《红楼梦》版本比较清楚,一般分为脂本和程高本两类;更为严格划分,应分为80回的“脂评本”、120回的混合本和120回程高本三个系统。

80回“脂评本”多题名“石头记”,带有“脂砚斋”等评语,现存12种。主要有甲戌本(16回,1754年)、己卯本(41回+2半回,1759年)、庚辰本(78回,1760年)、甲辰本(80回,1784年)、列藏本(78回)、戚序本(80回)、舒序本(40回)、郑藏本(2回)、卞藏本(10回,2006年新发现)。除戚序本为排印本外,其他都为手抄本。

120回的混合本前80回采用脂本,后40回属于120回的程高本,包括蒙府本(120回)、梦稿本(120回)两种。均为手抄本。

120回程高本都是排印本,主要有两种,其一是程甲本,1791年由程伟元和高鹗采用活字排印出版;其二是程乙本,是程甲本出版第二年(1792年)由程、高两人对程甲本修订后出版。

以上所有版本都出版了影印本,多数版本都有排印本。

《红楼梦》数字化目前已经完成其中主要的三种版本(甲戌本、庚辰本和程乙本)的简体字版,繁体字版将根据全部影印本制作。

三、古代小说数字化三种版本

3.1 大字库和汉字编码

古籍数字化中的一个基本问题是计算机汉字大字库和汉字编码问题。目前国内计算机的汉字字库一般采用 GBK 编码,即国家标准 GB13000.1,包含 27484 汉字,采用二字节编码。但 GBK 的 2.7 万汉字明显不适应古籍整理的需要,古代小说中有很多一般计算机字库中没有的异体字、俗体字。目前国际标准 ISO/IEC10646:2003(即 Unicode4.0)汉字编码有 70 195 个汉字,其中前面包括上述二字节的 2.7 万汉字,其余 42 711 字为四字节编码。

从字库容量看,采用 Unicode 4.0 编码明显要比 GBK 编码强。但目前一些计算机还采用 GBK 汉字,不支持 Unicode 4.0,但从长远考虑,汉字库应用 Unicode 4.0 编码是发展趋势,因此繁体字版将采用 Unicode 4.0 编码。

对于 Unicode 4.0 中没有的汉字,一般处理方法有以下几种:

- 造字:比较复杂,工作量大,并且在其他计算机上安装很麻烦;
- 偏旁组字:使用较多,但不直观,一个字要占用多个字位,对于检索等处理很困难;
- 图片:将原版中的字扫描成小图片,实现比较简单,但不协调。
- 以上处理方法都不理想,本系统解决的办法是利用其他字符(如□或■)代替。这种方法实现非常简单,虽然在文本中无法看到字的原貌,但字的原貌可利用图像版和图文对照去实现。

3.2 三种版本

古代小说版本数字化包括三种版本,两种文字版和一种图像版,两种文字版为“GBK 简体字标点修订版(简称“简体字版)”和“Unicode 繁体字原貌版(简称“繁体字版)””(图 4)。

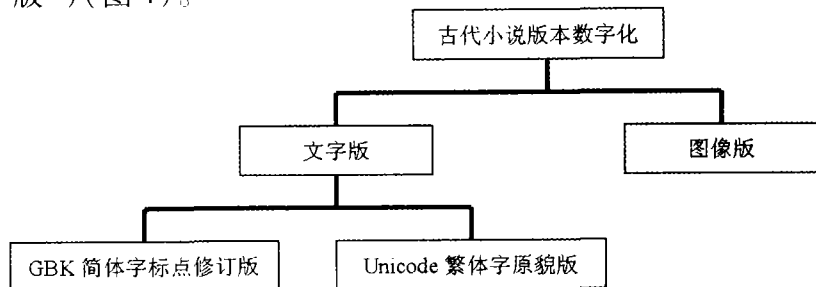


图 4 三种数字化版本

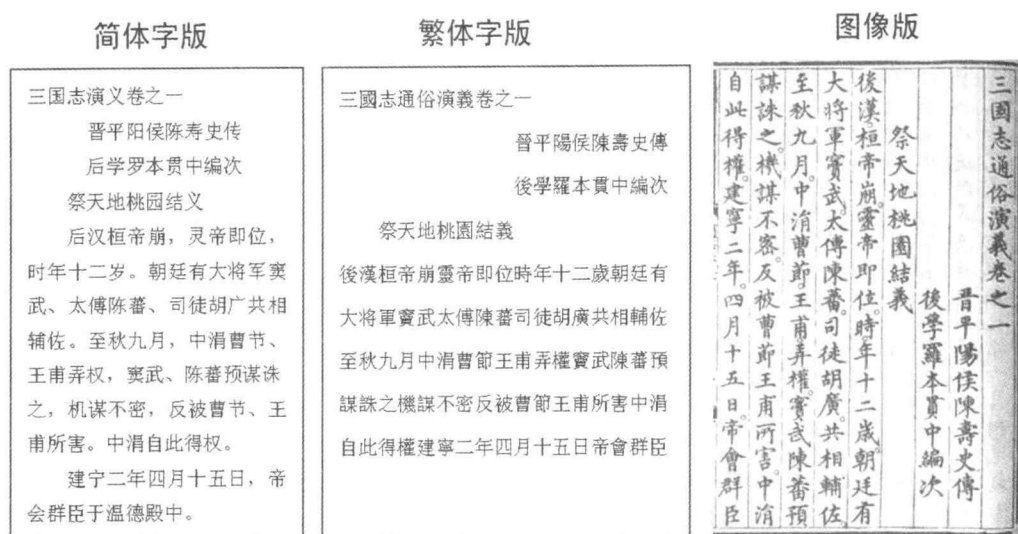


图5 简化字版、繁体字版和图像版(《三国演义》)

三种版本有很大差异,简体字版采用 GBK 简体字,分段,加标点,对于明显的错误作修订,主要用于阅读和版本比对。繁体字版采用 Unicode 繁体字,不分段,不加标点,文字保持原版原貌,对于错误不加任何修订,行款也与原版相同。图像版是扫描原书,完全保持了原版原貌。

3.3 两种文字版

古代小说版本数字化包括简体字版和繁体字版两种。

GBK 简体字标点修订版有以下特点:

- 采用简体字 GBK 字库;
- 一般分段、加标点;
- 一般根据排印本整理;
- 排印本为方便阅读,做了修订,已不是原版原貌;
- 主要用于一般阅读、检索和版本比对。

Unicode 繁体字原貌版有以下特点:

- 采用繁体字 Unicode 4.0 字库;
- 保留各种异体字、俗体字;
- 不分段、不加标点;

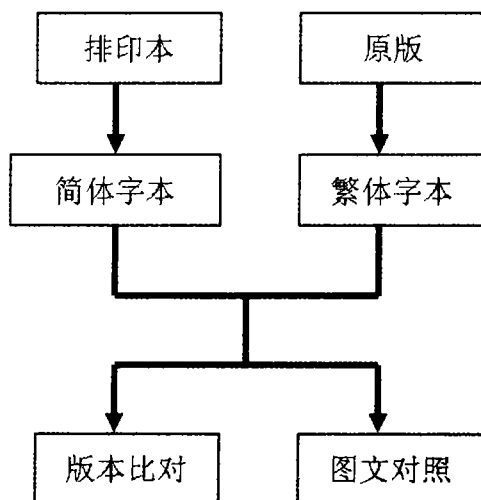


图5 两种文字版的整理

- 根据原版整理;
- 文字严格保持原版行款,但不要求完全保存原貌
- 对于任何错误都不作修订;
- 主要用于版本比对和图文对照。

制作两种文字版虽然会大大增加工作量,但有以下原因最终决定作两种版本:

(1) **阅读和检索**:简体字便于阅读、检索,而无法保存原貌。繁体字可以保存原版原貌,但不利阅读。

(2) **原貌和修订**:为一般阅读,最好对于明显的刻印错误做修订。但对于学术研究,应该严格保持原版原貌,不仅要保持原版行款,而且就是错误都要保留。但对于文字不修订,不利于一般阅读。

(3) **用途**:文字版有两个用途,即版本比对和图文对照。图文对照必须用繁体字,而版本比对中用简体字和繁体字都可以。

表 2 简体字标点修订版和繁体字原貌版

	简体字标点修订版	繁体字原貌版
字库	GBK	Unicode
字体	简体字	繁体字
排列	横排	横排、竖排
整理来源	排印本	原版
分段、标点	有	无
修订、改错	有	不修订、不改错
行款	无	与原版一致
目的	浏览、检索、版本比对	图文对照、版本比对、版本演化

3.4 异体字和俗体字处理

中国古代小说由于是通俗读物,刻印粗糙,因此有大量的异体字和俗体字,在简化字版中都作了处理,变成通行字。繁体字版的一个大问题就是如何处理计算机字库中没有的各种复杂的异体字、俗体字,有两种处理方式。

一种方式是和简体字版一样,全部转化为通行字。这种方法的优点是便于阅读和检索。不利之处是无法保持原版的原貌,而有些异体字和俗体字的原貌,

对于研究版本演化是很重要的。而要转化为通行字,就要求录入员要熟悉各种异体字、俗体字和通行字的对照,这对于一般不熟悉古文字的录入员,有很大困难。

另一种方式是尽可能保留所有异体字和俗体字的原貌,Unicode 4.0 有 7 万汉字,其中就有大量的异体字和俗体字。这种方法的优点是,可以尽可能保持原版的原貌,有些异体字和俗体字的原貌对于研究版本演化是很重要的。如《红楼梦》中有许多避讳字,对于版本研究非常重要。这种避讳字也是异体字,如全部转化为通行字,就丧失了版本研究价值。保留异体字和俗体字的不利之处是,大量异体字和俗体字不便于阅读,也无法统一检索。采取这种方法,要求录入员在输入时,要仔细在 Unicode 4.0 字库的各种字形中,选择符合异体字和俗体字的字形,因此要求录入员要非常仔细和小心。

比较两种方法。如果采取第一种方法,全部转化为通行字,再想复原版本原貌是不可能的。但如果采取第二种方法,先保留原貌,将来如果能制作一个异体字、俗体字和通行字的对照表,则可以很容易利用计算机将所有异体字和俗体字转化为通行字。当然,要制作这种对照表的工作量非常大。因此,第一种方法是不可逆的,而第二种方法是可逆的。

两种方法各有优缺点,有些学者主张采用第一种,有些学者主张采用第二种。考虑再三,还是决定采取第二种保持原貌的方法。主要考虑第二种方法,通过制作转换表,将来可能将异体字、俗体字转化为通行字,而第一种采用通行字后就无法再恢复原貌了。

表 3 异体字和俗体字处理方法

比较项目	转化为通行字	保留异体字和俗体字
原貌	丧失原貌,不可恢复	保留原貌
转化	不行	可以
阅读	方便	不方便
检索	容易	不方便
录入员	熟悉异体字、俗体字和通行字对照	熟悉大型字库各种字型

3.5 文件编目

古代小说一般分“卷”，“卷”下分“回(则、节、段)”。由于历史演变，一部小说不同版本的分回(则)可能不同，包括以下两种情况。

第一，回数不同，分回不同。一些版本的一回可能相对于其他版本的两回，甚至三回。《三国演义》没有这种情况，都是120回或240则。但在《水浒传》、《西游记》中，不同版本的回数可能不同，即某版本的一回对应于其他版本两回(甚至三回以上)。对于回目不同的版本，必须人工统一分回，统一编目，才可保证计算机版本比对进行顺利。

第二，即便回的数目相同，但不同版本的分回处也可能不同，这种情况在各种小说中都有，在比对时必须特别注意。

文本文件命名采用以下原则。文字版的文本文件是按照“回”连续编码，如《三国演义》嘉靖元年本第一、二回的文本文件命名为“嘉靖本1”、“嘉靖本2”等，叶逢春本第一、二回的文本文件命名为“叶逢春本1”、“叶逢春本2”等，依次类推。按照这种命名规则，《三国演义》120回就有120个文件，240则就有240个文件。



图8 图像版显示图

3.6 两种文字版的实施

两种文字版计划分两期实施。

1. 首先完成简体修订版，主要根据排印本整理，目前已经完成五部小说21种。

2. 再整理繁体原貌版：

- 字库要求采用 Uunicode4.0(四字节、十六进制)，包含 CJK 统一汉字扩充 B，即国际标准 ISO/IEC10646:2003，总计 70195 个汉字。如：驍:298F5、彰:291D5、羣:29166、艤:26A77、鱣:26A77、璣:24AA0

- 全部用繁体字,不用简化字;
- 根据原版整理,不分段、不加标点;
- 文字严格保持原版原貌,对于任何错误都不作修订;
- 异体字、俗体字和其他生僻字也保持原貌,Uonicoe4.0 CJK 统一汉字扩充 B 有的要使用 CJK 统一汉字扩充 B;
- 行款要与原版一致,每行末加一个“回车、换行”,另起一行,下一页加两个“回车、换行”;
- 批注:三种批注(文内双行批注、行间夹批、眉批)只录入“双行批注”,其余两种批注暂不作,以后有条件再作。文内双行批注用()括号表示。其他回前、回末批不再加特殊标记。

3.7 图像版

一般版本研究都利用文字版,但对于某些问题,还必须审阅原文。如版本文字的书体等,文字版无法表现,而在图像本中看得很清楚。

为看原文,最好制作文字原版,但要制作和原版一模一样的原版,就要解决:错讹字、异体字、俗体字、同音假借字、略字、漏字等一系列问题,情况十分复杂,几乎是不可能的。解决这种问题的最好方法是利用图像版。图像版可以保持原版原貌,比文字原版更直观和准确。图像版加上图文对照和检索后,可替代文字原版的许多功能。

图像版的制作一般利用扫描仪直接扫描输入计算机,目前一般采用 JPEG 压缩,形成 .jpg 文件,每页压缩后约几百 kb。图像可放大、缩小,可翻页查阅。

每幅图像编码规则采取“回 + 页数”。即第一回第一页编号为“1-1”,第二页编号为“1-2”,依次类推。

但要注意,古代小说一般分卷、回,逢新“卷”要再起一幅新页,而逢新“回”,则不换页,而是连续刻印下去。这样前一回最后一页和下一回第一页很可能在同一页上。如图 8 是《三国演义》嘉靖元年本第十则的开始,也是前一则(第九则)的最后一页。这样按照上述编码规则,作为第十则的第一页编码为 10-1,而按照第九则最后一页,编码为 9-29。也就是说,这类跨“回”的图像文件会出现两次,第一次是前一回的最后一页,第二次是下一回的第一页。

四、版本比对原理

4.1 计算机版本比对

在版本文本数字化的基础上,可利用计算机开展古代小说的版本比对。

中国古代小说一般都经过不断翻刻,屡次增删和修改,这些刻本的内容和形式不断发生变化,刻印过程中的疏忽也造成了大量的讹误。因此,研究中国古代小说必须对各种版本进行校勘,仔细比对各种版本之间的差异,探索它们之间的关系,研究各种版本的演化过程可以为进一步的深入研究打下基础。

版本比对是版本研究的基本工作,过去人工进行多个版本比对是非常繁琐的事情,极其费时、费力。在《红楼梦》版本方面,为比对各种版本,众多红学家“七易寒暑”,出版了《脂砚斋重评石头记汇校本》,汇集了今存 11 种脂评本和程甲本作对比校勘,为研究人员全面考察《红楼梦》的版本源流提供了方便。这种思路是对的,但纸本有很大局限性,应进一步进行版本数字化,并利用计算机实现任意版本之间的自动比对,将人们从繁重的手工劳动中解放出来。计算机版本比对包括所有数字化版本之中的任意数量版本之间的比对,即可实现任意两个版本(如《三国演义》毛评本和嘉靖本)任意回目之间的比对,也可实现多个版本(如《三国演义》毛评本、嘉靖本、李卓吾本等)任意回目之间的比对。数字化版本比对是进一步分析、研究版本强有力的工具。

4.2 版本比对原理

两个版本之间的文字只有“相同”和“不同”两种情况,而“不同”又包括“增、删、改”三种情况,因此合计为四种情况。

甲版本	增	删	改	相同
乙版本	删	增	改	相同

图 8 版本文字比较示意图

“增、删、改”三种情况的比对处理如下：

1)增:甲版本比乙版本增加若干文字,比对时在乙版本中要增加相应的空格。

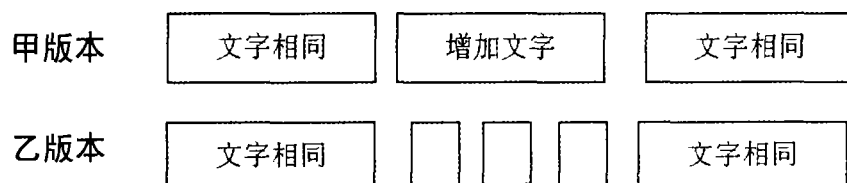


图9 “增”示意图

2)删:甲版本比乙版本减少若干文字,比对时在甲版本中要增加相应的空格。

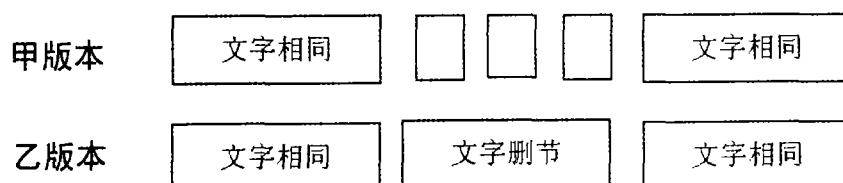


图10 “删”示意图

3)改:两个版本文字字数相同,但文字不同。比对时相同和不同的文字都对齐显示。

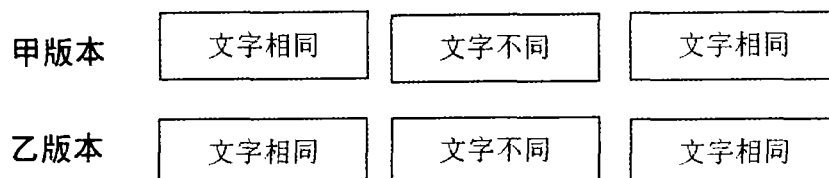


图11 “改”示意图

版本文字比对的方法实际和生物 DNA 比对的方法完全相同,所以版本比对的算法也借用了 DNA 比对的算法,利用 C++ 语言来实现。

4.3 多版本比对

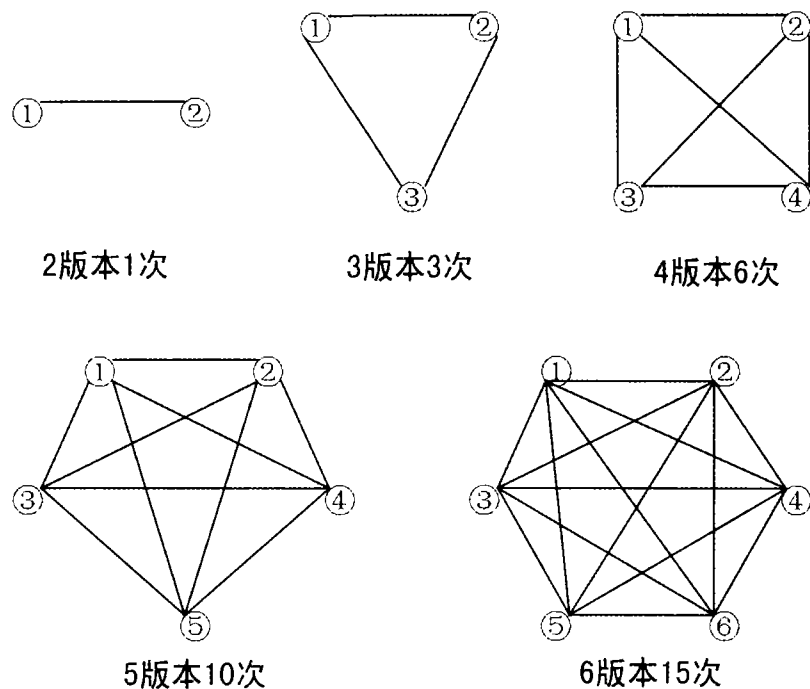


图 13 版本数和比对次数的关系

多版本比对, 比对次数会随版本数量增加而大量增加, 如图 13 所示。2 个版本要比对 1 次, 3 个版本比对 3 次, 4 个版本要比对 6 次, 5 个版本比对 10 次, 6 个版本比对 15 次, 7 个版本比对 21 次, 8 个版本比对 28 次。

五、逐行比对显示

5.1 逐行比对的特点

计算机版本比对显示主要有“逐行比对”显示、“分窗口比对”显示两种显示方式。

逐行比对是任选两个或更多的版本, 逐行、逐字进行比对。

逐行比对的显示方法:

- 每个版本显示一行;
- 所有版本的文字对应显示, 文字相同的对齐显示, 文字不同加空格;

- 版本名称采用缩写,如《三国演义》叶逢春本(叶)、黄正甫本(黄)、嘉靖本(嘉)、周曰校本(周)、李卓吾本(卓)、钟伯敬本(钟)、李渔本(渔)、毛本(毛)

逐行比对的特点:

- 研究人员直接利用上下行做逐字的比对研究,各种版本每个字的差异都很清楚;
- 标点符号比对可选保留或去掉;
- 可用光标或键盘进行下移或翻页,下移或翻页时,所有版本会自动相应地下移或翻页;
- 各种版本差异清楚,但某个版本显示不连续,阅读不清楚。

逐行比对的优缺点很明显:

优点:几个版本文字之间的差异很清楚;

缺点:每个版本文字连续性很差,阅读很困难。

以下以《三国演义》、《水浒传》、《西游记》、《金瓶梅》和《红楼梦》为例,显示逐行比对的结果。

5.2 逐行比对结果

例一、《三国演义》三个版本(嘉靖元年本、叶逢春本和黄正甫本)第一则逐行比对结果。

叶:后汉桓帝崩灵帝即位时年十二岁朝廷有大将军窦武太傅陈蕃 徒胡广共相辅佐至秋九月	
黄:后汉桓帝崩灵帝即位 年十二岁 大将军窦武太傅陈蕃司徒胡广共相辅佐至秋九月	
嘉:后汉桓帝崩灵帝即位时年十二岁朝廷有大将军窦武太傅陈蕃司徒胡广共相辅佐至秋九月中涓	
叶:宦官曹节王甫弄权窦武 陈蕃 谋诛 机谋不密反被曹节王甫所害 宦官自此得权建	*
黄:宦官曹节王甫弄权窦武 机谋不密反被曹节王甫所害 宦官自此得权建	
嘉: 曹节王甫弄权窦武 陈蕃预谋诛之机谋不密反被曹节王甫所害中涓 自此得权建	
叶:宁二年四月十 日帝会群臣 於温德殿中 却欲坐忽 狂风大作见一条青蛇从梁上	
黄:宁二年四月十五 日帝会群臣于 温德殿 忽 狂风大作见一条青蛇从梁上	
嘉:宁二年四月十五 日帝会群臣于 温德殿中方 欲 陞座殿角狂风大作见一条青蛇从梁上	
叶:飞下 约长二十余丈 蟠 於椅上灵帝惊倒武士急慌救出 文武互相推 倒 於丹墀者无数	*
黄:飞下 约长二十余丈 蟠于 椅上灵帝惊倒武士 救 起文武互相推 倒于 丹墀	
嘉:飞下来约 二十余丈长蟠于 椅上灵帝惊倒武士急慌救出 文武互相推拥倒于 丹墀者无数	
叶:须臾不见 此怪大雨大雷 降以 冰雹到半夜方 住东都城 中坏却房屋数千余间	*
黄:须臾不见 降下 雷 雨冰雹 半夜方止 东都城境 坏 屋数千	
嘉:须臾不见片时 大雷大雨降以 冰雹到半夜方 住东都城 中坏却房屋数千余间	

注:此处另有《水浒传》、《西游记》、《金瓶梅》三书各版本第一回逐行比对结果图表,限于本书篇幅有限,从略。下文在涉及分窗口比定时,亦从略了《水浒传》和《西游记》的比对结果。

例二、《红楼梦》三个版本(甲戌本、庚辰本和程乙本)第一回逐行比对结果,有标点。

甲:这东南一隅有处曰 姑苏 ,有城曰 阊门者,最是红尘中一二等富贵风流之地。这阊门外	庚:这东南一隅有处曰 姑苏 ,有城曰 阊门者,最是红尘中一二等富贵风流之地。这阊门外	乙:这东南 有 一个姑苏城, 城 中阊门 ,最是红尘中一二等富贵风流之地。这阊门外
甲:有个十里街,街内有个仁清巷,巷内有个古庙,因地方 窄狭,人皆呼作葫芦庙。庙傍 住	庚:有个十里街,街内有个仁清巷,巷内有个古庙,因地方 窄狭,人皆呼作葫芦庙。庙傍 住	乙:有个十里街,街内有个仁清巷,巷内有个古庙,因地方狭窄 ,人皆呼作葫芦庙。庙 旁住
甲:着一家乡宦,姓甄,名费,字士隐。嫡妻封氏,情性 贤淑,深明礼义。家中虽 无甚富贵,	庚:着一家乡宦,姓甄,名费,字士隐。嫡妻封氏,情性 贤淑,深明礼义。家中虽不 甚富贵,	乙:着一家乡宦,姓甄,名费:字士隐,嫡妻封氏, 性情贤淑,深明礼义。家中虽不 甚富贵,
甲:然本地便也推他为望族了。只因这甄士隐禀性恬淡,不以功名为念,每日只以观花 修竹,	庚:然本地便也推他为望族了。 因这甄士隐禀性恬淡,不以功名为念,每日只以观花 修竹,	乙:然本地 也推他为望族了。 因这甄士隐禀性恬淡,不以功名为念,每日只以观花种 竹。
甲:酌酒吟诗为乐, 到是神仙一流人品 。只是一件不足:如今年已 半百,膝下无儿,只有	庚:酌酒吟诗为乐, 到是神仙一流人品 。只是一件不足:如今年已 半百,膝下无儿,只有	乙:酌酒吟诗为乐,倒 是神仙一流人 物。只是一件不足: 年 过半百,膝下无儿,只有
甲:一女,乳名 英莲 ,年方三岁。一日,炎夏永昼,士隐于书房闲坐,至手 倦抛书,伏	庚:一女,乳名唤作英 菊,年方三岁。一日,炎夏永昼,士隐于书房闲坐,至手倦 抛书,伏	乙:一女,乳名 英莲 ,年方三岁。一日,炎夏永昼,士隐于书房闲坐, 手 倦抛书,伏

5.3 版本比对效果分析

版本比对的效果和以下两个因素有关。

第一,版本比对效果和版本的数量多少很有关系。

在上述五个例子中,版本比对的数量不同,两种版本比对的是《金瓶梅》,三种版本比对的是《三国演义》和《红楼梦》,四种版本比对的是《水浒传》、《西游记》。版本数量不同,比对的效果也不同。版本数量越少(两种),版本差异的细节越清楚;而版本数量增多(三、四种以上),版本差异宏观清楚,但细节差异就不太清楚了。因此应该根据需要,适当选择版本比对的数量。需要宏观分析版本差异,可以多选版本;但要细致分析版本之间差异,还必须选两个版本进行比较。

第二,版本比对的效果和小说版本的文字差异有很大关系。

版本差异越小,比对效果越好。如《三国演义》、《水浒传》、《西游记》和《红楼梦》大部分章节的文字差异不大,比对的效果也很好。而《西游记》的几种版本,开始的几回由于文字差异不大,比对效果还好。而越往后,由于文字差异很大,比对效果也就很差。

六、分窗口比对显示

6.1 分窗口比对

分窗口比对显示是在计算机屏幕上分别开多个窗口,分别同时显示多个版本进行比对。受计算机显示器屏幕大小限制,窗口比对显示一般以两、三个版本为宜。分窗口比对和逐行比对一样,几个版本文字采用对应显示,文字不同的用增加空格表示。当光标所在的窗口下移或翻页时,所有版本窗口都会相应下移或翻页显示(图 13)。

分窗口比对的特点:

- 逐行比对细节差异清楚,而分窗口比对各种版本的细节差异不如逐行比对清楚;
- 但分窗口显示中版本之间的宏观差异非常清楚;
- 分窗口比对虽然各种版本差异不太清楚,但每个版本显示连续,阅读清楚。

表 4 逐行比对和分窗口比对

项目	逐行比对	分窗口比对
细节差异	显示清楚	显示不如逐行比对
宏观差异	显示不清楚	显示清楚
版本文字连贯	文字不连贯	文字连贯
适用情况	版本文字差异小	版本文字差异大



图 13 《三国演义》分窗口比对

6.2 分窗口段落比对

除以上逐行逐字比对、分窗口比对外,还可以综合采用各种方式,“分窗口段落比对”就是另一种方式。这种方式分窗口显示,每句(或段落)对齐,而句(或段落)内部不再对齐。这种比对方式适应于文字差异比较大的文本比对,优点是整理比较简单,文字连续,缺点是句内没有做比对,文字差异不清楚。许多学者的研究中都常采用这种比对方式,如马幼垣整理《插增本简本水浒传存文辑校》等(表5)。

表5 分窗口段落比对(马幼垣整理三种《水浒传》简本《插增本简本水浒传存文辑校》)

新刊京本全像插增田虎王庆忠义水浒传卷之十七	新刊通俗增演忠义出像水浒传卷之十七	京本增补校正全传忠义水浒传评林卷之十七
第八十三回 宋公明奉诏破大辽 陈桥驿泪滴斩小卒	第八十三回 宋公明奉诏破大辽 陈桥驿泪滴斩小卒	宋公明奉诏破大辽 陈桥驿泪滴斩小卒
大辽国主起兵,侵占山后九州疆界,兵分四路而,劫虏山东、山西,抢掠河南、河北,各处申奏,请求救兵。 枢密童贯、太尉蔡京、高俅、杨戩纳下表章不奏。四个贼臣定计,奏陷宋江等。 殿前太尉宿元景奏道:“宋江方始归降百单八人恩同手足,死不相离,今又要害他。倘或泄漏,城中反变,将何解救?见今辽国兴兵,侵占山后九州所近县治,各处申表求救,屡次调兵征剿,贼势浩大,折兵损将,瞒着圣上不奏。以臣小见,正好差宋江等收伏辽贼,於国建功,实有便益。” 天子听罢,龙颜大喜,深责童贯等匿奏之罪,亲书诏敕,加宋江为破辽都先锋,其余诸将待建功封爵,就差宿元景敕去宋江军前开读。 宿太尉领敕出朝,迳到宋江行寨开读。 宋江等排香案,跪听宣读。	却说当年有大辽国主,起兵侵占山后九州边界,兵分四路而入,劫虏山东、山西,抢掠河南、河北,各处申奏,请求救兵,先经枢密院,然后得到御前。 枢密童贯、太尉蔡京、高俅、杨戩纳下表章不奏。四个贼臣定计,奏将宋江等众陷害。 殿前太尉宿元景奏道:“宋江方始归降,百单八人恩同手足,死不相离,今又要害他。倘或泄漏,城中反变,将何解救?见今辽国兴兵,侵占山后九州所近县治,各处申表求救,屡次调兵征剿,贼势浩大,折兵损将,瞒着圣上不奏。以臣小见,正好差宋江等收伏辽国之贼,实有便益。” 天子听罢宿太尉所奏,龙颜大喜,深责童贯等匿奏之罪,亲书诏敕,加宋江为破辽都先锋,其余诸将待建功封爵,就差宿元景亲赍诏敕,去宋江军前开读。 宿太尉领了圣敕出朝,迳到宋江行寨,开读诏敕。 宋江等排香案,跪听宣读。	却说当年有大辽国主,起兵侵占山后九州边界,兵分四路而入,劫虏山东、山西,抢掠河南、河北,各处申奏,请求救兵,先经枢密院,然后得到御前。 枢密童贯、太尉蔡京、高俅、杨戩纳下表章不奏。四个贼臣定计。奏将。 归降,百单八人恩同手足,死不相离,今又要害他。倘或泄漏反变,将何解救?见今辽国兴兵,侵占山后九州所近县治,各处申表求救,屡次调兵征剿,折兵损将,瞒着圣上不奏。以臣小见,正好差宋江等收伏辽国之贼,实有便。 天子听罢,龙颜大喜,深责童贯等匿奏之罪,亲书诏敕,加宋江为破辽都先锋,其余诸将待建功封爵,就差宿元景亲赍诏敕,去宋江军前。 宿太尉领了圣敕,迳到宋江行寨开读。 宋江等排香案,跪听宣读。

七、版本比对中的几个问题

7.1 按“回(则、节)”比对

古代小说计算机版本比对时,对于分回中的一些特殊问题,必须做特殊处理。

● “回”比对:

根据计算机比对算法,比对时必须按照“段落”比对。为简化比对,采用一“回”为一段落,即在一回内不再分段,而直接按照“回”进行比对。

● 120 回和 240 则:

同一部书不同版本的“回”(则、节)数很可能不同,已经完成的《三国演义》八个版本可分为两类:

120 回:毛本、李渔本;

240 则:嘉靖本、周曰校本、黄正甫本、叶逢春本、李卓吾本、钟伯敬本。

120 回中的一回对应于 240 则的两则,为方便比对,统一按照 120 回处理,一回一回进行比对。240 则版本与 120 回版本比对时,将两则合为一回进行比对。

● 分回(则、段):

有些古代小说不同版本的“回(则)数”虽然相同,但分回(则)处不同。在《三国演义》中嘉靖元年本和叶逢春本有两处分则不同,“志传”系列的繁本和简本的部分回(则)的分回(则)处也有不同,某版本回(则)末的文字在其他版本中移到下一回(则)中。分回(则)处的不同,导致分回(则)处的版本比对不理想。

● 回对应:

有些古代小说不同版本的“回数”不同,一些版本的一回可能相对于其他版本的两回,甚至三回。在《三国演义》中没有这种情况,但在《水浒传》、《西游记》中很多,某版本的一回相对于其他版本两回、甚至三回以上。

计算机版本比对是按照“回(则、节)”进行比对,因此分回很重要,分回不同,计算机无法进行比对。因此对于上述回目不同的版本,必须人工统一分回(或则),才可保证计算机版本比对进行顺利。

7.2 版本比对中的前后颠倒问题

版本比对中,由于多种原因可能出现前后颠倒问题。不同版本的叙述可能

发生前后颠倒的情况,如图 19 所示,甲版本叙述为 A、B 顺序,而乙版本叙述颠倒为 B、A 顺序,而文字完全相同。

在这种情况下,在计算机自动比对时就无法完全整齐地比对。如果前面对齐,后面就无法对齐;如果后面对齐,则前面又无法对齐。

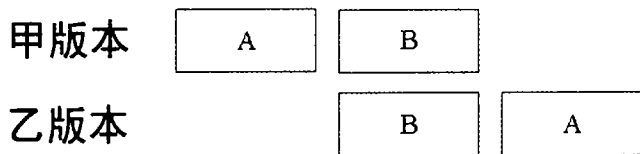


图 19 版本文字比较示意图(B 对齐)

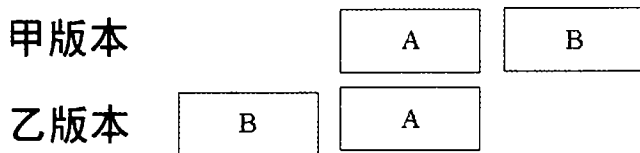


图 20 版本文字比较示意图(A 对齐)

如《三国演义》第一回刘备出场时介绍刘备,多数版本为“交游天下豪杰,素有大志”,而毛本改为“素有大志,专好结交天下豪杰”。语句颠倒,计算机无法完全对齐。

叶:	喜怒不形于色	好交游天下豪杰	素有大志
黄:		好交游天下豪杰	素有大志
嘉:	喜怒不形于色	好交游天下豪杰	素有大志
毛:	喜怒不形于色		素有大志专好结交天下豪杰

又如《三国演义》第一回介绍刘备身世称,多数版本称刘备曾“与公孙瓒等为友”,“刘元起见玄德家贫常资给”,而毛本又与其他版本的叙述刚好颠倒。这样比对时,只能按照最大匹配,将“刘元起见玄德家贫常资给”对齐,而“与公孙瓒等为友”就无法对齐。

叶:岁母使行学与同	宗刘德然辽西公孙瓒为友	刘德然	父刘元起	见玄德	
黄:岁	与同	宗刘德然辽西公孙瓒为友	德然	父刘元起	见玄德
嘉:岁母使行学与同	宗刘德然辽西公孙瓒为友	玄德	叔父刘元起	见玄德	
毛:	叔父刘元起奇其言曰此儿非常人也因见玄德				

叶:家贫常资给之元起妻	云各自一家何能	常尔	元起曰吾宗中有是	儿非常人也	中平元年涿郡
黄:家贫常资给之元起妻曰	各自一家何能济他	元起曰	宗中	此儿非常人也	中平元年涿郡
嘉:家贫常资给之元起妻曰	各自一家何能	常耳	元起曰吾宗中有	此儿非常人也	中平元年涿郡
毛:家贫常资给之	年				

叶:	
黄:	
嘉:	
毛:十五岁母使游学尝师事郑玄卢植与孙瓒等为友	

第三回叙述曹操灭宦官时也有类似的无法对齐情况。

叶:曹操一	时差	人救灭宫中之火	张让段珪
黄:曹操	令人救灭宫中	火	张让段珪
嘉:曹操一面	差	人救灭宫中之火	张让段珪
毛:曹操一面	救灭宫中之火	请何太后权摄大事遣兵追袭张让等寻觅少帝且说张让段珪劫	

叶:拥逼少帝及陈留王冒	突				
黄:拥逼少帝及陈留王	杀出後宫	门离城望北邙山而走逃难袁绍	请何太后权摄大事		
嘉:拥逼少帝及陈留王冒烟突		而走	袁绍请何太后权摄大事		
毛:拥	少帝及陈留王冒烟突火	出	后宰门离城望北邙山	逃难袁绍	请何太后权摄大事

又以曹操起兵讨董卓为例,其他版本介绍来投曹操的将领顺序为:乐进、夏侯惇、夏侯渊、曹仁、曹洪、李典,毛本改为:李典在乐进之后,在夏侯惇、夏侯渊、曹仁、曹洪之前。这样李典部分也无法对齐。

7.3 机械匹配与内容匹配的矛盾

由于计算机目前理解自然语言还有困难,只能机械地按照文字的“单字”进行最大匹配。这样可能发生计算机比对结果不合理的情况,即按照文字本身最大匹配的结果,与按照实际文字内容匹配结果不合理。如《三国演义》第九回叙述蒋雄救楚庄王,由于毛本叙述简化,也造成比对结果并不完全合理。类似例子还很多。

叶:绝纓	会正不知	戏爱姬者何人也	后庄王被	秦兵围住	见一大将杀入阵中	救出庄王
黄:绝纓	会		后庄王被敌人	围住	一将杀入阵中	救出庄王
嘉:绝纓	会正不知	戏爱姬者何人也	后庄王被	秦兵围住	见一大将杀入阵中	救出庄王
毛:绝纓之会	不	究戏爱姬			一大将杀入阵中	救出庄王
*						
叶:王见其人身带重伤	问之答	曰臣乃蒋雄				也昔
黄:王	问之	其将曰臣	蒋雄			也昔
嘉:王见其人身带重伤	问之答	曰臣乃蒋雄				也昔
毛:	之	蒋雄后为秦兵所困	得其死力相救	今貂蝉		

对于这类计算机比对的错误,可以人工进行修正,但工作量很大,因此目前暂时完全遵从计算机比对结果,而不对这类错误进行修正。

7.4 通用比对软件

计算机版本比对软件具有通用性,用户可以增加版本数量,而无须修改程序。

1. 增加版本:

增加版本包含两方面,一方面在一部小说中可以增加新的版本,另一方面,也可以增加新的小说。增加的版本必须遵守文本命名等规则。

2. 各种文本比较:

此系统是一种通用的文本比对软件,不仅可用于古代小说版本比对,而且可用于任何文本之间的比对,如各种文献之间的比对,文章之间比对等,这种通用的版本比对软件有一定的使用价值。

八、同词脱文

8.1 计算机分析同词脱文的意义

“同词脱文”是中国古代小说中经常发生的事情。小说在抄写或刻印中,当同一词出现两次而位置很近时,由于疏忽,很可能会将第二个词看成第一个词,而将两个词中间的文字遗漏掉,称为“同词脱文”或“串行脱文”。

同词脱文是小说研究,特别是版本演化分析的一个有力手段,在古代小说版

本研究中被广泛采用。在《三国演义》版本研究中,英国魏安用人工统计所收集到的 35 个版本,共发现 154 处“同词脱文”,以此对《三国演义》版本演化进行了详细分析,得出了《三国演义》版本演化图。另外,在目前《红楼梦》热点问题“程前脂后”的论战中,同词脱文也是一个重要的分析研究手段。有人也是利用人工方法,统计庚辰本和程甲本的同词脱文,共发现 120 多处。

传统人工进行多个版本同词脱文分析是非常繁琐的事情,效率极低,要准确统计和比较两个版本之间的同词脱文,是非常困难的事。而用计算机代替人工比较和准确统计两个版本之间的同词脱文,是非常容易的事,效率大大提高,在这方面,最能发挥计算机的优势。用计算机统计《三国演义》各种版本的“同词脱文”,其总数大大多于魏安人工统计出的 154 处。数字化同词脱文为学者研究版本提供了一个有利的工具。

8.2 同词脱文分析

和版本比对一样,计算机检索的同词脱文结果可以逐行显示(图 21)。

嘉靖本: 无脱文	今若取徐州,多留兵则不足用,少留兵则吕布乘虚寇之,是无兖州也。若徐州不得,
叶逢春本:有脱文	今若取徐州 不得,

图 21 同词脱文逐行显示示意图

和版本比对一样,同词脱文可以开窗口显示(图 22)。

嘉靖本:没有脱文	叶逢春本:有脱文
今若取徐州,多留兵则不足用,少留兵则吕布乘虚寇之,是无兖州也。若徐州不得,将军当安归乎?	今若取徐州, 不得,将军当安归乎?

图 22 同词脱文开窗口显示示意图

很明显,逐行显示同词脱文,比开窗口显示更清楚。
利用计算机可以不遗漏的查找到所有的“同词脱文”,但这些是否就全都是

真正的刻印中产生的“同词脱文”，而不是人工的增删，还需要经过人工仔细与对照原版分析，是否可能产生同词脱文。也不排除人工增删时，恰好出现了“同词脱文”的情况，这就是假“同词脱文”。计算机在不能实现自然语言理解的情况下会出现许多局限，这是其中一例。当然即便人工判别同词脱文产生的原因，也不是一件易事。

在完成同词脱文的查找后，还可以利用计算机，对两个（或更多）的版本之间的同词脱文，进行各种分析统计，为分析版本提供依据。

九、图文对照

9.1 图文对照的方法

研究版本有时需要比对图像版和文字版，即图文对照，这是一个非常重要的功能，利用计算机很容易实现。利用计算机进行图文对，就是在计算机屏幕上开两个窗口，分别显示文字版和图像版，研究人员可直接利用所开的窗口，做文字和图像的对照分析研究，图像和文字可同步滚动、翻页显示。

图文对照的一个目的是查看原版的原貌。前面曾介绍，在文字版中，许多异体字计算机无法处理，而利用一些符号（如□等）来代替。而要看到这些字的原貌，就可以利用图文对照，调出相应的文字时，就可以在图像窗口中看到原版的原貌。

图文对照最好做到图文严格对齐，即行款项同，每列（行）的字数与原版一样。而且当一页图像下移或翻页时，相应的文字也会同步滚动下移或翻页显示。为此制作繁体字版时，要严格按照原版的行款整理，在原版每列的结尾要另起一行，在每页的末尾加两个“回车、换行”标记。这样计算机软件在处理文字时，每逢一个“回车、换行”是另起一行，两个“回车、换行”是另起一行，并换一页。



图 23 图文对照

图像版与文字版比较最大缺点是检索困难,但在实现图文对照后,可以利用图文对照,对图像版进行间接的检索。

9.2 图像文字检索

有时要研究文字的原貌,需要对图像版的文字进行检索。检索时可利用图文对照系统,在这方面已有许多类似的先例,如国内一些单位开发出音韵学的名著《广韵》的电子检索系统和国外开发的碑帖拓片检索系统,就采用了类似的图像检索方式。

要实现图像文字检索,可以先在文字版中检索,检索出文字后,再利用图文对照系统,找到相应文字的图像

十、古代小说版本数字化的目的和作用

10.1 古代小说版本数字化的用户

古代小说版本数字化的主要用户可分为三类。

第一类用户是专业从事古代小说版本研究的学者,数字化和版本比对提供了一个人工很难实现的、更加快捷、方便的工具和方法,如逐行比对,可以用来仔细比对各种版本,对于研究版本演化有很大的帮助。对于版本研究的专业学者,要求是“精”和“深”,收入的版本数量越多越好。例如,对于专门研究《三国演义》版本的学者,《三国演义》的“演义”系统和“志传”系统(包括“繁本”和“简本”)都应该尽量收集全,目前《三国演义》八个版本还显不足。

第二类用户是古代小说一般研究人员,这些学者虽然并不专业从事版本研究,不像版本研究人员要求那样“深”,但要求“宽”。这些学者无需像版本研究学者那样细致了解各种版本的差异,但作为小说研究人员,也需要对于各种小说的各种版本的差异有个大致了解,有时由于研究的需要,也可能需要仔细了解某部小说的版本差异。对于这类学者,可根据需要分别选用逐行比对或分窗口显示,逐行比对可细致了解版本差异,而分窗口显示可以粗略、概括、快速了解各种版本的差异。

第三类是学习古典文学的学生,他们的要求可能比第二类一般的研究人员更低,作为学习古典文学的学生,也应该了解古代小说各种版本的差异,有时也可以利用这个系统写作论文。许多大学老师指导学生利用版本比对软件进行各种研究,写出了很多出色的论文。

要同时满足以上不同用户的不同要求,数字化的版本数量、质量都要大大增加,工作量非常大。

10.2 古代小说数字化的目的和作用

古代小说数字化一般的作用是为阅读浏览、检索查询、引用等,这是比较简单的应用。

古代小说数字化的最终目的是为研究服务,如果离开研究,只是为数字化而数字化,则小说数字化就没有动力,就没有持续发展的生命力。古代小说数字化如何与古代小说的学术研究结合,以促进古代小说研究,如何在古代小说研究中发挥作用,还要古代小说研究人员和计算机开发人员认真研讨。

古代小说版本数字化大大方便了版本研究,但真正的分析还必须由人工进

行,计算机不可能代替人工的分析研究(图24)。无论如何,古代小说数字化必定推进了小说的研究,这是不可否定的。忽视小说数字化是错误的,对数字化估计过高也是错误的,应该实事求是地评价小说数字化。

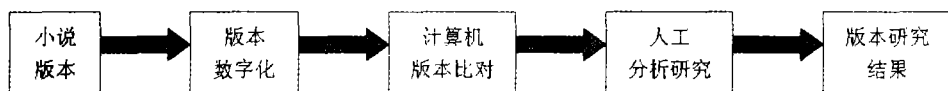


图24 从版本数字化到人工分析研究

版本数字化和版本比对只是古代小说数字化中的一个应用,主要体现在古代小说演化研究中,以《三国演义》版本研究为例,数字化在《三国演义》版本演化几个阶段的研究中都可以发挥作用。

第一、罗贯中“原本”原貌的研究。《三国演义》最初原本的原貌已经无法看到了,但利用数字化,可以对早期版本进行人工无法完成的细致分析,从而探讨原本的原貌。

第二、早期版本的演化研究。目前看到的《三国演义》早期版本分“演义”和“志传”两个系列,而“志传”系列中又分为“繁本”和“简本”两类。早期版本是如何演化的,是“演义”系列在前,还是“志传”系列在前,是“繁本”在前,还是“简本”在前;哪个版本更接近原本,这都是值得利用数字化研究的课题。

第三、后期版本的演化研究。从李卓吾评本开始,《三国演义》进入后期“评本”的发展阶段,这个阶段中各种版本是如何发展的,“志传”系列版本是如何逐渐消亡的,各种“评本”是如何兴起,而逐步取代各种“志传”本,都是可以利用数字化进行研究的课题。中国古代小说的发展都很类似,其他小说(《水浒传》、《西游记》、《金瓶梅》和《红楼梦》等)都有和《三国演义》类似的发展道路,都可以利用数字化进行研究。

古代小说版本数字化和古代小说版本研究的关系,类似体育和体育比赛(图25)。任何一种运动比赛理论上都有一个极限。如100米赛跑肯定有一个极限,这个极限是多少,没有人知道,但人要在7秒内跑完100米是不可能的。所以体育运动是没有终点的,但人类还在不断前进,赛跑速度的测量精度,也从0.1秒,到0.01秒,尽管没有终点,但体育运动还是在不断前进。

古代小说版本演化的研究和体育比赛很类似,都有一个理论上达不到的极限。虽然可能由于缺乏材料,可能永远也无法准确分清古代小说版本演化过程,更找不到古代小说原本的真实面貌,但古代小说版本的研究不是一定要研究出原本的真貌,只要研究有进步,就是成绩。所以,只要还可能前进,就要不断前进,要尽可能地不断逼近古代小说版本研究的顶点。关键在于,是否有进一

步深入研究的新线索、新方法、新思路。人工研究由于受版本数量多、文字量大、线索多等主观、客观因素影响,研究前进缓慢,甚至陷于停顿。而引入数字化后,通过前面分析,数字化提供了人工很难实现的功能,可能会大大加快版本研究的进度。虽然数字化肯定无法保证可以使版本研究达到最终的顶点,但会推动古代小说版本研究的进步,这是不可否认的。

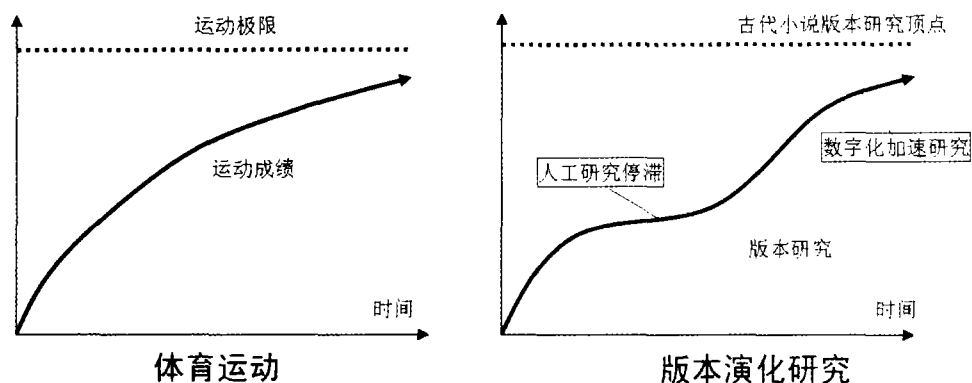


图 25 运动和版本研究示意图

古代小说版本演化研究也很象侦探破案。案件能否侦破关键看两点,第一,是否有足够的线索。目前版本研究中的线索还是很多,值得仔细研究。第二,是否有侦破手段。目前侦破手段在不断提高,许多案件过去无法侦破,但出现DNA检测手段后,许多过去的“死案”都可以侦破了。古代小说版本研究也很类似,过去只靠手工分析,因为工作量大,很难处理。而现在数字化后可以利用计算机进行分析,手段大大提高,提高的可能性也就增大了。

古代小说数字化研究除版本数字化外,在其他方面还有许多课题值得研究,数字化的应用领域还很宽。只要全面仔细地收集线索,利用数字化进行深入分析,还有许多工作可做,研究还可以继续深入。只要还有线索可查,只要有正确的方法和持之以恒的坚持,研究总会前进。

古代小说数字化研究也不是万能的,也有不成功、甚至失败的先例,在《红楼梦》研究中很值得探讨。八十年代开始国内外在《红楼梦》的作者分析、脂评研究、建立电脑数据库一些专题研究等方面都做了一些工作。

国外从八十年代开始出现利用计算机进行《红楼梦》研究,其中一个课题是利用计算机从用词频度方面分析《红楼梦》前八十回和后四十回,以判明《红楼梦》全书是否是由一人完成。早在上世纪五十年代瑞典著名汉学家高本汉就开创了从分析《红楼梦》的词语来探讨《红楼梦》后四十回作者的方法。上世纪八十年代美籍华人学者美国威斯康辛经济系教授赵冈先生应用计算机统计、分析

后四十回的用词规律,来分析《红楼梦》的作者问题。上世纪九十年代美国威斯康辛大学陈炳藻先生在周策纵教授的指导下,利用计算机从用词习惯、特点和字汇统计上探讨红楼梦作者问题。三人分析结果相同,即前八十回和后四十回为同一人所作。但这个结论与红学界普遍的共识,即前八十回和后四十回分别为曹雪芹和高鹗所作,有很大矛盾,并没有被红学界认可。

与国外研究相同,国内也开展了利用计算机从用词频度方面分析《红楼梦》作者问题。复旦大学统计系副教授李贤平在1987年采用计算机分析47个虚词在《红楼梦》中出现的频率,以此分析《红楼梦》的作者。他认为《红楼梦》为四人以上共同完成,前八十回是曹雪芹据《石头记》增删而成,并插入他早年的小说《风月宝鉴》,后四十回是曹雪芹亲友加工补写而成。深圳大学两位青年教师也利用计算机发现前八十回和后四十回在用词方面有明显“断层”,因此认为后四十回为另一人之手。这些研究在红学界引发一场学术讨论,上海师范大学孙逊和孙菊园教授以及华东师大陈大康等人对此提出批评,这些结论也没有得到红学界的认可。这场从海外蔓延到国内的利用计算机研究《红楼梦》作者问题,终于没有结果而终。造成这种结果的原因很多,由于各人选择用词频度各有自己的标准,所选择的词语很难保证合适和客观,选择的词频有限,信息量小,因此大大影响分析结果。

国内许多单位开发了多种《红楼梦》数据库。江苏省镇江市科委彭昆仑先生与南京工学院于1987年合作完成了“《红楼梦》数据库系统”,提供4154个项目检索,525名人物档案,24份统计表,总信息量5兆,内容十分丰富。1990年广东深圳大学也开发了“《红楼梦》多功能检索系统”,有30个专题,200多个项目,包括成语、修辞、饮食、服装、陈设、用具、诗词、戏曲、音乐、会话、爱情描写、心理描写、外貌描写、古代典籍等,内容十分丰富。但这些数据库似乎并没有在《红楼梦》研究中真正发挥什么大作用。

以上《红楼梦》数字化研究没有产生很大影响,这对古代小说的数字化研究是个很好的参考和借鉴,从中可以吸取如下教训。1. 数字化研究必须选择好课题,选题好,才有意义。2. 数字化研究的选题必须做到人工无法做到或很难做到,这样才有意义。版本比对、《红楼梦》前80回和后40回分析、《红楼梦》数据库这些都是人工很难做到的。但并不是只要人工很难做到的课题就一定有意义。3. 数字化研究的选题必须对研究确实有真正的价值,数字化研究必须得到大家认可,否则就是事倍功半,作用不大。《红楼梦》前80回和后40回计算机分析、《红楼梦》数据库由于对《红楼梦》研究没有实质的推动作用,其意义也就不大。4. 但选题是否真正有意义,有时必须要等研究结果出来才可以看清楚。因此,如何在研究结果产生之前就判断出其效益,就非常关键。

十一、结 语

1999年笔者开始从事《三国演义》数字化研究,得到了《三国演义》学会的大力支持,参加了1999年太原清徐第十二届《三国演义》学术研讨会到2004年四川绵阳第十七届《三国演义》研讨会的历届研讨会,每次会议上都在大会上发言介绍《三国演义》数字化研究的进展。近来由将数字化由《三国演义》扩展到《水浒传》、《西游记》、《金瓶梅》和《红楼梦》,由简化字版发展到繁体字版,不断前进。

2001、2003、2005年由笔者发起、组织在首都师范大学先后召开了三届(第一、二、四届)中国古代小说数字化研讨会,2004年由韩国中国小说学会组织在汉城召开第三次中国古代小说数字化研讨会,2006年在日本大东文化大学将召开了第五届中国古代小说文献与数字化研讨会,2007年8月再次由首都师范大学主办第六届中国古代小说文献与数字化研讨会。这几次专题研讨会大大促进了中国古代小说数字化的发展。

回顾1999年以来古代小说版本数字化的发展历程,非常感谢诸多专家对古代小说数字化的大力支持和无私帮助,其中包括国内专家沈伯俊、刘世德、段启明、陈翔华、周强、欧阳健、侯会等先生,日本学者金文京、中川谕和上田望先生等。此外还得到许多朋友的真诚帮助,包括国学网尹小林以及山东桑哲、上海李金泉以及天津邓宏顺等同志,古代小说数字化软件开发工作是由笔者的硕士研究生李东海完成的,对这些同志几年来对古代小说数字化研究的大力帮助和支持深表谢意!

【作者单位】 首都师范大学中国传统文化数字化研究中心