秦汉简牍语料库建设中基础语料的处理标准统一问题*

张再兴

【摘　要】秦汉简牍帛书文献具有多方面的研究价值。秦汉简牍帛书文献语料整理难度大，主要是：原始文献本身的面貌多种多样、各种原始释文文献的面貌很不一致、原始简牍与整理释文的不一致。因此在建设秦汉简牍帛书语料库的过程中，需要在数据单位、数据结构、数据处理方式、字库等方面进行统一加工，以最大限度地发挥其功效。同时，针对不同类型的文献，还应有针对性的处理方式，并采取数据分层处理机制。

【关键词】秦汉简牍帛书；语料库；处理标准统一

【作者简介】张再兴，博士，华东师范大学中国文字研究与应用中心教授，主要从事古文字信息化处理和古文字学研究。（上海　200062）

秦汉简牍帛书文献具有多方面的研究价值。但是秦汉简牍帛书文献数据量庞大，至今尚无完整、统一的语料库平台提供这类文献的检索，影响了其价值的充分实现。构建统一的语料库平台将便于秦汉简牍帛书文献的整体研究和内部比较，也便于满足字、词、句、篇多层次的研究需求。这一语料库平台的构建需要在保证科学性的基础上，满足文字学、汉语史、历史学等多学科领域的需求，满足语料库进一步深层次加工的要求。不过，秦汉简牍帛书文献的载体形制、内容类型以及整理现状都很复杂，建立统一的语料库平台实属不易。本文将讨论建立统一的语料库过程中基础语料的处理标准统一问题。

一

秦汉简牍语料库的建设是一个将简牍帛书图版加工成释文，再通过数据库提供给用户检索的转换过程。这一转换过程需要协调统一多种数据的不一致。这种数据的不一致具体体现在以下几个方面。

（一）原始文献本身的面貌多种多样

原始文献本身的面貌多种多样，如考古形态方面的清晰度、残损度、散乱度的不一致；简牍形制方面有简、牍、柧（觚）、檄、封检、削衣等的差异；文献内容方面有古书、官私文书、遣册等区别；书体方面有篆书、隶书、行书等区分。其中对语料库建设影响最大的当属各种简牍文献的书写行款的复杂多样。这种复杂性有些与简牍形制相关，如一枚简牍是否多行书写跟简牍的宽度相关，“宽 1 厘米左右的简一般是书写一行”①。有些则主要关于书写形式，如分栏、分段书写，这种情况对残断简缀合前后的读序有很大影响。

首先，简牍正文文字与标题、图表等内容之间有着巨大的差异。因此我们分别来看这三种内容。

1. 简牍正文文字行款的差异

主要有：

（1）行数。以一件完整的考古器物单位（残断者按不同单位计算）来检视，简牍正文文字大多数一

* 基金项目：本文为教育部人文社科重点研究基地重大项目《秦汉简语料库建设》（项目号：13JJD770029）阶段性成果。

① 程鹏万：《简牍帛书格式研究》，吉林大学 2006 年博士学位论文，第 141 页。

简一行，宽的木牍多数行，帛书则行数更多。个别也有在单行简中突然出现少量多行书写的情况。如睡虎地简《日书乙种》201简“不可取妻”，“妻”字写到了第二行。

（2）正反面书写。

（3）分栏。各种简牍文献都存在着分栏书写的情况，根据情形，栏数还有多有少。如放马滩秦简乙种《日书》常见分三栏书写。相应地，分栏书写的文字在简牍编联之后的读序也呈现出多种形式。如放马滩简甲种《日书》16—21简，上栏建除日，依照16到21简的读序，下栏接着还有收日、开日、闭日三个建除日分别在21简、18简、20简。16简、17简、19简下栏则是另一段文字。

（4）分写与接写。就同一种文献来看，一般各个意义单位之间分简书写。但是也有的直接写在同一简后面接写另外的内容。接写时有时会加分隔符，如睡虎地简《日书甲种》154简在“反枳日”的内容之后接写“毋以子丑傅户”，中间用●隔开。《日书乙种》46简第二栏“知旬六日毁”后面接写“五种忌日”，中间则用▬▬分隔。

就不同文献来看，同一种内容是否接写不同。如内容同为门，睡虎地简各门分简分栏书写；放马滩日书乙种则接写。

2. 标题的书写差异

标题的书写也有多种情况，例如：

（1）第一支简简首单行书写，睡虎地简中常见。

（2）第一支简简首分多行书写。如睡虎地简《日书乙种》196简上栏标题“穿户忌”，“忌”字在第二行。

（3）后面简的简首。如睡虎地简《日书乙种》“梦”部分从189简第一栏开始，标题“梦”在190简简首。“有疾”部分从181简开始，但是标题“有疾”却在183简简首，可能是由于181简简首有墨钉的缘故。

（4）最后一简简背。如睡虎地简《语书》，标题在15简简背。

（5）最后一简末尾。如周家台秦简131—244简“击行”两字放在最后，前有墨圆点，标明此是标题。①

（6）分两简书写。如《日书甲种》“盗者”部分标题，“盗”、“者”分别位于69简和70简简背首。

（二）各种原始释文文献的面貌很不一致

除了原始文献本身的差异外，语料库建设所依据的学界释文同样存在许多差异。

不同简牍帛书文献整理者所发表的原始释文的加工层次不同。如放马滩秦简只发表了没有断句标点的初步释文，而睡虎地秦简则发表了有标点、读法、注释及部分现代汉语翻译的释文。

而同样是标记了字的通假等文献读法的释文，处理的原则也不完全一致。这些不一致有些只是格式的不同，可直接统一。如重文、合文的处理，有些文献直接释作两个字，有些则加重文号“=”，并括注重文释文。有些则反映了认识的差异，需要根据具体情况加以统一处理。如读作“娶”的“取”字，有些文献括注“娶”，有些则不括注。

同一种文献释文内容，由于种种原因也存在着释文处理不一致的情况。有的是本应同义的字释读不同，如睡虎地秦简《日书》“病”篇某日病，某日有间，某日“酢”，甲种注“报祭”，乙种“有疾”则标注读为“作”，释为“起床”。还有的是因为断句不同。如《日书甲种》“稷辰”篇“正月以朔”与后面的气候词语“多雨”、“旱”等用逗号断开，而《日书乙种》“秦”篇中则不断开。

（三）原始简牍与整理释文的不一致

除了各简的编联及残断简的缀合外，典型的是书写行款与释读顺序不一致。例如，睡虎地简《法律答问》篇108简和109、110简的错简。

① 湖北省荆州市周梁玉桥遗址博物馆：《关沮秦汉墓简牍》244简注二，中华书局2001年版，第117页。

二

为了解决上述数据的不一致,在保证内容保真、处理科学、检索便捷原则的基础上,需要进行以下方面的统一工作,以便同一种文献内部统一,不同文献相同内容统一。统一还意味着符号含义的单一性和同一种情况符号表达的唯一性。如有残字的残简和无残字的残简,一般都用残断符号⊿表示,但是这两种情况并不一样。而残断的是多行简的话,如果只用一个符号表示,则更不能反映其原初的书写形式。

(一)数据单位统一

数据单位的设计牵涉文献的缀合、读序、编联等研究工作。简牍帛书的原始物理单位应当作为语料库数据的基础原始保存单位,而不是文献的意义单位。如简牍的条,帛书的幅。由于存在缀合等诸多问题,残断的简牍宜作为独立的单位。分栏、分面书写的简牍需作为独立的释读单位处理。

(二)数据结构统一

数据结构的设计影响语料库的最终使用功能。在满足数据检索、工具书编制、文本复原输出等要求的条件下,充分兼顾各种文献的特征,统一设计数据的结构。如简牍中的一些符号或作为内容的分隔标记,或作为句读的标记,需统一并加以标注。简牍释文的加工中,各种加工符号如标点等统一处理。

(三)数据处理方式统一

如果数据单位和数据结构的统一还是比较客观的技术处理的话,数据处理方式的统一则有着更高的学术要求。

除了重文、残字、补字、缺简、残简等符号标记方式的统一外,文字释读的统一是一项更为艰难的工作。

首先,由于原始整理者的出发点差异,字形隶定的原则并不一样。有的偏向用通用字形,有的偏向用隶古定。语料库设计应在可行的情况下以满足较高层次的要求为目标,即文字学的异体字研究为目标,确立文字的隶定原则。包括构件差异原则(如构件的更替或增减)、方位差异原则(如构件的移位)、缀符差异原则(如有无缀加符号及缀加符号类型的差异)等。另外,秦汉简牍文字已经有着明显的笔画区分,因此一些足以引起构件差异的笔画差异(如京京、广厂等)也应加以区别。

其次,读法标记的统一。除了前文所举的不同文献差异和同一文献不同地方的不一致外,还有如睡虎地秦简有些注释中说明的读法在释文正文中不加以标记。这种情况有些是出于整理者的慎重态度,有些则属于疏漏。

再次,尽可能以复杂形式的文献为标准,建立尽量细致完备的属性标记。相对于文献属性的缺失,属性标记的完整更有利于文献语料库的统一利用。

(四)字库统一

字库统一包括系统字库的字形统一和新造字的字形统一,主要目的是保证字形的唯一性。现有的系统字库由于大量异体字的存在,不同的整理释文往往采用不同的异体,也需要根据简牍的原始形体进行统一,如"爲"、"為"。同时,由于现有系统字库的限制,大量新造字更需要一个统一的确保字形唯一性的字库。

三

数据统一并不意味着形式上的完全一致,有些情况需要单独考虑。

一、数据处理的统一与学术分歧的关系处理。这是两个不同性质的问题。释读、编联等各方面的

学术分歧应当以特定的方式保存在数据库中并能随时呈现。

二、针对不同的文献内容类型，应采取相应的适当处理方式。如日书类文献中常见图表。表以简和栏结合组成表格，比较清晰，如睡虎地简《日书甲种》的"除"，将十二除和十二地支、十二月相配合。图则形态多样，如《日书甲种》的"人字"、"门"以及尹湾汉简的博局图等。此类文献无论是在显示形式还是用户检索上都有不同于普通文献文本的要求。图表往往需要多条简编联在一起，作为一个整体单位对待。甚至书写者也不是按简的顺序来书写的，如周家台简线图一的书写，学者认为是以扇面为单位由内向外书写。①而各种图表在具体使用上，也有着不同的阅读顺序，如有些按顺时针，有些按逆时针。而且作为实用占书，文献使用时的文本起点有着随机性。

三、采取数据分层处理机制。(1)原始数据与加工数据分不同层面。(2)加工数据分不同层面。(3)输出数据分不同层面。如简牍文献释文的输出分两种版本：一是各简直接隶定释文版，即只有字形隶定，无断句，无读法标记。二是整理释文版，即有读法及断句标记。

On Text-Processing Standards of Qin and Han Manuscripts Corpus Building

Zhang Zaixing

(Center for the Study and Application of Chinese Characters,
East China Normal University, Shanghai 200062, China)

Abstract: Qin and Han Dynasty Bamboo Slips and Silk Manuscripts Literature have a lot research value. However, Qin and Han Dynasty Bamboo Slips and Silk Manuscripts'original literature has many forms, the original interpretation literature is not consistent, and the original Bamboo Slips and organized interpretations are not consistent. Therefore, to maximize the effectiveness, when building the Qin and Han Manuscripts Corpus, it is necessary to unify the processing approach regarding data units, data structure and fonts. At the same time, for different types of literature, there should be different processing ways, and a hierarchical data processing mechanism could be adopted.

Key words: Qin and Han Dynasty Bamboo Slips and Silk Manuscripts; corpus; text-processing standards

① 黄儒宣：《〈日书〉图像研究》，中西书局2013年版，第49页。