

大徐本《说文解字》中 UCS* 所无字初步调查分析

楼兰

(华东师范大学中国文字研究与应用中心, 上海 200062)

摘要: 据初步统计, 字量只有 9000 多的大徐本《说文解字》有 33 字未包含于 UCS 中。通过对这 33 字的类型调查和成因分析, 我们发现汉字字符集的建设, 尤其是历史汉字字符集的完善, 是一个动态的过程, 需要进一步规范和调整。

关键词: 《说文解字》; 历史汉字; 字符集

中图分类号: H161 **文献类型:** A

0. 引言

建立历史字书的电子版, 是计算机信息化手段介入历史汉字研究的重要基础工作。几十年来, 从最初的 GB2312 到今天的 GBK, 我国在汉字内码存储方面获得了很大的进展。但是, 在建立历史字书电子版的过程中, 我们仍然面临需要造字的尴尬境地。《说文解字》(以下简称《说文》, 版本依据是大徐本) 是汉字历史上第一部汉字字汇专书, 我们发现 2000 余字不在 GBK 字符集中。因此我们在整理这部字汇的过程中采用了 UCS 国际标准字符集。该字符集中的汉字字符集(以下简称字符集) 共有 64288 个, 基本包含了《汉语大字典》和《康熙字典》中的所有字。然而, 据初步统计, 字量只有 9000 多的《说文》仍有 33 字没有包含于其中。这种现象的存在, 原因可能是多方面的。我们这里对这 33 字进行类型调查和成因分析, 以期对历史汉字的整理和字符集的建设与完善有所帮助。

1. 未收字的类型

根据这些字在《说文》中所处地位的不同, 我们把它们分为三类:

第一类是字头。共有 1 个。它是𠂔。

《说文·𠂔部》: “𠂔, 𠂔也。阙。”

第二类是在对字头进行结构说解时牵涉到的结构单位。共有 21 个。它们分别是𠂔、𠂔。

《说文·𠂔部》: “𠂔, 叉手也。从𠂔、𠂔。”

《说文·又部》: “𠂔, 分决也。从又, 𠂔象决形。”

《说文·目部》: “𠂔, 目围也。从目、𠂔。读若书卷之卷。古文以为丑字。”

《说文·𠂔部》: “𠂔, 大也。从𠂔𠂔, 𠂔亦声。阙。”

《说文·白部》: “𠂔, 别事词也。从白𠂔声。𠂔, 古文旅字。”

* UCS 是一个包括世界上各种语言的书面形式以及附加符号的编码体系。从 1993 年该标准的第一部分即 ISO/IEC10646.1 正式发布以来, 它一直在扩充发展。本文涉及的字符集, 其范围仅指目前 UCS 中所含的汉字字符集。它包括 20902 个 CJK 统一汉字, 6582 个扩展 A 和 36804 个扩展 B 汉字, 所收汉字总量达 64288 个。

《说文·肉部》：“胃，穀府也。从肉；囙，象形。”

《说文·亏部》：“亏，气欲舒出。与上碍于一也。亏，古文以为亏字，又以为巧字。”

《说文·丶部》：“主，镡中火主也。从𠂔，象形；从丶，丶亦声。” 臣铉等曰：今俗别作炷，非是。

《说文·木部》：“桒，槎识也。从木、𠂔。阙。《夏书》曰：‘随山桒木。’读若刊。”

《说文·贝部》：“赙，会礼也。从贝𠂔声。”

《说文·㝱部》：“㝱，病卧也。从㝱省，𠂔省声。”

《说文·页部》：“顶，颠也。从页丁声。𠂔，或从𠂔作。顛，籀文从鼎。”

《说文·长部》：“长，久远也。从兀从匕。兀者，高远意也。久则变化。亾声。𠂔者，倒亾也。” 臣铉等曰：倒亡，不亡也。长久之义也。

《说文·火部》：“𤇀，火飞也。从火，𠂔与𤇀同意。”

《说文·女部》：“嫫，有所恨也。从女𠂔声。今汝南人有所恨曰嫫。” 臣铉等曰：𠂔，古𠂔字。非声，当从𠂔省。

《说文·虫部》：“𧈧，马𧈧也。从虫、目，益声。𠂔，象形。《明堂月令》曰：‘腐𠂔为𧈧。’”

《说文·二部》：“凡，最括也。从二，二，偶也。从𠂔；𠂔，古文及。”

《说文·土部》：“埧，地也。从土𠂔声。”

《说文·𠂔部》：“𠂔，陋也。从𠂔，𠂔声。𠂔，籀文𠂔字。”

《说文·𠂔部》：“𠂔，𠂔也。从𠂔𠂔声。江夏有𠂔春亭。” 臣铉等案：《说文》无𠂔字，他字书亦无，此篇下有𠂔字，注云：江夏平春亭名。疑相承误，重出一字。

第三类是在对字头进行有关说解时牵涉到的另外的字。该类字，共有 11 个。它们分别是𠂔、𠂔、𠂔、𠂔、𠂔、𠂔、𠂔、𠂔、𠂔、𠂔、𠂔。

《说文·羽部》：“𠂔，乐舞。以羽𠂔自翳其首，以祀星辰也。从羽王声。读若皇。”

《说文·夂部》：“𠂔，越也。从夂从𠂔；𠂔，高也。一曰𠂔𠂔也。”

《说文·木部》：“𠂔，𠂔𠂔木也。从木𠂔声。”

《说文·木部》：“𠂔，𠂔也。从木𠂔声。一曰徙土𠂔，齐人语也。” 臣铉等曰：今俗作𠂔。

《说文·木部》：“𠂔，伐木余也。从木𠂔声。《商书》曰：‘若颠木之有𠂔。’”

《说文·石部》：“𠂔，以石刺病也。从石𠂔声。方𠂔切。”

《说文·豸部》：“𠂔，𠂔𠂔也。从豸𠂔声。”

《说文·火部》：“𠂔，束炭也。从火，𠂔省声。读若𠂔。”

《说文·虫部》：“𠂔，𠂔以注鸣。《诗》曰：‘胡为𠂔𠂔。’从虫𠂔声。臣铉等曰：𠂔非声，未详。”

《说文·车部》：“𠂔，直辕车𠂔也。从车具声。”

《说文·言部》：“𠂔，𠂔也。从言与声。”

2. 未收的原因

根据我们的调查分析，上述 33 字未见于字符集可能是由以下几方面原因造成的：

第一，传抄版本的不同造成字符集漏收。《说文》成书于东汉，至今已有近两千年的历史。现在所能见到的《说文》传本很多。各个传本之间存在一定的差异，比如部分字在其它版本的相应位置上另作它字。我们将大徐本和段玉裁《说文解字注》（下面简称段注）进行对比，发现字符集未收的 33 字中有 14 个在段注中写作或改作其它的字。大徐本作“𠂔”，段注作“𠂔”；大徐本作“𠂔”，段注作“𠂔”；大徐本作“𠂔”，段注作“𠂔”；大徐本作“𠂔”，段注作“𠂔”；大徐本作“𠂔”，段注作“𠂔”；大徐本作“𠂔”，段注作“𠂔”；大徐本作“𠂔”，段注作“𠂔”；大徐本作“𠂔”，段注作“𠂔”；大徐本作“𠂔”，段注作“𠂔”；大徐本作“𠂔”，段注作“𠂔”；大徐本作“𠂔”，段注作“𠂔”；大徐本作“𠂔”，段注作“𠂔”；大徐本作“𠂔”，段注作“𠂔”；大徐本作“𠂔”，段注作“𠂔”；大徐本作“𠂔”，段注作“𠂔”。以上出现在段注中的 14 字见于字符集中。我们可以推测，字符集在收字时，可能未考虑到各版本的差异性，由此造成漏收的情况。

第二，对汉字结构单位关系的不同认识导致字符集漏收。这主要是指对字头进行结构说解时牵涉到的结构单位。有一部分是为了帮助理解造字理据而在说解过程中产生的不成字结构单位。比如“𠂔”、“𠂔”、“𠂔”；有一部分是由于结构拆分而产生的不成字结构单位。《说文》中有些字能分析出一个独立的结构单位，但剩下的部分却丧失了独立性，它们不能独立用来记录语言。比如“𠂔”、“𠂔”、“𠂔”、“𠂔”。有一部分是由于篆形发生变化而产生的无理割结构单位。古汉字发展到小篆阶段，部分字形和甲骨文、金文相比较，已发生明显变化。虽然根据篆形仍可以分析为几个结构单位，但其中有一个或几个结构单位已变得不可解。比如“𠂔”、“𠂔”、“𠂔”、“𠂔”。以上 11 个不成字和无理割结构单位在字符集中未得到反映。那是不是不成字和无理割结构单位在字符集中都没有得到反映呢？我们可以注意到，字符集仍包含了《说文》相当一部分不成字的结构单位，如“、”、“、”、“、”、“、”、“、”等。汉字是由一定的结构单位按照层次组合而成。我们可以把汉字的结构分为多个不同层次。字符集到底收到哪一层的结构单位呢？字符集是否应收不成字和无理割性的结构单位呢？由于对汉字结构单位的不同认识，在这一问题上，字符集存在分类原则和收字标准不一致的地方。

第三，其它原因造成字符集漏收。这部分字主要是死字、异体字。比如“𠂔”“𠂔”、“𠂔”、“𠂔”、“𠂔”、“𠂔”。我们推测，可能是由于不同时期字符集规范制定者的语言文字观念不尽相同或是创建字符集时操作性的失误，造成了部分死字和异体字在字符集中未有反映。

3. 结论

通过对未见于字符集中的大徐本《说文》33 字的初步分析，我们发现汉字字符集的建设，尤其是历史汉字字符集的完善，是一个动态的过程，需要进一步规范和调整。这是因为：首先，汉字字符集漏收部分历史汉字与其现状相矛盾。据国家有关部门统计，社会常用汉字的数量仅为 2500 个，次常用汉字的数量为 1000 个，通用汉字 7000 个。汉字字符集之所以包容的汉字数量达上万个，这主要是存在相当数量的死字和异体字。漏收部分历史用字与这种现状相矛盾。其次，汉字字符集失收讹字不符合历史汉字实际。文字研究工作者利用信息化手段进行历史汉字的研究前提之一，就是字符集能反映历史字书的全部用字。一般来说，由于历代辗转传抄翻刻，历史字书的各个版本都必然避免不了存在一些错字、讹字。从校勘学的立场出发，我们应该对其进行匡误正讹。而从研究历史汉字的使用状况这一目的出发，

如果这些错字、讹字失收就等于丢掉了一部分重要的研究材料，不符合历史汉字的实际。因为从异体字的来源来说，很多异体字最早就是一些错讹字。最后，汉字字符集在收字时标准不一。由于学术界对于汉字结构单位理解的不一致，字符集在收字时候出现了前后矛盾的情况。这需要字符集规范制定者统一分类标准和分类原则，规范和协调字符集对不同层次的汉字结构的反映。综上所述，随着信息化社会的发展，历史汉字的整理研究，以及有关字符集的完善，需要有关语言文字研究者不断努力。

参考文献:

- [1]许慎《说文解字》(大徐本)，中华书局，1963年。
- [2]臧克和 王平《说文解字新订》，中华书局，2002年。
- [3]段玉裁《说文解字注》，上海古籍出版社，1988年。
- [4]汉语大字典编辑委员会《汉语大字典》，四川辞书出版社 湖北辞书出版社，1992年。

Preliminary research and analysis of some of the Chinese characters excluded out of UCS in “Shuo Wen Jie Zi”

LOU Lan

(Center for the Study of Chinese Characters and Their Application, ECNU, Shanghai 200062)

Abstract: Preliminary estimates show that 33 Chinese characters in “Shuo Wen Jie Zi” by Xu Shen which contains only over 9000 Chinese characters are not included in UCS . With the research of the types of these 33 characters and the analyses of this situation, we realize that the construction of Chinese characters set, especially the improvement of the diachronic Chinese characters set, is a dynamic process, requiring further regulation and adjustment.

Key words: “Shuo Wen Jie Zi” ; diachronic Chinese characters; collections of characters

收稿日期: 2003-11-10

作者简介: 楼兰(1982-), 女, 江西南昌人, 华东师范大学中国文字研究与应用中心硕士研究生。