中国语言生活绿皮书
国家语言文字工作委员会发布

YW B004

古籍 GUJI HANZI ZIPIN TONGJI

汉字字频统计

北京书同文数字化技术有限公司 编

商務印書館
THE COMMERCIAL PRESS

中国语言生活绿皮书

国家语言文字工作委员会发布

YW - B004

古籍汉字字频统计

北京书同文数字化技术有限公司 编

商務印書館

2008年·北京

图书在版编目(CIP)数据

古籍汉字字频统计/北京书同文数字化技术有限公司编. —北京:商务印书馆,2008
ISBN 978-7-100-05607-6

I. 古… II. 北… III. 古籍-汉字-使用频率-言语统计 IV. H087

中国版本图书馆 CIP 数据核字(2007)第 132150 号

GǓJÍ HÀNZÌ ZÌPÍN TǑNGJÌ
古籍汉字字频统计
北京书同文数字化技术有限公司 编

商务印书馆出版
(北京王府井大街36号 邮政编码 100710)
商务印书馆发行
北京民族印刷厂印刷
ISBN 978-7-100-05607-6

2008年7月第1版 开本 787×1092 1/16
2008年7月北京第1次印刷 印张 26½

定价:45.00元

目　录

前　言

1. 项目起源：2002 年 8 月，书同文数字化技术有限公司向国家语委申请了古籍汉字信息处理方面的攻关项目，当年 12 月得到批准。
2. 项目名称：古代汉语字频统计
   课题名称：计算机字库全汉字搜集整理及国际标准化研究中的课题
   （1）ISO/IEC 10646 CJK 汉字与《康熙字典》关联研究
   （2）中国古籍用字在 ISO/IEC 10646 CJK 汉字中分布研究
3. 项目与课题的合并：在项目执行过程中，由于“古籍用字在 ISO/IEC 10646 CJK 中分布研究”这一课题要依赖于字频统计的结果，所以与项目是合并进行、一并报告的。而“CJK 汉字与《康熙字典》关联研究”基本上是独立进行的。
4. 研究过程：
   （1）ISO/IEC 10646 CJK 汉字与《康熙字典》关联研究，是在长达 3 年的时间里，结合书同文公司的《康熙字典》电子版的开发与 ISO 汉字组 IRG 的工作进行的。其中，主要的过程是：
   - 扫描《康熙字典》，用软件切割抽取字头。
   - 将《康熙字典》字头按页码·字位入库，与 ISO/IEC 字码一一对应，进入 SuperCJK 数据库。
   - 利用数据库，反复排序、人工核对，清理“重见字头”和“遗漏编码字头”；其中包括：
     《康熙字典》中补遗字头的处理
     《康熙字典》中被 CJK 认同的字头的处理
   - 分别生成报表，见关联表“ISO/IEC 10646：2003 CJK 汉字与《康熙字典》关联表”正本与附件。

   （2）古籍字频和古籍 CJK 用字分布的研究，包含以下开发过程：
   - 《四库全书》和《四部丛刊》电子版开发（1997-2001）是工作基础。
   - 清理《四库全书》和《四部丛刊》电子版中的文本数据（滤除图形等信息），分别对 32000 字的 CJK＋编码汉字逐一统计出现率，然后将数据纳入 Access 数据库。
   - 代码映射，把 CJK＋中 EUDC 区中的某些自定义汉字、字符映射到 CJK 的标准区（如 CJK_B 和八卦等）。
   - 对汉字的出现字次、单字的相对覆盖率和累计覆盖率进行计算机处理，制图、制表，打印到纸张和 PDF 文件。反复校对。
   - 利用字频数据库进行古籍用字在 CJK，CJK_A，CJK_B，CJK_C 的分布的统计；同时也对原有的 Code Page 对古籍的贡献进行了分析，制表制图。
   - 在此基础上，又结合现代汉字字频数据，开发了“书同文查频”软件，可以随机查询

古今汉字单字/多字/关联字的字频及其累加覆盖率。

5. 项目成果：

(1)《大规模古籍字频统计及古籍汉字在 CJK 中的分布》

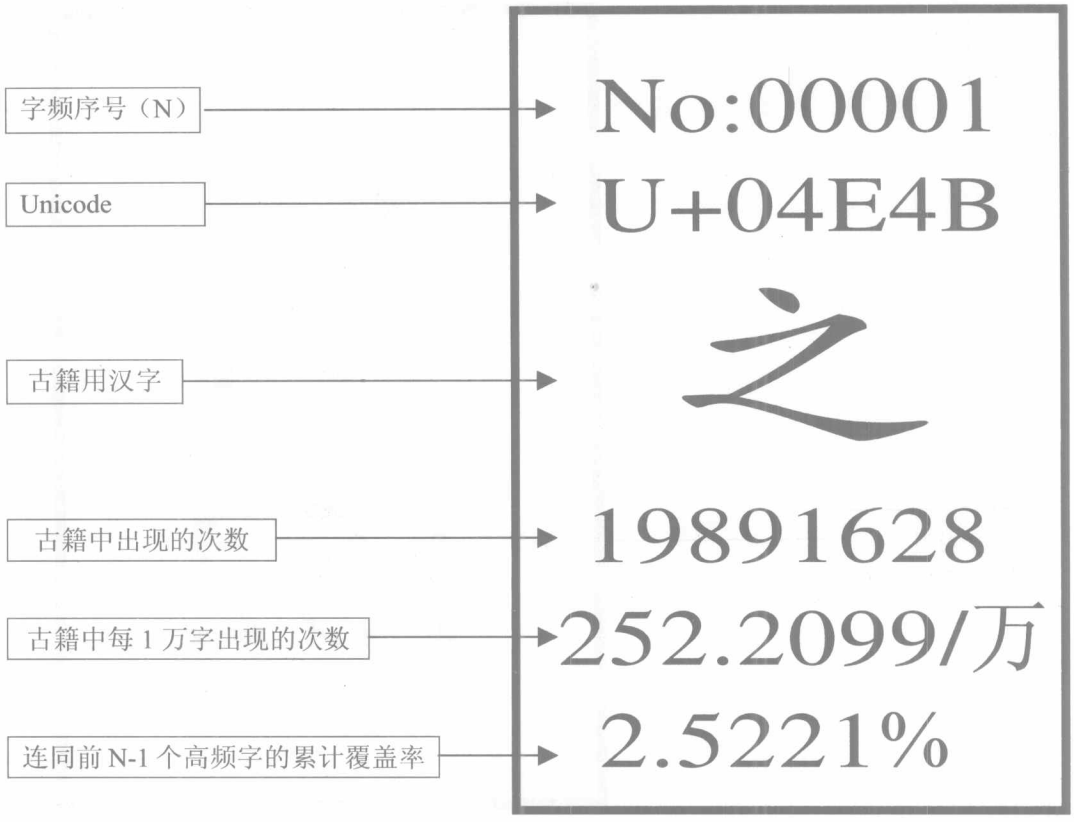
(2) ISO/IEC 10646:2003 CJK 汉字与《康熙字典》关联表

注：1. 本资料印刷部分，主要是本课题“项目成果”中的(1)部分，附赠光盘内容包括课题中的第(2)部分；

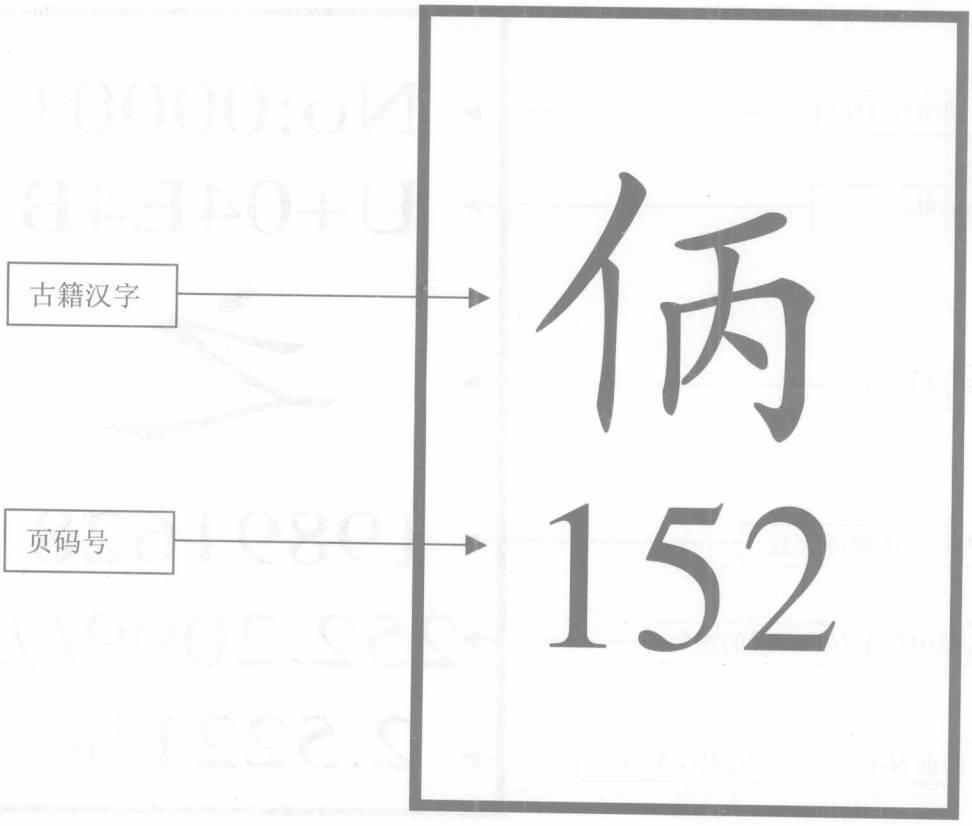
2. 在统计表甚低频字区，有 X 个阿拉伯字符和其它疑似错误的字，用方框表示。

凡　例

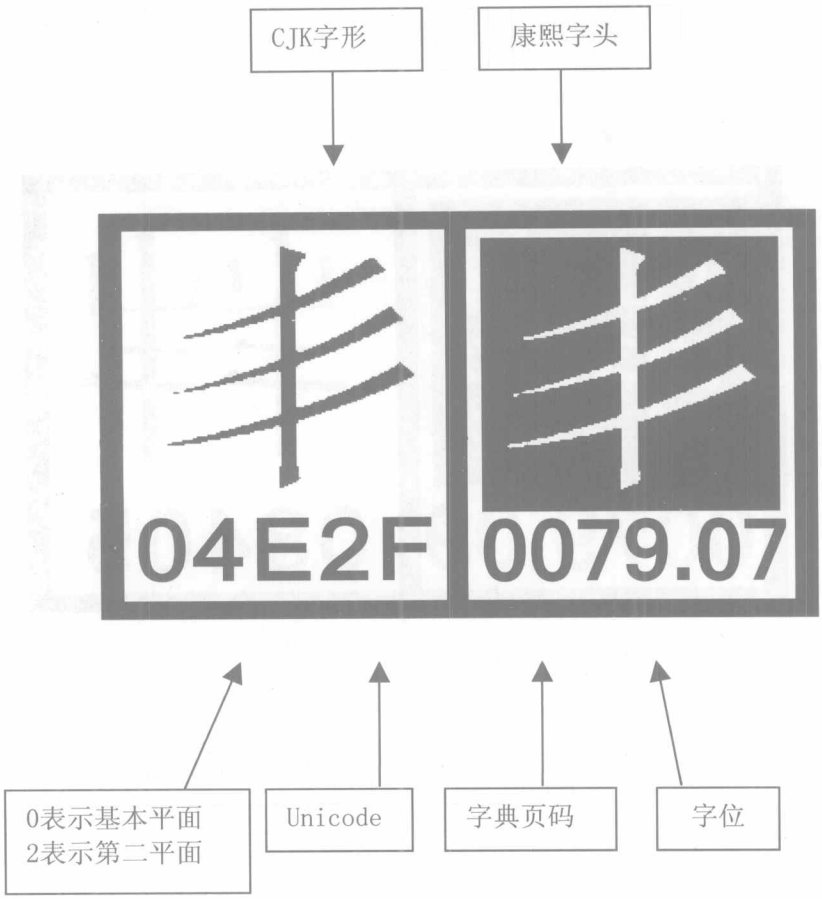
1. 古籍字频统计表凡例



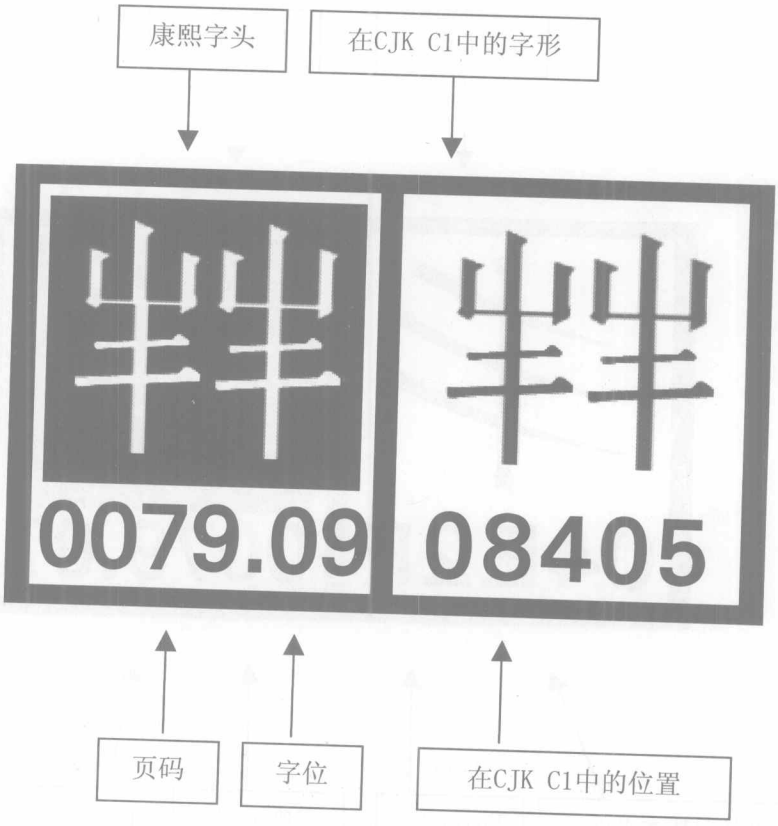
2. 古籍字频统计表索引凡例



3. ISO/IEC 10646:2003 CJK 汉字与《康熙字典》关联表凡例



4.《康熙字典》中即将在 CJK C1 中编码的汉字表凡例

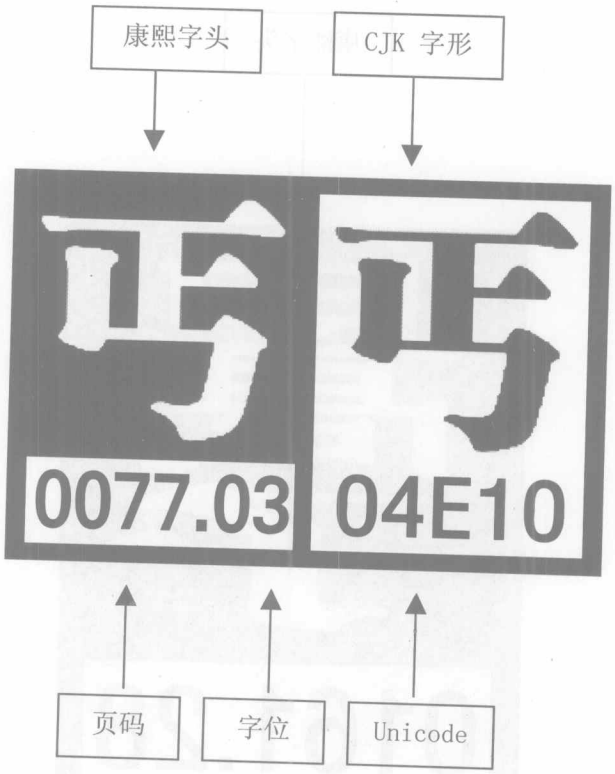


注：

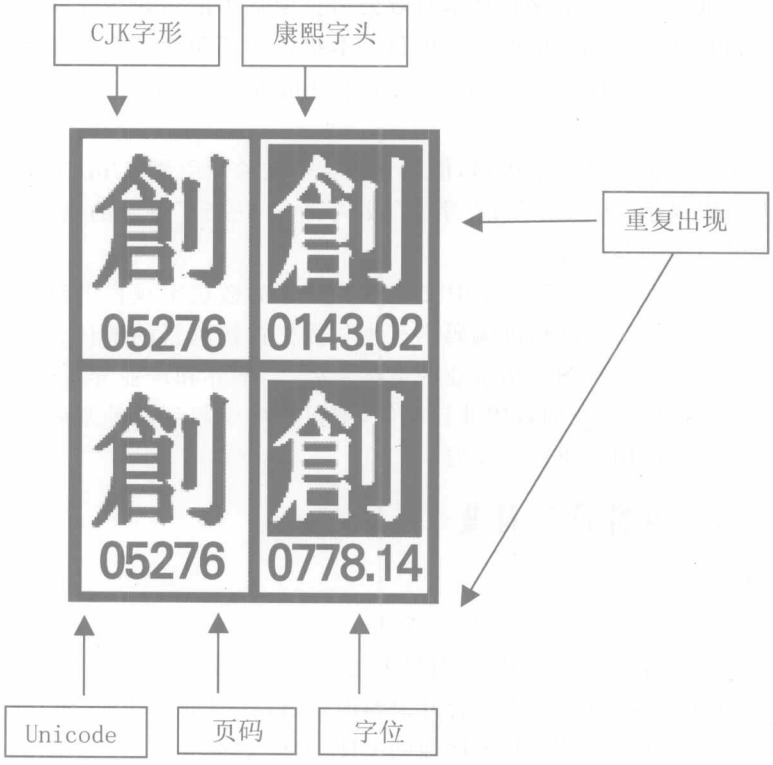
当时所指 CJK C1 的位置是依据 2004 年 7 月 ISO/IEC JTC1/SC2 IRG 版本。

5.《康熙字典》尚未编码的 36 个汉字凡例

6.《康熙字典》页码、字位序凡例



7.《康熙字典》重复字头一览表凡例



大规模古籍汉字用字统计报告

一、项目背景和概况

本项目是在北京书同文数字化技术有限公司长期从事的古籍数字化工作的基础上，由国家语委十五科技攻关计划资助的重点项目。本报告中还包含了另一个相关项目的成果，即“中国古籍用字在 ISO/IEC 10646 CJK 汉字中分布研究”。由于二者关系紧密，故一并报告。

这项研究基于八亿古籍汉字语料，借用书同文全文检索引擎(UniFTR 2.0)对语料中所出现的约三万编码汉字（接近于所谓“字头”或“字种”）进行了逐字的出现率（字次）统计，并对统计结果进行了初步的分析。

根据目前掌握的信息，迄今为止中文信息界所作的数亿字规模的汉字字频统计，都是基于现代汉语的。而基于国际标准编码字符集和数亿古籍语料的统计分析，此前尚未见报告。随着电子出版业和数字图书馆事业的迅猛发展，学术界和产业界对于古籍汉字的字频统计数据的要求日益迫切，我们希望并且相信，本报告可以起到某种基础性贡献的作用，有利于各项相关技术和应用的研究与发展。

二、统计对象- 语料及字符集

(一) 概述

本项目的基础语料来自文渊阁《四库全书》电子版和《四部丛刊》电子版。前者的汉字出现率，近七亿字次；后者近一亿字次；加起来接近八亿字次。尽管二者在使用率上尚有很大差异，但孰大孰小很难权衡，所以在合并语料时，未做任何加权处理，而是简单叠加。

前者于 1996—2000 年开发，基于增强的 CJK 编码字符集，CJK^+，用方正楷体表现；后者于 2000—2001 年开发，其字符集在 CJK^+ 的基础上稍有扩充，为编码汉字，用华天宋体字库表现。

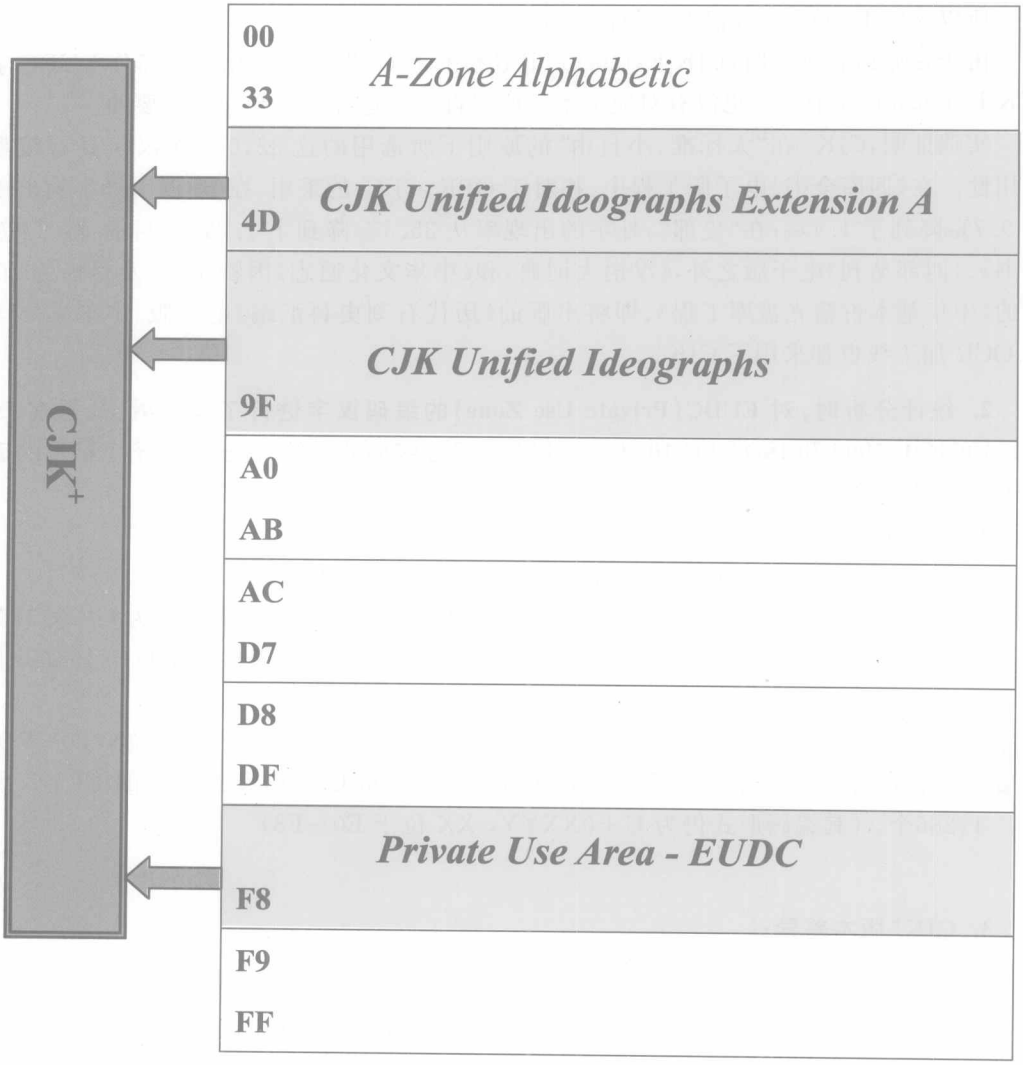
CJK^+ 的定义详见下节。

(二) 编码字符集

语料的编码字符集都是基于 ISO/IEC 10646-1:2000，UCS-2 或 BMP 模式。

遵循国际标准 ISO/IEC 10646:2003，它等效于工业标准 Unicode 4.0 和国家标准 GB 13000-2003（正在翻译中）。

1. 源语料基于 CJK^+，CJK^+ 的定义是：$CJK^+ ::= CJK + CJK\_A + EUDC$。



CJK 有 20,902 个标准编码汉字，CJK_A 有 6,500 个编码汉字，EUDC 则是在国际标准框架内专用区编码的汉字“Private Use Area - End User Defined Characters”。EUDC 有 6,400个码位，是精心选择的约 5,543 个汉字和符号，它们来自：

- 八卦-64 卦符号
- 古乐谱
- 《中华古汉语字典》外字
- 《四库全书》作者库外字
- 《四库全书》书目库外字
- 《四库全书》180 万条篇目（标题）外字
- 《四库全书》经-史-子-集出现率较高的外字
- 《中华文化通志》外字

● 《汉语大词典》中某些外字

所以这些 EUDC 字有相当高的代表性。

由于在原语料制作时，CJK Extension B 还没有颁布，况且还有相当一部分 EUDC 字在 CJK Extension B 中至今也没有对应关系。所以自定义这部分字是非常必要的。

实践证明，CJK$^+$ 在“大标准、小自由”的原则下所选用的这 32,000 个汉字具有很强的实用性。在《四库全书》电子版工程中，相对于 GBK，CJK$^+$ 的采用，在“经部”，外字的出现率从 9.7‰降到了 1.4‰；在“史部”，外字的出现率从 35.1‰降到了 1.2‰。目前，除了《四库全书》、《四部丛刊》电子版之外，《汉语大词典》和《中华文化通志》因特版，以及目标为 20 亿字的《中华基本古籍光盘库工程》、即将出版的《历代石刻史料汇编》电子版、中华书局语料库 OCR 加工线也都采用了 CJK$^+$。

2. 统计分析时，对 EUDC(Private Use Zone)的编码汉字进行了再映射，凡是在 ISO/IEC 10646-1:2000 和 ISO/IEC 10646-2:2001 中已经编码的字符，都已经给予了标准的而不是 EUDC 的编码。

- 有 3,755 个 EUDC 字映射到 CJK_B 中。编码为 U+2XXXX
- 有 82 个字可映射到 BMP。除八卦、六十四卦 72 个字外（编码形式仍为 U+027XX），另外有 5 个字在 CJK 中、5 个字在 CJK A 中，属于当初的错误重复编码。（其编码形式仍为 U+0XXXX）
- 有 320 字未来可以映射到 CJK_C 中。（其编码形式仍为 U+0XXYY，XX 位于 E0～F8）
- 在目前的标准编码字符集和未来的 CJK Extension C 中均没有对应的 EUDC 字有 1,286个。（其编码形式仍为 U+0XXYY，XX 位于 E0～F8）

3. CJK$^+$ 版本差异：

在《四库全书》电子版工程之后，应上海世纪出版集团要求，为《汉语大词典》因特版追加了 581 个自定义字，产生了新版本的 CJK$^+$ 字符集，用华天宋体字库显现。

这 581 个新追加的自定义字的 PUA 代码从 0EF6F 到 0F1B3。由于它们的出现频度相当低，所以对整个统计数据的影响微乎其微。

（三）文渊阁《四库全书》电子版语料

以下文字摘自文渊阁《四库全书》电子版出版说明：

《四库全书》是清代乾隆年间官修的规模庞大的百科丛书。它汇集了从先秦到清代前期的历代主要典籍，共收书三千四百六十余种。它是中华民族的珍贵文化遗产，也是全人类共同拥有的精神财富。

《四库全书》原抄七部，分藏北京故宫文渊阁、圆明园文源阁、沈阳清故宫文溯阁、承德避暑山庄文津阁、扬州文汇阁、镇江文宗阁、杭州文澜阁。后经战乱，今存世者仅文渊、文溯、文津三部及文澜本残书。

文渊阁《四库全书》是七部书中最早完成的一部，至今保存完好。自一九三四年

起，上海商务印书馆开始陆续影印文渊阁《四库全书》中的部分书籍，至一九八六年，才由台北商务印书馆将全书整套印出，题名《景印文渊阁四库全书》。过去半个多世纪，学术界从影印本《四库全书》得益良多，但从今天的角度看，影印本也存在明显的不足——体积大、书价高、保存难、检索不便；这些不足，影响并限制了该书的收藏、流通与利用。上海人民出版社和迪志文化出版有限公司有鉴于此，决心利用先进的数码技术将文渊阁《四库全书》电子化，为学术界提供体积小、售价低、易保存、检索快捷的电子版《四库全书》，并为推进中文信息电子化开辟新路。

“文渊阁四库全书电子版”以《景印文渊阁四库全书》为底本，由上海人民出版社和迪志文化出版有限公司合作出版，迪志文化出版有限公司、北京书同文电脑技术开发有限公司（现北京书同文数字化技术有限公司）承办全部开发制作工程。

电子化工程的重点是建立数据库和系统的技术开发。

数据库的建立：在国际标准的架构下，建立一个庞大的汉字信息数据库，是工程的第一步。为确保数据的齐备和准确，我们首先以数码扫描的方式录入全部二百三十多万页的原书图像，建立了原书图像数据库。然后利用先进的图像处理软件逐页检查，由计算机对原始图像自动分页、端正、去污，保证每幅图像的清晰度。

全文版数据的制作工程相当繁重，其校对工作更是极其艰巨。整个过程可分为三个阶段：(1)先对处理好的原文图像进行计算机切分、人工辅助纠错，提取每一个字的字迹图像；(2)再用清华大学计算机系人工智能研究室提供的多特定人规范手写识别引擎(OCR)，结合我们制作的超过七千字的 Unicode 版本的识别字典，把每个字迹图像识别成计算机的编码汉字，并给出每个字迹图像所可能对应的十个候选字及相关参数。解决百分之九十以上的录入问题；(3)然后用我们开发的“校得快”、“校得准”、“校得精”的三种“联机校对”软件，从不同的角度来进行五次无纸的数据校对工作。“校得快”在屏幕上显示字迹与其识别出来的汉字，一一对应、顺序校对，反复进行一、三校。“校得准”软件用于二、四校，以“交叉校对”方法打乱原文顺序，把所选页中同样的字聚集在一起，连同其所对应的字迹显示在屏幕上，从而使错字一目了然。“校得精”用于五校，它的特点是对全部数据再进行页对页、行对行的比对，将文本数据逐字逐句的和原文图像进行对照；并包括外字回填、一致性处理，实施全面检校。最后是专业校对，特别聘请专业工作者对数据作抽样校对。经此过程，建立起约七亿汉字的高质量的中文字符-字迹资料库。

参与技术开发的机构，除了迪志文化出版有限公司和书同文计算机技术开发有限公司以外，还有清华大学计算机系（负责 OCR 引擎开发），和北大方正电子有限公司（负责建立专用字库）。微软公司（北京）研究开发中心在平台技术等方面给予了有力的技术援助。

（四）《四部丛刊》电子版语料

《四部丛刊》是上个世纪初由著名学者、出版家张元济先生汇集多种中国古籍经典纂辑的。学者们公认此书的最大特色是讲究版本。纂辑者专选宋、元、明旧刊（间及清本者，则必取其精刻）及精校名抄本，故版本价值之高远在《四库全书》之上。多年来，该书一直深受文史工作者推崇，所收书常被用作古籍整理的底本。该书共计收书 477 种、3,134册、232,478页、近九千余万字。