

中国语言生活绿皮书

国家语言文字工作委员会发布



B006

中日韩

ZHONGRIHAN
CHANGYONG HANZI
DUIBI FENXI

常用汉字对比分析

北京书同文数字化技术有限公司 编

汉字



商务印书馆
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



B006

中日韩常用汉字对比分析

北京书同文数字化技术有限公司 编

商务印书馆

2009年·北京

图书在版编目(CIP)数据

中日韩常用汉字对比分析/北京书同文数字化技术有限公司
编. —北京:商务印书馆,2009
ISBN 978-7-100-06595-5

I. 中… II. 北… III. 汉字-对比研究-中国、日本、韩国 IV. H12

中国版本图书馆 CIP 数据核字(2009)第 032883 号

所有权利保留。

未经许可,不得以任何方式使用。

ZHONGRìHÁN CHÁNGYÒNG HÀNZÌ DUìBǐFENXī

中日韩常用汉字对比分析

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

北京瑞古冠中印刷厂印刷

ISBN 978-7-100-06595-5

2009年7月第1版

开本 787×1092 1/16

2009年7月北京第1次印刷

印张 34

定价: 68.00 元

《中国语言生活绿皮书》B 系列说明

《中国语言生活绿皮书》由国家语言文字工作委员会发布，旨在落实《中华人民共和国国家通用语言文字法》精神，提倡“语言服务”理念，方便社会共享政府拥有的信息，积极引导社会语言生活和谐发展，为构建和谐社会作贡献。

《中国语言生活绿皮书》分 A、B 两系列。A 系列是引导语言生活的“软性”规范。B 系列是中国语言生活的状况与分析，主要发布语言生活的各种调查报告和实态数据。国家向前发展的历史进程中，会遇到不少语言文字问题；世界语言生活风云万千，会对参与世界事务越来越多的中国发生各种影响。了解国内外语言生活状况，研究现实语言生活问题，制定科学的语言规划，对语言资源的保护与开发、保证语言生活的和谐与活力，具有十分重要的意义。

《中国语言生活绿皮书》B 系列是“实态”性质的：报告内容是实态的，语言数据的统计及其技术也是实态的。及时发布语言生活的实况与数据，就像发布水文监测、空气质量监测和气象监测数据一样，其作用不在于规范语言生活，而是为国家相关部门的决策提供参考，为语言文字研究者、语言文字产品研发者和社会其他人士提供语言国情服务。

《中国语言生活绿皮书》B 系列也可以视情况再分为自成系统的系列，如《中国语言生活状况报告》就是 B 中的一个分系列。

《中国语言生活绿皮书》是开放的。各年出版的《中国语言生活绿皮书》按 A、B 类各自连续编号，形成系列。发布的成果不局限于国家语言文字工作委员会的科研项目，也吸纳其他基金项目和企业、社会研究者的优秀成果。《中国语言生活绿皮书》的出版者也是开放的，欢迎各家出版社加入出版行列。《中国语言生活绿皮书》的出版内容和出版单位的选定，都遵循一定的遴选程序。

许嘉璐先生为《中国语言生活绿皮书》题字；教育部副部长袁贵仁同志，教育部副部长、国家语委主任赵沁平同志，国家语委咨询委员会副主任朱新均同志，都很关心《中国语言生活绿皮书》的编辑出版工作。相关课题组为《中国语言生活绿皮书》作出了贡献，一些出版单位和社会人士对《中国语言生活绿皮书》给予了支持与关心。在此特致谢忱！

教育部语言文字信息管理局

2008 年 6 月 15 日

顾问 许嘉璐 赵沁平

策划 教育部语言文字信息管理司

主编 张轴材

编者 北京书同文数字化技术有限公司

副主编：张福煜

数据库：张 虎

网络编辑：马 超

字库美编：薛 特

打印软件：魏永赳

开发协理：张弛宜

目 录

一、项目背景与目标	1
1. CJK 汉字统一编码的补充	1
2. 中文汉字教学的辅助	2
3. 中文信息技术开发的基础项目	2
二、对比对象与对比分析流程	3
1. 对比对象	3
2. 对比分析流程	6
三、字汇对比—覆盖率统计及其频率权重	12
1. 覆盖率	12
2. CY/HSK 覆盖/未覆盖各国家/地区字集诸区域在中国大陆现代报刊语料中的频率权重测算(数据表)	16
3. HSK 覆盖/未覆盖区域在中国大陆现代报刊语料中的频率权重图(交叉对比)	18
四、字形对比—字形异同统计及诸子集频率权重	20
1. J-J'、K-K'、T-T'、H-H' 的字形异同统计及其在中国大陆现代语料中的频率权重(数据表)	20
2. 各国家/地区之同形/微差/简化诸子集汉字比例(饼图)	22
3. 各国家/地区之同形/微差/简化诸子集汉字比例及权重贡献率(柱形图)	24
4. 各国家/地区之同形/微差/简化诸子集权重贡献率相对于子集大小比例的指数(柱形图)	25
五、简化模式的进一步分析	26
1. 香港小学用字 H-H' 简化异码分析	26
2. 台湾国小用字 T-T' 简化异码分析	26
3. 日本常用字 J-J' 简化异码分析	27
4. 韩国常用字 K-K' 简化异码分析	27
5. 各国家/地区字集简化方式之比例(交叉对比)	28
6. 各国家/地区字集各简化方式之频率权重(交叉对比)	29
7. 各国家/地区字集各简化方式之单字平均权重(交叉对比)	30
8. 两组重要的条件简化的细节	30
9. 大陆—台湾两组条件映射字的频率信息	33
六、凡例	35
1. 同形同码/微差同码字表(日本除外)	35

2. 简化异码字表(日本除外)	35
3. 同形同码/微差同码字表(中国—日本)	36
4. 简化异码字表(中国—日本)	36
七、对日韩港台汉字的特别说明	37
1. 关于台湾国小用字	37
2. 关于香港小学用字	39
3. 关于日本常用汉字	39
4. 关于韩国常用汉字	40
八、本报告的网络版	41
1. 网络版采用 B/S 体系结构(Browser/Server)	41
2. 查询类别的选择	41
3. 查询范围(子集)的选择	41
4. 待查汉字的输入	42
5. 浏览方式	43
6. 大字显示与动态笔顺	43
7. 察看说明与覆盖率统计、字形异同统计与频率权重等图表	44
8. 新旧字形大字对比图	44
9. 简繁异对照	45
10. 常见简繁错误	47
11. HSK 常用汉字	48
九、小结	49
十、技术注释	50
[注 1] 汉字认同的 XYZ 模型	50
[注 2] 关于汉字字形	50
[注 3] 关于“源字集分离原则”	51
[注 4] 关于“水平认同”与“垂直认同”	52
[注 5] 关于微小形差	54
[注 6] 关于汉字正形	56
十一、字形分类对照表	59
1. 基于台湾国小用字的大陆—台湾字形对照表	61
(1) [大陆—台湾] 同形同码字表 Y-T V4.0	61
(2) [大陆—台湾] 微差同码字表 Z-T V4.0	101
(3) [大陆—台湾] 简化异码字表 X-T V4.0	115
(4) [大陆—台湾] CY 独有字表 U-CY-T-T V4.0	145
(5) [大陆—台湾] HSK(汉办版) 独有字表 U-HSK0-T V4.0	147
(6) [大陆—台湾] HSK(北语版) 独有字表 U-HSK1-T V4.0	149
(7) 附: TW TTE Hanzi	151
2. 基于香港小学用字的内地—香港字形对照表	157
(1) [内地—香港] 同形同码字表 Y-H V4.0	157

(2) [内地-香港]微差同码字表 Z-H V4.0	185
(3) [内地-香港]简化异码字表 X-H V4.0	193
(4) [内地-香港]CY 独有字表 U-CY-H V4.0	215
(5) [内地-香港]HSK(汉办版)独有字表 U-HSK0-H V4.0	221
(6) [内地-香港]HSK(北语版)独有字表 U-HSK1-H V4.0	225
3. 基于日本常用汉字的中-日汉字字形对照表	227
(1) [中-日]同形同码字表 Y-J V4.0	227
(2) [中-日]微差同码字表 Z-J V4.0	245
(3) [中-日]简化异码字表 X-J V4.0	259
(4) [中-日]HSK(汉办版)独有字表 U-HSK0-J V4.0	275
(5) [中-日]HSK(北语版)独有字表 U-HSK1-J V4.0	287
4. 基于韩国常用汉字的中-韩汉字字形对照表	295
(1) [中-韩]同形同码字表 Y-K V4.0	295
(2) [中-韩]微差同码字表 Z-K V4.0	307
(3) [中-韩]简化异码字表 X-K V4.0	317
(4) [中-韩]HSK(汉办版)独有字表 U-HSK0-K V4.0	329
(5) [中-韩]HSK(北语版)独有字表 U-HSK1-K V4.0	341
(6) 附: Pronunciation Variants from KSX 1001: 1998	349
(7) 附: ISO/IEC 10646 Table 130-Row F9: CJK Compatibility Ideographs	353
十二、附: 原字表(Unicode 标注)	355
1 现代汉语常用字表	355
2. 汉语水平考试用字(汉办版)	385
3. 汉语水平考试用字(北语版)	409
4. 香港小学用字一览表	429
5. 台湾国小用字	459
6. 日本常用汉字表	501
7. 韩国科学和技术教育部指定汉字	519
后记	535

一、项目背景与目标

中日韩(CJK)各国家/地区的常用汉字表或教育用字表的颁布，均早于或平行于国际标准编码字符集 ISO/IEC 10646 的制定。因此，在这些常用字表中，往往都没有汉字的国际标准代码。由于技术条件和应用目的的限制，许多传统的汉字分析对比工作往往有两个缺憾：一是未能在大范围同时准确地描绘两个或多个国家/地区常用汉字的字形特征；二是未能指明两个国家/地区的汉字之间的对应关系，说明他们的编码状态、认同与否、是否存在简繁异的关系？抑或只有微小的笔形差异？这对于汉字对外教学和多元汉字文化的交流，以及汉字软件的开发与应用，都多少造成了一些困扰。

本项目是以中国《现代汉语常用字表》为基础、辅以汉语水平考试 HSK 字表，对照日韩和港台地区常用汉字或教育用字，采用 ISO 习用的汉字认同规则，进行大陆—台湾、内地—香港、中—日、中—韩的汉字对比，提供一套比较完整、准确的基础资料，作为 CJK 汉字统一编码的应用和补充，为汉字对外教学，特别是面向汉字文化圈的汉字教学提供有益的参考。

1. CJK 汉字统一编码的补充

中日韩汉字，属于同一文字体系，已是不争的事实。基于这一事实，从上个世纪九十年代初以来，经过各国文字专家与 IT 专家十余年的共同努力，本着求同存异的精神，在国际标准 ISO/IEC 10646 和工业标准 Unicode 的框架内，制定了一整套汉字认同甄别的规则，全面完成了中日韩汉字的统一编码 CJK Unified Ideographs 及 Extension A 和 B。目前，来自各个国家和地区标准的字符集、字表、字典中的逾七万汉字，已经得到了他们的国际身份证 ID 即国际标准代码。这是迄今汉字发展史上最大规模、最大范围的一次文字整理工作，也是汉字国际化的一个重要突破。随着 CJK 汉字在计算机、移动设备、因特网的实现，CJK 统一编码汉字将会日益显露其对于信息技术、语言教育、文化交流的深远影响。

上述 CJK 汉字统一编码，粗略地描述，是在一个所谓 XYZ 三维空间进行。（参见第 50 页[注 1]）这里，X 轴代表字义，Y 轴代表字形（Generic Glyph，或 Abstract Glyph），（参见第 50 页[注 2]），Z 轴代表具体的造型（文字专家习惯用“字样”和“字体”这样的术语来称各种字型）。CJK 汉字统一编码，就是把具有相同抽象字形的汉字聚集在一起赋予相同的代码；换言之，CJK 汉字的认同是在 Y 轴上进行的，而不是基于字义（X 轴）和具体造型（Z 轴）进行的。与此相关的，有两个悬而未决的问题：

(1) CJK 汉字认同之后，尚未进行简繁异体字或者新旧字形汉字在字义层次上（X 轴上）的关联标注。

(2) 大批具有微小差异的汉字被认同了，比如具有新旧字形的草字头、走之旁的汉字；

但是也有些本来应当认同的汉字，它们只在笔形上有微小的差异，或者 Z 轴上的微小变异，然而，为了兼容已有信息编码标准的需要，一个称作“源字集分离”(Source Code Separation)的例外处理规则，(参见第 51 页[注 3])使他们在国际标准中被分别编码了。比如，說一說，吳一吳，決一決，呂一呂在标准中都当作了“不同的汉字”，由于他们属于高频字，其分别编码的副作用不可小觑。

对于 CJK 汉字的应用和推广而言，这也是需要进一步完善的基础工作。换言之，迄今为止，CJK 统一编码做的是“水平认同”，而现在需要进一步做一些双语境、多语境的“垂直认同”的工作或“垂直标注”的工作。(参见第 52 页[注 4])

2. 中文汉字教学的辅助

汉字因在时空(中日韩，港澳台，教科书、报刊媒介、古籍)的不同而可能有各种不同的呈现(Presentation)。具体的汉字造型，具有很强的民族性、地域性和时代特征。而各种汉字造型往往以某一电脑字库(Font)为代表，不同字库在风格上、笔形上会有很大的差异，但都遵从各个国家/地区相应的规范。

为了高效地教好各类学生识读书写中国规范汉字，必须遵循“因材施教”的原则，按不同的学生类别对字形结构、规范笔画、规范笔顺进行教授，而不能千篇一律地照搬小学语文教材。实际上，教学对象无非是这两大类：没有汉字基础的对象(儿童，欧美学生)和有汉字基础的培训对象(日本、韩国、港澳台学生和受过港台汉字培训的外国人)。他们的基础和需求是有很大差别的，所以相应的教案也应细分。

对于有汉字基础的学生，显然没有必要一字一字地从头教起。教师的责任，首先要给他们以信心，告诉学生：他们已掌握的汉字在字形上有近一半与规范汉字是相同的，另有八分之一到四分之一仅有微小的差异，只有另外三分之一(强)是需要简化、正形或选用另一个异体字来代换。而所有的差异都有一定的规律性，不难掌握。

对初学和普通的书写，没必要苛求细微的差别；但绝不能因此而忽视汉字字形、笔形的差异。特别地，本资料所列举的大都是常用字(高频字)，对于跨语言环境的学生而言，这一点尤为重要。比如，一笔之差可以改变一个字的属性，可能使他们在电脑上不能输入自己的常用字(如說一說，吳一吳)，或者引起误解(如壳一壳)，或者在查字典时绕道其他的部首(如决一決)。当今，即使电脑的输入技术已经能相当容忍一定的书写偏差，但 OCR 或手写输入对笔形的变异和笔画的多寡还是相当敏感的。

本项目的目标，就是要方便师生们浏览和查阅这些汉字的异同，希望起到辅助的作用。初步的用户反馈是，这些资料不仅对学习中国大陆的规范汉字有益；反过来，对中国大陆的人士学习其他国家/地区的汉字文化也有帮助。

3. 中文信息技术开发的基础项目

事实上，本项目对于字库开发、多语境汉字键盘输入、手写识别、OCR 引擎开发、简繁精密转换、跨语境简繁异关联检索、辞书编纂等，也都有重要的参考价值。

本项目是由长期从事中日韩汉字统一编码工作（CJK Ideographs Unification）的国际标准化组织汉字组（IRG）前召集人张轴材提出并主持编著，在教育部语言文字信息管理司的支持与指导下，依托北京书同文数字化技术有限公司，开展中日韩大范围多边合作完成的。

二、 对比对象与对比分析流程

1. 对比对象：

(1) 现代汉语常用字表（本资料记作：CY）

《现代汉语常用字表》分常用字(2500字)和次常用字(1000字)两个部分。由中国国家语言文字工作委员会与国家教育委员会于1988年1月26日联合发布。在本资料中，V4.0版之后简称为“常用字表”，记作CY。V3.30版之前误称为小学用字，曾记作“XX”。

(2) 汉语水平考试用字（本资料记作：HSK）

国家汉语水平考试委员会办公室考试中心制定，经济科学出版社2001年6月出版的《汉语水平词汇与汉字等级大纲》（修订本）中“汉字等级大纲”（2905字）分为甲级字(800)、乙级字(804)、丙级字(590+11)和丁级字(670+30)。简称HSK（汉办版），或HSK。

在V4.0版对照表中分为甲、乙、丙、丙附录、丁、丁附录，其中丙附录标为“丙+”，丁附录标为“丁+”。

本资料V3.30版之前采用的是北京语言大学汉语水平考试中心编制、现代出版社2003年6月出版的《中国汉语水平考试大纲》（初、中等）中“HSK(初、中等)常用汉字”（2204字）。分为甲级字(798)、乙级字(808)和丙级字(598)。简称HSK（北语版）。

在V4.0版对照表中保留了“北语版”HSK的信息，甲、乙、丙、丙附录各等级字分别标为A、B、C和C+。HSK后不加版本说明时，指的是汉办版。

(3) 台湾国小用字（本资料记作：T）

《国语文教育丛书》四十三《国小学童常用字词调查报告书》2002年3月2版，编辑者：台湾地区教育部国语推行委员会(初版：2000年12月)。

(4) 香港小学用字一览表（本资料记作：H）

香港特别行政区政府教育局课程发展处中国语文教育组 2007 年编印之《香港小学学习字词表》中的《小学用字一览表》。分为第一阶段和第二阶段，记作“一阶”和“二阶”。

(5) 日本常用汉字表（本资料记作：J）

1981 年日本内阁告示颁布（1945 字）。其中：

学年别汉字配当表（=“教育汉字”）1958 年文部省“小学校学习指导要领”附录（881 字）。1968 年、1977 年、1989 年改订。

小学 1 年生	80 字	本资料记作：教—1
小学 2 年生	160 字	本资料记作：教—2
小学 3 年生	200 字	本资料记作：教—3
小学 4 年生	200 字	本资料记作：教—4
小学 5 年生	185 字	本资料记作：教—5
小学 6 年生	181 字	本资料记作：教—6
合 计	1006 字	

据查，国内常有人引用之《当用汉字表》系 1946 年日本内阁告示颁布（1850 字），但该版本字表已于 1981 年废止。

又，在本书排印过程中，2009 年 1 月 28 日，日本文部科学省文化厅文化审议会（原来是国语审议会）提出了新的《常用汉字表》，增加了挨、畏、岡等 191 个汉字，删除了銑、錘、勺、脹、匄五个汉字。鉴于该字表仍处于征求意见过程中，今后一两年内可能还有变动，因此本书尚未据此作相应调整。现将其增删部分的汉字附在下一页供读者参考。

(6) 韩国科学和技术教育部指定汉字（本资料记作：K）

1800 汉字于 2000 年 12 月颁布（其中 900 初中汉字，900 高中汉字）。

1972 年韩国亦曾颁布过 1800 汉字表。到 2000 年颁布时更换了其中 44 个字。颁布的部门也发生了变化：교육인적자원부 改为 교육과학기술부，由韩国教育和人力资源开发部改为韩国科学和技术教育部。经核对，900 初中汉字和 900 高中汉字中有一字重叠（球），故本资料 K 按 1799 字计算。

下页是 2009 年 1 月 28 日日本报纸刊登的常用汉字的增删表：

追加される漢字と削除される漢字

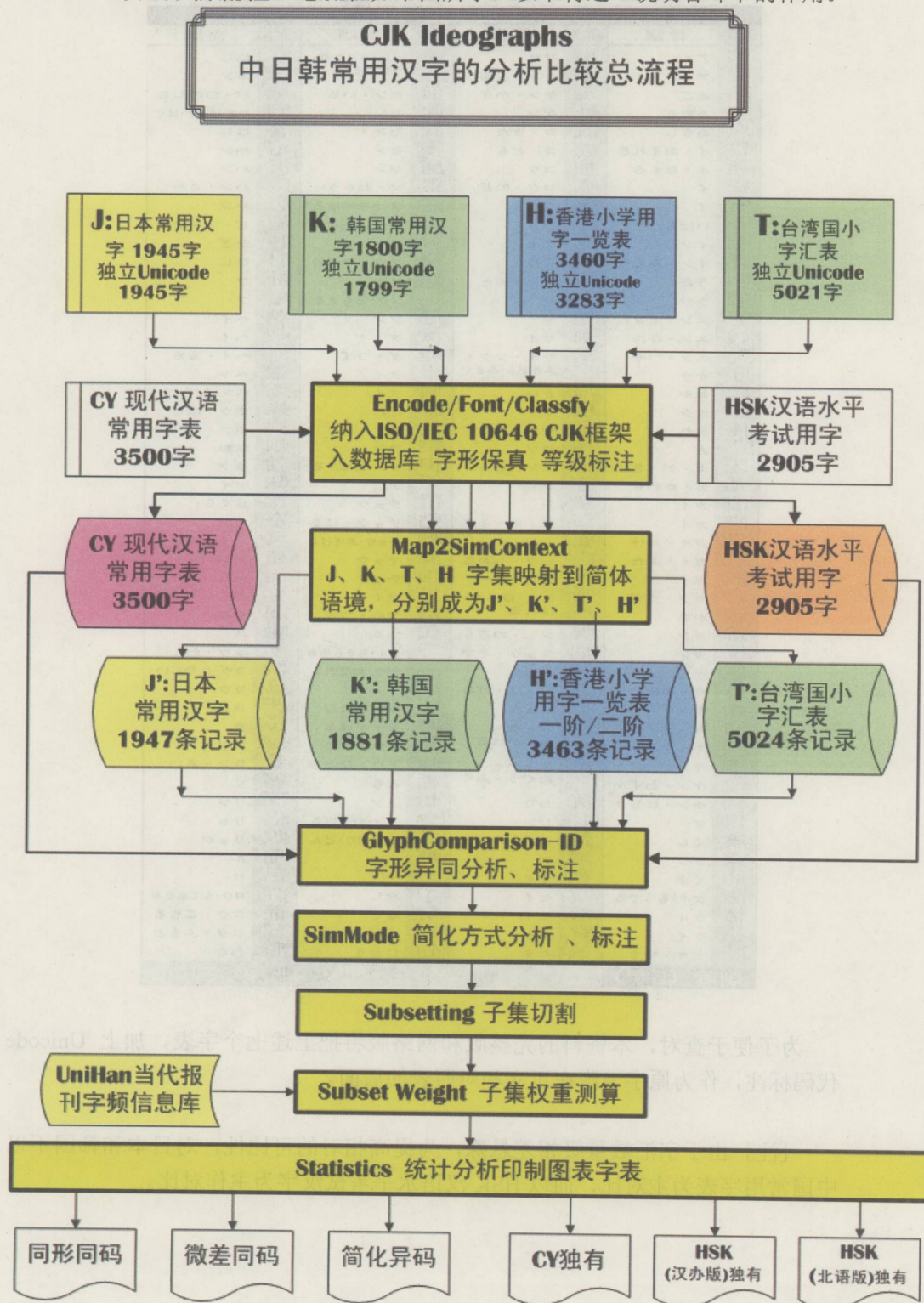
追加 (191字)

漢字	音訓	漢字	音訓	漢字	音訓	漢字	音訓
挨暖	アイ	桁拳	けた	羨	セン・うらやむ・うらやましい	虹掄	にじ
顎	アイ	ケン・こぶし		煎腺	セン・いる	鳬	ネン
宛	あご	ケン・かぎ		膳詮	セン	剝	バ・ののしる
風	あてる	ゲン		膳	セン	箸	ハク・はがす・はぐ
畏	あらし	コ・また		膳狙	セン	汜	はし
菱	イ・おそれる	コ・とら		迴	ゼン	斑	ハン
椅	イ・なえる	コウ		爽	ソ・ねらう	眉	ハン・さか
菓	イ	コウ・のど		曹	ソ・さかのぼる	膝	ハン
咽	いばら	ゴウ		瘦	ソウ・さわやか	肘	ビ・ミ・まゆ
淫	イン・みだら	ゴマ		踪	ソウ・ゾ	阜	ひざ
唄	うた	ゴロ		捉	ソウ・やせる	聿	ひじ
鬱	ウツ	コン・あと		遑	ソウ	卓	フ
怨	エン・オン	サ		汰	ソク・とらえる	聘	フ
媛	エン・ひめ	ザ		睡	ソン	敵	ヘイ
艶	エン・つや	サイ		堆	ダ	敵	ヘイ
旺	オウ	サイ・ソク・ふさがる・ふさぐ		戴	ダ・つば	蟻	ヘイ・もち
岡	おか	さい		誰	タイ	蜂	ヘキ
臆	オク	サク		旦	タイ	貌	ベツ・さげすむ
俺	おれ	サツ・セツ		綻	だれ	頰	ホウ・はち
哥	カ	サツ		緻	タン・ダン	勃	ボウ
牙	ガ・ゲ・きば	ザン・きる		耐	タン・はばかり	味	ボツ
瓦	ガ・かわら	シ		貼	チ	枕	マイ
潰	カイ	シ		嘲	チュウ	蜜	まくら
諧	カイ	ジ・え・えさ		課	チョウ・はる	寢	ミツ
崖	ガイ・がけ	しか・か		抄	チョウ・あざける	寞	メイ・ミョウ
蓋	ガイ・ふた	シツ・しかる		椎	チョウ	迺	メン
骸	ガイ	シツ		爪	チョク	閼	ヤ
葛	カツ・くず	シュ・はらす・はれる		鵲	ツイ	諭	ヤ
鏖	かま	ジュ・のろう		諦	つめ・つま	湧	ユ
韓	カン	シュウ・そで		溺	つる	妖	ユウ・わく
玩	キ	シュウ		填	テイ・あきらめる	瘍	デキ・おぼれる
伎	キ・かめ	シュウ・ける		妬	テン	沃	ヨウ・あやしい
鰐	キ	ショウ		賭	ト・ねたむ	拉	ヨウ
白	キユウ・うす	ショウ・あこがれる		藤	ト・かける	辣	ヨク
巾	キユウ・かく	ショク・ぬぐう・ふく		悻	トウ・ふじ	藍	ラッ
嘆	キン・わずか	しり		頓	ドウ・ひとみ	璃	ラン・あい
巾	キン・にしき	シン		食	とち	侶	リ
錦	グ	ジン		井	トン	僚	リッ
串	くし	ス		那	ドン・むさぼる	瞭	リョ
窟	クツ	すそ		奈	どんぶり・どん	瑠	リョウ
熊	くま	セイ		梨	ナ	瑯	ル
詣	ケイ・もうでる	セイ		謎	なし	呂	ロ
憶	ケイ	セキ		鍋	なぞ	弄	ロウ・もてあそぶ
隙	ゲキ・すき	セキ		句	なべ	寵	ロウ・こもる
		セキ			におう	麗	ロク・ふもと
						脇	わき
削除 (5字)				銑		鍾	
				勺		匄	
				服			

为了便于查对, 本资料的完整版和网络版将把上述七个字表, 加上 Unicode 代码标注, 作为原字表附在正文和对照表的后面。

[注] 由于字汇数量级相差悬殊, 为提高相对的可比性, 对日本和韩国不以中国常用字表为主对比, 而以 HSK 汉语水平考试汉字为主作对比。

2. 对比分析流程：总流程如下图所示。以下将逐一说明各环节的作用。



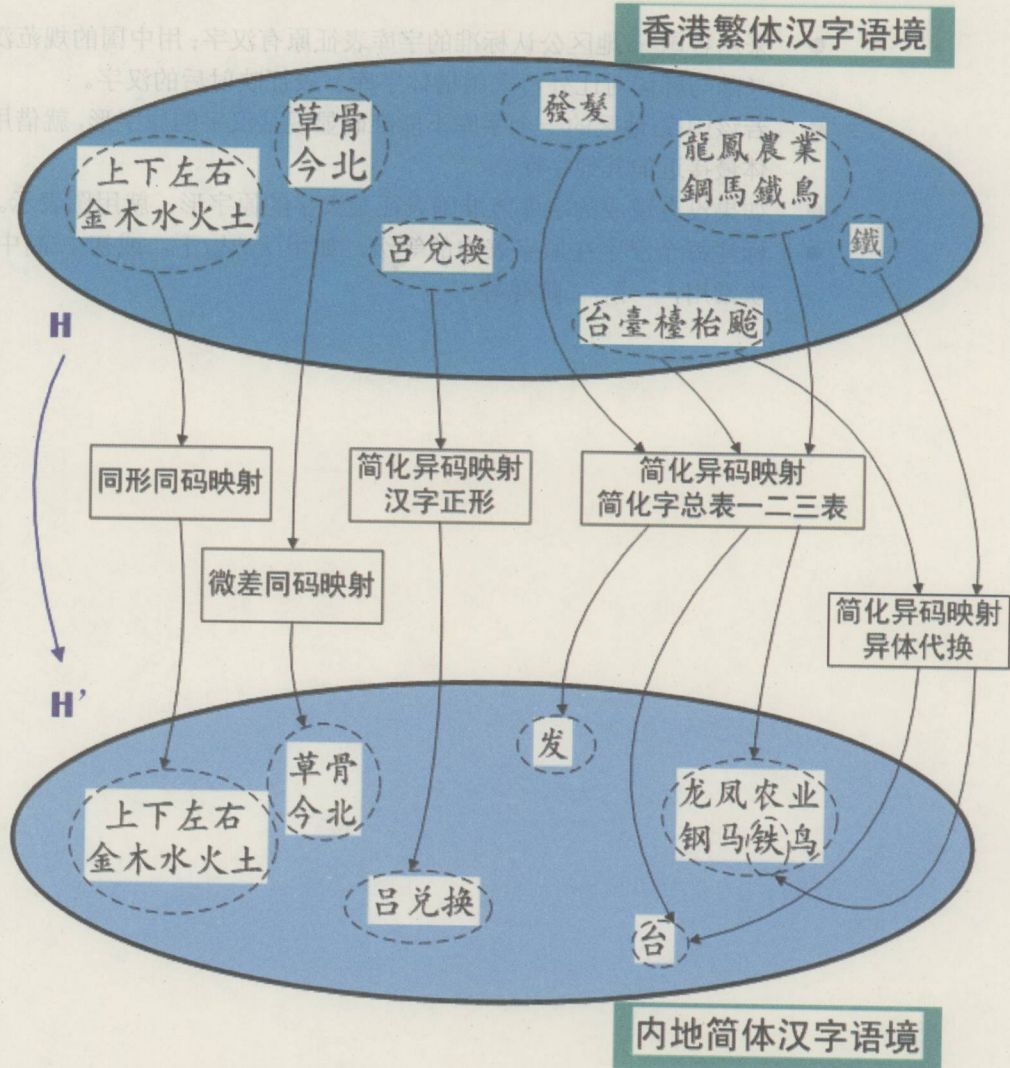
(1) Encoding/Fonting/Classify:

将各对比对象纳入ISO/IEC 10646 CJK框架，代码转换(如Big5—Unicode, TTE—Unicode, KSC — Unicode)、定码就位、进入数据库。形成J、K、T、H 字集库。在数据库中，记录每个原字表汉字的字形特征，实施“字形保真”和“等级标注”。

- 采用各国家/地区公认标准的字库表征原有汉字；用中国的规范汉字字形（除与韩国对比外皆采用楷体字库）表征映射后的汉字。
- 若该国家/地区的一个字库不能准确显现某汉字的原字形，就借用与其形体最接近的其他字库。
- 如果没有现成的字库能准确表征某汉字的原字形，就用图表示。
- 标注每个汉字在原字表中的等级，如甲/乙/丙/丁，或初中/高中，常用/次常用，一阶/二阶等等。

(2) MappingToSimContext 将J、K、T、H 字集映射到简体语境，分别成为J'、K'、T'、H'

CJK Ideographs 中日韩常用汉字的分析比较流程之一 字集向简体语境映射 以香港 H-H' 为例



从繁体语境向简体语境的映射，有的是“同码”的映射，有的则是“异码”的映射。对绝大多数映射而言，是一对一的关系；但也有多对一的情形；极少有一对多的关系。因此，映射的总结果，会使镜像字集J'、K'、T'、H'的独立Unicode编码汉字的个数略有减少。（但数据库的映射记录则会因极个别一对多关系而增加两三条记录数。）

T-T': 映射后, 集合字数从5021个收敛到4768个。

H-H': 映射后, 集合从3460(字样数)归并到3283个(字数), 最终收敛到H'的3098个(已排除重码)。

J-J': 映射后, 集合字数从1945个收敛到1920个。

K-K': 映射后, 集合字数从1799个收敛到1779个。

这一道映射程序, 是获得诸字集相对可比性的关键, 是整个对比分析工作的重要基础。

(3) Glyph Comparison-Identifying 字形异同分析、标注(Glyph-ID)。

根据汉字形态的变化, 在数据库中逐条标记同形同码、微差同码和简化异码的关系。

- 字形完全相同的, 代码不变, 标以“同形同码”记号。
- 字形有微小差异的, 代码不变, 标以“微差同码”记号。(参见第54页[注5])
- 导致转换异码的各种情形, 一律标作“简化异码”。

(4) SimMode

进一步对“简化异码”进行简化方式分析、标注其简化模式(SimMode)。实际上, 映射过程中的“简化”, 是一个广义的概念, 它包括下述各种不同的情况:

- 简化: 按《简化字总表》表一、表二、表三简化汉字。
- 异体代换: 按《异体字表》代换, 这些字可能习惯上被认为是简化汉字。如决—決, 異—异。
- 汉字正形(参见第56页[注6]): 按《印刷通用汉字字形表》正形造成的、习惯上被认作是简化字。如换—換, 爭—争, 内—內, 别—別。
- 异体简化: 既有异体代换, 又有简化的变换。如舉—舉—举。
- 条件简化: 较明显的一对多简化, 有的情况下不能简化。如乾—乾/干, 著—著/着。
- 条件异体代换: 有条件的异体代换。如珮—佩, 珮作人名时就不宜代换为佩。
- 正形简化: 汉字正形与简化交织在一起。如閱—閱—阅。
- 日文略字对应: 日文略字(简化字)对应中国的简化字, 如团—団; 也可能对应一个传承字, 如巢—巢。
- 日文国字对应: 日文国字(除略字外, 日文独有的汉字)对应一个 CJK C-Hanzi G 列下的一个编码汉字, 如桦; 还有的按字义亦可对应到一个简体字。如闘—斗, 働—动。
- 其他: 如可简化但尚未编码。

[注] 在图形坐标上, 以上分类分别缩写为: 简化, 条简, 正简, 异简, 异代, 条异, 正形, 日略, 国字。