第五届两岸四地中文数字化论坛发言

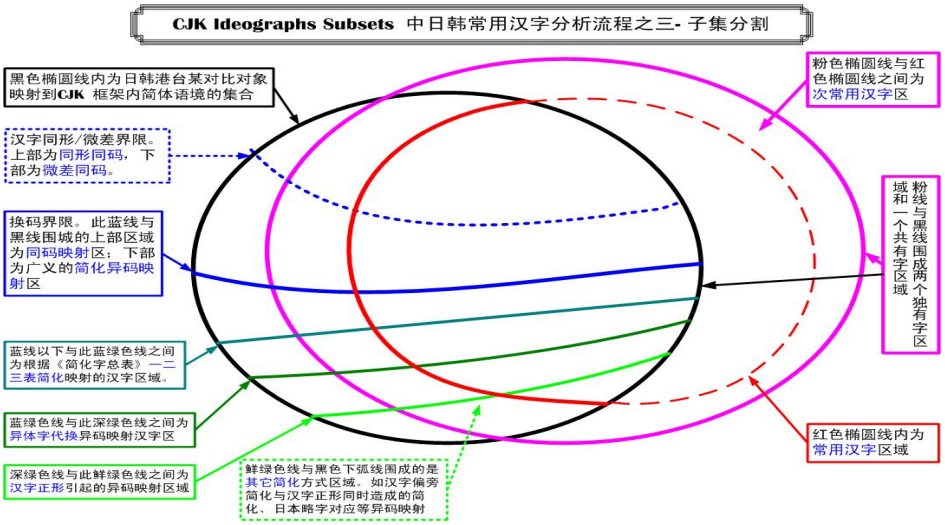
中文数字化，勿以初级而不为

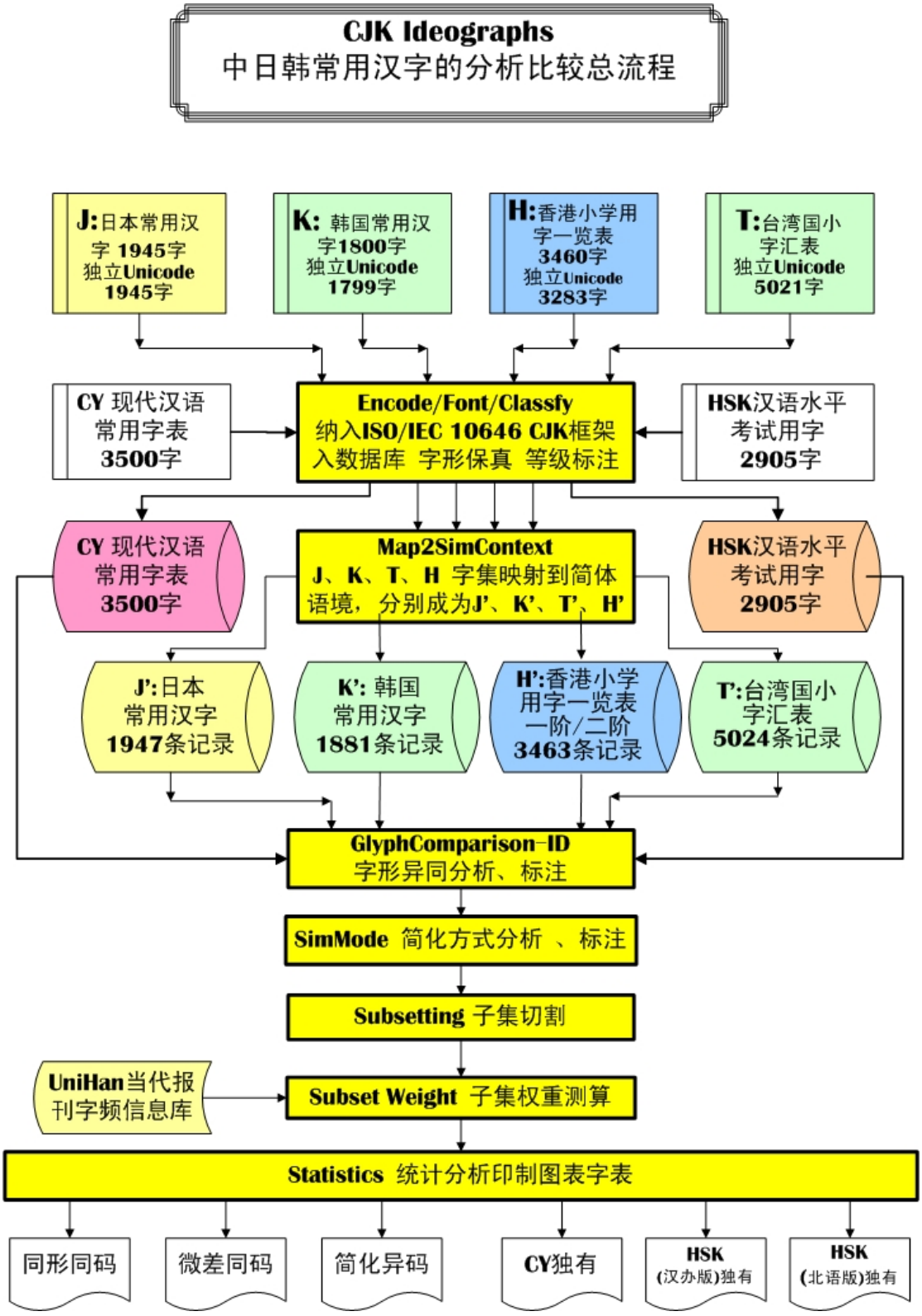
张轴材 北京书同文数字化技术有限公司
joe.zhang@unihan.com.cn

2007 年初第四届两岸四地中文数字化论坛（澳门）之后，我和我的团队，在包括两岸四地在内的众多朋友的帮助下，在中文数字化特别是汉字数字化的几个初级的、基础性的项目上做了一些工作。现汇报交流如下：

一、汉字基础项目

a) 中日韩常用汉字对比分析：该项目从汉字文化圈社会生活与教育界最常用的汉字入手，选中国大陆的《现代汉语常用字表》、《HSK 汉语水平考试字表》、中国台湾地区《国小字表》、中国香港特别行政区《小学生用字一览表》、日本《当用汉字表》和韩国文部省指定的汉字表为对比对象。将它们纳入 ISO/IEC 10646 的国际标准的框架下的数据库，分别进行覆盖率统计与字形异同的对比分析，并按照汉字简化、正形、异体代换映射的结果，分列出清晰的“同形同码”、“微差同码”和“简化异码”等多组对比字表，及其诸子集的频率权重统计图表。《中日韩常用汉字对比分析》作为一项处于语言文字与信息技术边缘的研究项目，由张轴材先生主持，依托北京书同文数字化技术有限公司，充分利用数字化技术，开展中日韩和港台地区的多边合作，“取得了具有实用价值的成果”。日前该项目通过了教育部语言文字信息管理司主持召开的鉴定会，获得了来自北京大学、北京师范大学和商务印书馆等教育出版界的专家的好评。根据鉴定组专家也提出的意见与建议，该项目组利用书同文公司的数字化资源进一步开发，完成了从 V3.30 版到 V4.0 版的大规模更新。目前已决定纳入《国家语言生活绿皮书》由商务印书馆出版。 见“CJK 求同询异”http://hanzi.unihan.com.cn/CoolHanzi/





b) 整合 CJK 汉字构件集：

目前业已成为国际标准或国家规范的各汉字基本笔画、部首、构件，之间，存在着复杂的关系，有的相互重叠，可能具有同形异码/异名，异形同名。本项目的人物就是梳理这些基本的构件，形成他们的并集，建

立各元素之间的映射关系(Mapping)。供从事汉字键盘输入、手写辨识、汉字教学、汉字字库等研究开发及标准化的工作者参考。

在该项目中，把笔画、部首、部件、构件，都视为广义的“构件”。他们的并集称作 CJK Component Set。广义的汉字构件集包括：

A. CJK Strokes in Unicode 5.0
B. CJK Radicals in Unicode 4.0
C. Kangxi Radicals in Unicode 4.0
D. GF3001-1997 构件集

文件可下载：http://hanzi.unihan.com.cn/CoolHanzi/#down_paper

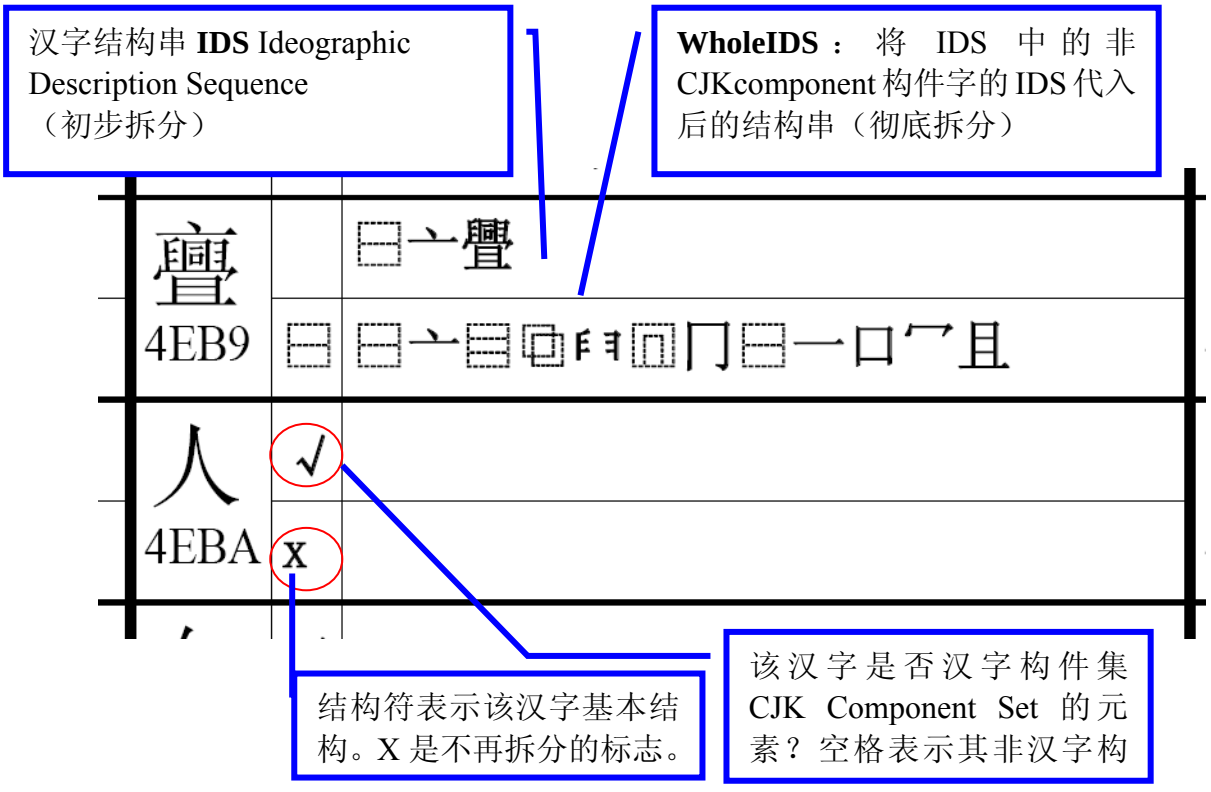
c） CJK 拆分序列 IDS：CJKDecomposed（文件可下载）

http://hanzi.unihan.com.cn/CoolHanzi/#down_paper

此前虽然有多人做过类似的项目，但遗憾的是一般都不提供公开的电子文档。同时，由于三个原因，汉字拆分的结果也往往有所差异：第一，依据的汉字构件集不同；第二，拆分规则不同；第三，一般电脑上的构件缺字编码不同（EUDC）。

本项目的主要目的，是在广义的汉字构件集和一个明确的拆分规则的基础上，提供一个汉字描述序列(IDS)的讨论文本。

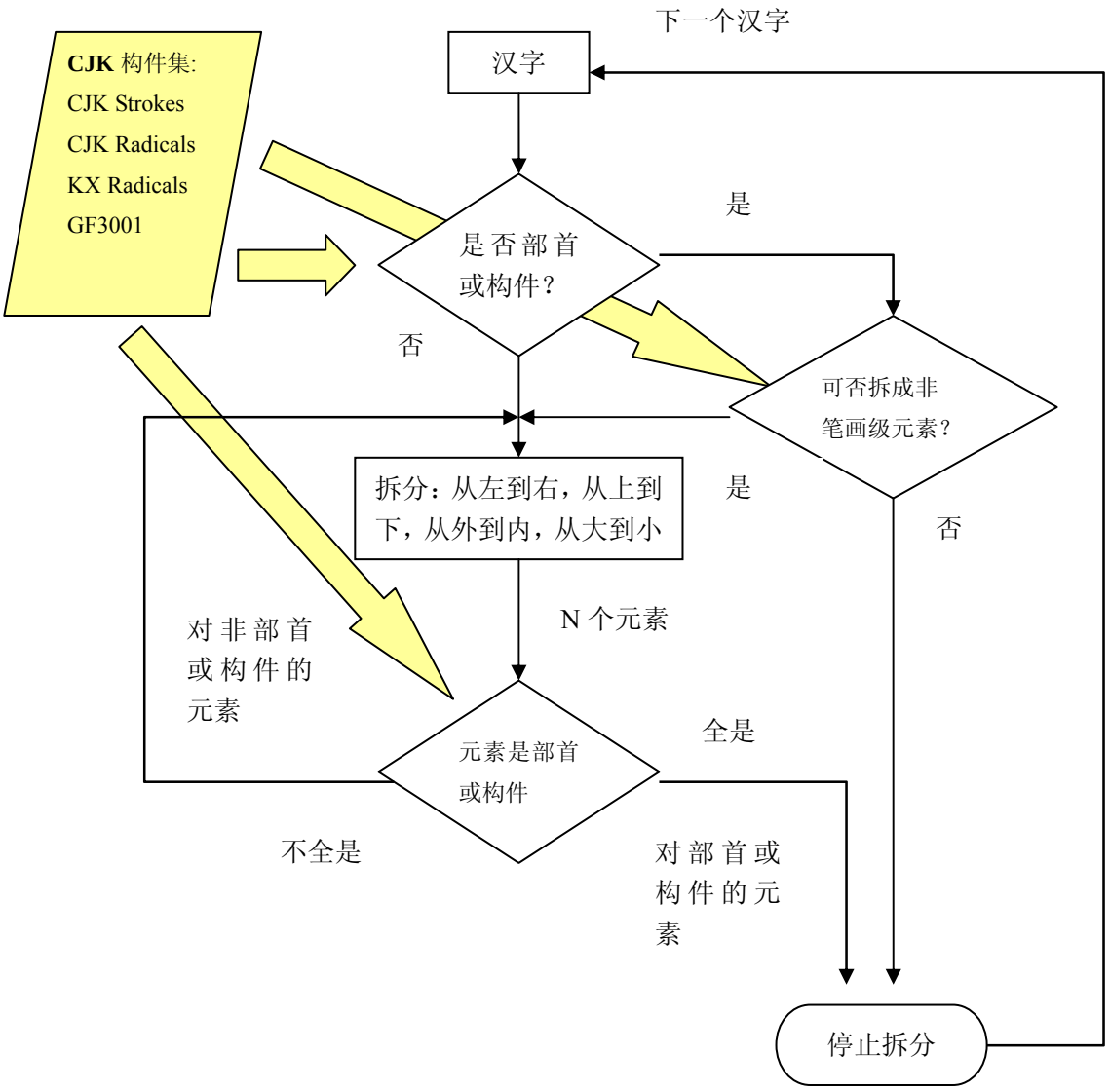
打印件凡例



. **CJKdecomposed**（1st Depth Decomposed），遵循下列拆分规则:

<Decomposing Rule>

A. If an element /component has no meaning, and ALL basic components composing it have no meaning, either, then stop decomposing the element.
B. If an element /component has no meaning, but SOME basic components composing it have meaning, then decompose the element..
C. If an element/ component has meaning, then stop decomposing it.
D. If an element/ component is a CJK Radical or Kangxi Radical., then stop decomposing.
E. If a Hanzi itself is a CJK Radical or Kangxi Radical, and more than one basic components defined by GF3001 compose the radical, then decompose the Hanzi.



二、汉字基础信息的网络服务- 书同文汉字网 http://hanzi.unihan.com.cn

a) 笔顺规则多文种版： 中日韩英法俄德葡
http://www.unihan.com.cn/coolhanzi/whr/#SR
b) 古籍字频网上查询： 基于《四部丛刊》《四库全书》的字频、覆盖率查询工具 http://hanzi.unihan.com.cn/Tools/Frequency/ 印刷本已作为国家语言生活绿皮书由商务印书馆出版。
c) 笔顺笔势动态图与跟随式笔顺生成器
d) 书同文 Web 工具 http://hanzi.unihan.com.cn/Tools/

- 拼音加调 给拼音加上音调字符，如将 "ni3 hao3" 转化为 "nǐ hǎo"
- 中西历转换 中国历朝纪年（年号）和公元纪年的相互转换，如〖唐〗太宗『貞觀』：公元 627-公元 649
- 干支纪年转换 传统的天干地支纪年和公元纪年的相互转换，如公元 2008 年【干支纪年：戊子年（鼠年）】
- 简繁转换 简体汉字和繁体汉字的转换，支持大篇幅文字格式无损转换
- 手写输入 在线手写输入模块，使用鼠标绘制轨迹，具有支持范围大，识别率高，反应速度快等特点

三、汉字学习与手机彩书的结合，以达到寓教于乐的目的。

http://caishu.unihan.com.cn

四、中文数字化技术，“云计算”小体验（另文）

a) 基于服务器的手写识别 Web 版-“云计算”的一种尝试
http://www.nciku.com
b) 手机汉字输入，字频词频的动态更新：基于 Android 平台的相关开发中的新体会。
c) 书同文汉字网中，多种字体的 On-Line Conversion

五、鸣谢

在上述项目进行中，多次得到了李宇明、王铁琨、傅永和、李大遂、松冈荣志、木村守、许其清、王晓明、韩秀英、金镇容、Dr. Jay Hyun、陈敏等各位专家的具体帮助和热情鼓励。在此我谨代表全体项目组成员表示衷心的感谢。

张轴材 2008-10-07
MSN: joezhang43@hotmail.com
北京朝阳区风林绿洲 F06-29C 100101