探讨古汉字使用IVS 编码的可行性

香港理工大学计算学系
陆勤

Ideographic Variation Sequence
异体字序列表示法简介

- 基本字：已编码的汉字
- 异体字选择符（variation selector，异选符）：
  U+E0100 至 U+E1EF(共240个)
- 使用方法：IVS
  IVS: 基本字+异选符
  IVS: => 对应一个字形，须经注册 => 异体字序列库
- 特点：
  - 与基本字的关联性
  - Combining sequence处理绝大部分系统都可以处理
  - 例： :) => ☺
  - 一个基本字可有最多240个异体字
  - 从“无序”编码到有关联的“次有序”编码

2

已有的IVS定义

- 日文: Adobe Japan-1 IVD
  - 开发到审定: 2006.12 - 2007.12
  - 字符总数: 14,645 日文汉字
- 使用方法
  - 用 'cmap' 子表— 14 表
  - 用两个 Unicode 字符 映射到一个字形标示符 GID (*Glyph ID*)
  - 基本字+ 异选符 = GID
  - 例: 辻(U-8FBB), U-8FBB + U-E0101 => 

3

已支持IVS的应用

- Adobe Acrobat 9.0版—PDF 文件
  - 以浮动板的形式提供不同异体字的选择
- Adobe Flash Player 10版
  - 通过 *flash.text.engine* 应用界面提供选项
- JustSystems(日本公司) 输入法提供异体字的选择

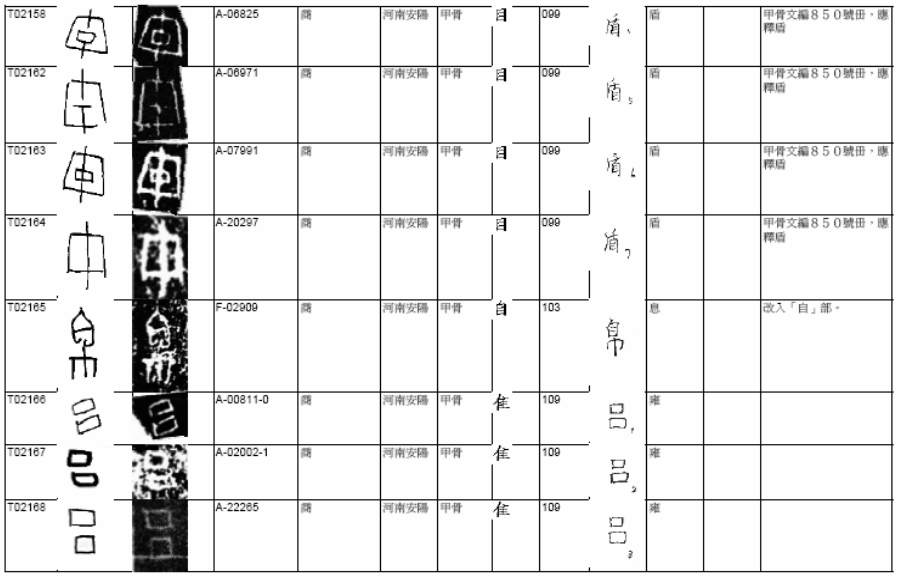
4

古汉字的特点(甲骨文)

- 异体字形多
- 异体字形需要得到保留
  - 至少不能用CJK的认同规则
  - 字形的研究,表达是甲骨文工作的重要部分
- 使用CJK的编码形式的优点和局限
  - 异体的关联需用附加的表提供
  - 需自行开发软件支持异体字的查找
  - 如编码已有某种顺序,新增符不可能尊重原序
  - 但有灵活性:一个字形可以和不同的字发生关联

5

古汉字例(IRGN1433)



T02158		A-06825	商	河南安陽	甲骨	目	099	盾、	盾	甲骨文編８５０號田，應釋盾
T02162		A-06971	商	河南安陽	甲骨	目	099	盾₅	盾	甲骨文編８５０號田，應釋盾
T02163		A-07991	商	河南安陽	甲骨	目	099	盾₄	盾	甲骨文編８５０號田，應釋盾
T02164		A-20297	商	河南安陽	甲骨	目	099	盾₇	盾	甲骨文編８５０號田，應釋盾
T02165		F-02909	商	河南安陽	甲骨	自	103	帛	帛	改入「自」部。
T02166		A-00811-0	商	河南安陽	甲骨	隹	109	品，	雕	
T02167		A-02002-1	商	河南安陽	甲骨	隹	109	品₂	雕	
T02168		A-22265	商	河南安陽	甲骨	隹	109	品₃	雕	

6

甲骨文用IVS的可行性

- 前提
  - 绝大部分异体字形组不超过240个
  - 绝大部分甲骨文工作者都有认同基本字的概念
    - 字形如与多个基本字可认同,可有一个稳定的主基本字
- 使用IVS的好处
  - 与现时甲骨文整合工作有一致性
  - 编码速度快,灵活性高
  - 对基本字的关联在编码层次上实现,有利不同应用程序的开发
  - 增加新字对排序的影响最小

7

可预见的问题

- 对不确定关联字的字形如何编码
  - 作为另类,分开编码
    - 事后关联的建立只能通过自定义表格来实现
  - 推迟编码指导可确定关联(IRG现时的做法)
- 结论:IVS是一个值得考虑的不错的选择

8