漢字智慧型編碼與應用研討會

缺字問題的回顧與前瞻

謝清俊

中華民國九十二年三月十七日

講 述 大 綱

- 缺字問題綜述
- 智慧型漢字編碼
  - 漢字構形系統概述
- 目前的成就
  - 本研討會的內容介紹
- 未來的課題
- 結語

缺字問題綜述 (續)

- 缺字是由於漢字交換碼中的字形不足引起的。
  - 工業標準的限制
  - 削足適履—錯誤引用西方文字的編碼模式
  - 歷史上字形的嬗變的問題
    - 不宜將數千年使用的各式各樣字形擠壓在一個字表中呈現
  - 交換碼設計者的自以為是
    - 不曾尋求文字學者與使用者的參與設計

缺字問題綜述 (續)

- 解決缺字問題的根本，在於解決現行漢字交換碼的根本缺失
- 我們解決缺字問題的方法，是遵從漢字構形的原理，對漢字字形的結構做制式表達與詳細的分析。一個字的字形結構式，是該字極佳的識別符號與工具；因為字形若不一樣，則字形結構必不相同；反之，字形結構若相同，其形亦必同。

缺字問題綜述 (續)

- 漢字的字集是一個開放性質的；也就是說，漢字的字集依古今的變異、專業與應用環境的差異等而有字數、字形、以及字音、字義上的變化。音義且不談，僅就字數而言，即已不適合作固定數量的限定；這與數量已定的西方語言的『字母集』，是不可以一概而論的。
- 然而，現行漢字交換碼的結構，卻仿照西方語言的字母集的結構來設計，這不能不說是『削足適履』。

缺字問題綜述 (續)

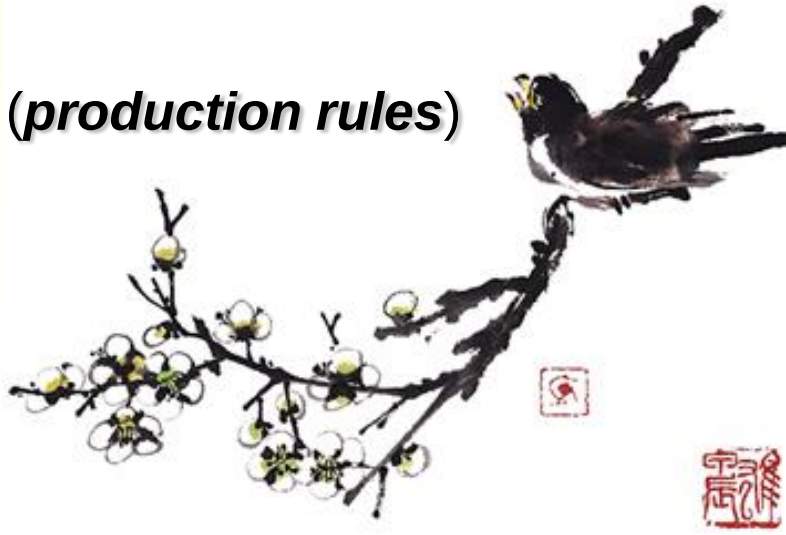
- 再者，語言是有生命的，任何活在當下的語言，都需要有創作新詞（word）的能力，否則必定無法配合社會變遷之需，而終將面臨被淘汰的命運。我們不時看見有英文新詞（字），就是個很好的例子。可是，在電腦中限定漢字的字數，就斬斷了漢語言電腦中創作新字（詞）的生命力，埋下了日後不能適用的因子。這個問題也不應等閒視之。
  - **智慧型漢字編碼系統也解決了這個問題**

缺字衍生的問題

- 大幅增加了資料登錄的工作
- 造字的管理不易
- 造字的空間不足
- 異體字造成文件檢索和處理上的困撓
- 造成資訊共享的障礙
- 降低了數位化文獻的真實度

漢字構形資料庫

- 漢字構形資料庫是一個表達漢字結構的制式系統
- 漢字構形資料庫的構成有：
  - 一群與字形相關的集合
    - **一些字集【 字集的數目不限 】**
    - **一個字根集【 共用 】**
    - **一個部件集【 共用 】**
  - 運算子(***operators***) 和運算規則(***production rules***)
  - 一個字形結構式的集合
    - **對應於每一個字形有一個以上的字形結構式。**



字形與構字式

- 目前，漢字構形資料庫中有 59,766 個楷書字形，乃下列字集中字形之聯集：
  - 《中文電腦基本用字》 8,528字形
  - 《五大字集》 13,053字形
  - 《中文大辭典》 49,416字形
  - 《漢語大字典》含簡化字總表 54,640字形
  - 《中央研究院補字集》 8,028字形

字形與構字式

除 **Window** 可提供的字型外，本系統尚有：

- 漢語大字典 **54,640** 個仿宋字形
- 漢語大字典異體字**12,208** 組，約**36,309**字形
- 說文小篆：
  - 《說文大字典》**11,100**個字形
- 金文
  - 預計本年度內完成

字的家譜與字根

- 構形資料庫中的 **59,766** 個字形，依其構形劃分為 **1,324** 個家族，每個家族為一樹狀結構，家族的領頭字即字根。
  - 此**1,324** 個字根又分為兩組：
    - 普及組有**801** 個字根，可處理**57,626** 個字的構形。
    - 罕用組有**523** 個字根，只處理**2,140** 個罕用字的構形。這些罕用字都是些楷化的古字，如楷化的金文。
  - 我們推薦流通的版本用普及組字根。

實 施

- 以『構字式』作為字形的識別碼
- 三個運算子：
  - 橫連、直連、包含
- 單運算子表達，再生式運作(***Recursion***)
- 以漢字構形資料庫管理字形知識
- 提供字形與構字式之對映

部件構字式、字根構字式與字根式

- 灕＝□□離
- 灕＝□□(离□隹)
- 灕＝□□((□□内)□(□□□))
- 灕＝□□(((□□凶)□内)□(□□(□□□)))
- 灕＝□□(((□□(山□□))□内)□(□□(□□□)))

- 灕＝□□□山内□□□

單運算部件構字式之例

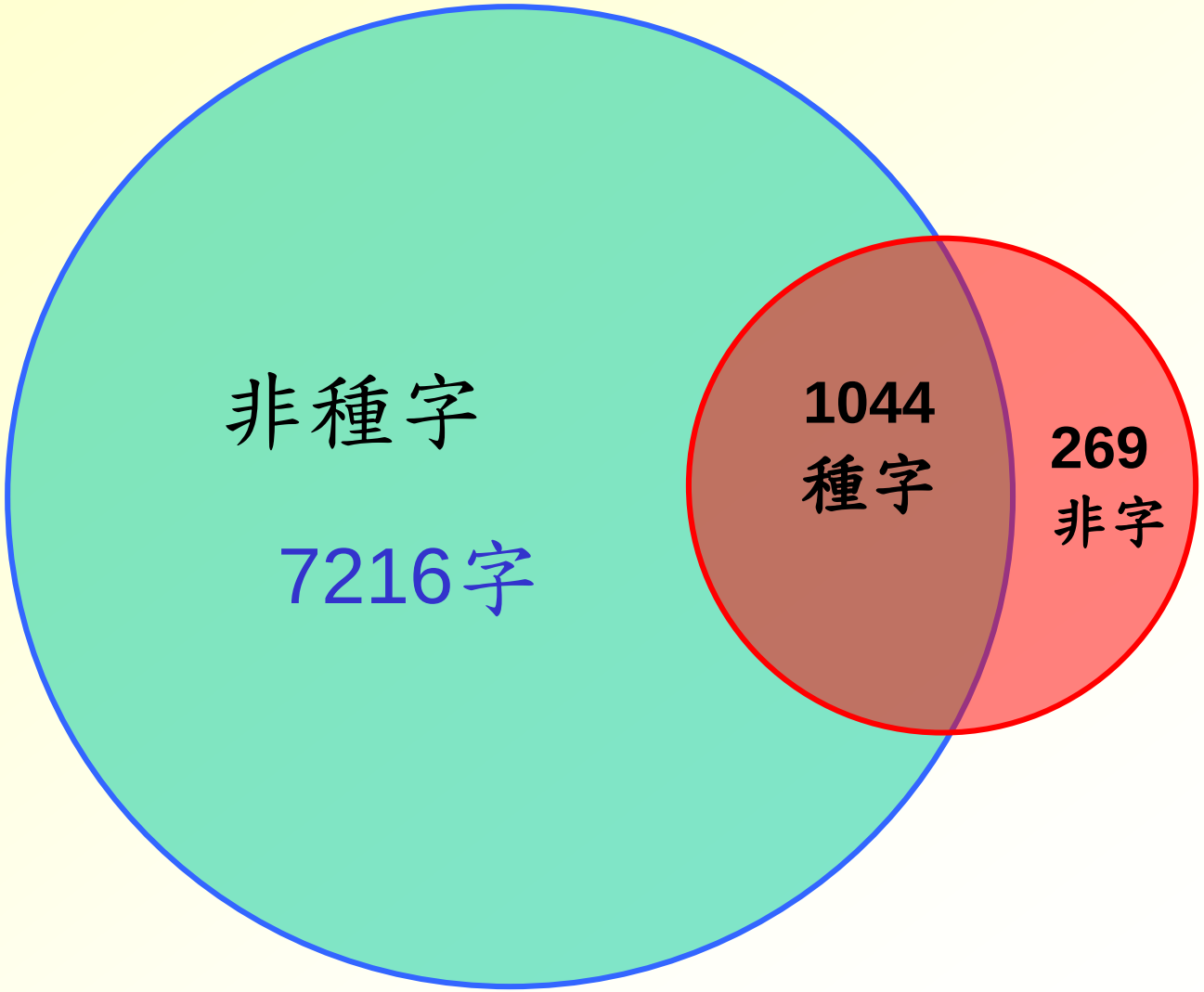
- 灕＝□□離
- 離＝离□隹
- 璃＝王□离
- 擒＝□□禽
- 噙＝口□禽
- 离＝□□凶
- 禽＝□□离

- 隹＝□□□
- □＝□□□
- □＝□□凶
- 凶＝凵□□

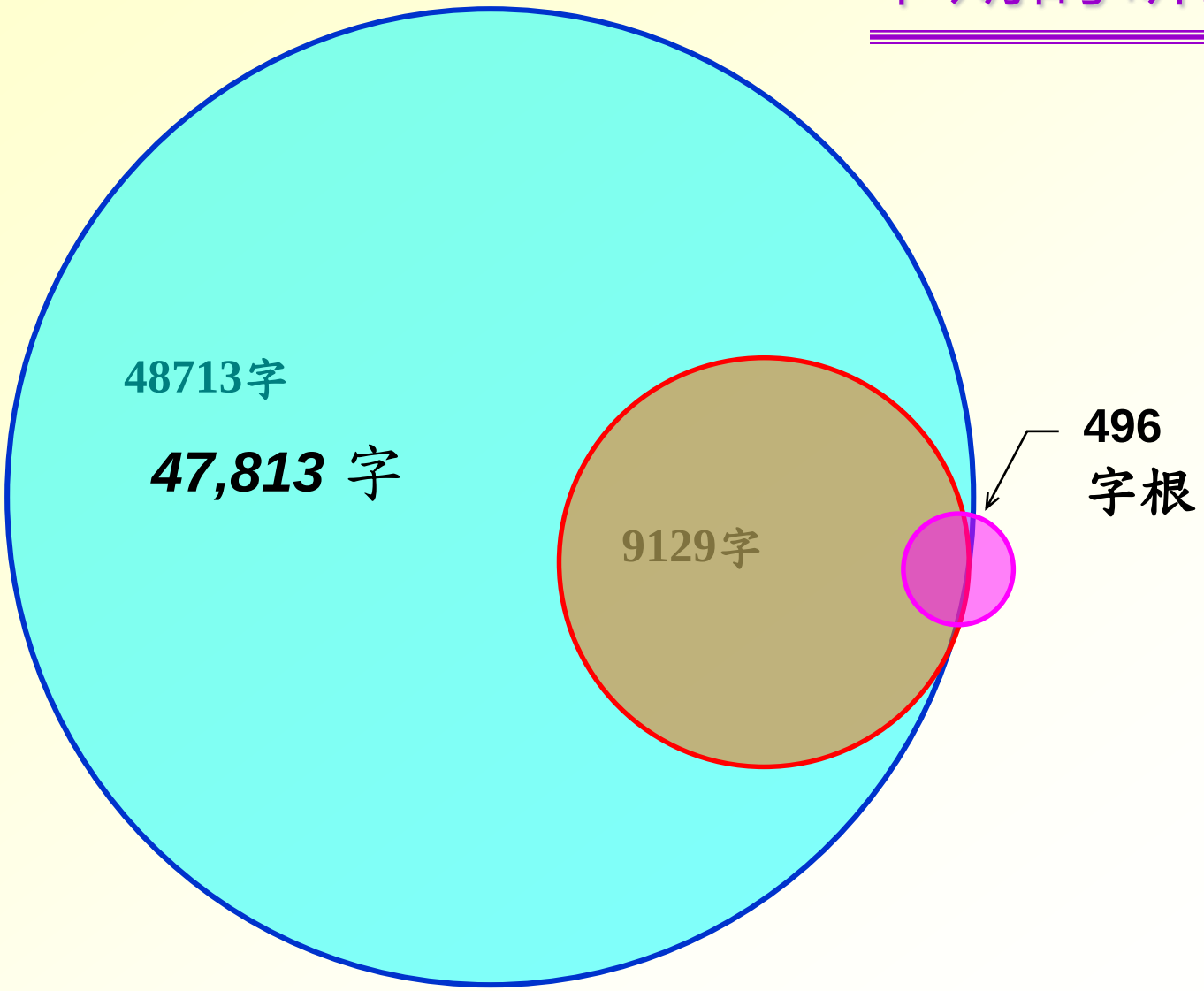
種字與終端字

- 在每一棵漢字家族樹中，有兩種字：
  - 一種是沒有孳生能力的，稱為『**終端字**』
    - 終端字位於家族樹的端末結點上。
  - 一種是有孳生能力的，稱為『**種字**』
    - 種字位於家族樹的非端末結點上。
- 種字的字形必定會出現在其他字形中，作為其字形的一部份
  - 終端字則無此性質
- 本系統的種字共 ***3361*** 個

林樹字集的種字 (1994年)



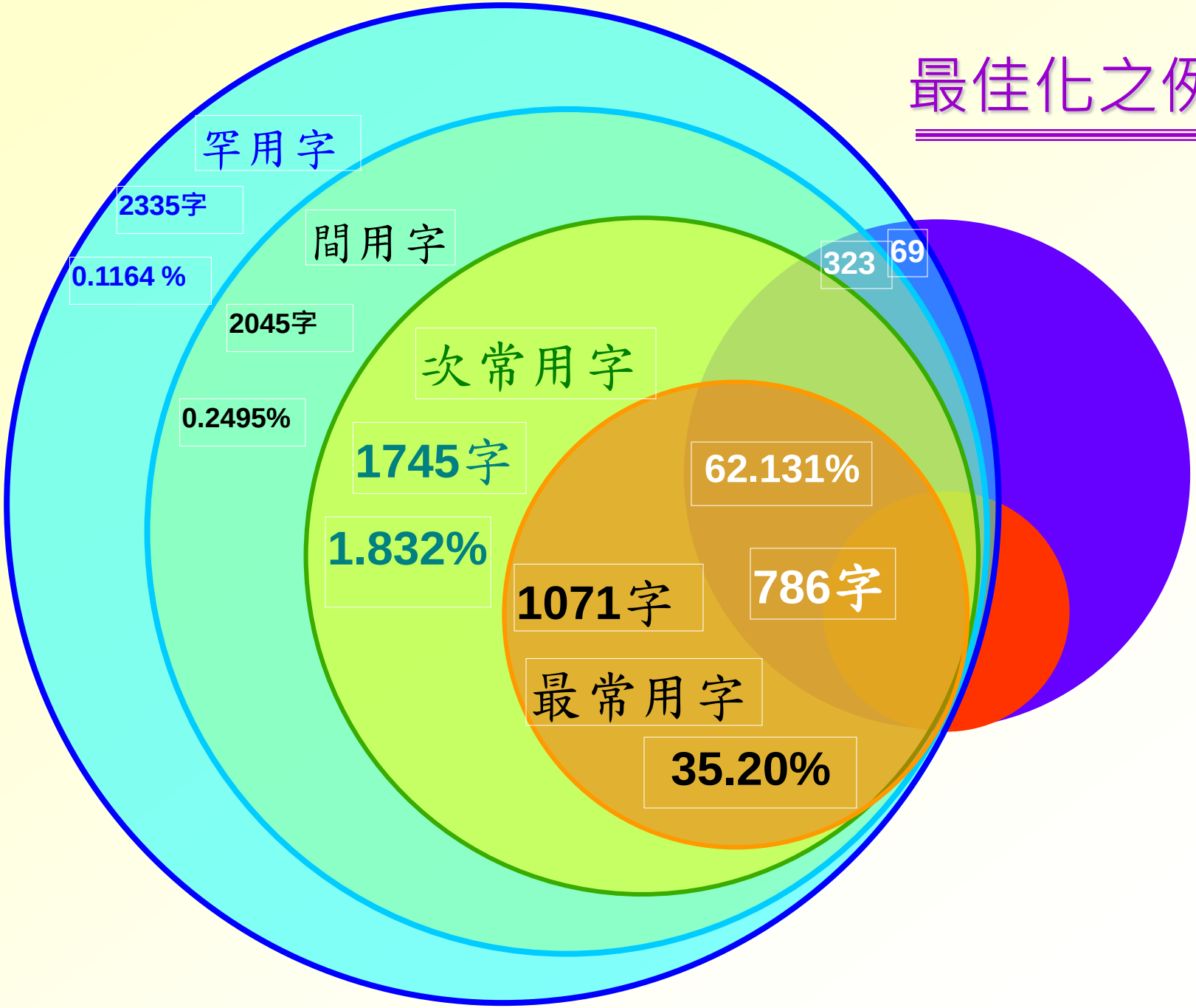
早期的研究



最佳化

- 將常用字納入
  - 約 2800 字,佔37% 的使用率
- 納入常用字後的系統
  - 約用 5200 個碼位
  - 直接用此碼的機率大於99% .
- *少於1% 的機會要用部件式來表示其他的字或新字.*

最佳化之例


罕用字
2335字
0.1164 %
間用字
2045字
0.2495%
次常用字
1745字
1.832%
323
69
62.131%
1071字
786字
最常用字
35.20%

字、部件、字根 與 其結構示意圖

字數

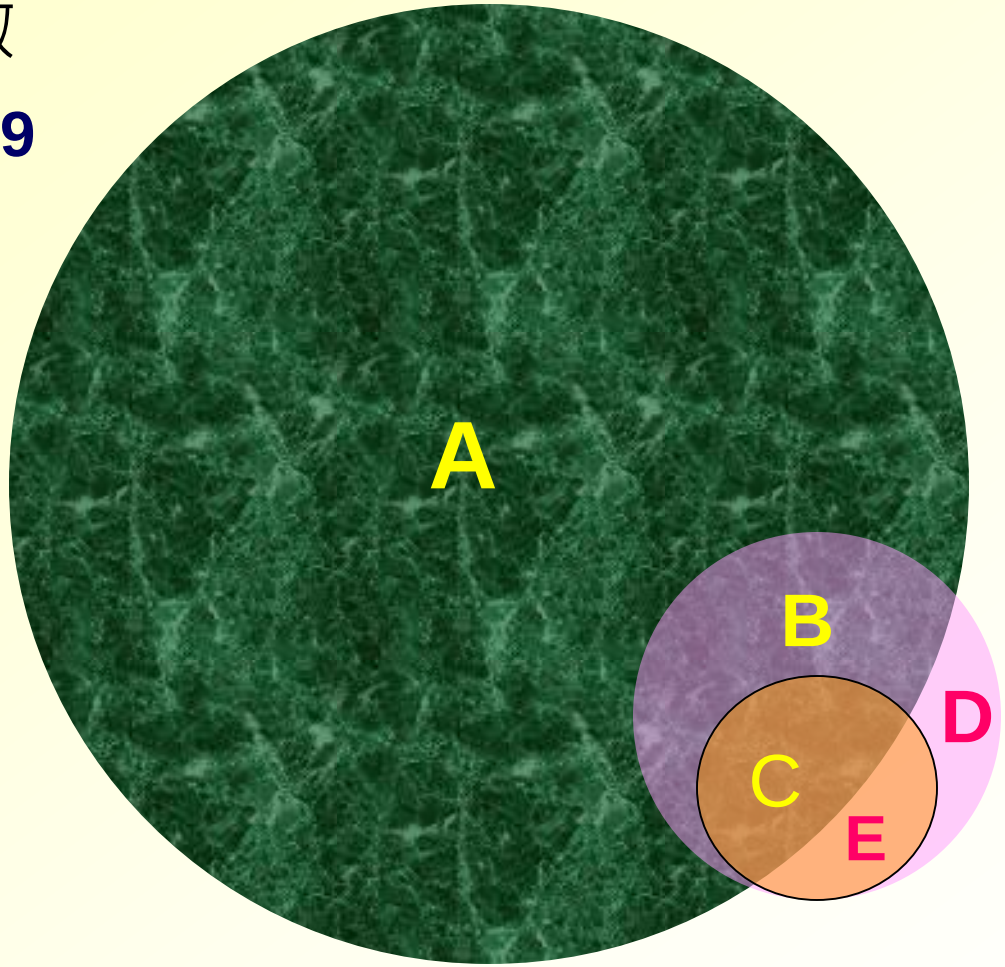
A: 51249

B: 3061

C: 330

D: 381

E: 994



漢語大字典

54640 字

A+B+C

部件集

4766 個

B+C+D+E

字根集

1324 個

C+E

D+E = 1375

個

字、部件、字根 與 其結構示意圖

字數

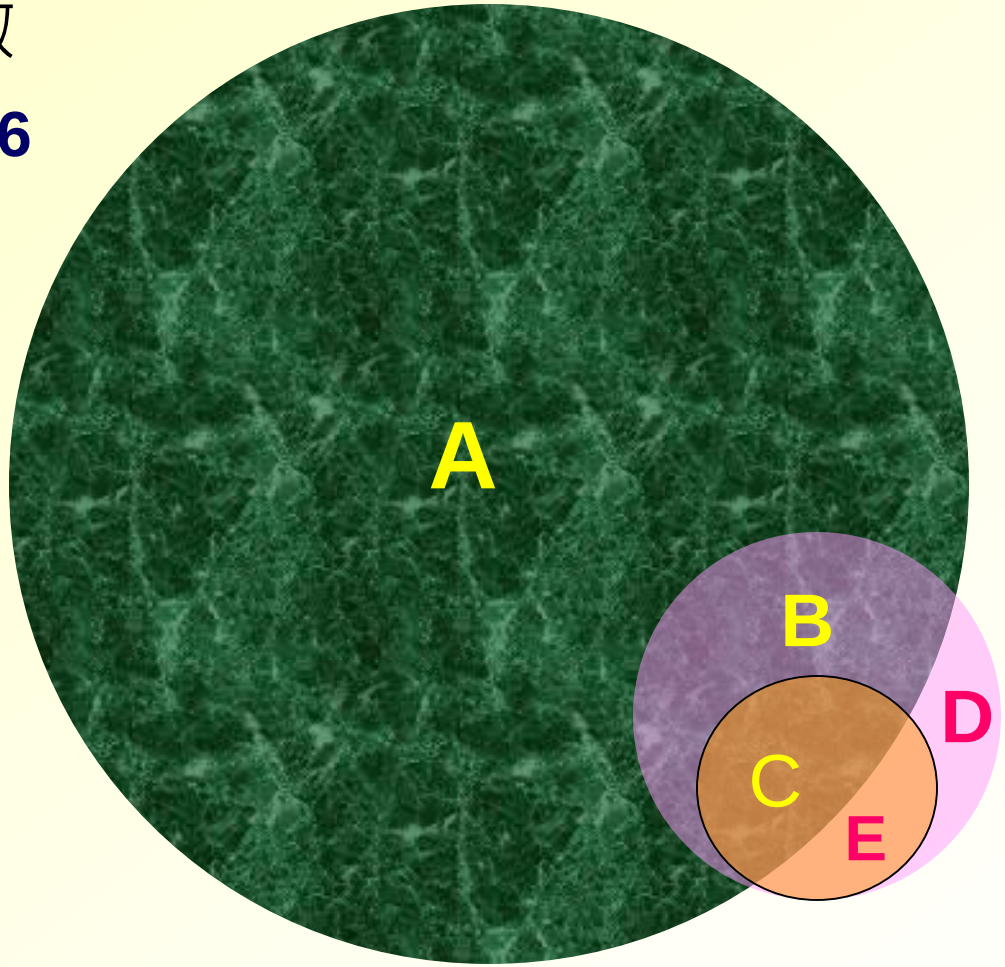
A: 11236

B: 1446

C: 371

D: 353

E: 430



BIG-5 碼
***13053* 字**
A+B+C

部件集
***2600* 個**
B+C+D+E

字根集
***801* 個**
C+E

D+E = 783 個

漢字的通用構字模式

- 以上漢字的構字模式是通用的，小篆、金文、甲骨文等亦可適用。
  - 小篆已建構完成，金文在建構中。
- 由於古今文字均採用相同的結構模式，是故古今文字的銜接就自然、容易多了。
- 此模式與任何一種交換碼細系統皆相容
  - 可應用於**Unicode**、**JIS**、**GB** 、**CCCII**等

此模式應可處理漢字生成後在時空上字形的變化

展　望

- 本研討會各報告已達到的新功能
  - 處理筆劃變異的異體字
  - 字形的自動產生器
  - 各個平臺缺字交換的標準
  - 世界通行標準的申請
- 以上功能的整合將使得智慧型漢字編碼系統更趨完善。

展　望

- 未來智慧型漢字編碼系統可能的發展
  - 作為發展數位化文字學知識庫的基礎
  - 促使歷史文字學的誕生與發展
    - 整合字形和字碼，建立漢字 ***Font*** 的時空結構
    - 在時空結構下，建立斷代的漢字語料
    - 有系統地整理斷代字形
    - 作為歷史語言學的基礎

展　望

- 文字語意學的發展
  - 以歷史文字學為基礎，解決異體字與文字語意的問題

結　語

- 一是古今字形與古今構字的銜接和對映
- 二是異體字形的表達和處理
- 使用者可以從現代字形直接聯繫到小篆與楷阿化的小篆字形，也可由小篆的構形直接查核現代字形。
- 使用者可不必具有任何文字學的知識，即可查閱到相當多古文字相關的知識和資料。

結　語

- 由於本系統的設計理念與目前的交換碼完全不同，所以本系統可以與使用任何交換碼的系統相容。換言之，任何亞東文字的處理系統，無論是中日韓越、無論是簡繁，都可以附加本系統作為澈底解決缺字的機制。

結　語

- 目前我們推出的是基於五大碼（**Big-5**）的系統，可將五大碼能處理的約一萬三千字形，立刻擴充到六萬以上，並擴及小篆和異體字等。更重要的是，這能力的擴大，並不需要佔用原交換碼的編碼空間，也不需要修改原來的應用軟體

報告完畢

謝謝

Thank you for your attention.