Here's a so-called Euler-Venn-Diagram that shows (1) how many CJK character (a.k.a. ideograph, sinograph, tetragraph, ...) codepoints in Unicode are considered 'important for daily use' (just under 10,000); (2) how many of those are used in each of the three important usage areas (i.e. © for Mainland China (PRC without the Hong Kong and Macau SARs), ① for Japan, and ① for Taiwan, Hong Kong, and Macau); and (3) how much overlap there is for each of these regions. The bars made up from sample characters in the lower part visualize the same data, but this time with bar heights instead of circle areas being proportional to numbers.

Just in order to stress it, a 'character' in this chart is equivalent to 'a Unicode codepoint', so for example 馬 and 马 count as two characters, and 關, 关, 関, 關, 關 count as five characters. Dictionaries will list 馬马 as 'one character with two variants' and 關关関關關 as 'one character with five variants', but that's not what we're counting here.

You can see at a glance that

- \* around 25%—one out of four—of all characters are common to all three regions,
- \* around 17%—one out of six—of all characters are unique to the PRC and Taiwan, respectively, while only 3% are unique to Japan,
- \* only 1% of all characters are uniquely shared by the PRC and Japan, the most frequent ones being 会来国学当内数声写将号参区. But observe that usage frequencies usually follow Zipf's Law (quote: "the frequency of any word is inversely proportional to its rank in the frequency table"); because of this, any one character near the top of a frequency ranking will appear vastly more often than any one near the bottom of the same list. As a result, although the outcomes of the PRC's and Japan's post-war efforts to simplify characters share but one out of a hundred character shapes, those shapes are still exceedingly often seen in daily life. Of the three regions—©, ①, ①—the Tregion is the one with the most balanced shared usages, the other two regions showing greater relative differences between their subsets shared with zero, one or two of the other regions. In a sense, therefore, the T region may be considered to be the 'most representative' or 'most average' one in terms of character usage.

这们说时为过对还发开经现样动从间长话实头问进车业两给电关见门语让场东别题书记觉师爱应报员边论张总处产认结风带乐难该亲马华资钱许务吗妈变请专计单办费传欢习广图转军视连尔远满读联红达导设术战队运买试选观识节讲义证轻级决统调刚线评领热权类历课组饭较兴脸则备阳飞谁约标终诉议谈规岁际错词谢众园团续罗馆虽质紧够济亚显绝脑简创户闻剧纪击举细责汉龙农刘钟陈丽营仅

萬國碼第一版 萬國碼第六版 萬國碼第八版 萬國碼漢字字形實用範圍視圖 萬國碼漢字通用字形圖 漢字文化圈通用字形圖 中國大陸 日本 韓台港澳 ©(j)(t) ©(J)(T) **1168** 

c1743这们说时为过对还发开经现样动从间长话实头问进车业两给电cj 134 会来国学当内数声写将号参区 cjz 2561 的人一中上要大在出以自他年可多家能生好本得日前子用方事知行c21352你她吧呢它怎每找啊跟孩您哪晚德份另黑步卡查怕啦假跑值呀j308 関気読経済説对悪実歲県続変発楽様労働価応権戾歷々験営帰与j21367 時間後題個現話間電無見資機動書場開車東記報為長過買結另21667 說沒麼嗎點樣關兩聽灣區內從歡號轉黨變單戰寫臺妳辦價貓黃© China 5,790 ① Japan 4,370 ① Taiwan, Korea, Hongkong, Macau 6947 ②,①,① Total 9,132 ❶ Unicode Version

Macau 6947 ©, ①, ① Total 9,132 ① Unicode Version 1.0 (1991) 20,914 ② Unicode Versions 5.2—6.1 (2009—2012) 74,617 ③ Unicode Version 8.0 (2015) 80,385

C...5790 glyphs J...4370 glyphs Z...6947 glyphs altogether, 9132 glyphs have a regional tag

Overview of Unicode CJK Characters by Regions

Shown here are (1) the numbers of characters deemed essential for general usage as reported by the ISO/IEC 10646 IICore initiative, consolidated into three regions: (C) Mainland China, ① Japan, ① (Taiwan, Hong Kong, Macau, Korea). Circled capital letters show sum totals; plain lower case letters show partial sums for codepoints reported only for (possibly overlapping) sub-regions (e.g. 'j' for characters only used in Japan, 'cj' for forms used only in the PRC and Japan, etc). (2) For comparison, the three outer circles marked **①**, **⑥**, and **⑧** show the growing number of CJK codepoints in Unicode through different versions.—All circles and overlaps are area-proportional to the resp. character counts; not shown is the small number of characters included in IICore, but missing from UCS VI. Sources: link\_open? http://appsrv.cse.cuhk.edu.hk/ ~irg/irg/IICore/IICore.htmlink\_close? , link\_open? http://www.ogcio.gov.hk/en/business/tech\_promotion/ ccli/download\_area/iicore\_compare\_utility.htmlink\_close? https://en.wikipedia.org/wiki/ link\_open? CJK\_Unified\_Ideographs#Unicode\_version\_historylink\_close?

----- Tumblr Comment

Here's a so-called Euler-Venn-Diagram that shows (1) how many CJK character (a.k.a. ideograph, sinograph, tetragraph, ...) codepoints in Unicode are considered 'important for daily use' (just under 10,000); (2) how many of those are used in each of the three important usage areas (i.e. © for Mainland China (PRC without the Hong Kong and Macau SARs), ① for Japan, and ① for Taiwan, Hong Kong, and Macau); and (3) how much overlap there is for each of these regions. The bars made up from sample characters in the lower part visualize the same data, but this time with bar heights instead of circle areas being proportional to numbers.

Just in order to stress it, a 'character' in this chart is equivalent to 'a Unicode codepoint', so for example 馬 and 马 count as two characters, and 關, 关, 関, 闊, 闊 count as five characters. Dictionaries will list 馬马 as 'one character with two variants' and 關关関闊闢 as 'one character with five variants', but that's not what we're counting here.

You can see at a glance that

- \* around 25% (¼) of all characters are common to all three regions,
- \* around 17% (1/6) of all characters are unique to the PRC and Taiwan, respectively, while only 3% are unique to

## 萬字一覽置換索引

## Japan,

- \* only 1% of all characters are uniquely shared by the PRC and Japan, the most frequent ones being 会来国学当内数声写将号参区. But observe that usage frequencies usually follow Zipf's Law (quote: "the frequency of any word is inversely proportional to its rank in the frequency table"); because of this, any one character near the top of a frequency ranking will appear vastly more often than any one near the bottom of the same list. As a result, although the outcomes of the PRC's and Japan's post-war efforts to simplify char-
- acters share but one out of a hundred character shapes, those shapes are still exceedingly often seen in daily life.
- \* Of the three regions—①, ①, ①—the ① region is the one with the most balanced shared usages, the other two regions showing greater relative differences between their subsets shared with zero, one or two of the other regions. In a sense, therefore, the ① region may be considered to be the 'most representative' or 'most average' one in terms of character usage.

'©nAf98',