インターネット時代の文字コード:
漢字コードの迷信を打破する！

宮澤 彰
国立情報学研究所

質問1

日本語は，世界で一番難しい言語（のひとつ）だ。

1) YES.
2) NO.

質問2

今の文字コードには漢字が足りない。

1) YES.
2) NO.

質問3

漢字はおおよそ何文字ある？

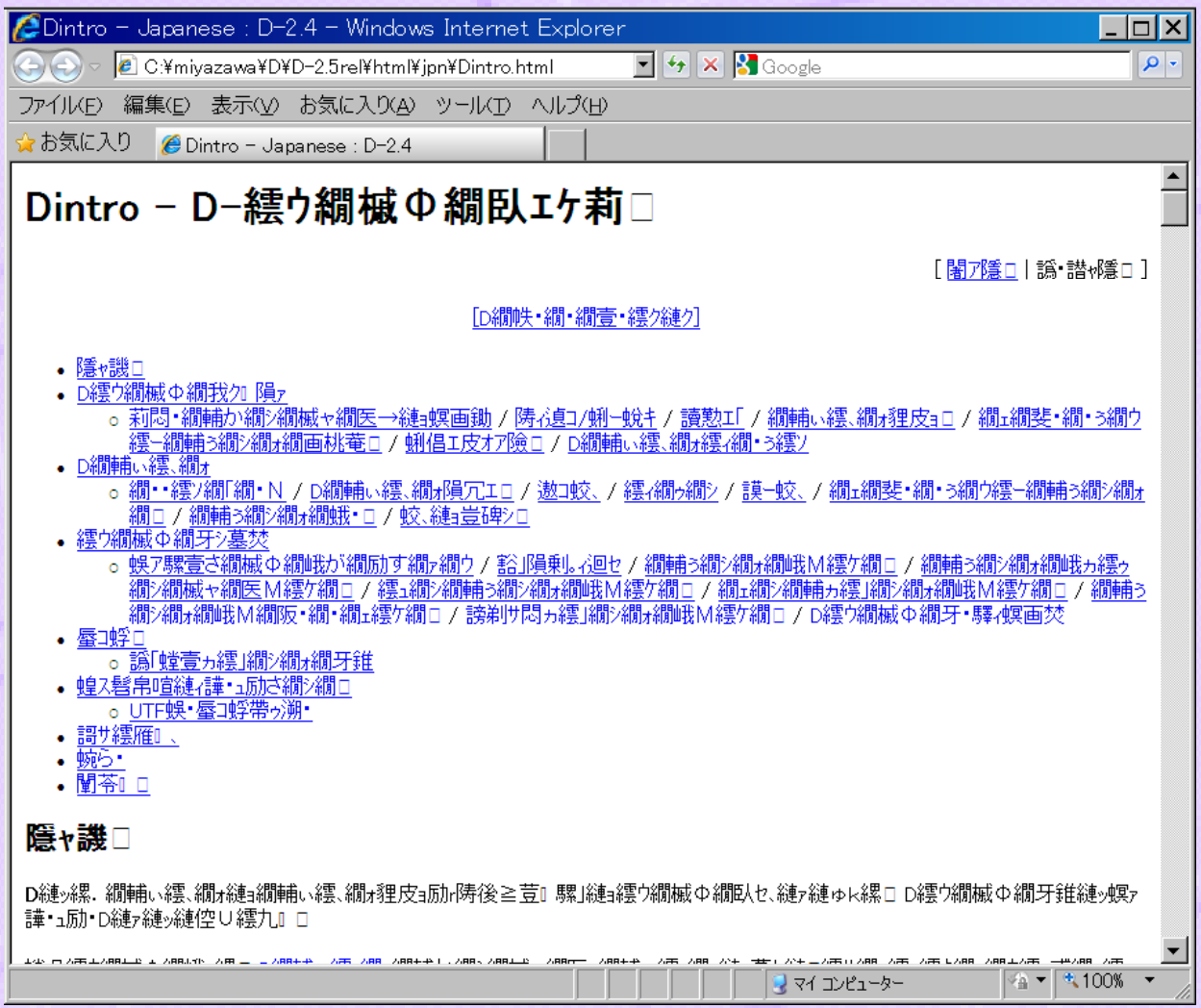
1) 2万字。
2) 5万字。
3) 8万字。

質問4

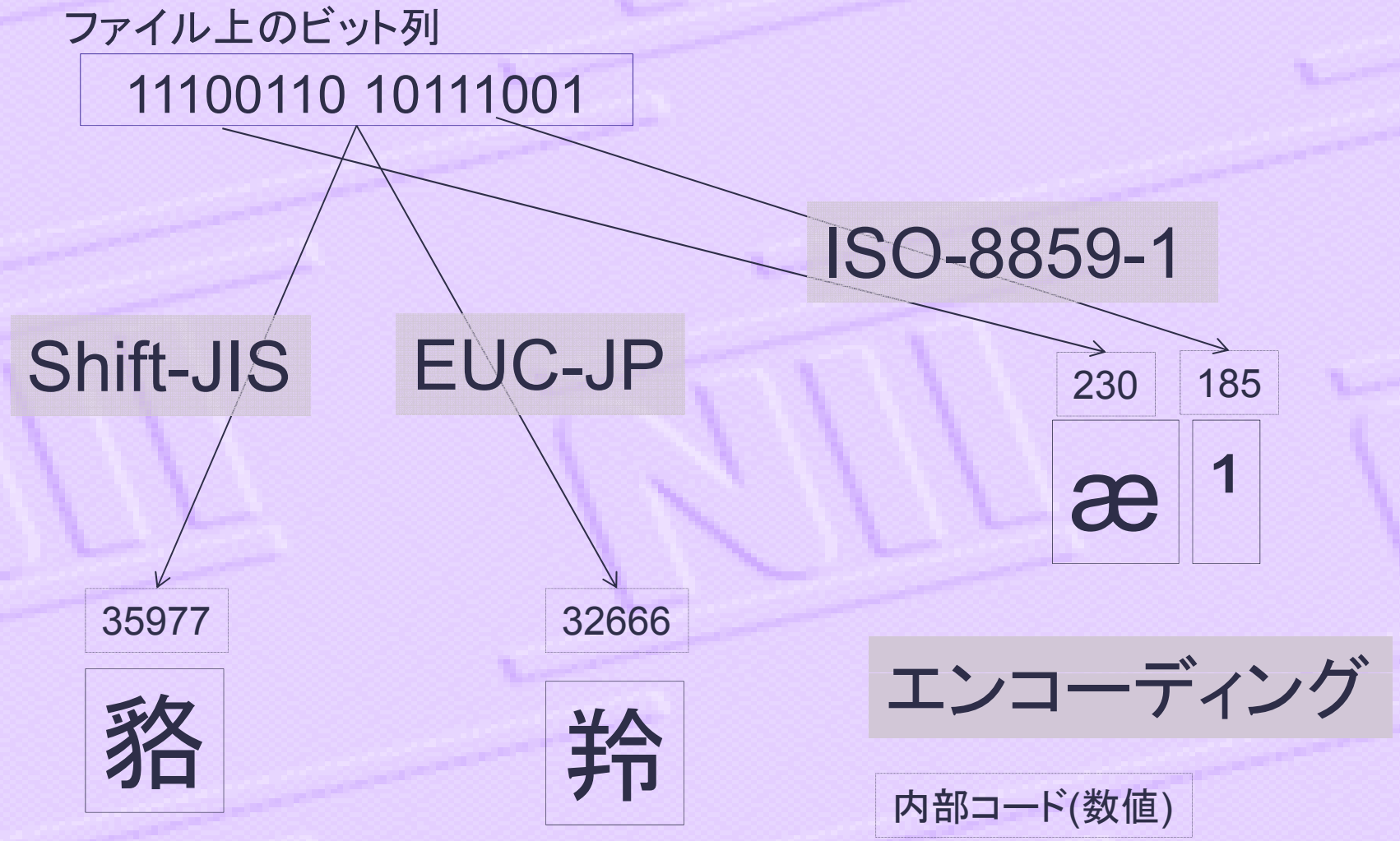
漢字には，意味がある。

1) YES.
2) NO.

文字化けしたウェブページ



コンピュータ内部の文字表現



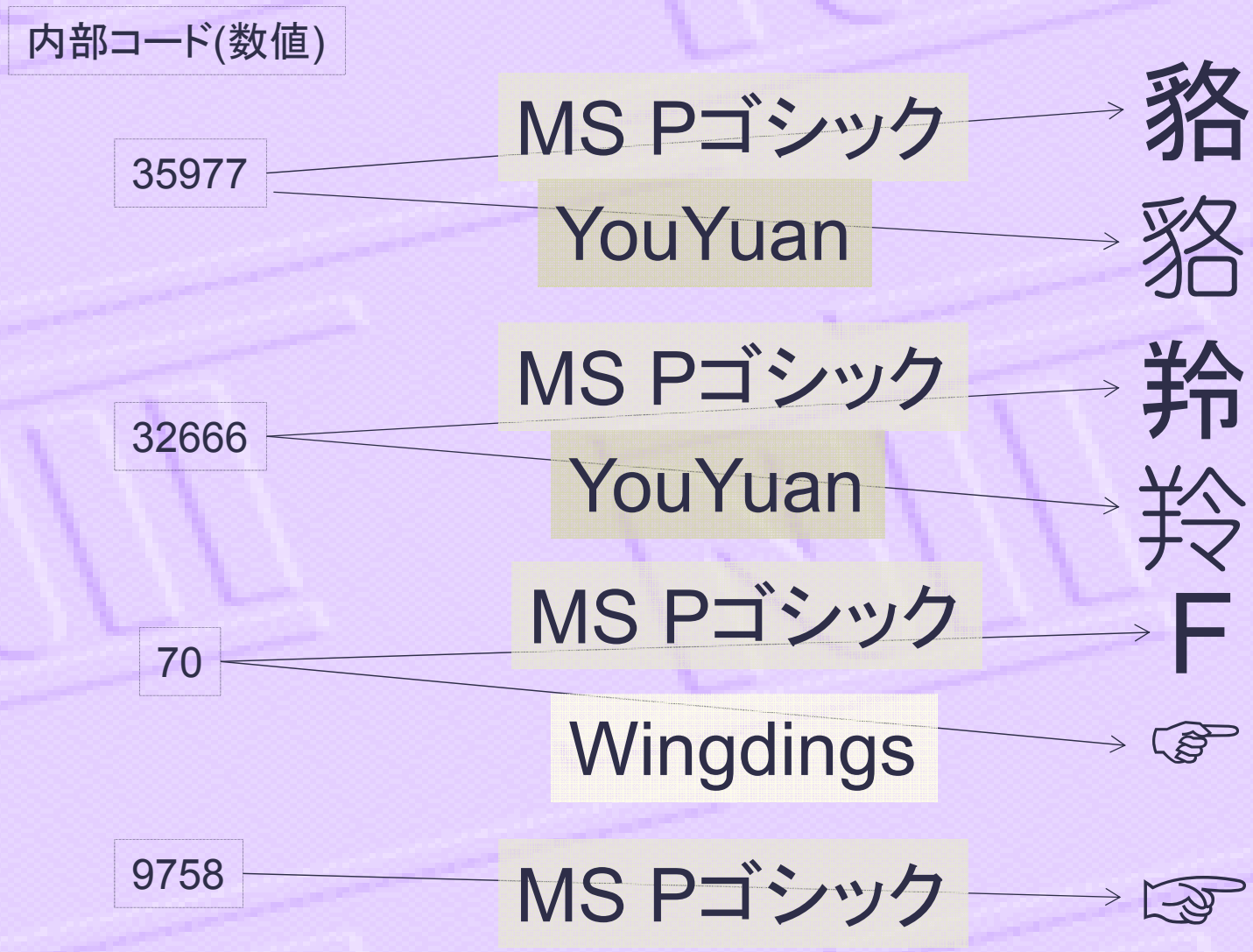
エンコーディング

- 多くの種類（ iconv 1.14 というプログラムによれば143種。別名も数えると425種。）
- 例えば

  ASCII, BIG5, GBK, EUC-KR, LATIN1, MacCryrillic, UTF-8
- 日本語（JIS X0208）だけでも

  EUC-JP, SHIFT_JIS, CP932, ISO-2022-JP, ISO-2022-JP-2, ISO-2022-JP-1
- 正確には，コード表とエンコーディングの差

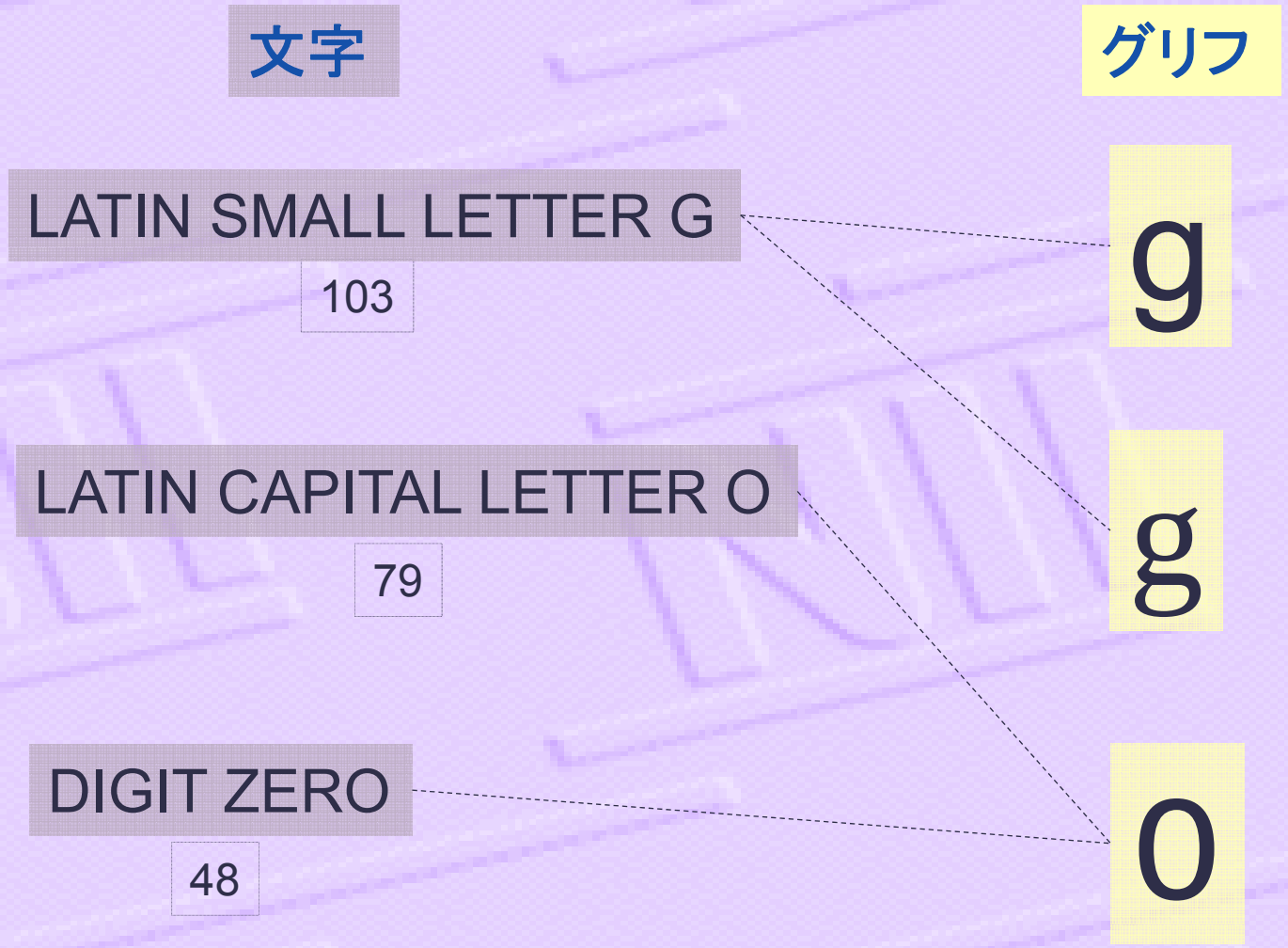
フォント


内部コード(数値)
35977
MS Pゴシック
貉
YouYuan
貉
32666
MS Pゴシック
羚
YouYuan
羚
70
MS Pゴシック
F
Wingdings
☞
9758
MS Pゴシック
☞

いろいろなフォント

Courier	abcdefg
Lucida Sans	abcdefg
Times New Roman	abcdefg
平成明朝 W7	**情報学**
Gungsuh	情報学
中ゴシックＢＢＢ	**情報学**
HGS教科書体	さいたま
HG明朝B	さいたま

文字とグリフ


文字
グリフ
LATIN SMALL LETTER G
103
LATIN CAPITAL LETTER O
79
DIGIT ZERO
48
g
g
0

いくつかの用語

- スクリプト:

  ラテンアルファベット，キリル文字，アラビア文字，漢字，ハングル，かな，…

- 話し言葉　と　書き言葉
- 書法

  書き言葉を表現するための規則の集合

日本語は，漢字，かなスクリプトを混用する書法を標準としている。

漢字とは？

- 元来が中国語用のスクリプトで，一文字が，中国語の単音節語に対応するという点が最大の特徴。（表意文字ideographではなく，表語文字logographと呼ぶことを主張する人々もいる。）
- 日本語では，音（当時の中国語の発音を日本語の音韻構造に移したもの）と，訓(中国語から日本語への標準的訳語）を混用しながら表記する書法が発達した。この中で，中国語の単音節語に対応しない漢字（国字）を創作することも行われてきた。

「漢字コード」の歴史

- 1978: JIS C 6226 情報交換用漢字符号系
  - 6349漢字（第1水準2965，第2水準3384）
- 1983: JIS C6226 改訂 ＝ JIS X 0208 (1987)
- 1990: JIS X 0212
  - +5801漢字
- 1993: ISO/IEC 10646-1 UCS
  - 21140漢字（統合漢字20902 + 互換漢字238 ）
- 2000: JIS X 0213:
  - + 3685漢字（第3水準1249，第4水準3390）

「漢字コード」の歴史

- 2000: ISO/IEC 10646 UCS
  - + 6583漢字
- 2003: ISO/IEC 10646 UCS
  - 計 71226漢字
- 2011: ISO/IEC 10646 UCS
  - 計 75616漢字

問題点：何が一つの「漢字」か?

野埜

沢澤泽

掻掻

姫姬

高髙

戸户戶

これらは，ISO10646 UCSでは，すべて別の文字。

問題点 続

葛葛

与与

芸(藝の常用字体)芸(ウン)芸(雲の簡体字)

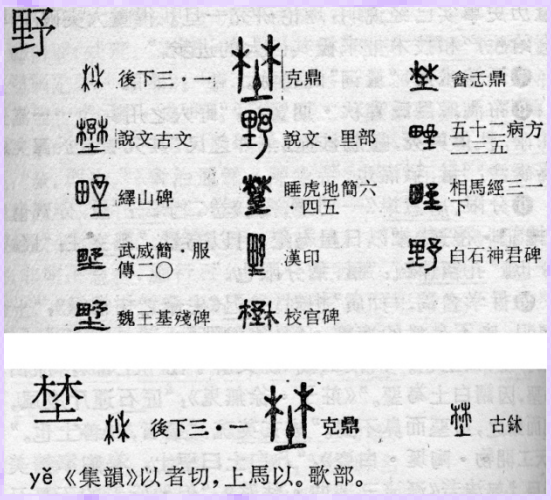
これらは，ISO10646 UCSでは，同じ文字。

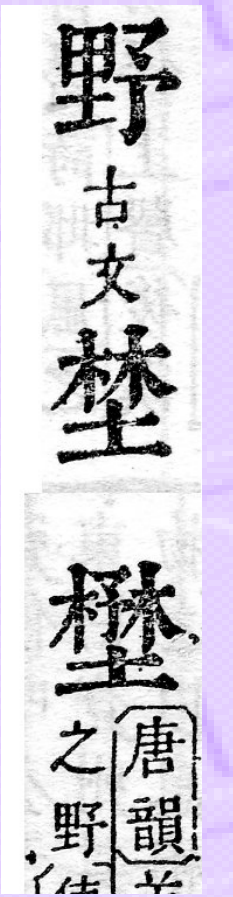
文字学

- 漢和辞典：諸橋轍次，『大漢和辞典』など
- 字源学的研究：白川静，『字統』，『字通』など

これらの文字学の成果は，前記の問題点を解決できるか？

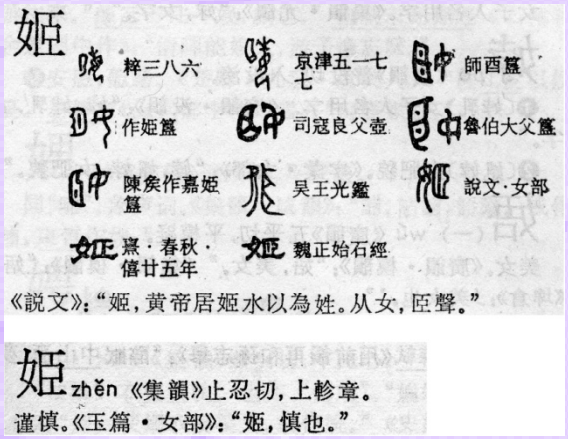
野埜





汉語　字通　大漢和　康熙

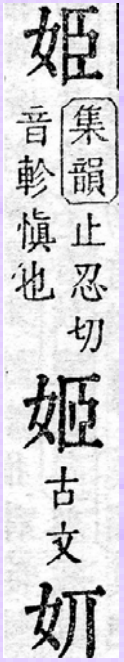
姬姬


姬 粹三八六 京津五一七七 師酉簋 作姬簋 司寇良父壺 魯伯大父簋 陳矦作嘉姬簋 吳王光鑑 說文·女部 熹·春秋·僖廿五年 魏正始石經

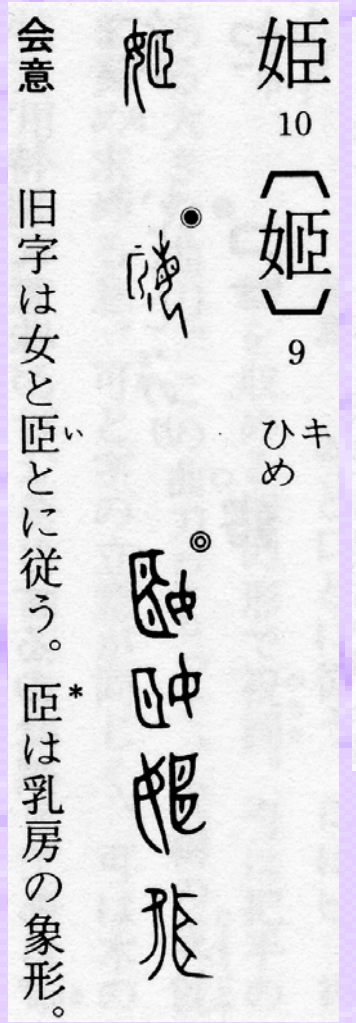
《說文》:“姬,黃帝居姬水以為姓。从女,臣聲。”

姬 zhēn 《集韻》止忍切,上軫章。謹慎。《玉篇·女部》:“姬,慎也。”

汉語

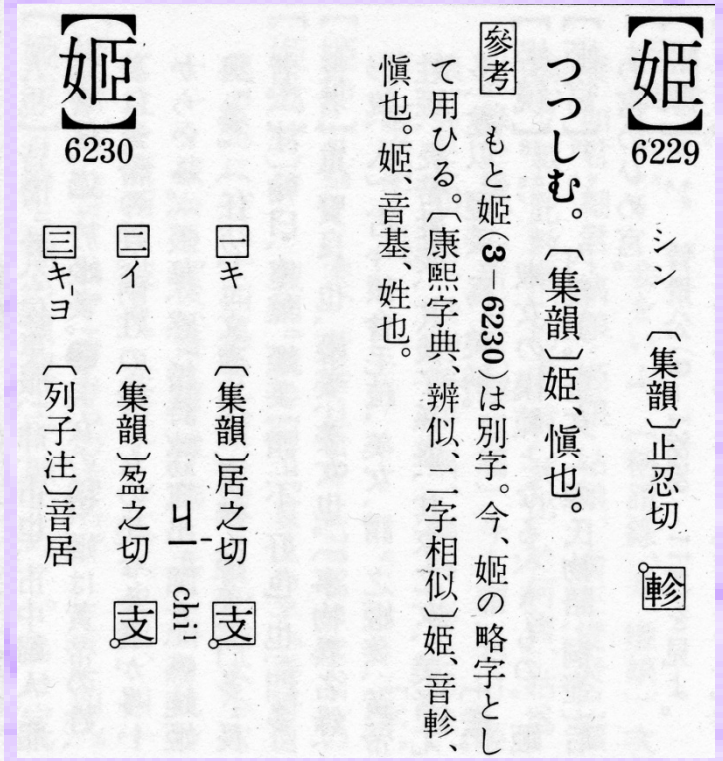

姬 集韻 止忍切 音軫 慎也 姬 古文 姒

康熙


姬 10 【姬】 9 キ ひめ

会意 旧字は女と匜(い)とに従う。匜*は乳房の象形。

字統


【姬】 6229 シン 〔集韻〕止忍切 軫

つつしむ。〔集韻〕姬、愼也。

参考 もと姬(3－6230)は別字。今、姬の略字として用ひる。〔康熙字典、辨似、二字相似〕姬、音軫、愼也。姬、音基、姓也。

【姬】 6230

㊀キ 〔集韻〕居之切 支 ㄐㄧ chi

㊁イ 〔集韻〕盈之切 支

㊂キ,ヨ 〔列子注〕音居

大漢和

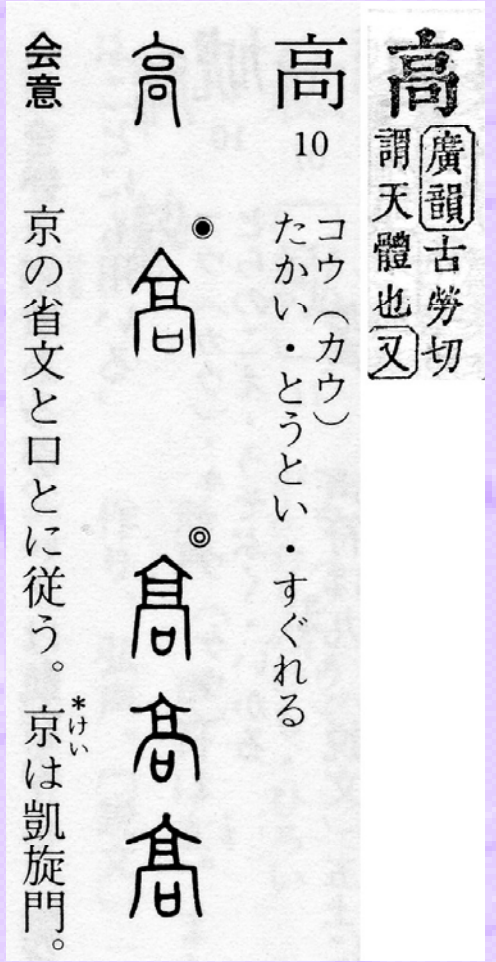
姫姬　用例

光明皇后杜家立成雑書要略
(国会図書館近代デジタルライブラリー) より

大書源
(二玄社)より

高髙

汉語　　大漢和

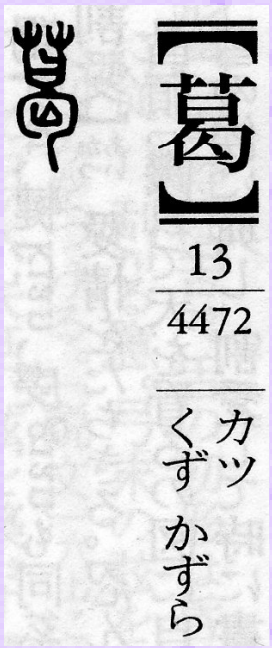


字統　　康熙

高髙 用例

大書源
(二玄社)より

葛葛



字通

北越奇談
(早稲田大学古典籍
総合データベース)より

大書源
(二玄社)より

文字学は

- 残念ながら，漢字の文字という単位の識別に対して一貫した答えは与えてくれない。
- 白川文字学は，その中で文字単位に対して一貫した立場と用語を貫いているが，「野」と「埜」とが同じ文字の異なる字形であるということになると，実用的立場からは受け入れ難い，という意見が多数である。
- 実際に書かれている文字は，字書には示されていない字体も多く使われている。

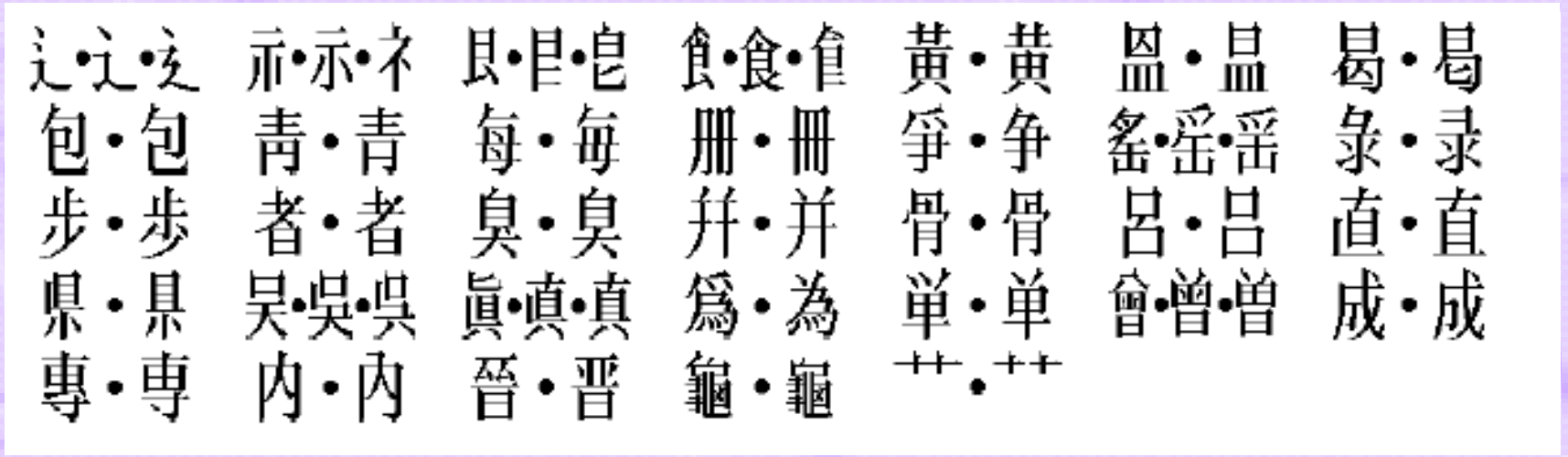
異体字

- 昔から「古文」,「古字」,「俗字」,「異体」などさまざまな呼び方で広く認められてきた「異なるグリフ」
- 楷書形への写し方の違いによってうまれる「異なるグリフ」
- 漢字簡素化など，書法を変更することによってうまれる「異なるグリフ」

字体の「包摂」

- 異なる字体（グリフ）を一つの文字（コード値）としてみとめること。
- UCSでは，付属書Sで，いわゆるunification rule（統合規則）を決めている。
- JIS X 0208は，1997版から包摂基準を持っている。

Unification ルール（Unifyの例示）



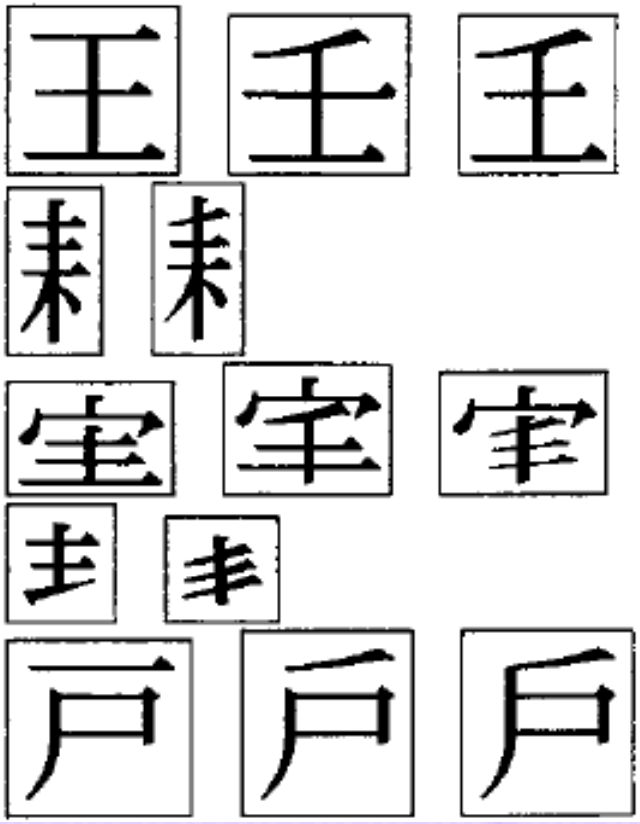
Unification ルール（Separateの例示）

JIS X0208-1997の包摂基準(部分)

a)　方向・曲直などの点画の性質による違い

— 入りの左右

包摂する部分字体



適用除外 (規定)

18-06 31-49
王, 壬

包摂の問題点

- 包摂規則より先にいくつかの漢字コード表があり，包摂規則を崩している場合がある。
- 形だけからの包摂規則は，字源から見れば問題となる場合がある。
- 言語におけるあらゆる規則と同じく，万人の納得する包摂規則は不可能であり，多数が認めるものとして決めざるを得ない。

  （例えば，一つの単語という単位でも同様の問題はある。）

結論として　1

- 文字は形ではない。

「康熙字典の[一]部には丁，丐，七…を録する。…字はその構造的な原理から離れ，その構造的な意味も捨てられて，ただ筆画の形式によって分離配列されている。そこにあるものはすでに文字ではなく，文字の形である。意味を失っている記号である。」

— 白川静，「字統の編集について」より

結論として 2

- しかしながら，漢字についてはどの単位で文字と考えるか，の社会的合意が弱い。

  現在のISO10646の包摂で不都合な場合は，アプリケーションソフト側で，（複数の文字を一つの文字として扱うための）正規化をおこなうか，一つの文字の中で異なるグリフを区別するためのフォントの切り替えをおこなう方法がある。

結論として 3

- 字形のわずかな違いを捉えて，異なる漢字である，としたがる社会的風潮は，データ作成のコストを上げ，知的資源の無駄遣いとなるばかりである。

  少なくとも，昭和初期くらいまでの日本人は，字体，字形の異なりに寛容であった。字体の包摂は，明文化しなくとも常識として通用していた。