

【マルチメディア文化論】

文字データの世界 (日本語文字)

2009年度 第7回

太田 信宏

1. コンピュータで漢字を扱うには

■ 漢字キーボード (漢字鍵盤) の時代



■ 日本語ワープロの誕生

東芝製 JW-10 昭和53年(1978年)

[JW-1、JW-10の紹介](http://www.asahi-net.or.jp/~SD5S-SKMT/wpsokki10.htm)

<http://www.asahi-net.or.jp/~SD5S-SKMT/wpsokki10.htm>

[写真で見る歴史的なコンピュータ](http://www.watch.impress.co.jp/pc/docs/article/20010313/ipsj.htm)

<http://www.watch.impress.co.jp/pc/docs/article/20010313/ipsj.htm>

■ かな漢字変換・ローマ字漢字変換の時代へ

変換精度の向上 AI変換 (Artificial Intelligence: 人工知能)

2. 漢字の数

<日常使用する漢字>

学習漢字 → 1006字

常用漢字 → 1945字

人名漢字 → 488字

<大漢和辞典の収録漢字>

およそ10万字

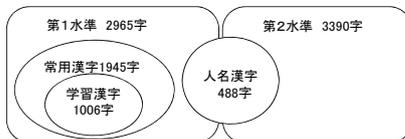
(多くの異体字あり)

劍, 劔, 劔, 劔, 劔, 劔, 劔など

<コンピュータの漢字>

◇ 第1水準 → 2965字

◇ 第2水準 → 3390字



常用漢字の改訂記事

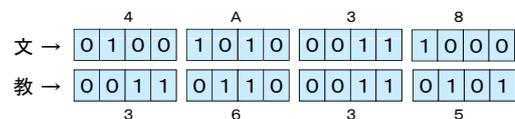
<http://www.asahi.com/national/update/0116/TKY200901160297.html>

3. 全角文字のコード

半角1文字は1バイト=8ビット → 256通り

全角1文字=2バイト=16ビット → 何通り?

(JIS漢字コードの例)



第1水準と第2水準の合わせて約6400字は、すべて異なる0と1の並びを持っている。

(文字を4桁のコード番号で識別) 文 → 4A38
教 → 3635

4. 欠字と外字

◇ 欠字 コードが存在しない漢字のこと。
(コンピュータに登録されていない文字)

◇ 外字 欠字や特殊記号に対して、あとから任意に0と1(のコード)を割り当てた文字。

メーカー外字 ... パソコンに内蔵されている外字
(機種依存文字)

高崎青弼 ① ② ③ I II III 絵文字、など多数ある。

ユーザー外字 ... 利用者が独自に作成した外字

◇ メールの外字使用はマナー違反。

使ってよい文字 → JIS漢字コード (第1水準+第2水準)

5. 文字コード例

	JIS漢字	シフトJIS	EUC	Unicode	種類
文	4A38	95B6	CAB8	6587	第1
教	3635	8BB3	B6B5	6559	第1
高	3962	8D82	B9E2	9AD8	第1
高	(967E)	(FBFC)	—	9AD9	外字
盆	4B5F	967E	CBDF	76C6	第1
桧	4930	95F4	C9B0	6867	第1
檜	5B58	9E77	DBD8	6A9C	第2
①	(2D21)	(8740)	—	2460	外字
I	(2D35)	(8754)	—	2160	外字

6. いろいろな文字コード

- ① JIS漢字コード(日本工業規格)
インターネットで利用可能な文字コード (第1水準+第2水準)
- ② シフトJISコード
マイクロソフト社が日本のパソコン用に開発した文字コード
- ③ EUCコード(拡張UNIXコード)
主にサーバー(パソコンの上位機種)で利用される文字コード

↓ **新たな提案**

- ④ Unicode(ユニコード)
世界中の文字を1つのコード体系にまとめるという壮大な構想。
英字、漢字(日本・中国・韓国・台湾)、ラテン、タイ、ハングル文字など
→ ①~③で表示できなかった多くの外字をカバー。



7. 文字コードの課題

- ① 欠字、外字の問題 (JIS漢字、シフトJIS、EUC共通)
 - ・過去の文献などをコンピュータに入力できない
- ② 略字体の代用ですむか？
 - ・森鷗外
 - ・高木さん→高木さん
- ③ Unicodeの課題
 - ・各国の漢字を統合したことの問題 → (例) 中国の「骨」という字
 - ・異体字の扱い(コードセパレート問題)

文字コード情報に関するサイト 「立命館大学」山田崇仁先生
(文字コード入門)
<http://www.shuiren.org/>



8. 外字の調べ方

Windowsヘルプで「文字コード表」を開く

