

インターネット応用特論

4. 文字通信:

文字コードと国際化

太田昌孝

mohta@necom830.hpcl.titech.ac.jp

<ftp://ftp.hpcl.titech.ac.jp/appli4.ppt>

参考書

- 太田昌孝、「いま日本語が危ない」、光芒社、ISBN4－89542－146－5、平成9年

文字とは？

- 言語を図形で表現するための要素
 - 表音文字
 - 表意文字

スクリプト(用字系)とは？

- 言語を文字で表記する体系
- 言語と文字を混同してはいけない
 - 「でいすいずあぺん」はかなスクリプトによる英語
 - “Koreha pen desu.”はローマ字(ラテン)スクリプトによる日本語

日本語の表記に用いる スクリプト

- かな(ひらがな、カタカナ、万葉仮名)
- 漢字かな混じり
- ローマ字(ヘボン、訓令、他)
 - 「まさたか」をフランス語風に書くと“massata
ka”
- その他、各地の文字による表音表記

デジタルとアナログ

- デジタルは、細部の違いを問題にしない
 - 雑音に強い
- 言語は（話し言葉も）デジタル
 - 声、歌はアナログ
- 文字もデジタル
 - 17文字があれば、極めて微妙な感情も伝わる
 - 書道はアナログ
- 問題は、どこまでの細部を見分けるか

文字コードとは？

- ある文字集合に含まれる文字からなる文字列のデジタル化の規則
 - 個々の文字にコード(数値)を振ることだけではない
- 文字集合の含む文字の数が問題
 - 多いと多数のビットが必用
 - 少ないと多様な文字が表現できない
 - 同種の文字の細かな違いも表現できない

バイト

- 本来は1文字を表すビット数の単位
 - 1バイト=8ビットというわけではない
- 4ビットバイトでは16文字を表現
 - 数字と記号(“ , . + - ”等)には十分
- 6ビットバイトでは64文字を表現
 - 英大文字と数字と記号には十分
 - 36ビットCPUで多用
- 7ビットバイトの例がASCII

多バイト文字コード

- 本来矛盾した概念
- 1バイトで1文字を表すと単純明快
- 1バイトのビット数が固定化すると
 - 1バイトでは文字が表現できなくなる
- 複数のバイトで文字を表すには
 - 連続するバイトで1文字(多バイト文字)を表現
 - バイト(制御「文字」)による状態の切り替え

ASCII (American Standard Code for Information Interchange)

- 米国標準の7ビットバイトの文字規格
 - 95文字(含空白)の(図形)文字集合
 - 英大小文字、数字、記号
 - 33の制御「文字」と混在
 - 国際規格であるISO646の米国版
- 英語を表現するには十分
- いろいろな意味で単純な文字コード
 - コンピュータ化が楽

ISO 646

- ASCIIと同じ構造をもつ
- 95文字のうち
 - 83文字は世界で共通
 - 12文字は国別に变えてよい
- 我が国のJIS X 0201 (JIS C 6220)は、ASCIIと2文字異なる
 - “\”が“¥”に、“~”が“—”に

ASCIIの(英語のラテン文字表記の)単純さ (1)

- 文字数が少ない
- 横書きのみ
- 片方向(左から右)のみ
- リガチャー(文字の形の前後関係による変化)が不要
- 文字同定とその字形に共通認識がある

ASCIIの(英語のラテン文字表記の)単純さ (2)

- 大文字小文字の対応が明確で規則的
 - 飾り文字もない
- 文字幅が一定
- 広く普及していてどこでもつかえる

ASCIIの文字数の少なさ

- 1バイトですべての文字を表現できる
- 多バイト文字や制御文字による切り替えは不要

英語のラテン文字表記は横書きのみ

- 漢字は縦書きが普通
- モンゴルスクリプトも縦書きのみ

英語のラテン文字表記は片方向 (左から右)のみ

- 横書きの漢字スクリプトは本来右から左に書く
 - 日本は明治時代に欧米にあわせた
- 縦書きの行は現在も右から左
 - 横書きは実は1行1文字の縦書き
- アラビアスクリプトは現在も右から左
 - ただし数字や英字が混在する部分は左から右
 - ネスティングもある

ASCIIにはリガチャーなし 英語のラテン文字表記には不要

- リガチャーとは
 - 文字の形の前後関係による変化
 - “fi”や“ffi”の“i”の上に“f”がかぶさることはあるが、なくてもいい
- アラビア文字やデバナガリ文字(インド)は、前後関係で字形が変わるのは当然
 - 活字文化が浸透しておらず、筆記体が残存

ラテン文字の文字同定とその字形に共通認識がある

- ラテン文字でも本来は
 - UとVはもともとと同じ文字 (BVLGARI)
 - WはUU (ダブルユー)
- 英語では区別は消失
 - AはA
 - aの書き方は各種あるが、、、

英語のラテン文字表記は大文字 小文字の対応が明確で規則的

- ラテン文字でも本来は
 - YはIJ
- 英語では由来は消失し、違う文字は違う文字

英語のラテン文字表記には飾り 文字も不要

- 飾り文字の由来
 - 中間的発音の表記のため
 - ラテンアルファベットの周りに別の字を小書き

ラテン文字は文字幅一定でいい

- 文字列のバイト数が表示幅に
- ラインプリンタも可能

ASCIIは広く普及していてどこでもデフォルトでつかえる

- 文字集合の指定の必要なし

まとめ

- 言語を図形で表現するのが文字
- 文字コードとは文字列をバイト列に変換する規則
- ビット数により文字コードには様々な制約がある
- ASCIIはいろいろな意味で「単純な」文字コード