

# 文字コード

---

はしもとじょーじ

# 計算機で文字を扱う

---

## 文字コード (character code)

コンピュータでは文字や記号ひとつひとつに固有の符号を割り当てている  
文字コードとはその対応づけのこと

コンピュータが扱うのは0と1だけ  
ひとかたまりの0と1の並びを文字に変換する際の**変換表**が文字コード

# 文字コード

---

## ASCII コード (アスキーコード)

- American Standard Code for Information Interchange
- 7 bit の文字コード(最大128文字)
- 英数字だけを使うならこれで間に合う

例 :      0100011      #                      0100100      \$  
            1000001      A                      1100001      a

# ASCII コード

| 文字         | コード |      | 文字         | コード |      | 文字           | コード |      | 文字          | コード |      | 文字       | コード |      | 文字       | コード |      | 文字       | コード |      |            |     |      |
|------------|-----|------|------------|-----|------|--------------|-----|------|-------------|-----|------|----------|-----|------|----------|-----|------|----------|-----|------|------------|-----|------|
|            | 10進 | 16進  |            | 10進 | 16進  |              | 10進 | 16進  |             | 10進 | 16進  |          | 10進 | 16進  |          | 10進 | 16進  |          | 10進 | 16進  | 10進        | 16進 |      |
| <b>NUL</b> | 0   | 0x00 | <b>DLE</b> | 16  | 0x10 | <b>SP</b>    | 32  | 0x20 | <b>0</b>    | 48  | 0x30 | <b>@</b> | 64  | 0x40 | <b>P</b> | 80  | 0x50 | <b>`</b> | 96  | 0x60 | <b>p</b>   | 112 | 0x70 |
| <b>SOH</b> | 1   | 0x01 | <b>DC1</b> | 17  | 0x11 | <b>!</b>     | 33  | 0x21 | <b>1</b>    | 49  | 0x31 | <b>A</b> | 65  | 0x41 | <b>Q</b> | 81  | 0x51 | <b>a</b> | 97  | 0x61 | <b>q</b>   | 113 | 0x71 |
| <b>STX</b> | 2   | 0x02 | <b>DC2</b> | 18  | 0x12 | <b>"</b>     | 34  | 0x22 | <b>2</b>    | 50  | 0x32 | <b>B</b> | 66  | 0x42 | <b>R</b> | 82  | 0x52 | <b>b</b> | 98  | 0x62 | <b>r</b>   | 114 | 0x72 |
| <b>ETX</b> | 3   | 0x03 | <b>DC3</b> | 19  | 0x13 | <b>#</b>     | 35  | 0x23 | <b>3</b>    | 51  | 0x33 | <b>C</b> | 67  | 0x43 | <b>S</b> | 83  | 0x53 | <b>c</b> | 99  | 0x63 | <b>s</b>   | 115 | 0x73 |
| <b>EOT</b> | 4   | 0x04 | <b>DC4</b> | 20  | 0x14 | <b>\$</b>    | 36  | 0x24 | <b>4</b>    | 52  | 0x34 | <b>D</b> | 68  | 0x44 | <b>T</b> | 84  | 0x54 | <b>d</b> | 100 | 0x64 | <b>t</b>   | 116 | 0x74 |
| <b>ENQ</b> | 5   | 0x05 | <b>NAK</b> | 21  | 0x15 | <b>%</b>     | 37  | 0x25 | <b>5</b>    | 53  | 0x35 | <b>E</b> | 69  | 0x45 | <b>U</b> | 85  | 0x55 | <b>e</b> | 101 | 0x65 | <b>u</b>   | 117 | 0x75 |
| <b>ACK</b> | 6   | 0x06 | <b>SYN</b> | 22  | 0x16 | <b>&amp;</b> | 38  | 0x26 | <b>6</b>    | 54  | 0x36 | <b>F</b> | 70  | 0x46 | <b>V</b> | 86  | 0x56 | <b>f</b> | 102 | 0x66 | <b>v</b>   | 118 | 0x76 |
| <b>BEL</b> | 7   | 0x07 | <b>ETB</b> | 23  | 0x17 | <b>'</b>     | 39  | 0x27 | <b>7</b>    | 55  | 0x37 | <b>G</b> | 71  | 0x47 | <b>W</b> | 87  | 0x57 | <b>g</b> | 103 | 0x67 | <b>w</b>   | 119 | 0x77 |
| <b>BS</b>  | 8   | 0x08 | <b>CAN</b> | 24  | 0x18 | <b>(</b>     | 40  | 0x28 | <b>8</b>    | 56  | 0x38 | <b>H</b> | 72  | 0x48 | <b>X</b> | 88  | 0x58 | <b>h</b> | 104 | 0x68 | <b>x</b>   | 120 | 0x78 |
| <b>HT</b>  | 9   | 0x09 | <b>EM</b>  | 25  | 0x19 | <b>)</b>     | 41  | 0x29 | <b>9</b>    | 57  | 0x39 | <b>I</b> | 73  | 0x49 | <b>Y</b> | 89  | 0x59 | <b>i</b> | 105 | 0x69 | <b>y</b>   | 121 | 0x79 |
| <b>NL*</b> | 10  | 0x0a | <b>SUB</b> | 26  | 0x1a | <b>*</b>     | 42  | 0x2a | <b>:</b>    | 58  | 0x3a | <b>J</b> | 74  | 0x4a | <b>Z</b> | 90  | 0x5a | <b>j</b> | 106 | 0x6a | <b>z</b>   | 122 | 0x7a |
| <b>VT</b>  | 11  | 0x0b | <b>ESC</b> | 27  | 0x1b | <b>+</b>     | 43  | 0x2b | <b>;</b>    | 59  | 0x3b | <b>K</b> | 75  | 0x4b | <b>[</b> | 91  | 0x5b | <b>k</b> | 107 | 0x6b | <b>{</b>   | 123 | 0x7b |
| <b>NP</b>  | 12  | 0x0c | <b>FS</b>  | 28  | 0x1c | <b>,</b>     | 44  | 0x2c | <b>&lt;</b> | 60  | 0x3c | <b>L</b> | 76  | 0x4c | <b>\</b> | 92  | 0x5c | <b>l</b> | 108 | 0x6c | <b> </b>   | 124 | 0x7c |
| <b>CR</b>  | 13  | 0x0d | <b>GS</b>  | 29  | 0x1d | <b>-</b>     | 45  | 0x2d | <b>=</b>    | 61  | 0x3d | <b>M</b> | 77  | 0x4d | <b>]</b> | 93  | 0x5d | <b>m</b> | 109 | 0x6d | <b>}</b>   | 125 | 0x7d |
| <b>SO</b>  | 14  | 0x0e | <b>RS</b>  | 30  | 0x1e | <b>.</b>     | 46  | 0x2e | <b>&gt;</b> | 62  | 0x3e | <b>N</b> | 78  | 0x4e | <b>^</b> | 94  | 0x5e | <b>n</b> | 110 | 0x6e | <b>~</b>   | 126 | 0x7e |
| <b>SI</b>  | 15  | 0x0f | <b>US</b>  | 31  | 0x1f | <b>/</b>     | 47  | 0x2f | <b>?</b>    | 63  | 0x3f | <b>O</b> | 79  | 0x4f | <b>_</b> | 95  | 0x5f | <b>o</b> | 111 | 0x6f | <b>DEL</b> | 127 | 0x7f |

# 日本語の文字コード

---

日本語は文字集合が大きい

- 2 byte のコード(最大65536文字)

参考：

塚本真也，日本人学生のための日本語教育  
日本で使われている漢字は約5万字  
常用漢字は約2000字

# 日本語の文字コード

---

面倒なことに複数の文字コードが使われている

junet (iso-2022-jp)

- JIS コードと呼ばれたりもする

EUC-JP

- UNIX での標準 (Extended Unix Code)  
EUC-KR, EUC-CN, EUC-TW などもある

Shift\_JIS

- Windows, Mac (OS9まで) での標準

UTF-8, UTF-16, UTF-32

# 文字化け

---

文字コードの指定の間違い

フォントの問題(機種依存文字など)

プログラムの問題

データの欠落や変質

# Unicode

---

## 世界中の文字をひとつの文字コードで表す

- 当初は 2 byte で文字集合を定義
- Unicode 2.0 以降では 4 byte
  
- Windows NT 系、Mac OS X での標準
- UNIX 系でも使えるようになってきた
  
- **使う文字を決めるのは誰なのかという問題**  
他人には同じに見えても、当人にとっては同じでないこともある

# 文字コードの相互変換

---

## 文字コードの変換ツール

- UNIX        nkf, iconv, ...
- Windows    QKC, KanjiTranslator, ...

## 多言語を扱えるエディタ

- UNIX        VIM, Emacs, ...
- Windows    VIM, Meadow, ...

# 改行コード

---

「**行の終端**」を表す文字(制御文字)

- CR (キャリッジリターン)
- LF (ラインフィード)

OS によって改行を表す方式が異なっている

- UNIX 系                      LF のみ  
  (Mac OS X は UNIX)
- Windows 系                CR + LF
- Mac (OS9 以前)            CR のみ

# まとめ

---

## 文字コード

- 計算機で文字を扱うための対応表

## 改行コード

- OS によって使っているものが違う