第25章　日本語とファイルサイズ

- ✓ 日本語文字のデジタル化
- ✓ Byte, KB, MB, GB
- ✓ 16進数
- ✓ パケット、通信速度

1. 日本語文字のデジタル化

日本語で使われる文字には、英数字、記号、ひらがな、カタカナ、漢字があり、全部合わせて 7 千文字程度ある。16 bit (2 Byte)の情報量があれば、全部で 65,536 通りの状態を表現できるため、それらすべてカバーすることができる。そのため 2 Byte 使用した日本語文字コード体系「**JIS(ジス)コード**」が作られた。また、JIS コードを改良した「**Shift JIS コード**」というコード体系もある。JIS は、通信分野で使われるのに対し、Shift JIS はパソコンで文字を保存したり処理したりするときに用いられている。

2Byte＝16bit＝日本語文字コード

Shift JIS コードで定義されている文字は数万文字あるため、ここに全部載せることはできないが、いくつか例を載せる(付録4参照)。

表1　日本語文字と 01 列の対応（Shift JIS）

すべての文字が 2 Byte (16 bit)で表現されている

	1000001010100000	東	1001001110001100	1	1000001001010000
あ	1000001010100001	洋	1001011101101101	2	1000001001010001
い	1000001010100010	英	1000100101110000	#	1000000110010100
う	1000001010100100	和	1001100001100001	α	1000001110111111
え	1000001010100110	女	1000111110010111	β	1000001111000000
お	1000001010101000	学	1000101001110111	γ	1000001111000001
カ	1000001101001010	院	1000100101000000	○	1000000110011011
キ	1000001101001100	0	1000001001001111	×	1000000101111110

2. [発展] 外字と機種依存文字

JIS コード(または Shift JIS コード)で何も定義されていない 01 パターンに、メーカーが独自に文字を割り振ることができる。これらの文字のことを**外字**と呼ぶ。携帯電話で使われる絵文字は外字を利用している[1]。それ故、異なる機種間やメーカー間で絵文字を使ったメールのやり取りができないことがある。

また、①㊤ミリ㏄㈱昭和Ⅱといった文字は、**機種依存文字**と呼び、特定の機種だけで使える記号であり、メールやホームページで使用してはいけない。

3. 全世界の文字のデジタル化

文字のデジタル化は、各国で行われており、それらの規格に互換性がほとんどない。英語は ASCII、日本語は JIS(または Shift JIS)、ヨーロッパ各国は ISO646、中国では CB2312、台湾では Big 5、韓国では KS X 1001 という具合に異なっている。このように各国が自国の文字をばらばらの規格でデジタル化してしまうと、複数言語の文字が混じった文章を扱ったり、全世界で情報を共有したりするには大変不都合である。

例えば、日本人向けのドイツ語の教科書を書こうとすれば、ある場所は、Shift JIS コードを使い、別の場所は ISO646 を使いといった具合になり、大変面倒であるばかりか、複数の文字コードを同時に表示できる OS やソフトウェアが必要となるなど煩雑である。

そこで、このような情報伝達上の問題や、古典、古語の研究促進のため、全世界・全歴史を通じて人類が作り出したすべての文字を同一規格でデジタル化しようとする試みがなされた。これによってできたコードを**Unicode**(ユニコード)という。Unicode は全世界の文字を 2Byte(16bit)で表現する。

Unicode は 2Byte で文字を表現するため、最大 65,536 文字が表現可能である。しかし、これでも全世界の文字を表現するには足りないため、4Byte(32bit)で文字を表す **ISO10646** が作られた。Windows や Macintosh では Unicode を扱えるが、まだ完全には普及していない。

4. 日本語のテキストファイル

日本語を入力して作成されたテキストファイルも、英数字と同じように、変換されて01列として保存される。メモ帳を使って「こんにちは」と書いてハードディスクに

[1] この外字機能を使って Windows でも携帯の絵文字を使えるようにするソフトがある。

保存すると、ハードディスクには Shift JIS コードの文字がそのまま保存される。そのファイルをメモ帳を使って開くと、その 01 列を Shift JIS コードだと認識して文字となってメモ帳に表示する。

英数字は ASCII コードによって1文字で1Byte、日本語文字は Shift JIS コードによって1文字が 2 Byte で表現されることがわかった。ここで、再び、実際にメモ帳を使って文章を書き、保存されたテキストファイルのサイズを調べてみよう。また、バイナリーエディターで中身を見てみよう。

5. 補助単位

ファイルのサイズを数えるのには、Byte という単位が用いられている。例えば、メモ帳で 100 文字の日本語を書いて保存すると、ファイルサイズは 200 Byte になる。10,000 文字の英数字を書いて保存すると、ファイルサイズは 20,000 Byte になる。

数が大きくなると、なかなか表記するものも難しくなるため、私たちが日常、距離や重さを表現するときに 1000 を単位に K(キロ)を使って表すように、コンピューターでも K(キロ)などの補助単位を使って表している（表2参照）。ただし、コンピューターの世界での単位の区切りは 1000 ではなく、1024 である。

1MB(メガバイト)
＝1024KB（キロバイト)
＝1024×1024B(バイト)＝1,048,576B(バイト)
となる。

表2 情報量の補助単位[2]

単位	記号	定義
バイト	B	8 bit
キロバイト	KB	1024B
メガバイト	MB	1024KB
ギガバイト	GB	1024MB
テラバイト	TB	1024GB

2 大きい補助単位は、 キロ（K） メガ（M） ギガ（G） テラ（T） ペタ（P） エクサ（E） ゼタ（Z） ヨタ（Y）と続く。IEC （国際電気標準会議） は、 2^{10}、2^{20}、2^{30} などの乗数を表す接頭語は IEC 60027-2 でそれぞれキビバイト （KiB）、メビバイト （MiB）、ギビバイト （GiB） としているがほとんど普及していない。

6. 日本語のテキストファイルサイズの計算

日本語文字は、JIS、Shift JIS、Unicode どれで表現しても 1 文字あたり 2byte の容量が必要となる。USB フラッシュメモリー1個の容量が2GB の場合、計算がしやすいように Byte に直すと、

$$2\mathrm{GB} = 2 \times 1024 \times 1024 \times 1024 = 2{,}147{,}483{,}648\ \mathrm{Byte}$$

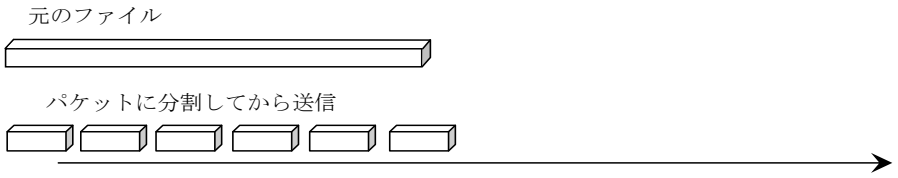
したがって、USB フラッシュメモリー1個に収納できる文章の量は、以下の式で計算できる。

$$2{,}147{,}483{,}648 \div 2 = 1{,}073{,}741{,}824\ \text{文字}$$

2GB の USB フラッシュメモリー1個には約 10 億文字の文章が保存できる計算になる。新聞は、1 ページ約 1 万字(1 行 11 文字×75 行×14 段程度、下の 3 段の広告を除く)なので、10 万 ページ分になる(これは計算上の理想値であって実際にはこれほど入らない)。

7. [発展] デジタル通信

デジタル情報を通信する際、コンピューターはそれを小さな**パケット**と呼ばれる塊に分割して送信する。



パケット通信の概念図

このように情報を一度に送らず、細かく分けて送る通信を**パケット通信**という。例えば、引越しなどで沢山の荷物を運搬しなければならない場合でも、一人で運搬できる程度の小さな小包に分けてから運搬するようなものだ。そもそもパケットとは小包の意味である。

インターネットや携帯のメールはパケット通信だ。パケットの大きさは使用する回線によって異なる。携帯電話の場合は 128Byte である。携帯電話のメール料金は通信されたパケットの量で決められている。およそ 0.1円／パケットである[3]。

---

[3] 2007 年 2 月現在の料金。各種割引があるので一概には言えない。パケット通信でのパケット単位は 128 バイトである。1 パケットで送れる文字は日本語全角の場合、64 文字という計算になる。パケットの中に「ヘッダ」と呼ばれる、宛先等の情報も含まれているので、実際には 1 パケットで送れる文字数は 64 文字よりも少ない。また、文字以外に文字の位置や色など画面表示に必要な指定情報、その他パケット通信に必要なデータも含まれているので、表示される文字以上のデータがやり取りされている。

8. 通信速度

インターネット接続は各種あり、通信速度が違う。通信のスピードは 1 秒間に通信できるbit 数で定義され、単位は bps　(bit per second)で表す[4](☞**第6章6節**)。

電話線とモデムを使って通信する場合の最高速度は、54kbps である。理論上、1秒間に 54000bit の通信が可能である[5]。

インターネットではサーバーからクライアントへのデータの流れ(下り)の方が圧倒的に多い。この通信量の非対称性をうまく利用した **ADSL**[6]方式がある。通信速度は幾つかの種類があるが、一般的には伝送速度は下り最大12Mbps、上り1Mbps 程度が理論値である。実際の速度はNTT の交換局からの距離などによって決まる。

ケーブルビジョンを使った通信では、会社やサービスによって速度は異なり一概には言えないが、例えばiTSCOM 社[7]やYCV[8]社の場合では、下り 1600Mbps のサービスが最高速度である(2017 年 2 月現在)。

光ファイバーケーブルを用いたNTT の **B フレッツ**や KDDI の au ひかりは、理論上は最大で1Gbps の通信速度が出る(2017 年 2 月現在)。

音楽CD1枚全部をインターネットで送信したらどのくらいの時間がかかるか計算してみよう。CD1枚に収められている音楽情報のサイズは曲によってまちまちだが 640MB として計算しよう。一度 bit に直してから計算するとよい(1MB＝1024×1024Byte、1kbps＝1000bps、1Mbps＝1000×1000bps)。

ADSL 12Mbps の場合:

640MB÷12Mbps=(640×1024×1024×8)÷(12×1000×1000)=約 447 秒

Bフレッツの場合:

640MB÷1Gbps=(640×1024×1024×8)÷(1×1000×1000×1000)=約 5.4 秒

上記の数値は単純計算をした理論値であり、実際はこの10倍以上かかる。

4　通信速度の場合には 1000 を基準に補助単位を使う。1kbps＝1000bps として計算する。kが小文字であることに注意。また 1Mbps＝1000kbps である。

5 54kbps は理論上の最高速度。現実には 36kbps 程度しか出せない。

6 Asymmetric Digital Subscriber Line(非対象デジタル加入回線)。最高速度下り 45Mbps/上り 3Mbps が登場してきている。

7 iTSCOM…イッツ・コミュニケーションズ株式会社。旧東急ケーブルビジョン。

8 YCV…横浜ケーブルビジョン株式会社。

演習

1 日本語文字コードの調べ方

1. 言語バーから IME パッド をクリックする。
2. メニューから[文字一覧]を選ぶ。
3. 調べたい文字の上にマウスをもってくるだけでそこに Unicode、JIS、Shift JIS の各コードが 16 進数で表示される。

4. この16進数を2進数に直す(☞**第23章9節**)。

2 自分の名前をデジタル化してみよう

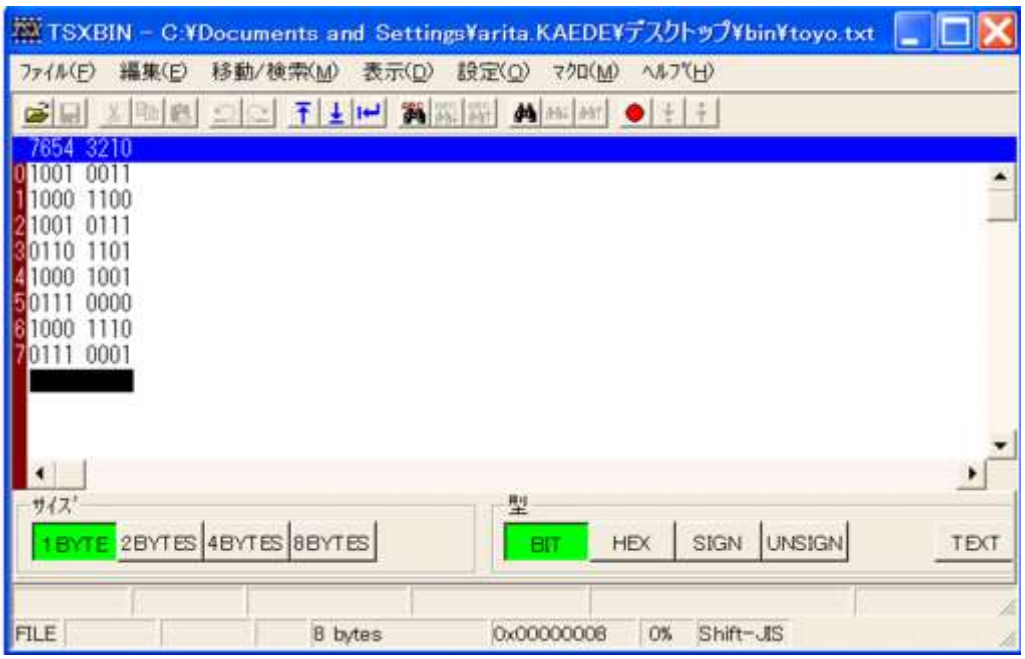
1. 自分の名前を漢字で書く。
2. 1文字ずつ上記の方法で調べ、16 進数で表記する。
3. この16進数を2進数に直す(☞第 23 章 9 節)。
4. これが「東洋」を Shift JIS コードでデジタル化した例である。

文字	16 進数	2 進数
東	938C	1001001110001100
洋	976D	1001011101101100

3 バイナリーエディターで文字入力しよう

演習2で調べた自分の名前の Shift-JIS をバイナリーエディターに入力してみよう。例えば、東洋英子の Shift-JIS は 938C 976D 8970 8E71 であるので、順に入力する。大文字小文字は区別しない。

入力後、 BIT ボタンを押して01を閲覧し、演習2と結果一致しているか確かめよう。



4 メモ帳で日本語を書いて、バイナリーエディターで閲覧しよう

メモ帳で、日本語文を書いて保存して、それをバイナリーエディターで閲覧する。（演習3の状態になっている場合は、バイナリーエディターを起動しなおすか、新規作成したほうがよい。）

5 日本語のテキストファイルサイズの計算

基礎情報科学のテキストは、およそ 1 行 37 文字、1 ページ32行、260 ページである。すべて文字で埋められたとして、1GB の USB フラッシュメモリーの何%になるか。枠を埋め、計算しなさい。

日本語 1 文字は、[ ] byte となる。

テキスト全体では [ ] byte ×[ ]文字×[ ]行×[ ]ページ

＝[ ] byte

＝[ ] KB

＝[ ] MB

＝[ ] GB となる。

USB フラッシュメモリーが1GBなので、 理論上、

およそ[ ]%を使用していることになる。

実際は絵が多量に入っているので、もっと大きい(☞**第26章1節**)。

6 通信速度の計算

自分のウェブページを立ち上げるべく HTML ファイルや画像ファイルを用意した。HTML ファイルは 50KB、画像は全部で 2MB のファイルサイズになった。1 秒間に 100Mbps の速度でプロバイダーと通信するとしたら、最も速くてどのくらいの時間でファイル転送が完了しますか。

転送速度は HTML ファイルも画像ファイルも同じである。

ファイル容量は

50KB+2MB＝(50 ×[ ]＋2 ×[ ] ×[ ])B

＝(50 ×[ ]＋2 ×[ ] ×[ ])×[ ]bit

＝ [ ]bit

転送時間は

(50KB+2MB)÷100Mbps

＝[ ]bit÷(100×[ ]×[ ])bps＝[ ] 秒