

談話室

プロッターによる漢字書き*

渋谷 政昭** 原文 子**

1. 情報処理の進歩に伴い、計算機の扱うデータはたえず多様化していく。特に人間と機械との対話が行なわれ、計算機利用者の層が広くなれば、当然より日常的な図形、記号、自然言語を外部言語とすることが望まれる。自然言語を用いるとなると、漢字を使用する日本語（および中国語、朝鮮語）は表音文字を使用する諸言語に比べて非常に不利である。しかし、漢字使用は今後二・三代などでは消滅しないだろうし、データ処理に対して漢字使用が決定的な障害であると断定するのも早急であろう。とにかく、情報処理の出力が人に理解されることを目的とする以上、読み易さという心理効果を見捨てることはできず、漢字の取り扱いをできるだけ技術的に解決するほかはない。

たとえば、我々はタイプライタの使用について、欧米人よりはなほ不利な状況におかれている。漢字タイプライタをいくら改善したところで、すべての人がそれを自由に利用するほど熟達することは期待できない。かなタイプライタは入力の実用としては一応の目的を果たすが、出力のためには十分でない。経済的に困難であってもかなタイプライタ → 和文漢訳プログラム → 漢文かな出力装置が、将来の我々のタイプライタであろう。

2. 漢文の入出力の開発は金物と平行して論じなければ無意味であろうが、ここでは X-Y プロッターが電子的となり十分高速となることを予想しておく。

この話の目的は、漢字の字形を計算機に記憶させておいて描かせる簡単な実験について報告し、実用的にはどれだけの容量と手数を要するかを概算することである。要点は誰でもすぐ気付くように、漢字の体系は木構造（見方によってはもう少し複雑なリスト構造）をなしているということだけであり、仕事の発端も木構造の実例を一つつくることにあった。

Backus 記号まがいの表わし方を用いると、漢字は

次のように定義できる。

<画の原図> ::= (点, ..., 点)

<縮図配置> ::= (a, b, g, h)

<画> ::= <縮図配置> <画の原図>

<文字> ::= <画> ... <画> | <縮図配置>

<文字> ... <縮図配置> <文字>

まず、基礎となる画としては折線で表わされる一画に限り、折線を指定するのに単位正方形内の点の座標を用いる。<縮図配置>とは、単位正方形に含まれ、その辺と平行な辺をもつ長方形であって、縦横の大きさ (a, b) と左上の角の座標 (g, h) で指定される。正方形そのものとなる「単位写像」も<縮図配置>に含めておく。一つの<文字>の中のいくつかの<縮図配置>の長方形は、互いに重なり合ってもよい。

上の定義は我々の画、筆順の概念に忠実な定義である。<文字>は漢字、かたかなといわゆる「部首」を含んでいるが、必ずしも漢和字典などに書かれている部首とは同一でない。この定義そのものの変形もいろいろ考えられるであろうし、上の定義に従っても<文字>の表現は一意的でない。この点についてはあとで論ずる。

漢字やかたかなを「書く」とは、字の番号（字の集合、たとえば当用漢字表における一連番号）を与えて、その各画の折線を表わす点の座標の系列を出力する（折線を描くサブルーチンと呼ぶ）こととする。

<文字>についての情報の記憶装置での表現は、各文字に対してリスト

<k: 構成文字の数> <(a₁, b₁, g₁, h₁): 縮図配置>

<m₁: 構成文字番号> ... <(a_k, b_k, g_k, h_k):

縮図配置> <m_k: 構成文字番号>

を用意する。ただし、m_j は<画の原図>の番号を示すこともある。<画の原図>についての情報の表現は、各画について

<l: 折線の端節点> <(x₁, y₁): 点の座標>

... <(x_l, y_l): 点の座標>

を用意する。

* Writing Chinese characters by a plotter, by Masaki Sibuya and Ayako Hara (Institute of Statistical Mathematics, Tokyo)

** 統計数理研究所

「書く」プログラムは、与えられた番号の<文字> (木の根) から出発して、構成する<文字> (木の節) をたどり<画の原図> (葉) にたどりついたらプロットする。もちろんこの探索には柵を用いることになり、柵には<文字>の番号、1段元から見たとき枝が何本あり、その何本目をたどっているかの3数値と、各段の<文字>の<縮図配置>をのせておかねばならない。

実験は実用性を無視して FORTRAN (水準 3,000 未満) を用い、既製のプロット・サブルーチンを利用して計、算、機およびそれに伴う言、十、竹、目、木、幾の字を描かせた。プロッタは CALCOMP MO DEL 560 R、八方に 5cm/sec. の速さで動く。プロットのサブルーチンを除いて FORTRAN で 60 ステートメントである (図参照)。

計算機 言十竹目木幾 計算機
言十竹目木幾 計算機

3. この実験に対して、種々の代案が考えられる。

(1) 一画を最終の単位とする必要はなく「部首」を単位とすればよい。つまり、上の説明での<文字>の表から基本的な部首を追い出して、<画の原図>の表と組合わせてしまう。各々の画についての縮図配置の情報は不要となるが、たとえば一の画が何度も出てくることになる。

(2) <縮図配置>の情報をこのように詳しくする必要はない。「へん」「つくり」「かんむり」「あし」などという分類で縮図配置は決まってしまう。漢字を分解したときの各部分の位置は 10 数とおりで十分のようである。ただ、同じ木へんでもつくりが複雑か簡単かで配分比を変える必要がある。<縮図配置>に実験では 16 ビット与えている。そのまま 12 ビットに減らしたのでは字がかなり不自然になるようだが、上の分類を用いれば数ビットに落とせるだろう。

(3) <画の原図>を慣用どおりの一筆に限る必要はない。たとえば、口は 3 画でなく 1 画とした方がよい。万年筆の書体が筆の書体と異なるように、新しくプロッタの書体と筆順を考えねばならない。しかし、慣用どおりだと読み易さのために単なる線でなく明朝活字にしたり、「永字八法」を組み込んだりすること

がより容易であろう。

4. 漢字の形状についての情報がどれだけの大きさとなるか。いうまでもなく、字を美しくしようとすれば情報を多くしなければならない。「単位正方形」を表わすビット数、画の折線の点の数、用意する部首の数 (林、森などの木の字はすべて同じ形とするかどうか) 等々。

ここでは「単位正方形」として 4 ビット四方を選び、一つの画の精しさは「はねぼう」と「ぼう」のみ区別しただけで、牛と「うしへん」は同一視し、木、竹などにおける画はすべて直線とする程度である。

当用漢字 1,850 字、かたかな 46 字、計 1,896 字について

(1) 画の原図の数、約 40; 画を表わす点の数、平均 3.6.

(2) 部首 (漢和字典にある部首で当用漢字とならぬもの) の数、55.

(3) <文字>を構成する部分の数、平均 2.5.

とし、必要なビット数を計算した。おおざっぱではあるが 160 キロビット (1 字あたり 84 ビット) である。

ある電光ニュースでは、漢字を 13×13 ビットで表わしている。

比較のために、3 の (1) の方法について試算してみたところ、約 2 キロビット余計に必要であるが、この方法では柵の平均の高さが 1 ないし 2 減少する。部首の数を 72 としたが、これで十分かどうか少々不安はある。

教育漢字に限ると、文字の数は減るが「部首」および「画の原図」の数はやはり同じくらい用意しなければならない。実験に用いた方法で計算すると、1 字あたり約 97 ビットに増加する。逆に中国での慣用約 8,000 字に増しても、1 字あたりのビット数は当用漢字の場合とほとんど変わらない。

実験における実際作業では、各々の文字と部首について縮図配置を適当に与える仕事に、かなり手間を要した。当用漢字全部を描かせるためには、ライト・ペンなど使用し、その作業を機械化しなければならない。また、漢字を指定するコードを、ここでは単に一連番号としたが、人が直接に指定する場合も考慮するならば中国語辞典で用いられているような種々の工夫が必要である。