
 報 告

台湾での第一回国際計算機科学会議の報告

—主とし漢字の入力について—

杉 田 繁 治*

1973年8月に台湾で開催された漢字の入出力を中心とする第一回国際計算機科学会議の報告を書くようにといわれた。すでにかなりの日数がたっているのと、読者にそれほど大きな情報を与える内容でもないような気もするが、一応形式的な報告と漢字の入力に関する論文を2,3紹介する。この方面に関心のなかった方の興味を引けば幸いである。

1. 会議について

会議名 : First International Symposium on Computers and Chinese Input/Output Systems

開催時期: 1973年8月14日~16日

開催場所: 中華民国 台湾 台北市 中央研究院

主催 : Academia Sinica, National Science Council R. O. C., China Institute of Electrical Engineering, Telecommunications Laboratory R. O. C.

協賛 : IEEE, U. S. A.

発表件数: 93 件

参加者数: 約 300 名

この会議のテーマは、急速に増加している中国語の科学技術情報の機械処理のために、高速で簡便な漢字の入出力方法を開発しようとすることを主目的にし、同時にミニコンピュータの各種応用を考えるというものであるが、対象論文の範囲としては、漢字の入出力およびその処理、画像処理、パターン認識、計算機システムと評価、データ通信、アルゴリズム、言語処理、その他応用など次に示す 11 の session があった。(カッコ内は chairman.)

- 1 Chinese Information Processing System (S. K. Chang)

- 2 Pattern Recognition (K. S. Fu)
- 3 Picture Processing (C. K. Chow)
- 4 Chinese Input Methods (T. Y. Kiang)
- 5 Machine Translation and Language Teaching (J. Mathias)
- 6 New Designs of Computer Systems (C. N. Liu)
- 7 Data Communication System (D. T. Tang)
- 8 Chinese Output Methods and Implementation (C. S. Hsieh)
- 9 Computer Systems Modeling and Evaluation (P. H. Kong)
- 10 Computer Applications (C. T. Chang)
- 11 Algorithms and Optimization Techniques (Y. Y. Chow)

各 session には 9 件位の発表があったが、プログラムについては通信学会電子計算機研究会資料 (EC 73-32, 1973 年 10 月 23 日) を参照されたい。会場は 1 つで (最後の session のみ parallel) 講演時間は 1 件当たり 15 分程度。参加者の大部分は中国人 (米国に住んでいる人も含む) で、純粹の欧米人は 12 名位、日本からは 6 件の論文が出され、13 名位が参加していた。発表は英語または中国語となっていたが、中国語で発表する人が多かった。Proceedings は 1973 年 12 月末までに発行される予定になっていたが 12 月末現在まだ出版されていない。

2. 漢字の入力について

漢字の計算機への入力、計算機からの出力については、我が国や米国においても今まで、かなりの研究がなされており、1971年には漢字の I/O に関する日米セミナーも開催されている。(Seminar on Input/Output System for Japanese and Chinese Characters: US-JAPAN Committee on Scientific Cooperation, January 1971.) またこれに基づき米国ハーバード大学

* 京都大学工学部情報工学教室

久野教授と東京大学藤村教授の間で漢字の I/O に関するプロジェクトを進めてこられている。また最近ではオンライン手書き漢字の認識や漢字 OCR などの研究も活発に進められており、漢字 I/O に関心が寄せられている。それと関連して漢字の標準コードを早く制定するよう各方面から希望されており、情報処理学会の漢字コード委員会でいろいろな案を検討されているようである。

漢字のコードを定めたり、鍵盤上に漢字を配列する時に問題となることは漢字の indexing 法である。即ち与えられた漢字がどこに配置されているかを、全体を憶えているのではなく比較的単純な規則によって確実に見出す方法である。音訓に基づく場合は漢字が読めないと使えない。偏や旁など部首引の場合どれを部首と考えるかに問題がある。例えば康熙字典式の部首別配列では「和」という字は「禾」偏ではなく「口」部である。これは歴史的な経過から来ているが、一般には分りにくい。これに類する例はたくさんある。そこで意味や歴史的経過を無視して、字形のみから index を求める方法が研究され、興味のある問題となっている。

漢字の入力法として考えられるものは

1. 鍵盤を用いて入力する方法、
2. オンライン手書き文字認識による方法、
3. 自動文字読取による方法

があり、会議では 1 についての発表が多かったので以下これについて述べる。

鍵盤を用いる方法を更に分類すると次のようになる；

- 1-a 漢テレ型入力……漢字と key が 1 対 1 対応。
- 1-b 字形分割入力……ストロークに分割、
Radical に分割、
特定の個所の形による。
- 1-c 発音による入力……1 字毎、
単語毎。

どのような方式で入力するかは、どのような場で、どのような使い方をするかによってそれぞれ適否がある。大量の情報を常に入力する必要がある場では、専門のオペレータによる 1 字 1 key 式の鍵盤方式で入力の方がかえって早い。少量のデータではあるが特に訓練しない人が計算機のタイプライタや電話機程度の端末などを介して入力する場合には、漢字を分割して少数の key で入力出来るようになっていることが望まれる。従って鍵盤上に漢字を配列する時の並べ

方 (indexing) と漢字の構造記述に興味を持たれている。

ここでは発表された論文の中から、(a)主として漢字の構造記述について述べたもの、(b)漢字の index 法について述べたもの、(c)変形四角号碼法による符号化について述べたもの、について紹介する。

(a) The Chiao Tung Radical System; C.C. Hsieh, Y.W. Hwang

この論文は約 500 の Radical を鍵盤上に配列してその組合せで漢字を入力する方式および Radical の配列の仕方について述べている。

まず漢字の出現頻度について少し述べている。8,532 字の漢字を出現頻度によって分類すると、1,857 字が非常にしばしば出現するクラス (A クラス)、2,068 字がしばしば出現するクラス (B クラス)、2,182 字がたまに出現するクラス (C クラス)、2,425 字がめったに出現しないクラス (D クラス) というように分類出来る。また 2,022,604 字のサンプルに対して 97.34% が A クラス、2.27% が B クラス、0.27% が C クラス、0.12% が D クラスに属しているということである。

この論文では漢字の出現頻度を考慮して漢字を Radical に分解するかしないかをきめ、平均として key の数とそれを押す回数が最適になるように工夫しようとしている。ここで Radical とは漢字の一部分であって、他の漢字の部分としても使用される構成要素 (字根) をいう。分解出来ない字の場合はそれ自体が Radical といわれる。いわゆる偏や旁、冠などはその一例である。

今かりに $n+1$ 個の key (即ち $n+1$ 個の Radical) があり各字に対して平均 s 回 key を押すとする。今 1 つの key 上の Radical を更に分解するとき、両者とも Radical として他の key 上にあれば、その key をなくすことが出来る。その代り key を押す回数が 1 回増加する。(オペレータ key の操作は除外して考える。) もしその字の出現する頻度が f 、全体が $\sum_i f_i$ とすれば $(n+1)s > n(s + f / \sum_i f_i)$ であれば、すなわち $f < s \sum_i f_i / n$ なる関係が得られる出現頻度の場合には分解することにする。このような考えに従って分解した結果 8,532 字の漢字に対して 496 個の Radical を得ている。表 1 はその Radical を出現頻度順に配列している。ここで最初の 5 個の Radical で 11.3%、最初の 25 個で 30%、最初の 50 個で 49%、最初の 100 個で 66.7%、最初の 300 個で 95%、最初の 368 個

表 2 構造記述用オペレータ

H		左右的結合 「行」, 「理」	Q		上下連合 「李」, 「尖」	R		右上包囲 「式」, 「句」
X		左中右三結合 「衍」, 「衛」	S		多面合囲 「國」, 「同」	G		左上包囲 「房」, 「屋」
V		上下的結合 「章」, 「呈」	T		上面加蓋 「字」, 「春」	I		中間挿入 「申」, 「垂」
Y		上中下三結合 「尋」, 「幸」	L		左下包囲 「遷」, 「趕」			

く、中央上部にまとめておく。数字および演算記号なども上方に配置しておく。他の Radical はその左上の筆形により “drop” (ノ, ㄣ, ㄥなど), “skim” (ㄣ, ㄣ, 禾など), “beamstroke” (マ, 子, 正など), “upright” (丨, 卜, 目など), “cross” (力, 才, 聿など) の5種類に分類し、グループ化して配列する。各グループ内ではストローク数、形の似たもの、出現頻度などを考慮して出来るだけ憶えやすいように配列する。またよく使われるものは太字にしておき注意を引くようにする。

また別の考え方では、上部、下部、左部、右部、中央、の Radical にグループ化し、鍵盤上でもそのように配置する。どのグループに属せしめるかはその位置で使われることが50%以上であるグループに入れる。あちこちで使われるがいずれも50%以下の場合には“中央”のグループに入れる。各グループ内では“drop”, “skim”, “beamstroke”, “upright”, “cross” に分類しておく。

また key の数を減らすために1つの key 上に8つの Radical を記入しておき、位置選択 key によってその内の1つを選ぶようにすることも考えている。この場合各 key の左上隅の Radical の場合には位置選択 key を押さなくてもよいようにしておく。このような Radical は全体の1/8にすぎないが、使用頻度は42%位になるとのことである。

(b) The Upper-right Corner Indexing System for Chinese Language and a New Chinese Character Encoding System; T. Y. Kiang

この論文は約8,000の漢字の中であるものを分割して約2,500の構成要素 (Radical) を抽出し、それを鍵盤上に配列し、入力する方法について述べている。こ

表 3 共通字根 (T. Y. Kiang)

(1) 左部 (61個)

王 土 才 走 耳 革 木 麥 車 豆 酉 石 歹 馬 齒 口
日 目 田 貝 足 虫 黑 巾 骨 山 巾 禾 未 牛 角 矢
イ 片 卩 身 彳 舟 金 食 豸 月 月 魚 牙 角 言 立
火 羊 米 丷 丷 丷 丷 子 弓 月 女 糸

(2) 上部 (16個)

冫 冫 冫 冫 冫 冫 冫 冫 冫 冫 冫 冫 冫 冫 冫 冫
厂 广 广 气 門

(3) 右部 (11個)

頁 戈 ㄣ 女 隹 鳥 彡 欠 卩 力 羽

(4) 下部 (8個)

土 木 貝 皿 衣 心 乚 糸

こで問題となることは Radical の indexing 法、つまり約2,500の Radical の中から必要なものをいかにして早く見出すかということに関する問題である。更にこの論文では実際の鍵盤の設計について述べている。

先ず漢字を大きく3つのクラスに分類する; (i) 共通 Radical を持たぬもの, (ii) 1つの共通 Radical と主部とからなるもの, (iii) 2つ以上の共通 Radical を含んでいるもの。共通 Radical とは表3に示すような約100種類の字形であって非常にしばしば出現する形のことである。約70%の漢字がクラス(ii)に属しており、これを分解することによって2,500位の基本要素にすることが出来る。クラス(i), (iii)のものは分解してもあまり効果がなく、符号化における手間が増えるばかりであるという。

さて漢字が与えられた時にどのようにして入力するかであるが、先ず表3の(2)上部および(3)右部の共通 Radicals 27個を憶えておき、与えられた漢字から

それらを取除く。その後残ったものに左部があればそれを無視する。こうして得た字形を後述する鍵盤上の第一行に集められた main parts の字形と一致するか探し、一致すればその key を押し、更につけ加えるべき共通 Radical があれば共通 Radical の鍵を押す。もしその字形が main parts の場所になればその字形の上部の形を index として第2～5列の key 上に配してある全体の字形と一致するものを探す。もし全体の字形とは一致しないが、ある共通 Radical をつけ加えれば与えられた漢字になるものがあればその key と共通の Radical の key を押し。以上の方法によって記憶する量が少なく、ちょっと練習すれば誰にでも容易に入力することが出来るということである。indices としては46個のものが有り、その下に sub indices が与えられており、各 indices に属する漢字の数は比較的均等になっており検索が容易であると述べている。

鍵盤の中央は8つのセクションに分かれ、各セクションには5行4列の key が配置してある。各 key 上には16個の字形が記入されている。共通 Radical のためには鍵盤上部に8つの key があり、各 key には16個の Radical が配置されている。ある字形を入力するには、その字形が記入されている key を押し、次にその key が属しているセクションの5行4列の key の内最初の4行4列の key を位置選択用 key として用いて、その字形の key 上での位置に相当する key を押し。こうして1つの字形を入力するには2回 key を押す必要があり、Radical 共通を持つ場合には4回 key を押す必要があるが、簡単な論理回路が組込まれていて自動的に16ビットのコードになって出力される。

(c) **Chiese Input-Output with Standard IBM Selectric Typewriter Terminal; E. W. Pugh, C. C. Tsao**

この論文はいわゆる四角号碼法と呼ばれる indexing 法を改良したものである。四角号碼法とは漢字の4隅の筆形を各々数値の0～9で表現し、4桁の10進数値で表現するもので、その筆形と数値コードを表4に示す。四角号碼法は key の数は少ないが、4隅の筆形の見方によって符号が異なり一義的にコードを定めるのはむづかしい。また異なった漢字に同一コードが対応することがあり、しかもそのような場合が非常に多い。

表 4 筆形とその符号

符号	筆 形	符号	筆 形
0	一	5	ㇿ (つらぬき)
1	一ノしゝ	6	口
2	丨ノ丨	7	ㇿㇿㇿㇿㇿㇿ
3	丶ゝ	8	ハ ヌ 人 一
4	十メ	9	小 ハ ヌ 人 一

例	四 角 法	九 角 法
口	6 0 0 0	6
琳	1 4 1 9	1 0 1 4 0 9 4 0 9
麻	0 0 1 9	0 0 1 4 0 9 4 0 9
床	0 0 1 9	0 0 1 4 0 9

そこでこの論文では漢字を Radical に分解し、各 Radical に対して四角号碼法によるコードを与え1～9桁の10進数値で表現しようとするものである(九角法)。即ち漢字を最大3つの Radical に分解し、各 Radical に対して、左上隅、右上隅、左下隅の3ヵ所(四角号碼法では更に右下隅も考える)の筆形を見て3桁の数として表現する。1つの Radical から成る場合、2つの Radical から成る場合などは残りの部分に対して0をつけ加えてもよいが、入力の場合には省略することが出来る。この方式では最大30ビット必要であり、32ビット word を使う場合には残りの2ビットを ambiguity を解消するのに用いる。この方式は単純な四角号碼法よりは異字同一コードのケースは当然少なくなるが、隅の筆形に対するコード化の ambiguity に関しては解決されているわけではなく、また Radical に分解する時のやり方については具体的には述べていない。

隅のコード化についてあいまいさのない方式を考えれば電話機などからも入力出来るので便利ではある。この点に関して更に改良が必要である。

3. その他

以上紹介したものは Radical に分解する方法であり、(a), (b)はどのようなものを Radical と考えているかに興味がある。その他発音記号によって入力し、同音異字の場合には CRT 上に表示して人間が選

扱するシステムについての発表もあった。この時1字毎ではなく単語・熟語単位で検索出来るように辞書を構成しつつあるものもあった。

漢字の出力についてはIBMのSELECTRIC TYPE-WRITER でドット表示を出力することも出来、漢字プリンターや合成法による CRT 表示などすでにいろ

いろな方法で出力出来るようになっている。

現在かなり漢字の I/O に対する関心と要求がもり上がりつつあるが、使用目的に応じた方式を具体的に鋭意研究すれば急速に収斂するような気がする。

謝謝

(昭和 49 年 1 月 10 日受付)