Vol. 25 No. 6　　情報処理学会論文誌　　Nov. 1984

漢字および漢字熟語の声形符号†

川口 喜三男†† 王 思鴻†††
井川 智††† 宇野 誠一††††

日本語漢字の読みの表記または中国語漢字の拼音表記と四角号碼を組み合わせた声形符号は，個々の漢字を識別し同定する能力が大である．本稿では，声形符号の適用範囲を，さらに漢字熟語および漢字仮名混り語をも含むように拡大する．また，欧米語と異なり，日本文および中国文の書法上の特徴である縦書きと横書きの双方とも取扱いの対象とし，そのいずれであっても熟語に対する符号の形部を一意に定められるように，四角号碼法の一変種として新四角号碼法を導入する．この声形符号を日本語および中国語に適用したとき，その識別能力がどのようであるかを，最悪ケースまたはそれに近いと考えられる二，三の実例を用いて論じる．

1. 序

漢字辞典の便利な索引手段として使用される四角号碼が，たとえば，中国語の拼音入力の補助情報としても利用しうることから，四角号碼を数学的に基礎づけるため，文献 1) では，四角号碼を伴う漢字の抽象的モデル—字形モデルと称した—が導入された．

字形モデルの特徴は，すべての漢字を 14 種の字形類に類別し，縦横2種の並置演算と6種の包摂演算を定義したことであった．漢字の四角（すみ）に付される番号は，必ずしも特定の角位置に固縛されるものばかりではなく，演算の結果，別の角位置へ移動するものもあった．たとえば，

厂厂^{8}人＝仄$_{8}$

では，8は‘人’の左上角から‘仄’の右下角へ移動する．その際，一般に，浮動点は，欠番があれば，第1（左上），第2（右上），第3（左下），第4（右下）の順で可能な限り第1角またはそれに近い角位置を占めようとする性質をもっていた．この角順序（第1角＞第2角＞第3角＞第4角）によると，たとえば，字形類  に属する^{2}次^{7}のように，第4角が欠番になり，8は第3角に繰り上がるといった現象が生じる．

一方，同じ文献 1) で言及された声形法による中国語漢字入力では，与えられた漢字 a（その四角号碼を $pqrs$ とすると，$^{p}_{r}a^{q}_{s}$）に対して四文字から成る符号 ‘$\alpha\beta ps$’（ただし，α は a の声母を表す記号，β は a の韵母を表す記号である）を入力符号として使用した．これを通常の英数字鍵盤上で実現すると，たとえば，

$^{5}_{-}$操$^{6}_{8}$（cao）に対して CE59

（Cは‘c’を表し，Eは‘ao’を表す）

この声形符号の形部に四角号碼の第1角と第4角番号が用いられた理由は，多くの字が偏と旁，冠と脚から成る構造をもつので，こうすることにより各対の両部分の情報を含むことができるからである．

一般に，中国文または日本文の構成単位である語は，漢字1字から成るものだけでなく，2字以上を含む熟語であることが多い（日本文では，さらに仮名が混じる）．そのため，入力の最小単位として漢字1字だけでなく2字以上から成る熟語をも扱えるようにするのはきわめて自然である．また，欧米の文章と異なり，中国文および日本文の書法上の特徴は，横書き（現在では両者ともに左から右へと書かれる）の外に縦書きという形式をもつことである．

そこで，本稿では，四角号碼の対象を単体としての漢字だけでなく，横書きおよび縦書きの熟語をも含むように拡大する．このためには，横書き熟語化（演算記号‘‖’を用いる）と縦書き熟語化（演算記号‘//’を用いる）の2種の新しい演算が導入されなければならない．たとえば，漢字‘帆’および‘船’から演算‖と//によって熟語‘帆船’および‘$^{帆}_{船}$’が得られるものとし，四角号碼の生成規則は，体系全体の複雑さを増大させないようにするため，それぞれ横並置演算｜と

† A System of Kanji Codes Making Use of Reading and Four-Corner Coding by KIMIO KAWAGUCHI-IZAWA (Department of Computer Science and Communication Engineering, Nagoya Institute of Technology), SIHONG WANG (Beijing Research Institute of Chemical Industry), SATOSHI IKAWA (Department of Computer Science and Communication Engineering, Nagoya Institute of Technology) and SEIICHI UNO (Research and Development Center, Technical Operations Headquarters, Sanyo Electric Co., Ltd.).

†† 名古屋工業大学工学部情報工学科
††† 北京化工研究院
†††† 三洋電機(株)技術本部開発研究所

縦並置演算／に対するそれと同じであるとする．すなわち，

$$\begin{matrix}1\\2\end{matrix}\text{帆}\begin{matrix}7\\1\end{matrix} \,/\!/\, \begin{matrix}2\\4\end{matrix}\text{船}\begin{matrix}8\\6\end{matrix} = \begin{matrix}1\\2\end{matrix}\text{帆船}\begin{matrix}8\\6\end{matrix} \quad \text{（横書きの‘帆船’）}$$

$$\begin{matrix}1\\2\end{matrix}\text{帆}\begin{matrix}7\\1\end{matrix} /\!/ \begin{matrix}2\\4\end{matrix}\text{船}\begin{matrix}8\\6\end{matrix} = \begin{matrix}4\text{帆}^{7}\\4\text{船}_{6}\end{matrix} \quad \text{（縦書きの‘帆船’）}$$

声形符号の形部として用いるのは第1角番号と第4角番号である．上記の帆船の例では，横書きの場合も縦書きの場合も同じく 46 である．ところが，多くの熟語では，第4角番号が互いに異なるという事態が生じる．たとえば，‘次女’は，縦書きの場合，‘$^{3}_{4}$姿$^{7}_{-}$’と同じく，$\begin{matrix}3\text{次}^{7}\\4\text{女}_{-}\end{matrix}$ となるから，第4角番号は0（欠番）であり，一方，横書きの場合，‘$^{3}_{1}$次女$^{4}_{4}$’となるから，第4角番号は4である．多数の熟語に対して声形符号の形部が縦書きと横書きとでは異なるということは，システムの実現の際に作成されるシステム内部の辞書に数多くの余分なエントリを設けなければならないことを意味する．その結果，同符号の語の数が増大して変換率の低下を招くことになる．さらに，声形符号の利用者にとっても大いに混乱の原因となる．辞書構造の複雑化とともにこれらはどうしても避けたいことである．そこで，本稿では，いかなる熟語に対しても，横書きであるかまたは縦書きであるかにかかわりなく，つねに同一の第1角および第4角番号を生成することが保証される新しい四角号碼法を導入する．この新四角号碼法では，角順序が 第1角＞第4角＞第2角＞第3角 となることが特徴である*．

漢字単体だけでなく，2字以上から成る漢語，さらには送り仮名を伴う（語尾変化がありうる）語をも扱うため，本稿では，声形符号の長さを固定しない．とくに形部については，対応する漢字の集合の大きさによっては2桁の 10 進数では不十分となり，4桁であるほうがよい場合もありうる．しかし，2字の熟語に対しては，2桁でおおむね十分であること，等が示される．さらに，声部の表記法についても，一つに固定しえない．中国語と日本語では異なるし，また，日本語でも，原文が現代仮名遣いによる場合と歴史的仮名遣いによる場合とでは異なるからである．

計算機で漢字を扱うとき，漢字の字種が極端に多く，また，漢字の字形が体系化しにくいため，入力時に特定の字を明確に指示する簡単なコーディング・システムが見いだしにくいのが最大の問題である[7]とされている．本稿の声形符号は，とくにこの問題のおおよその解決を意図したものである．

* 新四角号碼法は，漢字および漢語の計算機への入力の便宜のために考えられたものであり，四角号碼の元来の用途である漢字辞典索引の手段として旧来の四角号碼法に取って代わろうとするものではない．

2. 新四角号碼法

文献 1) で与えられた四角号碼法を一部変更することにより新四角号碼法を得ることができる．

改変の根本は，漢字の四つの角の筆形をとる順序，すなわち，角順序の変更である．旧四角号碼法では

第1角＞第2角＞第3角＞第4角

であった角順序が，新四角号碼法では

第1角＞第4角＞第2角＞第3角

とされる．新四角号碼法では，この順序に従って漢字の四角が取られ，対応する筆形番号が付されることになる．

ここでは，四角号碼作成法の細則を列挙することによらないで，各字形類[1]について新旧を対照させることにより新四角号碼法を定義する．

〔新旧四角号碼の対照〕

（1） 単 心 類

この類については新旧ともに同じである．

	旧	新
(字形類)	$i_{\circ}$	$i_{\circ}$
(漢字例)	$^{5}_{-}$中$^{5}_{-}$	$^{5}_{-}$中$^{5}_{-}$

（2） 二 心 類

類 $\overset{i}{\underset{j}{\circ}}$, $i\!\circ\!\!-\!\!\circ_{j}$, $\overset{i}{\circ}\!\!-\!\!\overset{j}{\circ}$ については新旧ともに同じであるが，類 $\overset{i}{\circ}\!\!-\!\!\overset{j}{\circ}$, $\overset{i}{\underset{j}{\circ}}$, $i\!\circ\!\!-\!\!\circ_{j}$ に属する漢字については変更があり，それぞれ記号 $\overset{i}{\circ}\!\!\diagdown\!\!\circ_{j}$, $\overset{i}{\circ}\!\!\diagup\!\!\circ_{j}$, $\overset{i}{\circ}\!\!-\!\!\circ_{j}$ で示される字形類を形づくる．とくに，$\overset{i}{\circ}\!\!\diagdown\!\!\circ_{j}$ の浮動点 i は第1角と第3角を浮動し，浮動点 j は第2角と第4角を浮動しうる．同様に，$\overset{i}{\circ}\!\!\diagup\!\!\circ_{j}$ の浮動点 i は第1角と第2角を浮動し，浮動点 j は第3角と第4角を浮動しうる．

	旧	新
a）(字形類)	$\overset{i}{\circ}\!\!-\!\!\overset{j}{\circ}$	$\overset{i}{\circ}\!\!\diagdown\!\!\circ_{j}$
(漢字例)	$^{2}_{-}$旧$^{6}_{-}$	$^{2}_{-}$旧$^{-}_{6}$
b）(字形類)	$\overset{i}{\underset{j}{\circ}}$	$\overset{i}{\circ}\!\!\diagdown\!\!\circ_{j}$
(漢字例)	$^{4}_{9}$木$^{-}_{-}$	$^{4}_{-}$木$^{-}_{9}$
c）(字形類)	$i\!\circ\!\!-\!\!\circ_{j}$	$\overset{i}{\circ}\!\!\diagdown\!\!\circ_{j}$
(漢字例)	$^{2}_{-}$白$^{6}_{-}$	$^{2}_{-}$白$^{-}_{6}$
d）(字形類)	$\overset{i}{\circ}\!\!-\!\!\overset{j}{\circ}$	$\overset{i}{\circ}\!\!-\!\!\overset{j}{\circ}$

(漢字例) $^{5}_{-}$戈$^{3}_{-}$ $^{5}_{-}$戈$^{3}_{-}$

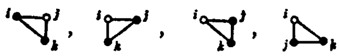
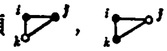
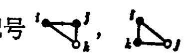
e) (字形類) 

(漢字例) $^{4}_{3}$寸$^{-}_{-}$ $^{4}_{3}$寸$^{-}_{-}$

f) (字形類) 

(漢字例) $^{4}_{-}$大$^{-}_{3}$ $^{4}_{-}$大$^{-}_{3}$

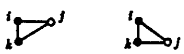
(3) 三 心 類

類  については新旧ともに同じであるが，類  に属する漢字については変更があり，それぞれ記号  で示される字形類を形づくる．

旧 新

a) (字形類) 

(漢字例) $^{2}_{2}$夕$^{7}_{-}$ $^{2}_{-}$夕$^{7}_{2}$

b) (字形類) 

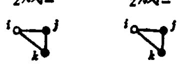
(漢字例) $^{4}_{9}$相$^{6}_{-}$ $^{4}_{9}$相$^{-}_{6}$

c) (字形類) 

(漢字例) $^{2}_{-}$失$^{5}_{3}$ $^{2}_{-}$失$^{5}_{3}$

d) (字形類) 

(漢字例) $^{5}_{2}$成$^{3}_{-}$ $^{5}_{2}$成$^{3}_{-}$

e) (字形類) 

(漢字例) $^{2}_{-}$片$^{2}_{2}$ $^{2}_{-}$片$^{2}_{2}$

f) (字形類) 

(漢字例) $^{6}_{2}$兄$^{-}_{1}$ $^{6}_{2}$兄$^{-}_{1}$

(4) 四 心 類

この類については新旧同じである．

旧 新

(字形類) 

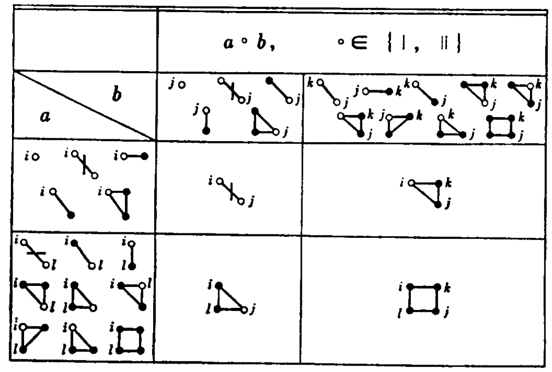
(漢字例) $^{2}_{3}$魚$^{7}_{3}$ $^{2}_{3}$魚$^{7}_{3}$

ここで述べた定義法は，与えられた漢字に対して，まず，旧四角号碼法によって旧四角号碼を求め，次に，上記の対照表を参照し，それが属する旧字形類に対応する新字形類から新四角号碼が正しく定められるということを意味する*．

* 実際には，この手順を踏まずに与えられた字から直接に新四角号碼を得る．馴れると瞬時にして得られる．われわれの実験[11]によれば，第1角および第4角番号の識別に限ると，馴れによって識別速度が飽和した時点で，7〜9秒/100文字，すなわち，0.07〜0.09秒/字であった．一般のキーパンチャの仮名入力速度 0.3〜0.6秒/タッチの1/8〜1/4の時間しか要しない．ちなみに，この時点で誤識別率は 0.01 以下であった．また，馴れるまでに，延べ7時間以上の訓練を行った．

最後に，新旧の対照表によってわかるように，差異は4桁の四角号碼を構成する各数字の占める桁位置が互いに入れ換わるだけであるから，新旧それぞれの担う情報に変わりはないことを注意しておく．

表1 横並置型演算
Table 1 Side by side concatenation operation.

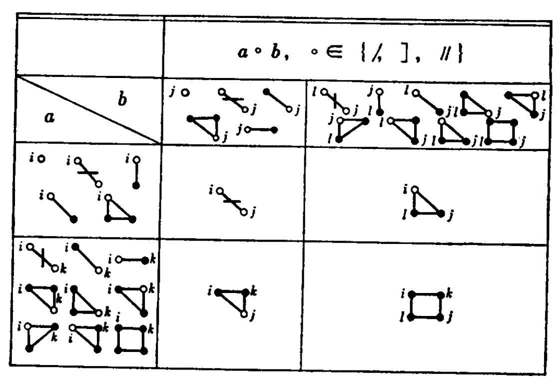


(例) 

($^{4}_{-}$大$^{-}_{3}$ ‖ $^{5}_{2}$成$^{3}_{-}$ = $^{4}_{-}$大成$^{3}_{5}$)

($^{2}_{3}$魚$^{7}_{3}$ | $^{8}_{-}$入$^{-}_{-}$ = $^{2}_{3}$魚入$^{-}_{8}$)

表2 縦並置型演算
Table 2 Above and below concatenation operation.



(例) 

($^{5}_{-}$損$^{6}_{8}$ // $^{2}_{-}$失$^{5}_{3}$ = ^{5}損^{6} / $_{5}$失$_{3}$)

($^{7}_{7}$匚$^{1}_{1}$] $^{6}_{-}$口$^{-}_{-}$ = $^{7}_{-}$回$^{1}_{6}$)

3. 演　算

一般に，漢字，および，それらから生成される横書きまたは縦書きの熟語は，その四角号碼 $ijkl$ を四角に添字として付加した英文字 ${}^i_k a^j_l$，または号碼の一部を省略した ${}^i_k a^j, \cdots, {}^i a^j, \cdots, {}^i a, \cdots, a_l$，または a などで表すものとする．とくに，欠番は - で表示する．

字または熟語を組み合わせて新しい字または熟語を生成する操作は，下記のように演算記号を用いて表現する．

2個の字をそれぞれ a, b と置くと，

（1） 横並置　$a|b$（第1項 a を偏，第2項 b を旁という）

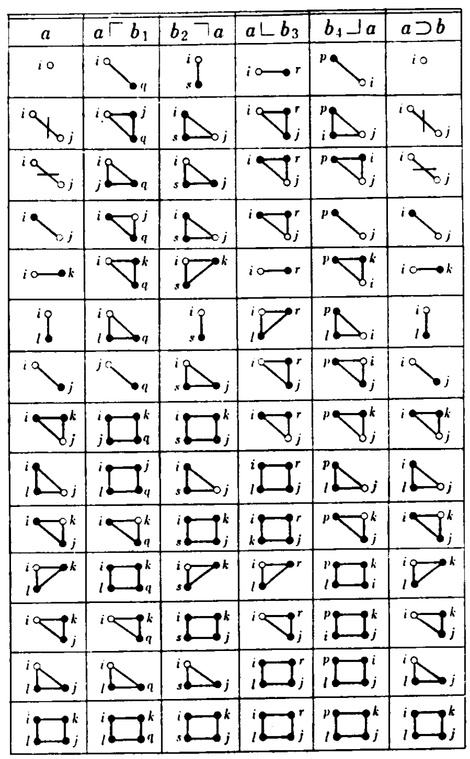
（例）　幵｜彡＝形

（2） 縦並置　a/b（第1項 a を冠，第2項 b を脚という）

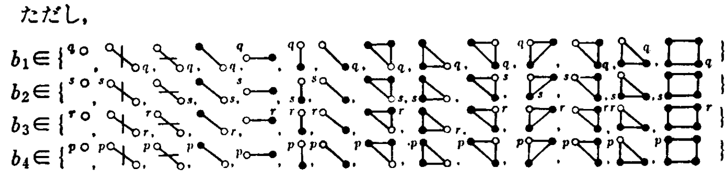
（例）　宀／子＝字

（3） 左上（からの）包摂　$a \ulcorner b$（第1項 a を

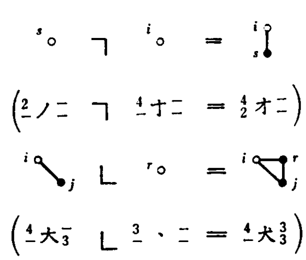
表3　包摂型演算
Table 3　Inclusion operation.



ただし，



（例）



垂という)

(例) 厂┌里=履

(4) 右上包摂 a ┐ b (第2項bを構という)

(例) エ┐弋=式

(5) 左下包摂 a ∟ b (第1項aを繞という)

(例) 廴∟聿=建

(6) 右下包摂 a ┘ b

(例) 氵┘十=斗

(7) 半包摂 a] b (第1項 a を構という)

(例) 匚]斤=匠

(8) 全包摂 $a \supset b$

(例) 冂⊃口=向

[注意] 半包摂と全包摂の相違については,文献1)を参照するとよい.

2個の字または横書きの熟語をそれぞれ a, b と置くと,

(9) 横書き熟語化 $a \| b$

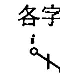
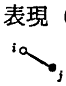
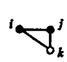
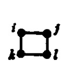
(例) 代表‖元=代表元

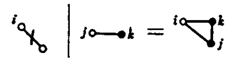
2個の字または縦書きの熟語をそれぞれ a, b と置くと,

(10) 縦書き熟語化 $a /\!\!/ b$

(例) 代表⫽元=代表元(縦書き)

上記の演算は与えられた字または熟語のすべての対に対して定義されるわけではないが,一般に字または熟語 a, b が演算∘によって字または熟語 c を合成するとき,$a \circ b = c$ と書く.

あらゆる字または熟語は,前章で示された 14 種の字形類に類別される.各字形類の表現(代表元)として記号 $i\circ$ (単心類),  , …,  (6種の二心類),  , …,  (6種の三心類), および  (四心類)が用いられる.とくに必要がないときには,各浮動点または固定点に付された筆形番号の一部または全部を明示しないことがある.これらの字形類の間の演算は,字形および四角号碼合成機能の類似性の存在から,横並置型,縦並置型,および包摂型の三つの群にまとめて,各類の代表元の間の演算という形式で**表 1〜3** に示される.たとえば,表1からは



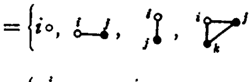
(^{8}从$_{\bar{6}}$|$^{5}_{-}$戈$^{3}_{-}$=$^{8}_{-}$成$^{3}_{5}$)

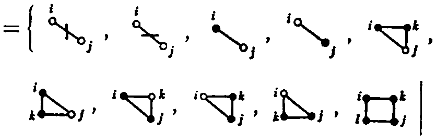
が読みとれる.

4. 対角号碼

前述の四角号碼の第1角番号と第4角番号が声形符号の形部に用いられる.とくにこの第1角番号と第4角番号を対角号碼と呼ぶ.対角号碼に関して次の諸性質が成り立つ.

字形類の集合を次のように定義する.

$C_1=\{i\circ,$  $\mid i, j, k \in \{0, 1, \cdots, 9\}\}$

$C_2=\{$  $\mid$

$i, j, k, l \in \{0, 1, \cdots, 9\}\}$

$C_0=C_1 \cup C_2$

また,演算の集合を次のように定義する.

$\Omega_1=\{|, \|, /, /\!\!/,], \ulcorner, \lrcorner\}$

$\Omega_2=\{\llcorner, \supset\}$

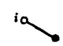
$\Omega_3=\{\urcorner\}$

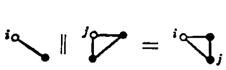
$\Omega_0=\Omega_1 \cup \Omega_2 \cup \Omega_3$

表 1〜3 から直接に以下の定理が得られる.

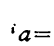
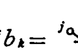
[定理 1] 1) 任意の ${}^{i}a \in C_0$ (第1角番号が i であるような C_0 の元),${}^{j}b_{-} \in C_1$ (第1角番号が j, 第4角番号が欠番であるような C_1 の元),$\circ \in \Omega_1$ に対して,${}^{i}a \circ {}^{j}b_{-} = {}^{p}c_q$ (${}^{p}c_q \in C_0$ (第1角番号が p, 第4角番号が q であるような C_0 の元))が成り立つならば,$p=i$, かつ $q=j$ である.

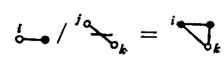
2) 任意の ${}^{i}a \in C_0$, ${}^{j}b_k \in C_2$, $\circ \in \Omega_1$ に対して,${}^{i}a \circ {}^{j}b_k = {}^{p}c_q$ が成り立つならば,$p=i$, かつ $q=k$ である.

(例) 1) ${}^{i}a=$  , ${}^{j}b_{-}=$  , $\circ=\|$ の場合,



であるから,${}^{p}c_q=$  , すなわち $p=i$, $q=j$ となる*.

2) ${}^{i}a=$  , ${}^{j}b_k=$  , $\circ=/$ の場合,



であるから,$p=i$, $q=k$ となる.

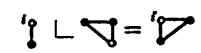
[定理 2] 1) 任意の ${}^{i}a_{-} \in C_1$, $b \in C_0$, $\circ \in \Omega_2$ に対して,${}^{i}a_{-} \circ b = {}^{p}c_q$ が成り立つならば,$p=i$, $q=-$

* 表1にある具体例を参照せよ.以下の諸例についても同様.

である.

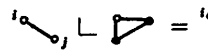
2) 任意の ${}^{i}a_j \in C_2$, $b \in C_0$, $\circ \in \Omega_2$ に対して, ${}^{i}a_j \circ b = {}^{p}c_q$ が成り立つならば, $p=i$, $q=j$ である.

(例) 1) ${}^{i}a_- =$  , $b =$  , $\circ =$ の場合,



であるから, $p=i$, $q=-$ となる.

2) ${}^{i}a_j =$  , $b =$  , $\circ =$  の場合,

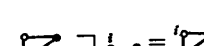


であるから, $p=i$, $q=j$ となる.

[定理 3] 1) 任意の $a \in C_0$, ${}^{i}b_- \in C_1$, $\circ \in \Omega_3$ に対して, $a \circ {}^{i}b_- = {}^{p}c_q$ が成り立つならば, $p=i$, $q=-$ である.

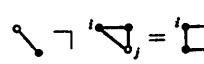
2) 任意の $a \in C_0$, ${}^{i}b_j \in C_2$, $\circ \in \Omega_3$ に対して, $a \circ {}^{i}b_j = {}^{p}c_q$ が成り立つならば, $p=i$, $q=j$ である.

(例) 1) $a =$  , ${}^{i}b_- =$  , $\circ =$ の場合,



であるから, $p=i$, $q=-$ となる.

2) $a =$  , ${}^{i}b_j =$  , $\circ =$  の場合,



であるから, $p=i$, $q=j$ となる.

とくに, 横書きおよび縦書きの熟語の対角号碼に着目すると, 定理1から次の系が得られる.

[系] 熟語の対角号碼は, 横書きであるかまたは縦書きであるかにかかわりなく一意に定まる.

(例) (横書き) 織 ‖ 女 $= {}^{2}$織女$_{4}$

(縦書き) 織 // 女 $= \dfrac{{}^{2}\text{織}}{\text{女}_{4}}$

この系は, 熟語の声形符号を考えるとき, その熟語を横書きで心に浮かべようと, あるいはまた縦書きで心に浮かべようと変わりはないことを保証している.

なお, 以上の諸定理を具体的な漢字に適用することは, 実際にはきわめて簡明であることを強調しておく.

5. 声　部

声形符号の声部には, 漢字, 漢語, ならびに漢字を含む語の '読み' を符号化したものが用いられる.

読みの正書法としては, 中国語の拼音表記法[3], 日本語のローマ字綴り方, 現代および歴史的仮名遣い, 等がある. これらに使用される記号の種類は, 漢字に比べるときわめて少ない. したがって, 用途に応じてこのようなものを声部に用いるのは適当である.

(1) 中国語の拼音表記は, 四声記号 (¯, ´, ˇ, ˋ) を伴うが, 広大な中国にあっては, 標準の四声は地域によって実情に合わないことがあるため, これは声部の表記からは排除するほうがよい.

拼音首字は, 声母 21 種 (b, p, m, f, d, t, n, l, g, k, h, j, q, x, zh, ch, sh, r, z, c, s) に 'y', 'w' とさらに 'a', 'e', 'o' を加えて, 26種が用いられる (ちなみに, 'i', 'u', 'v' を首字とする拼音表記は存在しない). また, 拼音後部は, 36種の韵母 (a, o, e, i(−i と i は同一視する), er, ai, ei, ao, ou, an, en, ang,

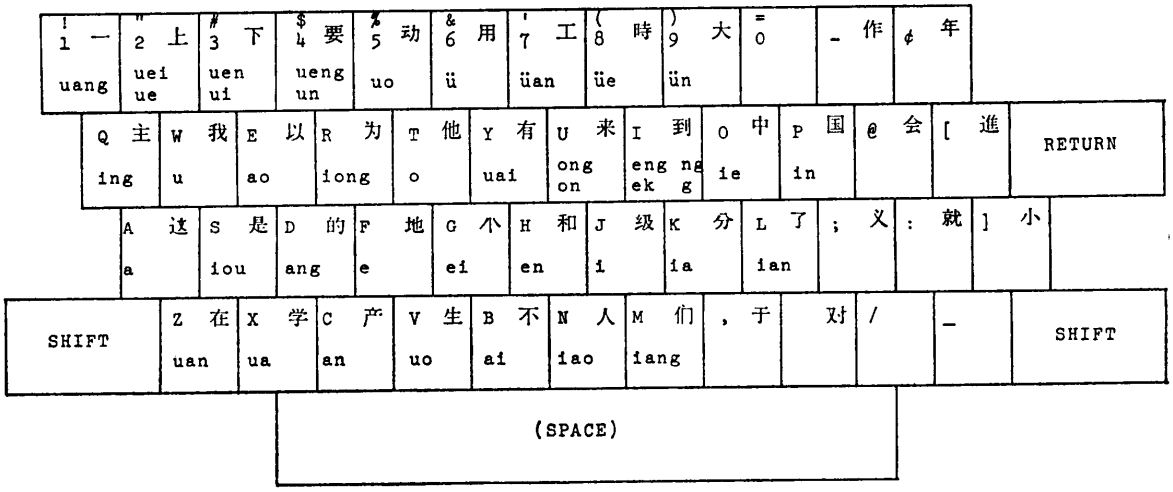


図 1 鍵盤
Fig. 1 Keyboard.

表 4 'こう' に対応する漢字の分布
Table 4 Distribution of Kanji characters for 'Kou'.

対角	こ			う			かう	対角	こ			う			かう
号碼	かう	こう	くわう	かふ	こふ	合 計	かふ	号碼	かう	こう	くわう	かふ	こふ	合 計	かふ
01	3	1	2			6		53		1				1	
02	9		1		1	11 (3)		54	3					3	
03	3	1	3		1	8 (2)		55	1	2		1		4 (1)	
04	7					7 (2)	1	56		2		1		3	
05	1					1		57				1		1	
06	1	1				2		58	1					1	
08	2	1	1			4		59	1					1	
11	4	5				9 (2)		60		1				1 (1)	
12	6	1	1			8 (2)		61	2	1	4	1		8	
13		2	1			3		62	4	1				5	
14	3	1				4 (2)		63	1	1				2	
15	2					2 (2)		64	5	1				6 (1)	
16				1		1		65	1	2		2		5 (1)	
18	2	2	3			7 (2)	1	66	1			2		3	
21	7	4	7			18 (3)		67		1				1	
22	16	6	1			23 (5)		68	1	1	1			3	1
23	5	3	1		1	10 (3)		69	1					1	
24	11	1				12 (1)		71	6		2	1		9	
25	3			1		4		72	1	1				2	
26	2			1		3 (1)		73		2	1			3	
27	1					1		74	2	1				3 (1)	
28	1		1			2		75	5			2		7 (1)	
29			1			1		76		1		1		2 (1)	
31	3	1	3	1		8 (2)		77	1					1	
32	2	2				4		78	1	2	1			4 (1)	
33		2	3			5		81	3	1	2	2		8	1
34		1				1 (1)		82	4	2		1		7 (1)	1
35	2	2				4		83	4	4	2			10 (2)	
36	4			2		6	1	84	3					3	
38	1	1	1			3 (1)		85		1		1		2	
41	7	2	4	3		16 (2)		86		1				1	
42	10	6	1	1	1	19 (1)		88			2	1		3	
43		2				2		91	2	2	7			11 (3)	
44	10	3				13 (5)		92	2					2	
45	1	3		2		6		93	3	1				4	
46	1	3				4		94	2					2	
48		1	5			6 (1)		95	1					1	
49	1					1		96				1		1	
51	4	2	1			7 (2)		98		2				2	
52	1	1	1			3 (1)			198	96	64	30	4	392 (60)	6

ただし，合計欄の括弧内は常用および人名用漢字数.

eng, ong, ia, iao, ie, iou, ian, in, iang, ing, iong, u, ua, uo, uai, uei, uan, uen, uang, ueng, ü, üe, üan, ün) が使用されるが，さらに，'on', 'ue', 'ui', 'un', 'ng', 'g' をも加え，42 種とする.

拼音首字および後部記号の合計は 68 種であるが，実用上は，これらを入力用鍵盤の各鍵に割り当てればよい．割当ての実際例を図 1 に示す．なお，鍵盤上に示されている 44 の漢字は，最も使用頻度の高い字を 1 打鍵で入力するためのものである.

(2) 日本語のローマ字綴り方には，訓令式，日本式，標準式，その他があるが，いずれも声部の表記法として使用することが可能である．仮名による場合

は，現代仮名遣い，または歴史的仮名遣いが用途* に応じて使用可能である．

ローマ字綴り方は，ある程度は音韻表記の性格をもち，またある程度は現代仮名遣いの字ごとの符号変換となるように設計されているので，現代仮名遣いとともに，記号列自体が担っている情報量が豊富でないため，一つの記号列に対応する漢字（語）の数が多いという欠点がある．この点は，歴史的仮名遣いによればある程度は改善される．たとえば，表4に示されるように，'こう' で表される漢字は，文献 2) では，392個**に上るが，歴史的仮名遣いで表記すれば，'かう' が198個，'こう' が 96 個，'くわう' が64個，'かふ' が30個，'こふ' が，'恋ふ' のように送り仮名の部分をも含めると，4個というように細分される．なお，このときには，'かふ' には上記の他に '沽ふ' などの6個を追加しなければならない．

しかしながら，結局，どの表記法を用いるかは，用途によって決めるべきである．

6. 声形符号の識別能力

声形符号による入力は，漢字1字，2字以上から成る漢語，また，漢字仮名混り語を単位として行う．

（1） 漢字1字の入力

a） 中国語の場合，形部に旧四角号碼の第1角および第4角番号を用いる声形符号によって，第1級および第2級漢字6,763字を含む数千字のなかから目的の漢字が 0.95 以上（技術文献，'計算机操作系統' の一部をテキストとして用いた場合，0.990）の確率で認識（変換）されることが文献 1) で示されている．本稿では，熟語に対してその対角号碼を横書きか縦書きかにかかわりなく一意に定めようとする立場から，新四号角碼を用いることを仮定している．しかし，その結果，たとえば，'到' と '刀'(拼音は両者とも 'dao') の対角号碼は旧四角号碼法によれば10と12であったのが，両者とも 12 になって区別を失う．しかも，使用頻度は，文献 10) によれば21，629，372字中に '到' は88,008回（第28位），'刀' は 11,500 回（第436 位）出現するから，両者とも高く，その上，両者とも熟語を作らずに単独で出現することが多い．このような新たな不都合も生じうるし，また，旧四角号碼法で対角号碼が同じであったものが新四角号碼法では分離されて改善されるといった例も見いだされる．そこで，実際の文章によって識別能力を調べると，文学の例（'十渡絮語' から）では，756 文字の文章に対して誤変換数 10，変換率 0.987，また，政治文献の例（'中华人民共和国憲法' から）では，498 文字の文章に対して誤変換数 10，変換率 0.980 という良好な結果が得られた．したがって，識別能力という観点だけから判断すると，新旧四角号碼法のいずれを用いてもほぼ同等であり，熟語に対して横書き縦書きの別なくその対角号碼を一意に定められるという点では，新四角号碼法でなければならないということがわかる．

b） 日本語の場合，声部に字音を用い，形部には2桁の対角号碼あるいは必要とあれば4桁の四角号碼を使用する．

最悪のケースとして字数が最も多い（392 字）'こう'（表 4）を例にとる．392字の全部を対象とする場合，4桁の四角号碼を形部に用いるならば，'こう 2192' によって字 '絎' が一意に決定されるというように識別能力は向上する．

普通は，用途に応じてもっと少ない字数の部分集合に対象が限定されよう．たとえば，常用および人名用漢字（60字）に限るならば，2桁の対角号碼を用いても，'こう 42' により '考' が一意に決定されるというようによい識別能力が保てる．

さらに，字の使用頻度のばらつきの大きいことを利用して，同符号の字を使用頻度順に配列することにより識別能力をより向上させることができる．

また，対象となる漢字数が多い場合でも，歴史的仮名遣いを声部に使用してよい場合は，2桁の対角号碼を形部に用いても識別能力を高水準に保てる．392字（および常用・人名用漢字 60 字）を対象としたときの適中率* をそれぞれ示すと，現代仮名を用いる場合，従来の漢字指定方式（A）では，0.0026 (0.0167)，声形法（B）では，0.201 (0.550) であるのに対して，歴史的仮名を用いる場合は，Aで 0.0128 (0.0833)，Bで 0.426 (0.750) である．常用・人名漢字を歴史的仮名遣いを用いて声形符号で入力すると 75.0% の適中率を示し，一方，これを現代仮名を用いて漢字指定方式で入力すると適中率は 1.67% に低下する．

（2） 2字以上から成る漢語の入力

たとえば，'普遍妥当性' という5字から成る語を '普遍'，'妥当'，および '性' に分けても，それぞれが一般に通用する語，すなわち，辞書に記載されている

* たとえば，第二次大戦以前の文章を入力する場合は歴史的仮名遣いで思考しつつ入力するのが自然であろう．

** 俗字等の異体をどう扱うかで数が変わり，個数を定めることにも困難が伴う．

* 入力符号を入力して目的の漢字が得られる確率．

表 5 'shi' を第1音節とする2音節熟語の分類
Table 5 Classification of disyllabic words with 'shi' as the first syllable.

k_{shi}	$p(k_{shi})$	$k_{shi}\times p(k_{shi})$	第2音節拼音
19	1	19	shi
14	1	14	ji
11	2	22	li, yi
9	2	18	zhi, wu
7	2	14	yan, zi
6	8	48	以下省略
5	10	50	
4	18	72	
3	16	48	
2	38	76	
1	93	93	
	合計 191	合計 474	

ただし，拼音を同じくする語が k 個存在するような2音節の拼音は類 k に属するといい，とくに第1音節が α であるとき，類 k のサブクラス k_α に属するという．サブクラス k_α に属する拼音の数を $p(k_\alpha)$ で表す．したがって，$k_\alpha\times p(k_\alpha)$ はサブクラス k_α に対応する語の個数である．たとえば，shi-shi に対応する語は全部で19個ある．

表 6 熟語 'shi shi'
Table 6 Homonymous words for 'shi shi'.

shi shi		shi shi	
熟語	対角号碼	熟語	対角号碼
失勢	22	史实	53
失实	23	事实	
失时	24	史诗	54
失事	25	事事	55
实施	31	时势	62
适时	34	时式	64
视事	35	时时	
实事		时事	65
世事	45	食事	85
誓师	52		

語である．このように漢語は2字(および残余の1字)の語に分離することができる*．もともと，熟語は2字で造られるのが基本であるからである．しかし，たとえば，'神無月' というような和語については，2語に分離しないまま扱うほうが無理がない*．しかも，まったく同じ読みをもつ語が他に存在しないのが普通であるから，形部が空のままでも識別できることが多い．

次に，2字から成る漢語について考察を進める．

a) 中国語の場合．文献 3), 4), 5) によれば，2音節（2字）から成る熟語では，'shi' という音節で始まるものが最も多い．この 'shi' に対応する首字の数

表 7 'こう' を第1字とする熟語の分類
Table 7 Classification of words whose first character is 'kou'.

$k_{こう}$	$p(k_{こう})$	$k_{こう}\times p(k_{こう})$	第2字仮名表記	$k_{こう}$	$p(k_{こう})$	$k_{こう}\times p(k_{こう})$	第2字仮名表記
47	1	47	しょう	15	2	30	がい，ふ
43	1	43	し	14	4	56	きょう，だいしょ，ほう
39	1	39	こう				
31	1	31	とう	13	4	52	いん，てんしょく，よう
30	1	30	せい	12	11	132	以下省略
29	1	29	き	11	9	99	
28	2	56	しん，そう	10	7	70	
27	2	54	かん，せん	9	8	72	
26	2	52	えん，てい	8	8	64	
24	2	48	じょう，りょう	7	13	91	
22	2	44	か，じ	6	14	84	
21	1	21	ちょう	5	19	95	
20	3	60	い，けい，しゃ	4	30	120	
19	1	19	きゅう	3	21	63	
18	2	36	ぎ，ち	2	51	102	
17	5	85	かい，しゅげん，じん，ひ	1	355	355	
16	2	32	どう，ばい		合計 586	合計 2,211	

ただし，欄見出し記号の意味は表5と同じである．

* 文献 3) に記載されているすべての語を調べた限りでは，一部の固有名詞を除くと，すべてそのように分解できる．

* 文献 6) には '神無' ('かみな', 'かんな', または，'かみなし') は記載されていない．'無月' は 'むげつ' として記載されているが，'神' と '無月' (むげつ) に分離するのは少々強引な感じがする．

は全部で57であり，熟語の総数は613語に上る．このうち，2音節語は474，3音節語は64，4音節語は75を数える．474語の2音節語には，191種の第2音節が含まれる（表5）．第2音節が'shi'である'shi shi'という熟語が最も多く，表6に示すように19語がある．次に多いのは'shi ji'という語で14語がある．

最も多い'shi shi'という音節をもつ熟語に対して声形符号の識別能力を見ると，表6に示されているように，'shi shi 35'，'shi shi 53'および'shi shi 64'にはそれぞれ2語が対応するけれども，その他の語は一意に定まる．これは非常によいということができる．

表8　熟語'こうしよう'
Table 8　Homonymous words for 'Kousyou'.

こうしよう			こうしよう		
熟語	歴史的仮名表記	対角号碼	熟語	歴史的仮名表記	対角号碼
高姓 康正	かうしゃう	01	綱掌	かうしゃう	25
高尚† 康尚 高商	かうしゃう	02	行省 紅晶	かうしゃう こうしゃう	26
講誦 交鈔 交渉†	かうしょう かうせう かうせふ		行賞	かうしゃう	28
			洪鐘	こうしょう	31
高昌 高唱†	かうしゃう	06	考証† 黄鐘	かうしょう くわうしょう	41
高声	かうしゃう	07	好尚	かうしゃう	42
高蹤 講頌 交睫	かうしょう かうせふ	08	咬傷 甲匠 口誦	かうしゃう かふしゃう こうしょう	62
交床	かうしゃう	09	口承 哄笑†	こうしょう こうせう	63
巧匠 工匠 工商	かうしゃう こうしゃう	12	降将	かうしゃう	74
			厚相	こうしゃう	76
			厚賞	こうしゃう	78
巧笑	かうせう	13	公証	こうしょう	81
工廠†	こうしゃう	14	公傷	こうしゃう	82
行粧 後証	かうしゃう こうしょう	21	公相 公妃†	こうしゃう	86
翱翔	かうしゃう	22	公称† 鉱床	こうしょう くわうしゃう	89
行障 後章	かうしゃう こうしゃう	24	斯う為よう	かうしよう	

ただし，漢字体は常用漢字による．また，† 印は文献12)に記載されていることを表す．

b）日本語の場合．現代仮名遣い'こう'で始まる熟語は，文献6)によれば，全部で2,211語に上る．このなかで，2字熟語は1,879語，3字熟語は229語，4字熟語は57語，5字熟語は8語あるが，漢字と送り仮名から成る語も35語含む．現代仮名遣いによってこれらを分類すると，表7のようになるが，このなかで'こうしよう'という熟語が最も多く，47語があり（さらに'斯う為よう'という語を加えると，48語），次に多いのは'こうし'で43語がある．

'こうしよう'と現代仮名遣いで表記される語は最も多いが，これに対する声形符号の識別能力を見ると，表8に示されるように'こうしよう 02'には6語，'こうしよう 08'と'こうしよう 62'には3語が対応している．その他には2語または1語が対応している．しかし，個々の語を検討すると，日常的に用いられる語は少なく，したがって，用途によって一部のものしか対象とならないため，実際には良好な識別能力を発揮することになる．

もしも，声部に歴史的仮名遣いが用いられるならば，対象がかなり広くても識別能力は大きい，'こうしよう'47語（および，日常語8語[12]（表8中の†印付きの語））を対象としたときの適中率をそれぞれ示すと，現代仮名を用いる場合，漢字指定方式(A)では，0.021 (0.125)，声形法(B)では，0.574 (0.875)であるのに対して，歴史的仮名を用いる場合は，Aで0.191 (0.750)，Bでは0.851 (1.000)である．47語を対象として歴史的仮名遣いで声形符号を入力すると85.1%の適中率であるのに対し，8語だけを対象としても現代仮名遣いで漢字指定を行うときは12.5%の適中率しか得られない．

さらに，'こうしよう'に対する上記の結果と'こう'に対する前述の結果を比較することにより，声形符号がとくに熟語に対してその良好な識別能力を発揮するということがわかる*．

（3）漢字仮名混じり語

日本語には特殊な表記法として漢字仮名混じり語がある．これらに対する声形符号は，たとえば，'明らか'という語に対しては漢字部分の対角号碼を形部に用いて'あきらか 62'というように書くことができる．同じ字訓をもつ漢字の数はあまり多くないのが普通であるが，それでも'あきらか'に対応する漢字は文献2)によれば50字に上る．これらは，声形符号

* 一般に仮名漢字変換法は，漢字1字に対してよりも熟語に対して高い識別能力を示す．

表 9 ‘あきらか’という字訓をもつ漢字
Table 9 Kanji characters for ‘Akiraka’.

あきらか		あきらか	
対角号碼	漢　字	対角号碼	漢　字
01	亮	61	昱, 晃, 晥
02	彰	62	昉, 明, 昺, 晑
04	章	64	昕, 晈, 暭, 曈
12	了	65	晟
18	耿	66	昌, 昭, 晤, 晶
22	的	68	顕, 顯
23	奐	69	杲, 暸
24	皎, 叡, 皦	75	闡
26	皓, 皛	76	冏
41	旭	92	灼, 炯, 炳
42	彬	93	煥
43	爽	94	焯, 焞
46	晢	96	炤
53	悊	98	熒
56	啓, 晢	99	燦

注) 四角号碼を用いると，啓と晢が同番号 5206，昺と晑が 6022，灼と炯が 9782 をもち，他はすべて互いに相異なる番号をもつ.

の形部によって表9のように識別することができる．字音に比べれば対応する字数が少ないため，字訓を声部に用いた声形符号の識別能力はより良好である．実際には，用途に応じて一部の字しか対象にならないから，さらによい．たとえば，‘あきらか’に対して{明, 顯}（多くの国語辞典では‘明’しか記載していない）だけに限定すると，‘あきらか 62’と‘あきらか 68’によってそれぞれ一意に決定される．

（4） 総合的識別能力

まず，一般日常語の集合の例として既存の単語辞書[12]の収録語 27,857 語（ただし，読みは現代仮名遣い）を調べた結果を述べる．辞書の見出し総数は 21,834 で，これはそのまま漢字指定方式での入力コード数となる．声形符号の数は 27,369 である．入力コード当りの語数はそれぞれ 1.28 語および 1.02 語となる．声形符号を用いれば，ほとんどの場合に1語に絞られることがわかる．同音（同符号）語総数も声形法を用いることにより 9,807 語から 955 語へと約 10 分の1に減少し，同音（同符号）語を構成する語の平均数は 2.91 語から 2.04 語へと減少する．したがって，識別率* は，漢字指定方式では 0.648 にすぎないが，声形法では 0.966 と著しく向上する．

次に，実際の文章小林秀雄著「人生について」の一部分を調査した結果を示すと，識別率は，漢字指定方式で 0.655，声形法では 0.983 であった．これは前述の結果とほぼ一致している．

以上から，漢字指定方式に比較して，声形法は著しく高い同音語識別能力を有しているということができる．

7. 結　語

日本語または中国語の文を計算機で処理するときにまず生じる問題は入力をどのようにして行うかである．

文献 7）では，計算機が漢字を扱うときの本質的な問題は入力において発生し，それは

ⅰ） 字種が極端に多いこと

ⅱ） 字形が体系化しにくいこと

にあり，その結果，特定の字を明確に指示する簡単なコーディング・システムができないとしている．

本稿で取り扱った声形符号は，この問題の解決を目差す一つの方法であるということができよう．

日本語の仮名漢字変換方式には漢字単位と単語（熟語）単位の変換があるが，とくに単語単位の変換では，変換率の向上のためのおもな問題点として

1） 分かち書きの方法

2） 同音異義（字）語の処理

が挙げられている（文献 8）等）．

1）に関しては，声形符号の実用向き実現[11] によれば，たとえば，「容易に習得でき，変換率が高い」という文は「3ようい2に1しゆうとく4でき，0へんかん 80 りつ4が0たか2い」と打鍵入力される*．

2）に関して，声形符号による入力では，字数の多い熟語をなるべく2字単位に切り離して扱うことにより，計算機内の辞書の膨大化を回避しつつ，さらに形部の助けによって識別能力（変換率）の向上を図っている．とくに同音異字語の問題については不完全とはいえほぼ解決されているといえる．

なお，補助情報を利用する仮名漢字変換法としてはこれまでにも文献 9）等があり，とくに文献 9）では字形情報を用いている．字形の分類法に関しては，これら従来の方法に比べ，声形法はさらに厳密かつ精密なものになっている．

* 単語辞書に収録されている語が入力符号により一意に識別される割合，すなわち $\left(1-\frac{同音語総数}{単語総数}\right)$.

* 前章までは，‘容易’の声形符号を‘ようい 32’というように表していたが，実用上の符号は‘3ようい2’とする．

参考文献

1) 川口喜三男，王 思鴻：漢字四角号碼の代数構造と‘声形法’による中国語漢字入力について，情報処理学会論文誌，Vol. 24, No. 4, pp. 521-530 (1983).
2) 諸橋轍次，渡辺末吾，鎌田 正，米山寅太郎：新漢和辞典，改訂版，大修館，東京 (1982).
3) 北京外国语学院英语系≪汉英词典≫编写组编：汉英词典，商务印书馆，北京 (1981).
4) 吉林大学汉日词典编辑部：汉日词典，吉林人民出版社，长春 (1982).
5) 新华词典编纂组编：新华词典，商务印书馆，北京 (1981).
6) 新村出編：広辞苑，第二版補訂版，第四刷，岩波書店，東京 (1979).
7) 高橋秀俊：特集「日本文入力法」の編集にあたって，情報処理，Vol. 23, No. 6 (特集：日本文入力法)，pp. 516-517 (1982).
8) 牧野 寛：カナ漢字変換入力法，情報処理，Vol. 23, No. 6 (特集：日本文入力法)，pp. 529-535 (1982).
9) 白鳥嘉勇他：特徴分類形かな漢字変換方式，電子通信学会論文誌，Vol. J 64-D, No. 8, pp. 773-779 (1981).
10) 汉字查频小组：汉字频度表，新京新华印刷厂，北京(1977).
11) 井川 智，川口喜三男：声形符号を用いた日本語漢字入力方式，昭和 59 年度電子通信学会総合全国大会講演論文集，分冊6，1510 (1984).
12) 沖電気(株)編：KJISYO. SYS, OKI IF 800 model 50 システムディスケット内 (1983).

(昭和 58 年 12 月 5 日受付)
(昭和 59 年 5 月 15 日採録)