

<研究報告>漢字字画による日本語漢字入力支援システムの提案と実装

メタデータ	言語: Japanese 出版者: 公開日: 2007-06-29 キーワード (Ja): キーワード (En): 作成者: 張, 小剛, 高, 建斌, 高橋, 勇, 黒岩, 丈介, 小高, 知宏, 小倉, 久和, ZHANG, Xiaogang, GAO, Janbin, TAKAHASHI, Isamu, KUROIWA, Jousuke, ODAKA, Tomohiro, OGURA, Hisakazu メールアドレス: 所属:
URL	http://hdl.handle.net/10098/827

漢字字画による日本語漢字入力支援システムの提案と実装

張 小剛* 高 建斌** 高橋 勇*** 黒岩 文介*** 小高 知宏*** 小倉 久和***

Japanese Character Input Auxiliary Method by Character Strokes

Xiaogang ZHANG *, Janbin GAO **, Isamu TAKAHASHI ***
Josuke KUROIWA ***, Tomohiro ODAKA *** and Hisakazu OGURA ***

(Received August 11, 2003)

The input system of the Japanese mainly is constructed on the kana character input or Japanese Roman alphabet input by Japanese Pronunciation. This system is difficult to operate for the foreign Japanese beginner who does not understand Japanese pronunciation, especially the pronunciation of the kanji's. In this paper in order to overcome the difficulty, we propose the system which support the kanji strokes input auxiliary method (the JCSIM). We built the Japanese input system using this JCSIM method. We performed the experiment to evaluate this system by three foreign Japanese beginners. The results showed that the time to input Japanese sentence was decreed by using this JCSIM proposed here compared with the normally Roman alphabet inputting method.

Key Words : Japanese Kanji, Input System, Kanji Strokes, Foreigner Japanese Beginner

1. はじめに

現在、コンピュータへの入力装置としてキーボード、ライトペン、タッチパネル、音声入力システムなど様々あるが、まだ、キーボードが最も有力な入力装置である^[1]。パーソナルコンピュータとインターネットの普及によって、日本人が英語をキーボード入力する機会が増加する一方で日本国内の外国人が日本語をキーボード入力する機会も増加している。また、日本語教育の現場でも、外国人が利用するための日本語ホームページの開設、あるいは、E-mail やイントラネットを利用した日本語作文指

導などが、外国人が日本語をキーボード入力することを前提として実施されている^[2]。

ところで、ほとんどの日本語入力システムは、キーボードで誤入力した場合への対応など「入力支援」機能を充実させて、使いやすくしている。しかし、外国人留学生にとってはこの支援機能は役立っておらず、外国人留学生に対する入力支援システムはまだ未開発といつてよい^[2]。

キーボードを用いた日本語入力方式には、主にカナ文字入力とローマ字入力がある^[1]。漢字は入力された漢字の読みカナを変換して入力するのが主流である。文字の読み方によるこの二種類の入力方式は、日本語の読み方が分かる人にとっては簡便であるが、日本語の読み方、特に日本語の漢字の読み方が分からない外国人日本語初心者にとっては大変難しい操作である^[3]。

漢字を含む言語を母語としない外国人留学生は、漢字や漢字単語の意味を覚える際に読みもいっしょに覚えなければならない。日本語文章を入力する際には、意味と読みの両方とも分かっている語だけ使うことになる。しかし、たとえば中国語を母語としている外国人留学生は日本語を勉強したことがなく

*大学院工学研究科情報工学専攻

**大学院工学研究科システム設計工学専攻

***知能システム工学科

*Information Engineering Course, Graduate School of Engineering

**System Design Engineering Course, Graduate School of Engineering

***Dept. of Human and Artificial Intelligent Systems

でも、日本語のほとんどの漢字や漢字を含む単語の意味が分るので、読みを知らなくてもそのような単語を使うことができる。このような漢字と単語をコンピュータに入力する場合、これらの留学生には漢字単語の意味が分かっているのに入力できないという問題が生じる。

一方、現在いろいろな日本語電子辞書が開発されている。外国人留学生が未知の単語の意味を調べるにはこれらの電子辞書がよく役立つ。しかし、そのためには、やはり、その読みが分からなければならない。留学生の多くは読みが分からない単語の意味を調べるとき、まず紙製の日本語辞書(漢字辞書)で読みを調べてから、次に読みを電子辞書に入力して、単語の意味を調べる。しかし、これは手間がかかる^[4]。われわれは、読みが分からなくても簡単に漢字を入力できる補助方式を検討してきた。

以上のような観点から、漢字圏の日本語初学者にとっては読み方が分からなくても少数のキー入力で、漢字を効率的に入力できる日本語入力方式が必要となる。

漢字の形は漢字の一つの重要な特徴である。特に漢字に馴染んだ中国人にとって漢字の形は漢字の最も理解しやすい特徴である。漢字の読みは予め覚えていないと、読みがいつまでも分からない。ところで、少し漢字の常識を持っている入力者は漢字の構成要素の解析ができる。漢字に馴染んだ中国人でなくても、少し漢字の常識を持っていれば、他の国の人でも漢字の形を理解できる。ここでは漢字の形で漢字を入力する方式を採用し、それによる入力補助を検討する。

本報告では日本語の読み方によらない、少数キーを用いる日本語漢字字画入力補助方式 JCSIM(Japanese Character Strokes Input Method) 法を提案し、提案システムの評価実験を行った結果を報告する。

2. JCSIM 入力方式による漢字の入力

2.1 漢字の形による入力方式

中国でよく利用されている漢字の入力方式は「ピンイン」(読み方)入力方式であるが、それ以外では「五筆字型」という入力方式がある。これは漢字の部首を用いた入力方式である。漢字の部首を用いた入力方式には次の利点がある。

- 一つの入力コードに二つ以上の漢字が定義されることは殆んどないので、入力者は候補から入力したい漢字を選択する手間が不要

である。入力者はディスプレイを見なくても、入力できる。

- 熟練すると、入力速度はかなり速い。速い人は一分間に二百以上の漢字を入力できる。

しかし、次のような問題点がある。

- 部首が多いので、覚えにくい。中国漢字の部首は百八十九個^[5]、日本語漢字部首は約二百四十個^[6]ある。入力者はこれらの部首を定義したキーを全て覚えなければならない。入力者に負担がかかる。
- 長時間使わないと忘れやすいから、また使う時にはもう一度覚えなおさなければならない。

文書処理を職業としている人にとっては少々覚えにくくても、熟練すると速く入力できる入力方法が良い方法である。しかし、たまに文書処理をするような人にとってはあまり有効ではないと考えられる。

漢字字画の種類は部首よりかなり少ない。そして、字画の構造も部首より単純で、分類しやすい。入力に漢字の字画を用いることにより、初学者にもわかりやすい入力方式とすることができる。JCSIM(Japanese Character Strokes Input Method) 方式はこの方式による日本語入力システムの補助システムとして用いる。

2.2 JCSIM 入力方式の概要

JCSIM 方式では、全ての漢字字画を六種類に分け、一つずつ、六つのキーを対応させる。入力する時、漢字を字画に分析して、漢字の書き順によって一つずつ字画と対応するキーをタイプすることで漢字を入力する。同じ入力コードに二つ以上の漢字が対応する場合は、それらの漢字に番号をつけて表示し、入力者は番号を入力して、漢字入力を終了する。

読みが分かっている漢字を入力するにはやはりローマ字とかな変換などの読みによる入力方式が便利である。JCSIM 方式は、補助入力方式として、読みが分からない漢字を入力する場合のみ使うこととする。

2.3 入力に用いる字画

入力に用いる字画を選択するために各字画の出現頻度についての調査を行った。EUC コードの漢字 282 文字を字画ごとに分析して各字画の出現頻度を求めた結果を表1に示す。出現頻度が高い字画「イチ」、「タテボウ」、「ヒダリハライ」と「ハネボウ」を入力に用いる字画

表1 字画出現頻度調査表

字画名称	字画	出現頻度
イチ	一	502
タテボウ	丨	348
ヒダリハライ	ノ	401
ミギハライ	ㄨ	303
ハネボウ	┘	321
テン	ㄨ	215
オレ	乙	205
ソノタ	ほか	101
合計		2369

表2 漢字字画定義表

字画名称	字画	定義
イチ	一	左から右への水平的な字画
タテボウ	丨	上から下への垂直的な字画
ヒダリハライ	ノ	右上から左下へまたは左下から右上への字画
ミギハライ及びテン	ㄨ	左上から右下への字画
ハネボウ	┘	字画“丨”の終点にはねがある字画(はねの方向と無関係)
ソノタ	ほか	上記以外のすべての字画

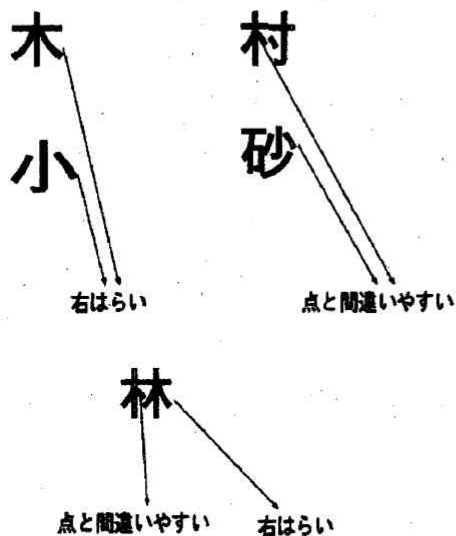


図1 漢字字画区別の例

として採用する。出現頻度が低い字画「オレ」と「ソノタ」を合併して「ソノタ」とする。

字画「ミギハライ」と「テン」は分類しにくい場合が多い。図1の例のように、漢字「木」と「小」の字画「ミギハライ」は漢字「村」「砂」の中では「テン」と間違われやすい。また、漢字「林」の右の「木」の最後の字画の正しくは「ミギハライ」であり、左の「木」の最後の字画も同じであるが、後者は「テン」と間違われやすい。ところで、字画「ミギハライ」と字画「テン」の出現頻度はともに低いのでこの二種類の字画を合併する。

これら六種類の字画の定義を表2に示す。

以上の六種類の字画に対し、表3のように、一つずつキー「S」、「D」、「F」、「J」、「K」、「L」を割り当てた。これらはキーボード中段左手の方の三つと右手の三つのキーであり、一番使いやすいキーである。

表3 漢字字画と入力キーの対応

字画名称	字画	キー
イチ	一	S
タテボウ	丨	D
ヒダリハライ	ノ	F
ミギハライ及びテン	ㄨ	J
ハネボウ	┘	K
ソノタ	ほか	L

入力者に負担をかけないために、JCSIM方式で一つの漢字を入力するための入力字画数を最大五画としている。入力方法は、字画数によって次の三つの場合に分ける：

1. 五画未満：漢字の書き順により、実際の字画キーをすべて入力してから、スペースキーを入力する。
2. 五画：漢字の書き順により、実際の字画キーをすべて入力する。
3. 五画以上：漢字の書き順により、最初の四画のキーを入力してから、最後の二画を入力する。

漢字の画数が五画以上の場合では初めの五画を入力するのではなく、初めの四画と最後の字画を選ぶが、その理由は同じ漢字コードに対応する文字を減らすためである。例えば、「言」、「設」、「記」、「計」の字画コードは初めの五画でコードすると、全て同じである。JCSIM方式では最初の四画と最後の二画で漢字を入力するから、「言」、「設」、「記」、「計」の字画コードはそれぞれ「一、一、一、一、一」、「一、一、一、一、ソノタ」、「一、一、一、一、丨」となり、同じ字画コードの漢字の数を減らせる。

2.4 入力規則

表4 入力字画例

	第一画	第二画	第三画	第四画	第五画
小	丿	ノ	丶		
倉	ノ	丶	一	ソノタ	一
研	一	ノ	丨	ソノタ	丨
究	丶	ノ	ソノタ	ノ	ソノタ
室	丶	ノ	ソノタ	一	一

表5 入力キー例

	第一画	第二画	第三画	第四画	第五画
小	K	F	J	スペース	
倉	F	J	S	L	S
研	S	F	D	L	D
究	J	F	L	F	L
室	J	F	L	S	S

同じ字画コードに多数の漢字を持つ場合は、すべての候補漢字に番号をつけて入力者に表示する。入力者は番号を選ぶことで入力したい漢字を選ぶ。

表4と表5は「小倉研究室」を入力する際の入力字画例と JCSIM 法による入力キー列である。

2.5 JCSIM 方式における漢字コード

同じ字画コードは多数の漢字を持つ場合がある。調査によって、現在 JCSIM 方式に基づいて、作成したシステムの候補数と字画コード数の状況を表6に示す。

3. JCSIM 方式に基づく入力支援システム

3.1 システムの概要

JCSIM 方式に基づく提案システムは C 言語で作成した。更に「UNIX」上のテキストエディタ「Emacs」で利用可能にするために、インタフェースを「Emacs Lisp」で作り、「Emacs」に組み込んだ。入力者はコントロールキーを押しながら、キー「x」と「o」を押すことで、「Emacs」のフロントエンド(かな漢字変換システム)から本システム(JCSIM 入力方式)に切替える。本システムから「Emacs」フロントエンドへの切替えは一文字(漢字)の入力が終わった時に、自動的に行なわれる。本システムは図2に示すように「漢字字画コード生成」、「漢字検索処理」、「字画コード定義辞書」、「多漢字出力と多漢字出力入力者選択」と「単漢字出力」からなる。システム各部における処理の詳細を「システム構成」で述べる。

表6 候補と字画コード状況図

候補数	字画コード数
1	568
2~9	645
10~19	114
20~29	44
30~39	16
40~49	13
50~	12

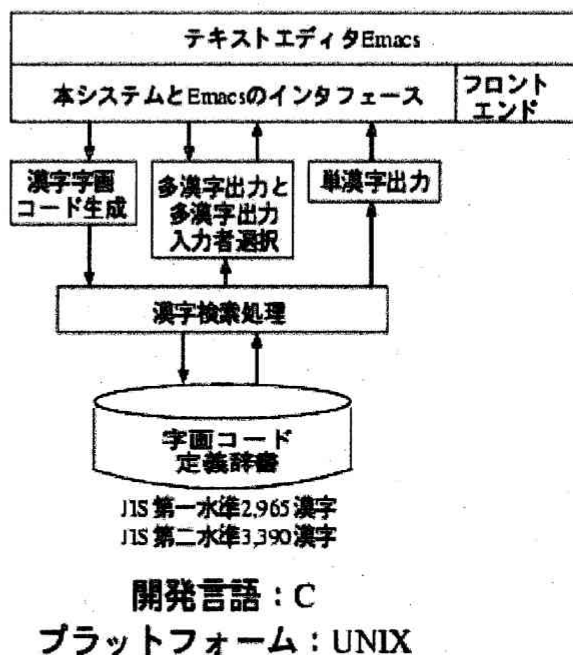


図2 JCSIM 方式にもとづくシステム構成

3.2 システム構成

(1) 漢字字画コード生成

漢字を検出するために、漢字の一文字ずつの字画コードが必要となる。「漢字字画コード生成」では、入力者からの入力が一漢字の字画コードになるかをチェックする。「JCSIM 方式」では一漢字の完全の字画コードはローマ字五つまたはローマ字四つ以下の最後にスペースを付けたものの二種類がある。これは漢字字画コード生成の基準となる。

(2) 漢字検索処理

漢字検索処理では字画コード定義辞書とのインタフェースを定義する。これは漢字を検出するための字画コードと検出された漢字の通路といえる。また、

表7 EUCコード漢字アドレス

```
#define A16  '¥xb0'
#define A17  '¥xb1'
#define A18  '¥xb2'
#define A19  '¥xb3'
#define A20  '¥xb4'
#define B01  '¥xa1'
#define B02  '¥xa2'
#define B03  '¥xa3'
#define B04  '¥xa4'
#define B05  '¥xa5'
#define F(u, v) *(b+n)=u; (b+n+1)=v;
```

表8 字画コード定義辞書

s	F(A16, B76)
ss	F(A38, B83)
sss	F(A27, B80)
sssss	F(A16, B66)
sssss	F(A21, B35)
sssss	F(A21, B45)
sssss	F(A22, B64)

「漢字検索処理」では、「多漢字出力と多漢字出力入力者選択」、「単漢字出力」のインタフェースを定義する。漢字字画コードに対する処理と検出された漢字に対する処理を行う。

(3) 字画コード定義辞書

「字画コード定義辞書」は人手により作ったものである。最終の出力漢字をEUCコードから呼び出すため、字画コード定義辞書には検索用インデックスとEUCコード漢字アドレスの組が記録されている。表7に示すように、字画コード定義辞書では予めEUCコード漢字アドレス(行と列二つの部分からなる)とEUCコード漢字アドレスを保存するポイントを定義し、辞書の中でEUCコード漢字アドレスの定義した文字列を使ってEUCコード漢字アドレスを表記し、EUCコード漢字アドレスを定義した保存するポイントに保存する。検索用インデックスは「漢字字画コード生成」で生成した字画コードである。「字画コード定義辞書」にはJIS第一水準2,965漢字と第二水準3,390漢字を登録した。これにより、一般的な日本語文章の全ての漢字は本システムで入力できると考える。

(4) 多漢字出力と多漢字出力入力者選択

同じ字画コードは多数漢字を持つ場合がある。この場合では入力者の選択機能が必要となる。「多漢字出力と多漢字出力入力者選択」は入力者の選択機能を提供する。同じ字画コードに定義された全ての漢字候補(予めEUCコード漢字アドレスでEUCコードから漢字を呼び出す)に番号をつけて入力者に表示する。更に、入力者からの番号選択に対応する漢字を出力する機能を持つ。

(5) 単漢字出力

字画コードは単漢字を持つ場合ではEUCコード漢字アドレスでEUCコードから漢字を呼び出し、出力する機能を持つ。

3.3 システムの支援機能と利用法

JCSIM方式に切替えるには、「Emacs」上のフロントエンドでコントロールキーを押しながら「x」を押した後で「o」を押す。読みが分からない漢字を入力する際にJCSIM方式を使うので、入力者は多数の場合ではフロントエンドで日本語文書を入力することを考えて、入力者に便利を図るために、JCSIM方式で一つの漢字を入力し終えたら、システムは自動的にフロントエンドに戻る。

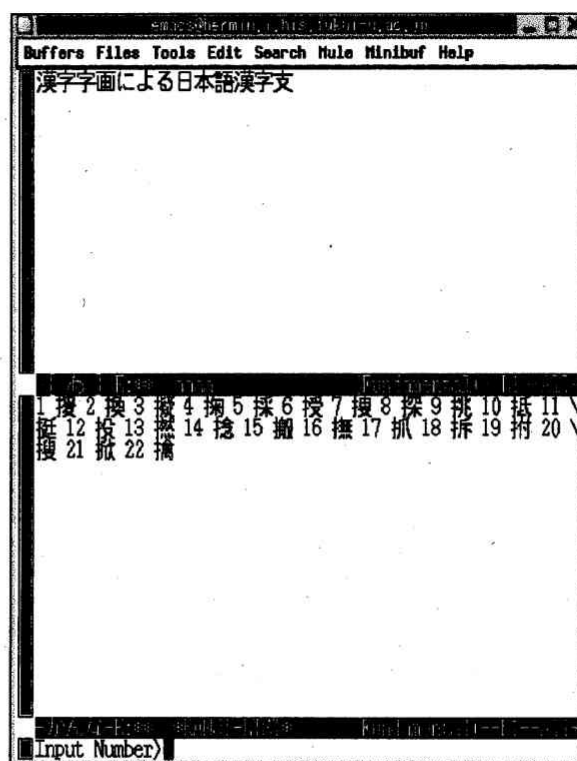


図3 入力画面1

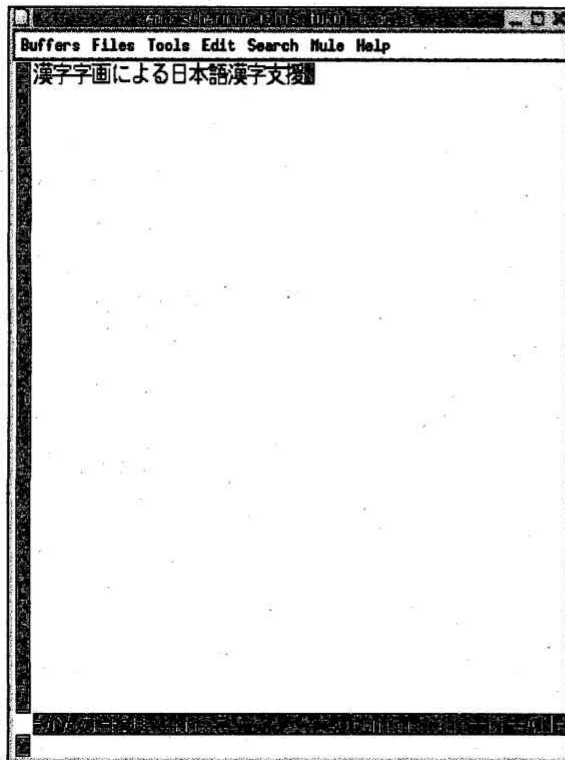


図4 入力画面2

入力された字面コードが多数の漢字を持つ場合は図3のように「Emacs」のバッファは自動的に二つに分かれる。上のバッファは元のエディタ画面を表示する。下のバッファは番号をつけた候補を表示する。入力者が番号を入力して漢字を選んだら、図4のように下のバッファはなくなる。エディタを表示する上のバッファは元のサイズに戻る。システムは「Emacs」のエディタ状態に戻る。

4. 実験と評価

提案システムの有効性を確認するため、評価実験を行った。

4.1 実験データ

入力者は、外国人日本語初心者であり、実験を行う前には JCSIM 法の使用経験がない三人の留学生である。入力者のデータは表9に示す。入力者Aは日本在留期間と日本語学習時間が一番長く、日本語レベルも一番高いと考えられる入力者である。また、入力者Cは日本在留期間と日本語学習時間が一番短く、日本語レベルも一番低いと考えられる。

実験用文は 100~200 文字以内の3種類の文とした。実験に関する詳細は表10に記す。文F1からF3まで漢字比率が増えている。特に文F3は日本の古文で、漢字のみの文になっている。

表9 入力者データ

入力者	専門	日本にの在留期間(月)	日本語学習時間(時間)
A	材料開発	22	約 1,200
B	機械	10	約 350
C	情報処理	4	約 300

表10 実験用文データ

文番号	文源	総字数	漢字数	仮名数	漢字率
F 1	工業日本語	132	50	82	38%
F 2	日本建築史	181	103	78	57%
F 3	史籍雑纂	117	117	0	100%

日本語入力システムは2種類用意した。入力システムRはJCSIM方式を組み込んでいないローマ字入力システムであり、入力システムJはJCSIM方式を組み込んだシステムである。入力はA、B、Cの順に行った。各入力者はまず入力システムRで文F1を入力してから入力システムJで文F1を入力する。次に入力システムRで文F2を入力してから入力システムJで文F2を入力する。F3についても同様繰り返す。実験JCSIM方式で漢字を入力するのは、読みが分からない漢字のみとした。

4.2 実験結果と実験評価

各入力者はJCSIM方式を組み込んだ入力システムJで文を入力する時、読みが分からない漢字をJCSIM方式で入力する。入力の時JCSIM方式で入力した漢字の数とその利用率を表11に示す。図5はJCSIM方式の利用率の漢字率との関係を示す。

表11 JCSIM方式利用率

入力者	文番号	JCSIM方式で入力した漢字数	利用率
A	F 1	4	8%
	F 2	12	12%
	F 3	22	19%
B	F 1	8	16%
	F 2	19	18%
	F 3	38	32%
C	F 1	15	30%
	F 2	45	44%
	F 3	76	65%

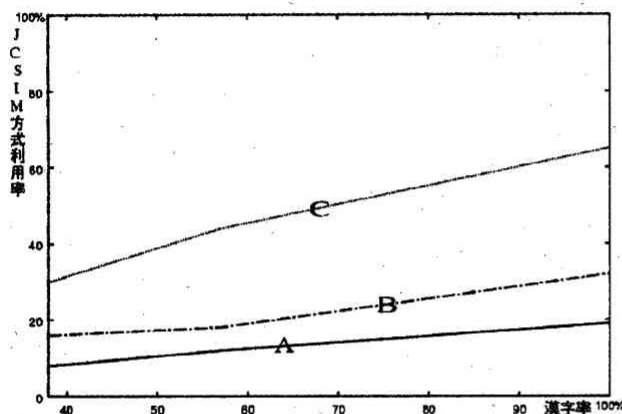


図5 JCSIM方式の利用率図

図5の横軸は入力文の漢字率で、縦軸はJCSIM方式利用率である。曲線A, B, Cはそれぞれ入力者A, B, Cが三つの入力文を入力した時のJCSIM方式の利用率である。日本語レベルが一番高い入力者AはJCSIM方式の使用率が一番低く、日本語レベルが一番低い入力者Cが一番高い。また、漢字の存在率が高い文章F3でいずれの人もJCSIM方式の利用率が増加している。

JCSIM方式を利用することによる入力時間軽減率を表12に示す。

図6の横軸は入力文の漢字率で、縦軸は入力時間軽減率である。曲線A, B, Cはそれぞれ入力者A, B, Cの入力時間の軽減率である。日本語レベルが一番高い入力者Aは入力時間軽減率が一番低く、日本語レベルが一番低い入力者Cは入力時間軽減率が一番高い。また、漢字の存在率が高い文章F3ではいずれの人もJCSIM方式での入力時間軽減率が増加している。

5. 考察

5.1 評価実験での課題

評価実験から日本語文章を入力するにはJCSIM方式が入っているローマ字日本語入力システムを使うことで入力時間を軽減することができ、入力文の漢字率が高ければ高いほど、入力時間の軽減率も高く、入力者の日本語レベルが低ければ低いほど、入力時間の軽減率が高いという結果を得られた。しかし、入力者によってJCSIM方式はまだいくつかの問題が存在していることがわかった。これらの問題を解決するため以下のような検討を行った。

表12 JCSIM方式による入力時間軽減率

入力者	文番号	ROMA ONLY	ROMA JCSIM	軽減率
A	F 1	27(分)	25	7%
	F 2	55	48	13%
	F 3	40	28	30%
B	F 1	34	27	21%
	F 2	49	38	22%
	F 3	36	24	33%
C	F 1	7	6	14%
	F 2	17	11	35%
	F 3	20	13	35%

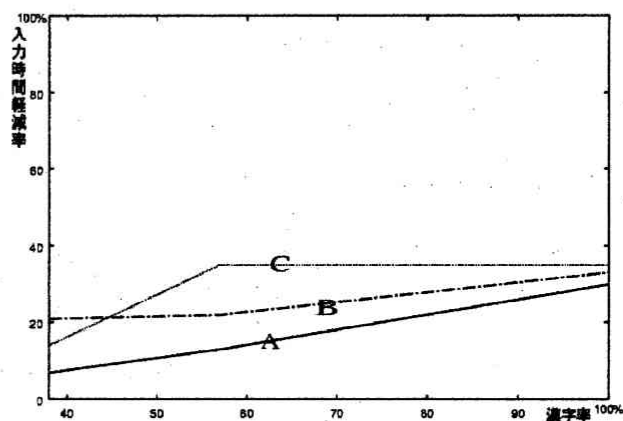


図6 入力時間軽減図

(1) 入力者によって漢字の書き順が異なる問題

前に述べたようにJCSIM方式は漢字を書き順により、字画に分ける。字画を用いて字画コードを生成して、字画コードで字画コード定義辞書にて漢字を検出する。同じ漢字としても、書き順が違えば、漢字の字画コードも違う。違った字画コードで正しい漢字を検出できない。図7のように同じ漢字「左」は書き順「一」、「ノ」、「一」、「丨」、「一」で書く人もいるし、書き順「ノ」、「一」、「一」、「丨」、「一」で書く人もいる。間違った後者の書き順^[7]では正しい漢字「左」を入力できない。しかし、日本人でも間違った筆順で漢字を書く場合が多い。外国人日本語初心者は日本語漢字の書き順はもっと間違いやすいと考えられる。

この問題を解決するために複数の書き順の字画コード定義辞書を作成した。間違いやすい書き順でも漢字を入力できるようにした。前の例では、字画

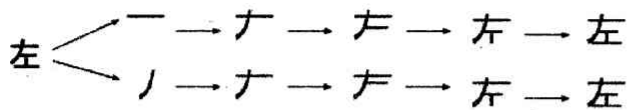


図7 書き順例

コード「一, ノ, 一, |, 一」と「ノ, 一, 一, |, 一」どちらでも漢字「左」を入力できるようにした。

(2) 漢字候補が多数存在する問題

同じ字画コードに多数の漢字が対応している場合には、候補が多くでてきて、入力者はその中から入力したい漢字を探すのがたいへんである。この問題を解決方法の一つとして、一つの漢字を入力するため、入力字画数を増やす方法を提案した。しかし入力字画数が増えれば入力者に負担をかける。入力字画数が多ければ多いほど、入力者への負担が重い。入力字画数を何画にすればキーをタイプする負担と候補から入力したい漢字を探す負担の両方がバランスよく軽減されるかを検討する。

5.2 今後の課題

JCSIM 方式に基づくシステムは「UNIX」上のテキストエディタで実行できる。しかし、外国人留学生日本語初学者は文章を入力するために、よく使うのは「Windows」上の「Word」などのワードプロセッサソフトウェアである。多くの外国人留学生日本語初

学者に JCSIM 方式を使ってもらうためには、それらのソフトウェアとのインタフェースを作り、JCSIM 方式が「Windows」上の「Word」などで実行できるようにすることも今後の課題である。

参考文献

- [1] 尾関雅則, “第 32 回全国大会にあたって一日本語入力方式について”, 情報処理, Vol. 27, No. 5, pp493-497, 1986.
- [2] 土屋順一, “外国人のための日本語キーボード入力支援システムの母語別カスタマイズ”, 電気通信普及財団研究調査報告書第 15 号, 98-01044.
- [3] 張小剛, 高建斌, 高橋勇, 黒岩丈介, 小高知広, 小倉久和, “漢字字画による日本語漢字入力補助方法”, 情報処理学会第 65 回全国大会, 4 分冊, 3Q-4 (2003. 3)
- [4] 白井英俊, 荒木卓也, “日本語入力手段としての漢字モード”, 計量国語学, Vol. 14, No. 6, pp. 239-251, 1984.
- [5] 中国漢易ソフト開発株式会社: <http://www.1han.com/hanease/menzhai/dict.htm>
- [6] 三重大大学教育学部学校教育講座: http://www.schl.edu.mie-u.ac.jp/comp_ed/2002/yokoi/kanzi5.htm
- [7] 江守賢治, 漢字筆順ハンドブック, 三省堂, 1982.