

論文

日本語文入力用カナ漢字変換システムの試作*

木村 健** 遠藤 安彦** 小橋 史彦**

Abstract

This paper reports a Japanese sentence input system "IROHA-1 (Ideograms input Routine by Homonymic Analysis)", which translates Kana sentences into Kanji-Kana mixed sentences. Procedures of the translating algorithm consist of word separation processes and homonym analysis processes. Main features of our method are syntactical and semantical analysis between several clauses in a sentence.

And this system is on-line system, so it's easy to operate and versatile for various experiments.

1. まえがき

電子計算機の利用が市民生活の身近まで浸透してくるにつれ、自然言語処理の機械化が強く望まれるようになってきた。しかし、日本語*** は表音文字と表意文字を混用する、漢字の種類がきわめて多い、音節構造が単純で同音語が生じやすいなどの特徴があり、欧米語に比べ処理の機械化が遅れている。これらの問題は特に日本語文を情報処理システムに入力する場合に障害となり、経済的で操作し易い入力手段が望まれている。

近年は、カナまたはローマ字で日本語文を入力しそれを何らかの方法で漢字かな混じり文に変換するカナ漢字変換方式が注目されている。特に、言語学的興味やシステム構成の柔軟性、将来の発展性が大きいなどの理由で、ソフトウェアによりカナ漢字変換を行う試みがいくつかなされてきた。

ソフトウェアによるカナ漢字変換方式は早くから九大・栗原らによって始められた。栗原らは簡単な分かち書きを施した文節について、カナ漢字変換単語辞書(以下、辞書と略称)を使って構文解析並びに意味解析を行う基礎的な手法を提案するとともに、辞書作成の

ためのデータを収集整理した^{1),2)}。その後、沖電気は九大と共同研究を行い、電子計算機を使ったカナ漢字変換システムを作成した³⁾。また NHK 技研の相沢らは九大・沖の研究成果に基づき、適用範囲をニュース文に絞ったシステムを作成した⁴⁾。このシステムでは、変換に必要な情報を分かち書き単位分析表と呼ぶテーブルに収め、全体をコンパクトに構成したことが特長である。

カナ漢字自動変換を実現するには、カナ文を語単位に区切る手順と同音語の中から適正な1語を選ぶ効果的な手続きを開発せねばならない。それには、日本語に関する理論的な研究と膨大なデータの収集をした上で、変換プログラムと辞書を作る必要がある。そこで、筆者らはまずある程度の性能を備えた実験システムを使いながら変換手法の検討と辞書の最適化を図ることとした。以下先に報告した IROHA-0⁵⁾ の経験に基づき作成した IROHA-1 につき報告する。

IROHA-1 におけるカナ漢字変換法は基本的には九大・沖の延長であるが、同音語解析において関連語情報や多レベルの意味分類情報の利用など従来試みられなかった多文節にわたる構文解析・意味解析を行うこと、そのための辞書を新たに作成したこと、及び実験効率の向上と実用性を高めるためオンライン方式としたこと等が大きな特長である。なお、対象とする日本語の範囲は約3万語の日常語とし、人名・地名等の固有名詞、学術用語等の専門語、古語・雅語等は除外し

* A Japanese sentence input system by Kana-Kanji translation by Ken KIMURA, Yasuhiko ENDO and Fumihiko OBASHI (Musashino Electrical Communication Laboratory, Nippon Telegraph and Telephone Public Corporation).

** 日本電信電話公社武蔵野電気通信研究所

*** 範囲を書きこむことばに限定する。

た。

2. システム構成

2.1 ハードウェア構成

ハードウェアは当研究所内で共通的に利用される計算機設備を使用したもので、大形の中央処理装置と小形の端末制御計算機を 4,800 BPS のデータ通信回線で結び、さらに後者から 1,200 BPS のデータ通信回線で利用者宅内の画像端末に至るオンライン・システムである (Fig. 1 参照)。画像端末は蓄積管式 CRT ディスプレイ装置で英数字キーボードを持ち、最高 4 台まで接続可能である。利用者と IROHA-1 システムとの情報交換はすべて上記画像端末を介して行われる。

2.2 ソフトウェア構成

ソフトウェアは全体として中央処理装置側のオペレーティングシステムの下で動作するが、利用者から見れば日本語処理専用のコマンド制御の下に動作する。中央処理装置側は主としてカナ漢字変換処理と各辞書ファイル・文章ファイルに関する処理を担当し、端末制御計算機側は編集・表示を主とした端末制御、ローマ字カナ変換など各種コード変換、中央のオペレーティングシステムと日本語処理コマンドシステムとのインタフェース等を担当する (Fig. 2 参照)。

2.3 システム動作と制御

利用者は画像端末のキーボードからコマンド形式で処理の開始・終了、カナ漢字変換、結果の編集、表示、ファイル処理等を指示する。例えば、カナ漢字変換を行わせたい場合は、TRANSLATE コマンドを発したのち、後述の方法で文節単位に分ち書きしたローマ字文* をキーボードから入力すればよい。計算機内部での変換終了後の結果は DISPLAY コマンドで CRT 画面上に表示して見ることができ、変換誤りを発見したときとか表記を変更したいときは EDIT コマンド

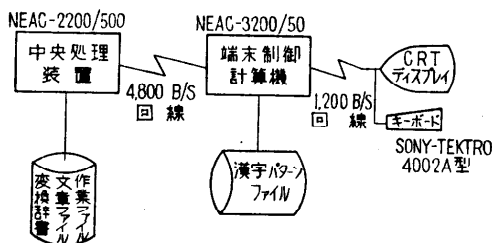


Fig. 1 The hardware configuration

* 使用した入力キーボードにカナ文字がなかったことによる。ローマ字はヘボン式・割合式・日本式のいずれでもよくまたこれらを混合したのもよい。

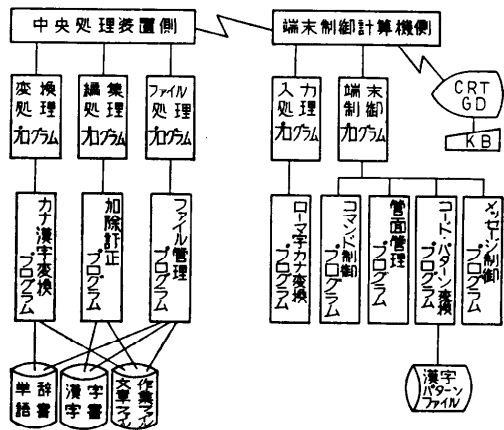


Fig. 2 The software configuration

によって加除訂正等の編集手続きを行う。画面上の文章はやはりコマンドによってハードコピー装置または文章ファイルに出力できる。

利用者はこれらの処理過程を通じ CRT 画面上から必要な情報を得る。すなわち、CRT 画面は入力文の監視、変換後またはファイル内の漢字かな混じり文の表示、システム制御のための通知メッセージの表示等に使用される。

3. カナ漢字変換機能

カナ漢字変換を行うということは文字の列を単語に置き換えるということであり、このことは会話における聞き手の役割をシミュレートすることを意味している。会話における聞き手は話し手の発した音素列を変換して意味を認識する。この変換を行う過程で言葉のあいまいさの原因である同音語の判定が重要な問題となり、さまざまな情報を利用して同音語の処理を行うことになる。

カナ漢字変換の観点から見ると、その過程は大きく単語の分離と同音語の解析の2つからなることができる。単語の分離とはカナ文を漢字かな混り文に変換するため単語単位に分割することをいい、分かち書き法、最長一致法を用いて行う。また、同音語の解析とはカナから漢字に変換する際に生ずる同音語の中から適当な語を選ぶことをいい、文法情報、意味情報等を用いて行う。以下、これらを中心に本システムで用いたカナ漢字変換手法について述べる。

3.1 分かち書き法

カナ文を分かち書きなしで入力すると次の例でもみられるようなあいまいさが生じ、カナ文を単語単位に

分離する場合に困難を伴うため、カナ漢字変換においても分かち書き入力が必要がある。(プログラムにより自動分かち書き処理を施す方法も考えられるが、同音文* 処理法が確立していない現状では分かち書き入力が実際的である。)

例. ヒルルspanニコイ

昼 留守 晩に來い
昼 留守番に來い

現在一般に行われている分かち書き法として次の3つのものがある。

- (1) 単語単位によるもの
- (2) 文節単位によるもの
- (3) 単語と文節の折衷によるもの

カナ漢字変換等の機械処理に分かち書き法をとり入れる場合、規則が簡単でだれもがすぐ覚えらるること、分かち書き回数が少なく打鍵速度が上がることなどの条件が要求される。これらのことを考慮し、(2)の形式を主体とする以下の入力規則を設けた。

- (1) “文節” 単位に分かち書きとする。

文節 = 自立語 + n · 付属語

(ただし、 n は非負整数)

- (2) 連語は単語単位に分けて入力する。
- (3) 数字、記号、句読点類は分かち書きをしなくてもよい。

3.2 最長一致法

カナ漢字変換の第1過程として、分かち書きして入力されたカナ文について、機械翻訳等で用いられている最長一致法を利用して文節内の単語の分離を行う。最長一致法は辞書の見出し語との比較により文節単位内で最も語長の長い自立語を検出して、文節を自立部と付属部に分離する方法である。もし、付属部をいくつかの付属語に分離でき、かつ、各単語間の接続が可能ならば、上記の分離が適当であったことになる。しかし、そうでない場合には再び最長一致法を適用して次に長い自立語を検出し、上の検査を行う。この繰り返しにより入力文を単語単位に分離できる。

例. コノヨウナ

× この世 う _____ な
名詞 助動詞 “う”
接続不可

○ この ような _____
連体詞 助動詞 “ようだ”
接続可

上記の方法で単語の分離ができれば、第2過程として次に述べる同音語の解析を行う。しかし、単語の分離するには同音語の解析が必要なことが多く、また、同音語を解析するには単語の分離を行わなければならないので、実際には両者を平行して処理することになる。

3.3 同音語の解析

同音語解析の対象となる単位は基本的には単語である。ここでは単語のもつ文法的情報と語彙的情報の2種の情報に着目して解析を行う。文法的情報はさらに形態的情報と構文的情報に分けられる。前者は活用形など単語自身の語形であり、後者はその単語と関連している語の語形から得られる情報で文脈の情報ともいえるものである。また、その単語が一定の構文上の成分になりうることも構文的情報の一種であろう。語彙的情報とはその単語の意味であり、通常の国語辞書から得られる情報に相当する。ここではこれらの情報を利用した同音語解析法を示す。

- (1) 形態的情報

単語間の接続関係の規則性⁶⁾を利用する。例えば、助詞、助動詞は直前の単語の品詞、活用形を指定し、文節、文の切れ目を明確にする性質をもつ。活用語尾は各品詞、活用型で定まっており、語幹の品詞を推定できる。

例. カケば 書け(仮定形)+ば(助詞)
掛(連用形)+ば(助詞)

- (2) 構文的情報

文節間の修飾関係を利用する。例えば、助詞は属する文節に連用修飾の性格あるいは連体修飾の性格を与えるものがある。また、後にくる動詞などの種類を限定する場合もある。活用形において連用形は連用修飾、連体形は連体修飾の働きがある。

例. 服を キセル 着せる(動詞)
煙管(名詞)

- (3) 関連語情報

他の単語と関連して現われる語がある。例えば、単語や語句が結びついて各単語自体にはない意味を表わす慣用語、慣用語がある。また、類義語や反対語などは文章中に並立の形で出現することがある。

例. イッセンを 画す 一線(を画す)
一戦(を交える)

- (4) 意味情報

文中に意味概念の密接なつながりがある語があるかどうかにより判断する。国立国語研究所の分類語彙

* 意味は異なるが形の同一な文(例参照)。

Table 2 Connection matrix for auxiliary verb

前につく語	うご	ず	し	せ	そ	た	た	た	な	か	べ	ま	み	よ	ら	る
後につく語		体	未	未	未	未	未	未	未	未	未	未	未	未	未	未
うご																
ず																
し																
せ																
そ																
た																
た																
な																
か																
べ																
ま																
み																
よ																
ら																
る																

未=未然形, 用=連用形, 終=終止形, 体=連体形, 仮=仮定形, 幹=語

図った。

単語辞書, 文法辞書および編集処理 (後述) で使用する漢字辞書の検索法, 記憶容量を Table 3 に示す。

4. 編集・制御機能

4.1 制御コマンド

本システムを利用者側から見ると, 機能の選択・入出力操作・ファイルの登録抹消等の制御はすべて Table 4 で示すコマンド群で制御される。これらのコマンドまたはサブコマンドは利用者が CRT 端末のキーボードから打鍵入力する。この場合, 全文字を打たなくとも上位 3 字だけ打鍵すればよい。入力したコマンドに対する応答は, CRT 画面へのメッセージ表示または端末機内蔵の警告器の鳴動によってなされる。

4.2 編集機能

3. で述べたカナ漢字変換法では現在のところ変換率 100% の実現は困難である。そのため漢字かな混りに変換された文章に対し誤変換のチェックと修正が必要

Table 3 Dictionaries

種別	見出し数	検索法	記憶容量	用途
単語辞書	28,000 (語)	インデックスシーケンシャル検索	3 (Mch) (磁気ディスク)	自立語の検出
文法辞書	270 (語)	2分割法	13 (Kch) (コア内)	文法照合
漢字辞書	3,200 (読み)	ラディックスツリー検索	1.5 (Mch) (磁気ディスク)	文の加除訂正

となる。今回試作・実験したシステムは誤変換された箇所を字単位で修正する機能をもつ。

編集は漢字辞書と CRT 端末を用いて行う。漢字辞書は自動的に漢字かな混り文に変換した文章に誤りがある場合, システム利用者が CRT 端末を介して文の加除訂正を行うためのもので, 漢和辞典に相当する。なお漢字辞書は当研究所数値計算研究室の杉森・吉井¹⁰⁾が作成したものを利用した。

編集機能には次のものがある。

(1) 削除機能

当機能は利用者が変換された文章に対し不必要と思った箇所, あるいは誤字を削除したい場合に用い, DELETE コマンドで指定する。

(2) 挿入機能

当機能は文に対し, 文あるいは語・字を加えたいときに用いるもので, INSERT コマンドで指定する。

(3) 置換機能

当機能は文に対し, その中の文字を置換したいときに用い, REPLACE コマンドで指定する。この場合置き換える文字としては文中の他の文字, または漢字辞書より検索してきた文字を用いる。

(4) 追加機能

当機能は文の末尾にさらに文章をつけ加えたいときに用いるもので, ADD コマンドを用いる。

(5) 漢字検索機能

上記の(2), (3), (4)手順で用いる置換用の漢字を漢字辞書から取り出す機能で, SEARCH コマンドで指定する。

Table 4 Control commands

コマンド	制御内容
CONNECT	利用開始, システムオープン
CLOSE	利用終了, システムクローズ
TRANSLATE	カナ漢字変換要求
EDIT	編集要求
*SEARCH	漢字検索要求
*DELETE	文字消去要求
*INSERT	文字挿入要求
*REPLACE	文字置換要求
*ADD	文章追加要求
*CONTINUE	再編集要求 1
*NEXT	同上 2
HOLD	ファイル登録・抹消・保存要求
DISPLAY	表示要求
*NEW	次頁表示要求
*RETRY	再表示要求

*で始まるものはサブコマンド

4.3 表示機能

2. で述べた通り本システムを使用する場合、情報交換はすべて CRT 端末を介して行われる。すなわちシステム側からの情報は CRT 画面上に表示される。

(1) 入力時における表示機能

カナ漢字変換するための入力はローマ字表記の日本語文である。その理由は端末装置にカナキーボードがないためである。この場合、入力文をローマ字表記で表示するよりも、カナ表記で表示した方がチェックしやすいし、変換プログラム側でもカナ表記に直っている方が変換しやすい。そこで、端末制御計算機側でローマ字表記の入力文ををひらがなに変換して CRT 画面上に表示するエコー処理機能を設けるとともに、かなに変換したものを中央処理装置側へ送出する。

(2) 編集時の表示機能

当システムの入出力装置は蓄積型の CRT 端末を使用している。このためフリッカーがなく、表示が鮮明、比較的安価であるという利点がある。しかしこの端末はライトペンやリフレッシュメモリ等は備えていない。編集時における表示機能はこれらの欠点を補っている。

誤変換された文章を加除訂正するには一度端末に表示してから行う。このときリフレッシュメモリがないので Fig. 3 のような表示形式とした。現在、一画面に表示できる字数は横 25 文字縦 15 行、計 375 字/面である。変換率を約 95% と考えた場合に単純に計算すると、20 字で 1 字の誤りがあることになる。漢字が誤って変換された場合には送りがない等の誤りも生じるため、漢字かな混り文を表示する部分を 3 行、漢字を検索して表示する部分を 10 行とした。その他座標表示部分を横 1 行縦 1 列設けた。

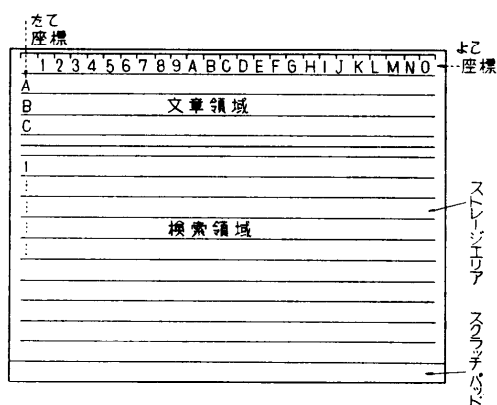


Fig. 3 A display layout of editing area

(3) 出力時の表示機能

カナ漢字変換機能、編集機能により作成された日本語文(漢字かな混り文)を画面に表示する機能である。

表示する場合には入力時に指定したファイル名、パスワード、ユーザ ID を入力しなければ表示できない。これによりファイルの機密保護が行える。

4.4 ファイル処理プログラム

本プログラムは後述の制御情報ファイルを基に各種ファイルの入出力処理、登録、抹消、内容保護等に関する処理を行う。CONNECT と CLOSE コマンド以外のコマンドからファイル処理に関するパラメータが制御情報ファイルに書き込まれる。

ファイルには次のものがある。

- (1) 制御情報ファイル
- (2) 変換文蓄積ファイル
- (3) 編集文蓄積ファイル
- (4) 漢字検索蓄積ファイル
- (5) 単語辞書ファイル
- (6) 漢字辞書ファイル
- (7) 文法辞書ファイル

(1)は(2),(3),(4)のファイルを制御する情報を蓄積するものである。

(2)は変換した漢字かな混り文を蓄積する。

(3)は(2)のファイルを編集した結果を蓄積する。

(4)は(6)の漢字辞書ファイルから SEARCH コマンドで指定した漢字を検索した結果を蓄積する。

(5),(6),(7)は前述の辞書をファイル化したものである。

上述のファイルは中央処理装置側のプログラムで用意するものであり、端末制御計算機側では漢字パターンファイル(約3,600字種)をもっている。これには CRT 端末に文字を表示するためのデータがベクトル方式でドラムに格納してある。

5. 結果と考察

このシステムの評価実験の一例として、日刊紙 3 日分の中から 25 記事(9分野)を選んで実験した結果を Table 5(次頁参照)に、変換例を Fig. 4(次頁参照)に示す。入力データ量は約 7,000 字(2,100 文節)であり、変換処理速度は 2.5 文節/秒である。また、正しく変換できなかったものについて、原因別の割合を Table 6(次頁参照)に示す。以上の結果から、

- (1) 誤り原因①~④のうちの多くは比較的単純な誤りであり、辞書単語の整備、最長一致法の修

IKIGAI RONGA HAYATTEIRU. KOUHUKUTOHA NANIKA.
SHIAWASENO SHIHYOUTOHA NANDAROU. ITTAI NAKIGA SHINNO
YUTAKASANANOKA. HIMOZIKATTA SENGONO.NAN NEWKANKAGA
IBUKURONO ZIDAIDATTATOSURUTO, SONO NOTI KEIZAI SEITYOUNO
ZIDAIDATTATOSURUTO, SONO NOTI KEIZAI SEITYOUNO ZIDAIWO
HETE, IMA KOKORONO ZIDAINI SASIKAKATTANODEHANAIKAROUKATO
OMOU. KEIZAI SEITYOUKAKEDEHA KOUHUKUNIHANARENAI,
TOIUKOTONI WATAKUSHITATIHA KIGA TUITA. ISSEINI, SELKATU
ZIKKANTOSITE SOREVO SATOTTA.

入 力 文

生き甲斐論がはやっている。幸福とは何か。幸せの指標とはなんだろう。一体なにが新の豊かさなのか。ひもじかった戦後のなん年間かが胃袋の時代だったとすると、その後経済成長の時代を経て、今心の時代に差掛かったのではないだろうかと思う。経済成長だけでは幸福にはなれない、ということにわたくしたちは気が付いた。一斉に、生活実感としてそれを悟った。

出 力 文 — 部分は変換誤り

Fig. 4 An example of translation

Table 5 Translation program efficiency

処理法	変換率		字単位(%)	文節単位(%)
	文類	情報		
①	文類 頻度	情報 情報	92.9	82.7
②	①+ 連意	連語 意味 情報 情報	93.7	84.1

変換率

$$\text{字 単 位} = \frac{\text{正しく変換できた字数}}{\text{投入文章の総字数}} \times 100$$

$$\text{文 節 単 位} = \frac{\text{正しく変換できた文節数}}{\text{投入文章の総文節数}} \times 100$$

Table 6 Examples of mistranslation and error-rates

原 因	例 (正→誤)	割合 (%)
① 辞書に単語が登録されていない	端木→たんまつ	13.8
② 最長一致法によるもの	対し→大使	11.2
③ 文法処理によるもの	1歩→1ぼ	14.3
④ 関連語、複合語処理によるもの	起源説→紀元節	12.5
⑤ 意味処理によるもの	延々と…唱歌→炎々と…消火	25.4
⑥ 頻度処理によるもの	占拠→選挙	17.4
⑦ その他	各行(各銀行の略語)→角綱	5.4

正(最長単語内の同音文節発生の有無を単語辞書情報として追加し、短い自立語を含む同音文節の適応性も検査する)、連濁、転音、接辞などの処理を追加することにより、誤りの約1/3は改善できる。

- (2) 意味処理が十分に効果的に働いたとはいえない。これは構文上の結びつきを意味のつながりとしてうまく利用できなかったこと、意味分類

として木構造の意味表現を単純な形で利用したため言葉が本来もつ多方面な意味的役割が除外されたことなどによる。このため、単語の連想的な結合関係を網目構造の意味表現として記述すること、助詞・接辞における意味的情報を利用すること、構文上の単語間のつながりを意味処理に利用することなどの方法があげられる。これらの処理は、同音語の発生が他品詞に比べて圧倒的に多く、かつ、形態的情報の少ない名詞の場合に有効な方法となる。

その他、編集手法、画面表示形式、ファイル登録形式などについては定量的に評価するに至らなかったが、オンライン・システムとして簡易型画像端末を使用しての操作面等について、一般利用者にも容易に使用できるシステムとすることができた。

この間、端末制御計算機と画像端末間の回線を延長し、日本電信電話公社データ通信展示センタ(霞ヶ関)や第2回日米コンピュータ会議展示会等で実演を行い、オンライン構成の操作性・信頼性を確認した。

なお、このシステムの主製造言語は COBOL で約 3,500 行の規模である。

6. む す び

以上報告した IROHA-1 システムは現在、所内の漢字情報処理業務への適用を検討中である。固有名詞処理の導入、名詞を主体とした意味解析の高度化等変換手法の改良、語彙の選定・情報内容の再検討等による辞書の最適化、および簡易型漢字プリンタ接続による端末機能の拡充化を図って実用システムの実現に進みたい。

最後に、辞書類製造に協力頂いた構造計画研究所味生氏、オンライン関連のプログラミングで協力頂いた日本電気の兼依氏、各種資料の提供とご助言を頂いた国立国語研究所の田中・中野両氏、種々ご指導頂いた九大田町教授他関係者の方々、その他ご指導ご協力頂いた当研究所内各位に厚く感謝する。

参 考 文 献

- 1) 栗原, 黒崎: 仮名文の漢字混り文への変換について, 九大工学集報, Vol. 39, No. 4, pp. 659~664 (1967).
- 2) 栗原, 稲永: カナ漢字変換(1), 九大工学集報, Vol. 42, No. 6, pp. 880~884 (1970).
- 3) 松下, 山崎, 佐藤: 漢字かな混り文変換システム, 情報処理, Vol. 15, No. 1, pp. 2~9 (1974).

- 4) 相沢, 江原: 計算機によるカナ漢字変換, NHH 技術研究, Vol. 25, No. 5, pp. 261~298 (1973).
- 5) 木村, 吉井, 小橋: 日本語情報投入実験システムについて, 昭和 47 年信学会全国大会.
- 6) 松村: 日本文法大辞典, p. 1054, 明治書院, 東京 (1971).
- 7) 国立国語研究所: 分類語彙表, p. 362, 秀英出版, 東京 (1964).
- 8) 国立国語研究所: 電子計算機による新聞の語彙調査(I), p. 342, 秀英出版, 東京 (1970).
- 9) 金田一他: 新明解国語辞典, p. 1248, 三省堂, 東京 (1972).
- 10) 杉森, 吉井: 日本語情報投入システムにおける漢字処理に関する一考察, 昭和 48 年信学会全国大会.

(昭和 51 年 2 月 23 日受付)

(昭和 51 年 4 月 5 日再受付)