HOKKAIDO UNIVERSITY

| | |
|---|---|
| Title | 研究者向き日本語ワードプロセッサKKH IIの開発 |
| Author(s) | 栃内, 香次; 伊藤, 太亮; 荒木, 健治 他 |
| Citation | 北海道大學工學部研究報告, 119, 119-126 |
| Issue Date | 1984-02-15 |
| Doc URL | https://hdl.handle.net/2115/41846 |
| Type | departmental bulletin paper |
| File Information | 119_119-126.pdf |

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

北海道大学工学部研究報告
第119号（昭和59年）

Bulletin of the Faculty of Engineering,
Hokkaido University, No. 119 (1984)

# 研究者向き日本語ワードプロセッサKKH IIの開発

栃内　香次　伊藤　太亮　荒木　健治　鈴木　康広　永田　邦一

（昭和58年9月30日受理）

# Researcher Oriented Japanese Word Processor KKH II

Koji TOCHINAI, Taisuke ITOH, Kenzi ARAKI, Yasuhiro SUZUKI and Kuniichi NAGATA

(Received September 30, 1983)

## Abstract

An improved version of the researcher oriented Japanese Word processing system using a small and user adaptive Kana-Kanji translation dictionary, KKH II is reported.

The performance of the present version of the system are discussed based on the text processing experiments. And important factors necessary to increase the performance are pointed out.

Based on the considerations mentioned above, improvements such as:

1) the automatic homonym selection to reduce manual operation,

2) modifications to reduce keyboard operation errors, and

3) the use of a common Kanji word dictionary to accumulate all Kanji words used in a user group,

are proposed to be realized in the KKH II.

The performance expected to be increased in the KKH II is also discussed.

## 1．はじめに

計算機による日本語文書処理の一般化に伴ない，種々の入力方式が出現している。かな漢字変換はそれら多種多様な入力方式のなかでも，

1）通常の計算処理と全く同じタイプライタ型鍵盤を使用することができ，

2）専業タイピストはもちろん，素人でも容易に使うことができる，

といった特長を有し，広く普及している。

先にわれわれは，研究者が自分の専門分野に関する論文等を自ら作成するために使用することを主目的とし，作成対象分野を限定して変換辞書の小容量化，使用者への適応化を可能とするかな漢字変換方式を提案し，実用システムの試作を行った[1),2)]。現在このシステムは著者らの研究室において論文，講演予稿，テキストなどの作成に利用され，使用経験の累積とシステムの性能評価を行っている。その結果，より高い性能を得るためには，

2）同音語の選択をなるべく人手によらず，システム内で自動的に行えるようにすること，

2）入力仕様を改善して，鍵盤操作を合理的にし，操作ミスを減少させること，

電子工学科　電子機器工学講座

3）未登録語の登録をより容易に行えるようにすること，

などが必要であることがあきらかとなり，これにもとづいて現在実用システム第2版の開発を進めている。

本論文はこれら現在までに判明した問題点と，それに対処するために第2版で行っている手法について述べるものである。なお，われわれはこのシステムをKKHと称し，現在開発中の第2版をKKH IIと称しているので，以下本論文でもこの名称を使用する。

## 2．KKHの性能評価

### 2．1　KKHの構成概要

KKHは本学大型計算機センターのHITAC M—280H／200H上に作られており，その概要は以下のようになっている[1)~3)]。

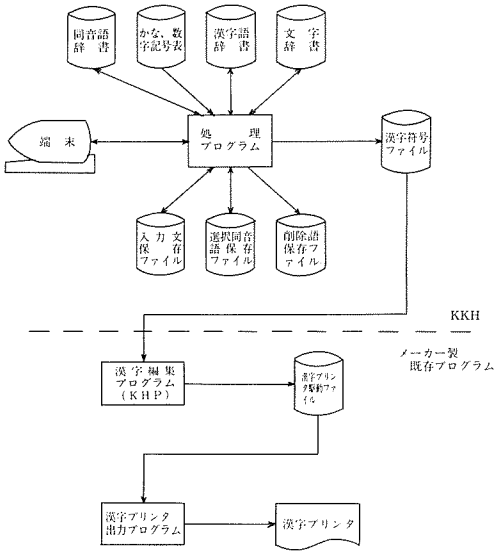
1）文の入力はセンター内外の通常のTSS端末から行われ，出力はセンターに設置されているレーザビーム漢字プリンタに行われる。

2）入力文はローマ字表記とし，大文字，小文字の使い分けによって漢字とかなの区別を行う。

3）かな漢字変換のための辞書は，小容量（2500語）で個々の使用者ごとに作られ，使用分野ならびに使用者に適応するように構成されている。

4）文の入力中に辞書に未登録の語（新出語という）が出現したとき，それを容易に登録できるように，個々の漢字を収録した文字辞書を別に設け，これを用いて語を合成できるようにしている。

以上の概要を図1に示す。




図－1　KKHシステムの構成

### 2．2　辞書管理

上述のように，KKHでは変換辞書を動的に構成し，個々の使用者，分野に適応させて変換性能を向上させるようにしている。以下，辞書管理方式の概要を述べる。

1）辞書の容量は2500語（ただし，このうち5語分は制御情報）である。

2）収録されている各語ごとに，頻度カウンタ $n$，履歴カウンタ $h$，の2つのカウンタを設けてある。

3）一つの文献を入力するためにシステムを起動したとき，すべての収録語について

$$h+1 \rightarrow h$$

とする。

4）文の入力中，辞書に収録されているある語が参照されたとき，その語について

$$n+1 \rightarrow n,\quad 0 \rightarrow h$$

とする。

5）新出語が出現したとき，その語について

$1 \rightarrow n,\quad 0 \rightarrow h$

とし，

（i）辞書に空きがあれば直ちに登録し，

（ii）空きがなければ，$h/n$ を計算し，その値が最大の語の中から1語を選んで削除し，そのあとに登録する。

したがって，3），4）により参照頻度が小さく，長期間使用されない語ほど $h/n$ は大になり，そのような語が5）で選ばれることになる。このアルゴリズムは $n$ に関して敏感であり，$n$ がすこし大きくなると極めて長期間削除されないことになる。なお，語は頻度順に配列され，4）により $n$ が変化したときは語の入換えを行っている。また，登録された語が短期間で削除されることがないように，この他にパラメータ $h_0$，$n_0$ を設け，$h-h_0 \leqq 0$ のものは削除の対象とせず，さらに $n \leqq n_0$ のものは $n_0 \rightarrow n$ とおきかえて，$(h-h_0)/n$ が最大のものを選ぶようにしている。$h_0$，$n_0$ は使用者ごとに設定することができ，実際の使用では $h_0=10$，$n_0=2$ 程度に選ばれている。

## 2．3　変換性能の評価

KKH を使用してひとつの文献を入力するとき，ある割合で新出語，同音語が出現する。これらは入力時点で使用者による登録，あるいは選択操作を必要とする。また，変換結果（漢字かな混り文）にはいくつかの誤変換された語が含まれ，修正する必要がある。これ以外の語は人手を要さず，自動的に正しく変換される。また，漢字以外のかな，その他の文字，記号はそのまま漢字符号に変換される。以上により，次に示すように変換率を定義し，性能指標とする。

- 入力された文字総数を $C$ とする。
- そのうち，漢字語の語数を $K$，漢字の字数を $C_K$ とする。
- 新出語の語数を $Wn$，その字数を $C_{Wn}$ とする。
- 同音語の語数を $Wh$，その字数を $C_{Wh}$ とする。
- 同音語の語種（異なり語数）を $Wh_1$，その字数を $C_{Wh1}$ とする。
- 誤変換された語数を $We$，その字数を $C_{We}$ とする。

これらにより，以下に示す2種類の変換率を定義する。

- 語単位正変換率　$R_W=1-(Wn+Wh+We)/K$　(1)
- 文字単位正変換率　$R_C=1-(C_{Wn}+C_{Wh}+C_{We})/C$　(2)

ただし，$C_{Wn}$，$C_{Wh}$，$C_{We}$ が得られていない資料があり，その場合は1語あたりの平均漢字字数をあらわす $C_K/K$ を用い，

$$R_C'=1-(Wn+Wh+We)\cdot\frac{C_K}{K}/C \qquad (2')$$

とする。

実際にいくつかの文献を入力した結果，ひとつの文献の中では同音語のうち1種のみが出現する場合が極めて多いことがわかった。そこで，KKH には，1回選択された同音語をその選択に固定し，以後は内部で自動的に選択する機能を用意している。これを用いると，人手による同音語選択回数はほぼ同音語の語種，$Whs$ に等しくなる。よって，この場合は上記2種の変換率は次のようになる。

- 語単位正変換率*　$R_W^*=1-(Wn+Whs+We)/K$　(3)
- 文字単位正変換率*　$R_C^*=1-(C_{Wn}+C_{Whs}+C_{We})/C$　(4)

または　$R_C^{*'}=1-(Wn+Whs+We)\cdot\frac{C_K}{K}/C$　(4′)

いくつかの分野について入力実験を行った結果，平均 $R_W^*\simeq 85\%$，$R_C^*\simeq 95\%$ という結果が得

られ，実用上ほぼ十分な性能に達している[4]。

### 2．4　問題点

上述の入力実験結果，およびこれまでの使用経験から，以下に示すような問題点があきらかになった。

1）新出語，同音語，誤変換語のうち，同音語が最も多い。すなわち，上記(3)式において，

$$Wn/K \simeq 5\,\%,\quad Whs/K \simeq 10\%$$

程度となる。なお，誤変換語は極めて少なく，0.5～1％程度である。

2）実際に入力操作を行った感覚からは，上記$R_C^* \simeq 95\%$は実用になる下限で，十分満足できるためには97％以上が必要のようである。このとき，$R_W^* \geqq 90\%$である。

3）削除語を選出する際に$h/n$が最大のものを選ぶ現行方式では，過去に頻繁に使われてその後全く使われない語がなかなか削除されず，辞書の利用効率を低下させている。

4）上記各変換率の定義式には含めていないが，鍵盤操作ミスによる誤りが，入力文字数の1％程度存在し，このうち大文字，小文字の誤り，すなわちシフト操作のミスが相当数を占めている。

## 3．KKH IIの構想

### 3．1　システム構成

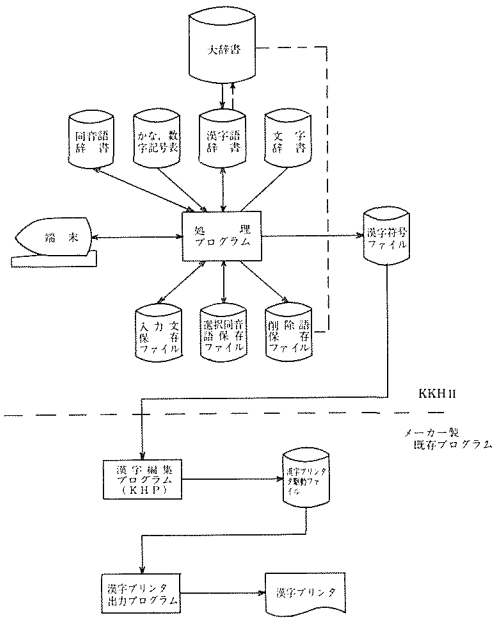
KKH IIは上述の諸問題に対処する機能を現行KKHにつけ加えたもので，図2にシステム構成の概要を示す。KKH IIと現行KKHとの主な相違点を以下に示す。

1）複数の使用者で共用する「大辞書」を設け，各使用者の変換辞書の和集合を収録する。

2）新出語が出現したときは大辞書を検索し，大辞書にあればそれを登録する。大辞典にもない場合は「真」の新出語として現行と同じ方式で登録する。

3）変換辞書の管理方式をあらため，削除語の選出は履歴のみによることとする。

4）同音語とその前後にある文字との組を記録する同音語辞書を設け，これを利用して同音語の自動選択を行う。



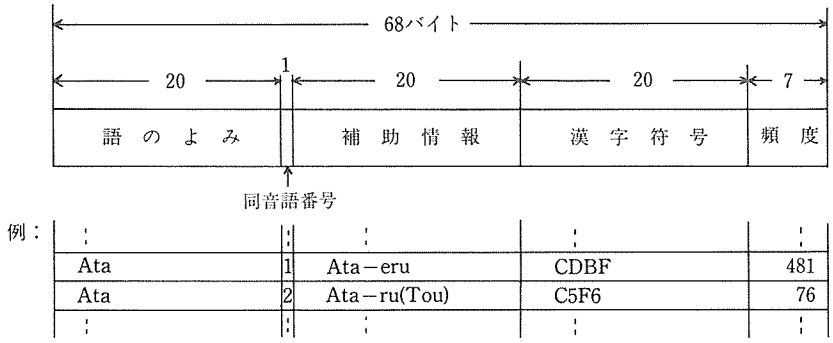

図－2　KKH IIのシステム構成

### 3．2　大辞書

KKHでは，各使用者が個別に変換辞書をもち，各自適応させてゆく。したがって，ある使用者の辞書に未登録の語が，他の使用者の辞書には登録済である場合も少なくない。そこで，これらを一括して複数使用者からなるひとつのグループ間で共用される辞書を設け，自分の変換辞書にない語でもこの共用辞書にあれば簡単に登録できるようにした。これを大辞書と称する。

各使用者の持つ変換辞書は入力中頻繁に検索されるので，主記憶上で動作させる必要があるが，大辞書の方は検索頻度が小さく，容量は大きくなるので，ファイル上におくことにしている。ま

た，大辞書中の語を各使用者の変換辞書に登録する際は自動ではなく，使用者に確認を求めてそ，のうえ登録するようにして，誤登録を避けている。図3に大辞書の構造を示す。




図－3　大辞書の構造

### 3．3　変換辞書の管理方式

前述のように，現行KKHでは削除語を選出する際に $h/n$ が最大のものを選ぶようしている。しかしながら，ある期間使用した後に変換辞書の内容を調査した結果，この方式には以下のような問題点があることがわかった。

1）ひとつの語の平均出現間隔に比べて履歴カウンタ $h$ の増加の度合が大きく，新出語として登録された語が次に出現するより以前に削除される場合がある。

2）逆に，短期間に多数回出現した語は，その後相当長期間一度も参照されなくても辞書内に残存し，辞書の利用効率を低下させている。

先に述べたように，パラメータ $h_0$，$n_0$ を適切に設定することにより，これらの問題を若干軽減することができるが，本質的には解決できない。そこで，KKH IIでは履歴 $h$ によって選出することとし，以下のような方式を採用した。

1）変換辞書中で，各漢字語は履歴 $h$ の小さいものから順に配列され，システム起動の際，すべての語について　　$h+1 \rightarrow h$ とする。

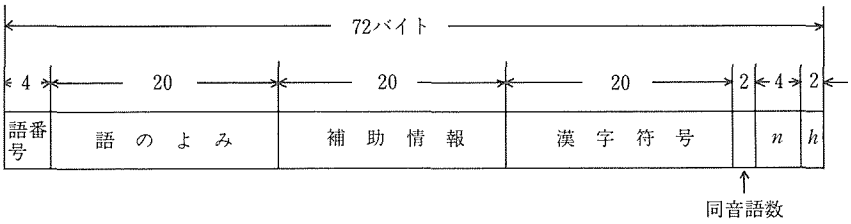
2）辞書に登録されているある語が参照されたとき，その語について

$0 \rightarrow h$，$n+1 \rightarrow n$ とする。

3）新出語は，大辞書中にあれば使用者の確認を経て登録し，なければ現行KKHと同じ方式[4]で登録する。このとき変換辞書が一杯ならば，$h$ が最大の語のグループをしらべ，その中で $n$ が最小のもの1語を選んで削除する。

4）削除された語は一旦削除語ファイルに保存しておき，一定期間ごとに大辞書の内容と比較して，必要ならば大辞書に書込む。

この方式では，ある語が過去にいかに高頻度で使用されても，その後使用されなければやがて $h$ が最大のグループに入り，いずれ削除される。しかし4）によりシステムから完全に削除されることはない。なお，$h$ は最大99までとし，それ以上は一定としている。また，変換辞書の構造は現行のものを多少修正し，図4に示すようにしている。




例：

| | | | | | | |
|---|---|---|---|---|---|---|
| 569 | Zi | Mozi | BBFA | 3 | 257 | 0 |
| 614 | Tanmatu | Terminal | C3BCCBF6 | 1 | 167 | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 1077 | Kairo | Circuit | B2F3CFA9 | 1 | 62 | 99 |

図－4　変換辞書の構造

### 3．4　同音語の自動選択

KKH の使用経験によれば，入力された漢字語のべ語数の 20%以上が同音語をもつ。同音語を最初の選択に固定することにより，手動による選択回数をほぼ同音語の語種程度に減らすことができるが，それでも約 10%はある。同音語の選択をシステム内で自動的に行うことができれば，この値をさらに減らすことができる。同音語の自動選択方法としては，語を意味属性によって分類し，意味の関連によって選択する手法が報告されているが[5]，各語について詳細な属性分類が必要であり，新出語を次々と登録してゆくわれわれの方式には不適当である。そこで，われわれは同音語とその前後にある文字の組を登録しておき，その比較によって選択する方法を用いることにした。この方法の概要を以下に示す。

1）ある文の中でよみwをもつ同音語 $w_1$，$w_2$ が，各々 $x_1\,w_1\,y_1$，$x_2\,w_2\,y_2$ という形であらわれているものとする。ここで $x_1$，$y_1$，$x_2$，$y_2$ は空白を含む任意の文字である。

例：　……が多い……，……は大き……

2）もし $(x_1, y_1) \neq (x_2, y_2)$ がつねに成立するなら，それにより $w_1$，$w_2$ を識別できる。

現実には，漢字語の前後につく文字は種々あり，また上記2）がつねに成立するとはいえない。しかし，以下に示すようにして，同音語の相当数について，いずれであるかを推定することができる。いま，$w_1$，$w_2$ が各々 $x_{1i}\,w_1\,y_{1i}$，$x_{2j}\,w_2\,y_{2j}$ $(i, j = 1, 2, \cdots)$ という形であらわれているものとする。

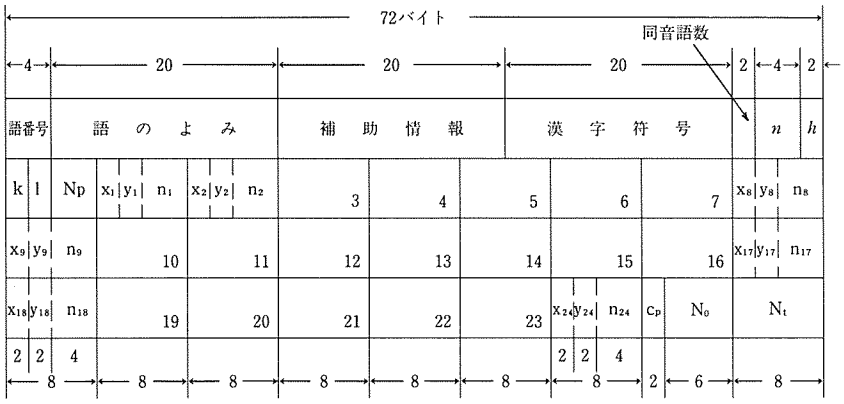
1）各同音語について $(x_{1i}, y_{1i})$，$(x_{2j}, y_{2j})$ の組とそれぞれの頻度を記録する。これを同音語辞書という。

2）入力文中に x w y という3つ組が出現したとき，wについて同音語辞書を検索する。

3）(x, y)が同音語辞書に記録されている $(x_{1i}, y_{1i})$，$(x_{2j}, y_{2j})$ のいずれかと一致し，かつその組の頻度が一定値をこえているとき，それにしたがってwは $w_1$ または $w_2$ のいずれかであるとして自動選択するとともに，当該の組の頻度に1を加える。

4）一致しているがその組の頻度が設定値に達していないときは使用者に選択を促し，それにかたがって当該の組の頻度に1を加える。

5）一致する組が存在しないときは使用者に選択を促し，それにしたがって新たにあらわれた(x, y)の組を同音語辞書に登録する。




k：同音語番号　　l：登録されている(xy)組の個数(1～24)

$N_p$：次の同音語番号のものへのポインタ

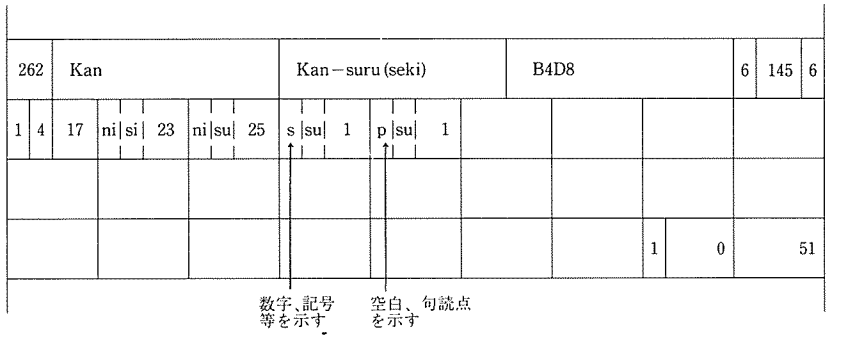
$x_1$ $y_1$ $n_1$：$x_1$…語の前の文字，$y_1$…語の後の文字，$n_1$…頻度

$C_p$：(x,y)組以外のものの出現頻度

$N_0$：登録されている(x,y)組以外のものの出現頻度

$N_x'$：この同音語の出現頻度合計

例：




図－5　同音語辞書の構造

6）（x，y）＝（$x_{1i}$，$y_{1i}$）＝（$x_{2j}$，$y_{2j}$），すなわち同一の3つ組が$w_1$，$w_2$のいずれにもあるときは，両者の頻度を比較し，その比が一定値をこえていれば大きい方であるとして自動選択し，当該の組の頻度に1を加える。

7）頻度の比が設定値に達していないときは使用者に選択を促し，選択されたほうの組の頻度に1を加える。

以上のアルゴリズムにより，特定の前後文字組で多数回出現した同音語を自動的に選択することができる。なお，x，yの分類は，かなの場合は各々の文字であるが，漢字，記号その他については各々「漢字」，「記号」というように一括して扱っている。なお，1個の同音語にあたり登録できる（x，y）組の種類は24個までとしている。図5に同音語辞書の構成を示す。

この方式は現行KKHにも組込まれて一部試用されており，同音語のべ語数の50%程度を自動選択できた例がある[6]。

## 4. おわりに

使用者ごとに個別に編成された小容量の変換辞書を用いる，研究者向きかな（ローマ字）漢字変換システムの開発について述べた。上述のように，本システムは現在その第1版（KKH）が稼動しており，その使用経験からあきらかになった問題点を解決するため，第2版(KKH II) の開発が進められている。しかしながら，未だ残されている問題点もいくつかある。

そのひとつは，KKH の実際の使用において，鍵盤のシフト操作ミスに起因するエラーがかなり多いことである。この問題は鍵盤操作の練習によりある程度改善されるが，本質的にはシフト操作の少ないような入力仕様にすることによって解決すべきである。KKH，KKH II では，シフト操作，すなわち大文字，小文字により漢字とかなの区別を行っている。それゆえ，将来は漢字，かなの区別，その他の区切りに際し，人為的な区切りを行わず，いわゆるべた書きのまま入力する方式を採用すべきであると考えられる。われわれは現在，べた書き入力についても基礎的な検討を進めており，KKH II の次の版で実現したいと考えている。

次に，漢字プリンタには通常の鍵盤にない多数の文字(ギリシャ文字，ロシア文字)，記号がある。これらを入力するためと，漢字プリンタを制御するための種々の情報を表現するために，鍵盤上にある文字，記号の組合せを用いているが，これらについてもシフト操作を要するものが多く，また指使いの難しい組合せがある。これらの一部は KKH II で改良されているが，さらに今後の検討に待つところも多い。

また，KKH II に組込まれる新しい機能，制御方式には，システムの完成後相当多数の文献入力を行い，その結果にもとづいて評価すべきものが多い。これらについては別な機会に報告したい。

## 謝　辞

KKH の使用経験の蓄積，および KKH II の仕様検討にあたり，種々御協力，御討論いただいた研究室の各位に感謝します。

## 参考文献

1）栃内香次，斎藤　康：情報処理学会論文誌，Vol.24，No.2，pp.209－213（1938）
2）斎藤　康，岡沢好高，栃内香次，永田邦一：北大工学部研究報告，No.108，pp.43－52（1982）
3）岡沢好高，斉藤　康，栃内香次，永田邦一：情報処理学会第24回全国大会，5G－1（1982）
4）岡沢好高，栃用香次，永田邦一：北大工学部研究報告，No.116，pp.79－86（1983）
5）牧野　寛，木澤　誠：情報処理学会論文誌，Vol.22，No.1，pp.59－67（1981）
6）伊藤太亮，栃内香次，永田邦一：情報処理学会第27回全国大会，2H－6（1983）