解　説

——文字認識技術への応用(2)——

# 印刷漢字認識の現状†

坂井邦夫††　渡辺貞一††

## 1. まえがき

日本語情報の電子計算機処理に対する要求と関心は，最近とみに高まりを見せている．日本語情報処理における最大の問題の一つが入力の困難性にあることは広く知られており，漢字を含む日本語文字のパターン認識が，この問題の解決に役立つものと期待されている．中でも，印刷漢字を読取る装置（以下，印刷漢字OCRと呼ぶ．）の役割は，毎年$10^{12}$〜$10^{13}$の割合で紙に記録された形で発生される大量の日本語データ，そして過去十数年にわたって蓄積された膨大な量の日本語データのうち，活字で印刷されたものを読取り，電子計算機にファイルを作成（日本語情報のコンピュータ入力）することである．電子計算機に日本語情報を入力することを高速・高精度に行えるようになれば，従来はあまりにもコスト高につくという理由から見送られてきた新しい日本語情報サービスが可能となってくる．

本稿では，印刷漢字認識の現状技術を実用化の観点から概観し，合わせて今後の課題を述べる．

## 2. 印刷漢字の認識方法

漢字認識の研究は古くは1960年代に米国で始められており[1]，その後の多くは漢字使用国の我が国において1970年以降活発に行われている[2),3)]．漢字は英数字と対比すると以下のような特質があり，これらはすべて漢字認識を困難なものとしている要因である．

（1）字種が多い．少なくとも2,000種が常用されており，英数字の数十倍にもなる．

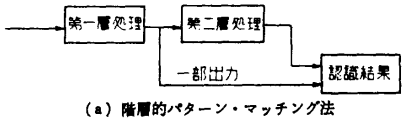
（2）形状の複雑さの分布範囲が広く，大多数のものは英数字に比べて格段に複雑である．

（3）部分パターンを共有する文字が多く，構造がきわめて類似した文字対が存在する．

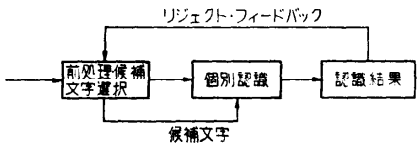
（4）文字の大きさや字体（たとえば明朝体，ゴシック体など）が多種・多様である．

以上の困難性を克服するための基本的なアプローチは，①候補文字選択をあらかじめ行い，見かけ上の認識対象文字数を減少させて効率良く認識する，②候補文字群の中の互いに類似した文字対を正確に弁別できる高精度の認識方式を開発する，③前処理を強化して入力データの多様性を吸収するなどである．ここで漢字認識とは，いうまでもなく漢字，平仮名，片仮名，英数字，記号から成る混合文字セットを認識することである．

これまでに発表された漢字認識法[4)〜6)]のほとんどは上記①の意味での階層構造認識方式を採用している．また，認識段階では，パターン・マッチング法（重ね合せ法）が用いられることが多い．印刷漢字認識の方法論として代表的なものには，図-1に示す2つの方法がある．まず，階層的パターン・マッチング法[7),8)]では候補文字選択を粗い分解能のパターンを用いたパターン・マッチング法により行う．第一層では入力パターンを縦横50点ほどの二値量子化パターンで表わし，これを圧縮して縦横8×8点で深さが4ビットの多値量子化パターンに変換する．この圧縮パターンと同次元の辞書パターンを全認識対象文字について用意しておく．入力パターンの圧縮パターンと全辞書パターンとを重ね合わせて類似度を計算し，類似度の上




(a) 階層的パターン・マッチング法




(b) 特徴抽出による大分類法

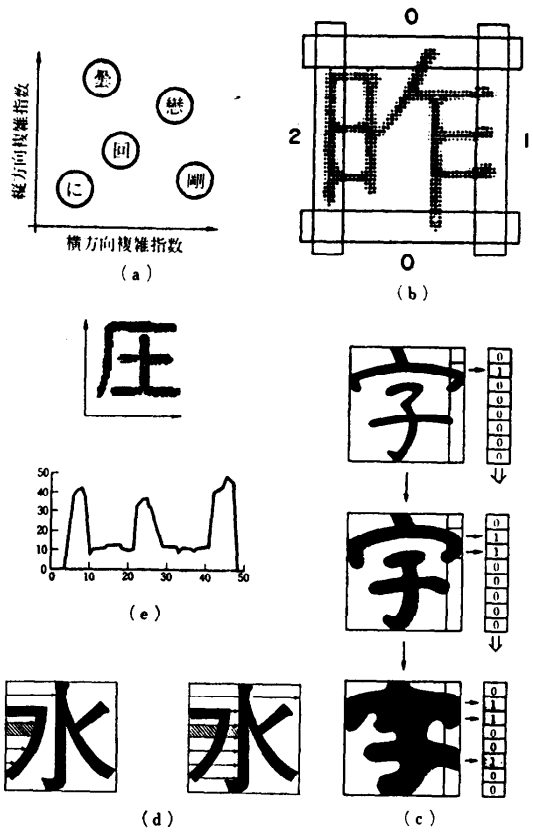
図-1　印刷漢字認識の方法論

† The State of the Art in Printed Chinese Character Recognition by Kunio SAKAI and Sadakazu WATANABE (Toshiba Research and Development Center).

†† 東京芝浦電気(株)総合研究所

位に入ったものを候補文字として出力する．候補を一つに絞れた場合は，その文字を認識結果とする．第二層では圧縮前の二値量子化パターンと，このパターンと同次元の別の辞書パターンとの照合を候補文字についてのみ行い，最終的な認識結果を出力する．この方法の長所は，パターン・マッチング法を候補文字選択と認識に共用しているため，圧縮の程度が少ない場合には比較的小数の候補文字が得られる．しかし，各層ごとに別個の辞書パターンを用意する必要があること，総当りの照合が必要なことなどが欠点である．次に，特徴抽出による大分類法[9)~11)]では，何らかの特徴を用いて，あらかじめ全認識対象文字を複数の類似文字群に大分類しておく．入力パターンについても同種の特徴を抽出し，このパターンがどの類似文字群に属するかを決定し，その類に含まれる文字群を候補文字とする．認識は先と同じく候補文字についてのみ行う．認識結果が読取拒否となる場合は，候補文字選択部にフィードバックをかけて候補文字選択からやり直す．この方法の長所は，前処理時間中に特徴抽出・候補文字選択を行うことにより高速な認識システムを組むことができる点にある．ただ，リジェクト・フィードバックをかけられるためには，個別認識の方法が相当信頼のおけるものでなくてはならない．

候補文字選択のための特徴としては，図-2 に示すような方法が提案されている．複雑指数[12),13)]は，文字パターン全体の文字線密度によって定義される量である．漢字パターンが主として縦方向と横方向の線分によって構成されることから，縦横2方向に分けて計算される．この量は文字の位置と大きさには無関係の無次元量である．四辺コード[12),13)]は漢字パターンの上下左右の四辺が，線幅変動雑音に対して安定でかつ分類情報を多く含むことに着目した方法である．四辺の検査領域内の文字線分の量を0，1，2の3値に量子化し，任意の文字を4桁3値の符号で表わす．帯パターンの時刻変化[10)]は，文字パターンを人工的に順次太めることによって，四辺コードの時間的変化をとらえる考えに立っている．ペリフェラル・パターン[11)]も周辺特徴を用いる方法である．文字を囲む4つの外接枠を8分割し，各枠部分から最初に文字線に出合うまでの領域面積（斜線部）の割合を計数する．同じく2度目に文字線に出合うまでの面積の割合を計数し，計 64 次元のベクトルを求めて分類特徴とする．周辺分布は縦および横方向の軸に文字パターンを投影し積算して得た波形を用いる方法である．候補文字選択法の評価基準は信頼性，安定性，効率，柔軟性，簡潔性の5点である．相補的な複数の特徴を組合わせることにより，これらの要請を満たせることが確認されている[9)]．




(a) 複雑指数
(b) 四辺コード
(c) 帯パターンの時刻変化
(d) ペリフェラル・パターン
(e) 周辺分布

図-2　大分類特徴の例

印刷漢字の認識段階にパターン・マッチング法，とりわけ類似度法が主として用いられる理由は，多数の読取文字種に対して辞書の自動設計が可能なこと，装置化が比較的容易なこと，入力パターンの局部に加わる雑音に対して一般に安定であることなどがあげられる．しかし，文字位置や文字線幅の大幅な変動に対しては，類似度の低下をもたらすため，これらの雑音に対する対策を講じておかないと，高精度の認識を行うことはできない．たとえば，重ね合わせ位置を何度かずらして類似度の最大値を求めるなどの方法がとられていた[14)]．これに対して複合類似度法は[15)]，文字パターンに混入する雑音成分が入力文字パターンと独立では

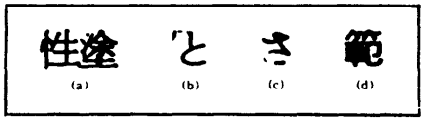
なく強い相関関係を持っていることに着目し，両者の位相関係を理論的に解明することによって，雑音に対して本質的に安定化を行える方法である．

類似文字対の識別は，漢字認識の中でも最もむずかしい問題の一つである．これに対しては，差分マスク・パターンを用いた部分パターン・マッチング法や対判定荷重相関法などが試みられてきた．これらの方法はマスクや荷重の決定に人手を要し，パターン・マッチング法本来の良さを相殺してしまう恐れがある．混合類似度法[16]は複合類似度理論を発展させたもので，類似文字対に対する弁別能力をさらに向上させる特長がある．類似文字組を与えることにより，辞書パターンを自動設計することが可能である．

## 3. 印刷漢字 OCR の具体例

以上に述べた印刷漢字認識の方法のうちのいくつかは実際にハードウェアが設計され，その有効性確認のための実験が行われている．日立で試作した OCR[14]は不動産登記申請書読取を想定したもので，読取精度向上のための後処理を組込んでいる．入力データとなるものは，一定の大きさの和文タイプ活字である．富士通[10]も OCR を試作しているが，特に用途は明らかにされていない．また，大型プロジェクトパターン情報処理システムでは，特許公報，文庫本，単行本，和文タイプ原稿などの実データを読取り，入力から修正までを一貫して行える印刷漢字 OCR (委託先東芝)[9),12),13)]が開発され，昭和55年10月には約1カ月間の公開実演が行われている．ここでは，最近開発されたこの大型プロジェクトの印刷漢字 OCR を代表例として少し詳しく紹介する．

印刷された日本語の文書は，文中に使用される文字の種類，文字サイズ，字体，印字品質，紙質などが多種多様である．日本語の OCR に対する規格は現在の所，全く存在せず，印字品質一つに限っても，図-3に示すような低品質文字が入力されることも希ではない．規格や制限を前提条件とすることができない以上，OCR 側の方に入力データの多様性を吸収し得る機能を持たせる必要がある．




(a) 隣接文字間の接続
(b) 印字の際の汚れ
(c) 文字線のかすれ
(d) 文字線のつぶれ

図-3　印刷漢字 OCR の入力データとなる低印字品質文字の例

表-1　印刷漢字 OCR の性能諸元

| 形　式 | ページ式 |
|---|---|
| 取扱用紙 | 普通紙，OCR 用紙<br>A 4～A 6 |
| 書類様式 | フリー・フォーマット<br>縦組および横組 |
| 読取文字 | 最大 2,176 字種<br>8～12 ポイント<br>任意字体（辞書切換） |
| 読取速度 | 100 字／秒<br>4 枚／分 (A 5, 10行の場合) |
| 修正・編集 | 日本語ワードプロセッサ |
| 出　力 | フロッピ・ディスク，通信回線 |

試作された印刷漢字 OCR (プロトタイプ・モデル)は，ハードウェア・コストを低下させ，実装を容易にし，実用的な漢字入力装置に必要な機能・性能を実現するために，以下の設計思想に基づいて設計されている．(表-1 に基本的な性能諸元を示す．)

(1)　**読取文字種**　日本語の文章では 6,000 種以上の漢字が使用されているが，使用頻度からみると，2,000 種の文字で約 99% をカバーする．ハードウェアの規模とスループットを勘案して，約 2,000 種を読取文字種としている．ただし，文字種の組替えは任意である．

(2)　**読取文字サイズ，字体**　通常の日本語文書は 8～12 ポイント* の活字で印刷されている．大きさの正規化を施すことにより，この範囲内の文字を一つの辞書で読取れるようにしている．また，字体に関する制限は無く，明朝体，ゴシック体をはじめとする任意のものを読取れ，混在もさしつかえない．

(3)　**読取領域の指定**　頁ごとに，読取領域指定を素早く簡単に行えるよう，タブレットを用いた即時フォーマット指定方式および機構走査制御方式を開発している．

(4)　**読取速度**　読取速度は修正・編集に要する時間とつり合いがとれていなくてはならない．この場合にシステムとしてのスループットは最も高くなり，かつ，無駄が無いからである．修正端末での処理時間，読取精度（後述）を考えると，100 字／秒 の読取速度は十分この要請を満たしている．

(5)　**修正・編集**　日本文を入力データとする限

* 1ポイントは 1/72 インチ角の大きさをいう．

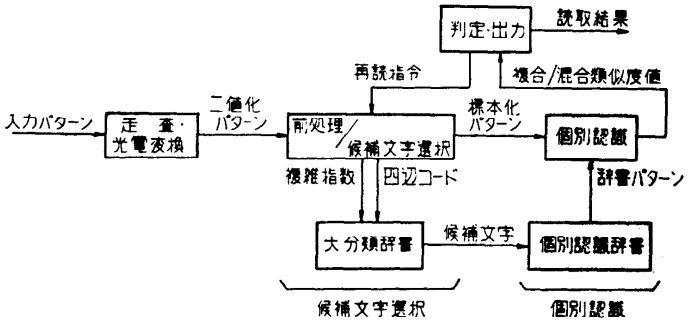



図-4　印刷漢字 OCR の認識方式ブロック図

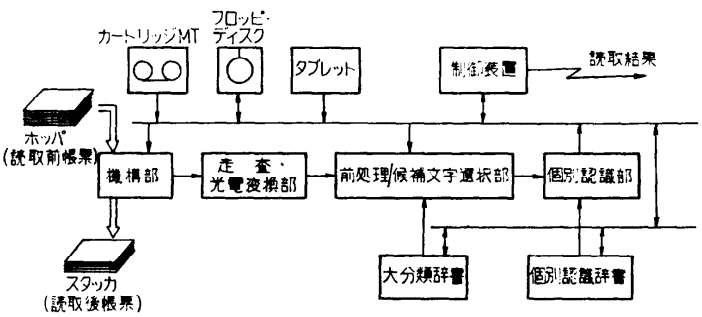



図-5　印刷漢字 OCR のハードウェア構成

り，その編集には日本語に対するワードプロセッシングの機能が必要となる．読取結果の確認，誤読文字や読取拒否文字の修正，出力形式の変更などを行える日本語ワードプロセッサを OCR 本体に接続している[17]．読取と修正・編集の機能は完全に独立しており，並列に動作できる．

図-4 は印刷漢字 OCR の認識方式ブロック図である．前処理部と一体化された候補文字選択部が個別認識部の前段に配置されている．入力パターンについて計算された複雑指数と四辺コードが大分類辞書検索のキーとなり，2つの方法を併用した場合の候補文字が個別認識辞書パターンの格納アドレスの形で順次出力される．個別認識部では，各候補文字ごとに読出された辞書パターンと前処理の結果としての標本化パターンとを照合し，複合類似度もしくは混合類似度を計算する．判定・出力部では類似度の最大値と次最大値とから判定を行い，読取拒否の場合は候補文字選択部に再読を指令する．再読の場合は複雑指数もしくは四辺コード単独での候補文字について個別認識をやり直す．最後に全候補文字についての判定が行われ，認識結果の文字コードが出力される．

この二段階の階層構造認識方式の利点は次の通りである．

（1） 候補文字選択を行うことにより，認識対象文字数を実効的に減少できる（約 1/20～1/30）ので，認識速度の高速化をはかれる．

（2） 個別認識部から候補文字選択部へのフィードバックにより，候補文字選択の一時的な誤りを回復できるので，認識精度を下げることなく効率的な認識を行える．

（3） 簡単かつ有効な特徴を用いた候補文字選択法，理論に裏付けられた高度な個別認識法の採用により，高速・高精度の認識を小規模のハードウェアで実現できる．

図-5 は印刷漢字 OCR のハードウェア構成である．

機構部のホッパには読取帳票を一枚ずつ手挿しで供給する．OCR 用紙の場合は自動給紙も可能である．この部分は 8 ビットのマイクロ・コンピュータにより制御される．走査・光電変換部は 1,728 ビットの CCD ライン・スキャナを 2 本用い，約 60 $\mu$m の解像度を出している．サンプリング・ピッチは 56.4 $\mu$m であり，8 ポイントの漢字が 50×50 点の量子化パターンとして得られる．前処理部は，走査して得られた一行分のイメージから一文字分のイメージを取出す検出切出部と，位置および大きさの正規化を行う正規化部と，候補文字選択のための特徴抽出部から成っている．この部分は専用ハードウェアとマイクロ・プログラムによるファームウェアとで構成されている．候補文字選択部は専用ハードウェアである．候補文字の作成に要する時間は 4 ms である．個別認識部も積和 LSI を並列に用いた専用ハードウェアである．類似度の計算に要する時間は 1 候補文字当り 40 $\mu$s である．類似度のソートや答の決定はマイクロプログラムによるファームウェアで行われる．個別認識辞書は 256 k ワード（1 ワードは 36 ビット）の IC メモリである．カートリッジ磁気テープを辞書パターン・メモリのバックアップに用いている．制御装置は 8 ビットのマイクロ・コンピュータであり，システム・プログラムと辞書パターンのロード，フォーマット制御情報の作成，読取結果の編集と答ファイルの作成，修正編集端末との交信を受け持つ．

図-6　印刷漢字 OCR の全景（写真）

機構走査部，前処理部，候補文字選択部，個別認識部はパイプラインで動作し，制御装置は一行分の答を約1秒ごとに受取ることができる．

図-6 に装置の全景を示す．一般に OCR の性能は，文字認識の速度，精度，自由度・柔軟性の3点において特徴づけられる．試作 OCR の文字認識速度は人間の入力速度の 50 倍以上高速である．また，システムのスループットも人手による入力のそれよりも1桁近く高い．精度は通常の OCR の規格に適合する品質の文字に対しては，正読率 99.9% 以上，前記の実データに対しても 99% 以上である．

また，普通紙に両面印刷されたさまざまな大きさの文書が扱え，その中の任意の部分を指定して読取れること，縦または横に配列されたさまざまな大きさの文字を一つの辞書で読取れることなど，従来の OCR にはなかった機能を持っている．

## 4. 応用分野と将来展望

印刷漢字 OCR の応用分野は冒頭にも述べたように，日本語情報のコンピュータ入力にある．漢字入力装置[16]と呼ばれる現用の機器のほとんどすべて(図-7)は人間がキーを叩く操作に拠っており，英文タイプライタに比べて入力速度が遅い．また大部分のものは習熟したオペレータを必要とするなどの欠点がある．したがって，蓄積された大量の日本語データを入力するとなると，コストと速度の両面から採算がとれないことになる．

本稿で述べたような印刷漢字 OCR が大量データの高速・集中・一括入力に用いられるならば，従来はあまりにも高価になるとして諦められていた入力が可能となる．

たとえば事典や文庫本・単行本などを再版する際の

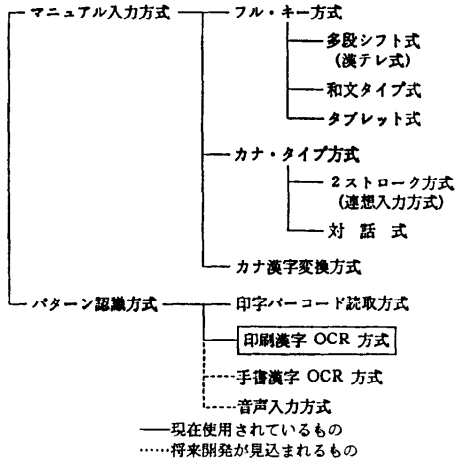



図-7　漢字入力方式の系統的分類

入力，特許公報や技術文献の読取りによる日本語情報データ・バンクの作成，名簿や各種申請書の読取りによるファイルの作成などが考えられている．

今後の課題としては，OCR 側の問題，OCR を使用する側の問題，規格や精度の問題がある．第一に OCR 側の問題としては，よりコスト／パフォーマンスの良い装置を開発することである．人手による漢字データの入力（約1.8円／字）よりも価格的に有利であることが実用に供されるための必要条件となる．第二に OCR を使用する側の問題としては，OCR の可能性と限界を正しく理解し，人間と機械とが補い合うようなマン・マシン・システムを構築することが肝要である．第三に規格や制度の問題は日本語情報処理全般にわたる大きな問題である．このうち，OCR に関連する事項としては，漢字プリンタのドット文字パターンを標準化しようという動きがある．英数字・片仮名字形の JIS 化が現用 OCR の普及を促進したことを考えると，印刷漢字パターンの規格化も漢字 OCR の実用化にとって重要な課題の一つであるといえよう．

## 5. ま と め

文字認識は，いわゆるパターン認識と呼ばれるものの中で，最も古くから研究されかつ実用化の進んだ分野である．過去 20 年間にわたる研究成果の積み重ねと，昨今の半導体技術の進歩に支えられて，印刷された日本語文書を高速・高精度に読取る OCR が開発された．ここで用いられている主要な技術は次の通りである．

（1）　多数の文字パターンを効率良く安定に大分類

し，小数の候補文字を選択する技術

（2） 候補文字の中の互いに類似した文字を正確に識別することのできる高精度の文字認識技術

（3） 入力文字列から一文字のパターンを正確に切り出し，また印字の汚れや文字線の切れ，かすれ，つぶれなどの雑音に対して影響されにくい光電変換・前処理技術

（4） 読取用紙上の任意の領域を指定して読取ることを可能とする柔軟なフォーマット・コントロール技術

（5） これらの技術を実際の装置の中に組込み，コンパクトな日本語入力 OCR としてまとめ上げた文字認識ハードウェア技術

なお，本稿では触れなかったが，文字認識の最後の難関を突破すべく，手書漢字認識の研究もすでに開始されている．漢字 OCR をパターン認識による日本文の機械入力としてみなすならば，手書漢字と印刷漢字を同一の装置で読取れることが適用分野を広げるために必要である．また，文字認識技術の高度化という立場で考えるならば，手書と活字を同一原理で認識する統一的な方式を開発し，その有効性を実証すべきであろう．今後の研究成果に期待したいと思う．

## 参 考 文 献


1) Casey, R. and Nagy, G.: Recognition of Printed Chinese Characters, IEEE Trans., Vol. EC-15, No. 1, pp. 91-101 (1966).
2) 長尾：漢字情報処理システムを考える，エレクトロニクス，Vol. 17, No. 1, pp. 42-45 (1972).
3) 通産省工業技術院，大型プロジェクト・ニュース (1973).
4) Stallings, W. : Approaches to Chinese Character Recognition, Pattern Recognition, Vol. 8, No. 2, pp. 87-98 (1976).
5) 日経エレクトロニクス編集部：印刷，手書き漢字の認識技術を展望する，日経エレクトロニクス，No. 154, pp. 42-59 (1977).
6) 増田：日本語文字読取り装置，電子通信学会誌，Vol. 63, No. 7, pp. 719-723 (1980).
7) 山本他：階層的パターン・マッチング法による漢字認識の実験，電子通信学会論文誌，Vol. 56-D, No. 12, pp. 714-721 (1973).
8) 中野他：周波数領域での階層的パターン整合法による漢字認識，電子通信学会論文誌，Vol. 58-D, No. 2, pp. 94-101 (1975).
9) Sakai, K. et al.: An Optical Chinese Character Reader, Proc. of 3 IJCPR, pp. 122-126 (1976).
10) 藤田，中西，宮田：帯パターンの時刻変化を用いた印刷漢字認識方式，情報処理，Vol. 17, No. 12, pp. 1098-1104 (1976).
11) 梅田：マルチフォント印刷漢字認識のための粗分類，電子通信学会論文誌，Vol. J 62-D, No. 11, pp. 758-765 (1979).
12) 森，坂井：2,000 字種を 100 字/秒で読む印刷漢字 OCR の開発，日経エレクトロニクス，No. 172, pp. 102-128 (1977).
13) 坂井他：印刷文字認識システム，大型プロジェクト・パターン情報処理システム研究開発成果発表会論文集，pp. 29-45 (1980).
14) Fujisawa, H. et al.: Development of a KANJI OCR: An Optical Chinese Character Reader, Proc. of 4 IJCPR, pp. 816-820 (1978).
15) 飯島，森：人間の識別能力に迫る OCR, ASPET/71，日経エレクトロニクス，No. 30, pp. 66-80 (1972).
16) 飯島：混合類似度による識別理論，信学技報，PRL 74-24 (1974).
17) Kawada, T. et al.: Linguistic Error Correction of Japanese Sentences, Proc. of COLING 80, pp. 257-261 (1980).
18) 日本電子工業振興協会：日本語情報処理の調査研究 (1979).


（昭和 55 年 12 月 8 日受付）