サブ・パタンの導入による文字辞書の構造化

益田千冬、中川正樹、高橋延匡(東京農工大学　工学部　数理情報工学科)

１．はじめに

ワード・プロセッサの出現は、それまでの文書作成の形態に革命的変化をもたらした。文書の保存ができる、同じ文書を何度も呼び出すことができる、修正が容易にできる、しかもきれいな出力が得られる、という利点からその普及は目覚ましく、オフィス・オートメーションの発展に一層の拍車をかけた。しかし、これほどの普及にもかかわらず、現状では使い勝手の面で問題点が山積しているのも事実である。入力方式を考えても、日本文を入力するのに優れた入力方式が確立していない。現在は仮名漢字変換方式が主流であるが、もともと欧米文化の産物であるキーボードで漢字を入力しているので、日本語の特徴が充分に生かされていない、不自然な入力方式である。

容易でかつ自然な日本文入力方式として、我々はオンライン手書き文字認識システムの研究を行なっている。この研究プロジェクトはJOLIS(Japanese On-Line Input System) として1979年に始められた。現在、JOLIS-1 の構築および評価を終え、JOLIS-1 後継システムJOLIS-2 の構築段階にある。

JOLIS-1 の評価により得られた知見の一つに、漢字の“へん”や“つくり”などのある部分パタンを誤った筆順やつづけ字で書く習慣があると、その部分を含むすべての文字が認識されない、ということがある。これに対処するために、漢字の構造を認識に反映させることを考えた。ただし、構造を認識手続きに反映させるには、パタン中のエラーに対処した構文解析を行なう必要がある。パタン認識の構造解析的手法において、構文中のエラーに対処するためには、stochastic syntactic parsing が提案されている。しかし、この手法は計算時間がかかるため、現段階では実現が困難である。そこで表現(representation)として文字辞書に反映させ、認識はトップ・ダウンのマッチングを行なうことを試みた。漢字には、“構造を有する”という特徴がある。この特徴を生かして文字辞書を作成することにした。この、漢字を構成する部分を“サブ・パタン”と名付けた。

本報告では、サブ・パタンの導入によりJOLIS-1 の文字辞書を構造化したこと、および、その辞書を用いた認識実験と結果について述べる。

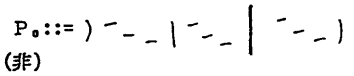
２．サブ・パタンの導入

２．１　サブ・パタン導入の目的

JOLIS-1では、それぞれの漢字をストローク・コードの列で表現している。漢字の変形(つづけ字等)に対しては、そのストローク・コード列を辞書に追加登録することで対処している。しかし、これでは辞書の増大を招いたり、他の文字に認識されてしまうなど、漢字の変形をカバーするのには必ずしも効果的でない。そこで、漢字の“へん”や“つくり”といった構造に注目して辞書を構造的に記述することを考えた。これによって、同じ“へん”や“つくり”を持った漢字は、その部分の情報を共有することができ、記述の無駄を省くことができる(図１参照)。また、これにより、つづけ字や筆順誤りに統一的に対処することも可能となる。さらに、サブ・パタンによる辞書の構造化により、辞書更新の手間を省くと同時に、半自動生成も可能となる、と考えられる。

数千という文字に対する大量のデータを辞書という形で登録する作業には多大な労力を必要とする。この作業を少しでも軽減すること、さらに辞書の管理を統一的に行なうことが要求されている。特に、開発中の実験システムでは、そのツールとしての辞書更新プログラムが重要となる。この要求に応えるような辞書管理システムを作るためにも、辞書を構造的にするメリットが期待できる。以上、サブ・パタン導入の主な目的をまとめると次の項目が挙げられる。

(1) 辞書の構造化
(2) つづけ字(一定の略字)、筆順誤りへの統一的対処
(3) 辞書更新、管理に要する手間の減少



悲 ::= P。 ＋ 心

俳 ::= イ ＋ P。

図１　部分の共有

２．２　サブ・パタンの概念

サブ・パタンの導入による辞書の構造化によっていくつかの効果が期待できる。本節では、そのサブ・パタンをどのように定義づけるか、について述べる。

日本語では、画数の変動が大きく、しかも、その種類が数千にも及ぶ“漢字”という文字がある。しかし、その数千の漢字が全く別々の特徴を持つパタンではなく、ある部分である共通の特徴をもつことがわかる。本来、漢字は“へん”や“つくり”といった部首から成り立っている。その基本となる部首は、320 ほどあり、すべての漢字はその組合せにより構成されている。“へん”、“つくり”、“かんむり”にはそれぞれ意味と読みがあり、それが漢字の意味及び音を決定している。このために、初めて見る漢字でも意味や読み方がわかる例が少なくない。我々は漢字を“一つの表意図形”としてではなく、“表意図形の集まり”として無意識に見ているのではないか。そしてこのことが、漢字を書くときにも反映しているのではないか、と思われる。

手書き漢字の形は書く人によって千差万別である。文字(文章)を長い間書いているうちに、変形が加えられている。筆の流れに従って２本の線が１本になったり、略されたりしている。また、文部省の定めた筆順ではなく、各個人の都合のよい筆順になっているのも一つの変形と考えられる。しかし、一般的にしろ自己流にしろこのような変形にもかかわらず、どういう漢字であるかを認識することができる場合が多い。これは、漢字が変形したにもかかわらず、その基本構造を保っているためではないだろうか。具体的には、“へん”、“つくり”といった構造の単位内での変形が多く、“へん”、“つくり”自体はその特徴を残しているためではないか。

以上の考えに基づき、オンライン手書き文字認識におけるサブ・パタンは、

「“へん”、“つくり”、“かんむり”を基本とし、漢字の変形を吸収しうる範囲のストローク・コードの集まりを一つの単位とした漢字の構成要素」

と考えた。

２．３　サブ・パタン設定の方針

JOLIS-1では、平仮名81文字、片仮名81文字、漢字996文字、記号4文字の計1162文字を認識対象としている。そこでまず、漢字996文字を基に、サブ・パタンを設定することにした。第一版としてのサブ・パタンの設定は以下の方針に基づいて行なった。

① 漢字の“へん”、“つくり”、“かんむり”等の部首を基本とする。
② つづけ字による“まとまり”を考慮する。ストロークがつながっている一まとまりを一つのサブ・パタンとする。
③ 一文字当り２～４個のサブ・パタンで構成する。
④ 漢字の構造を反映させ、辞書の更新、拡張を容易にできるように設定する。

方針①、②はサブ・パタンの概念に基づいている。③は、サブ・パタン設定の目的の一つとしての“文字辞書の構造化”を考慮している。低画数文字を細かく分けてしまうと、(1) 今までのストロークがそのままサブ・パタンとなり、構造的にすることによりかえって検索の手間が大きくなること、(2) 高画数文字に対してそのサブ・パタンを適用してもつづけ字等への対処にならない、といった弊害が予想されるので、分割しすぎることがないようにした。④は、認識対象文字の拡張に伴うサブ・パタン辞書の追加をなるべく押さえるためである。サブ・パタンはJOLIS-1 後継システムにおいて、漢字を記述する最小単位であるという考えから、一度設定したサブ・パタンの変更・追加は最小限に留めたい。

サブ・パタン導入の目的の一つとして、つづけ字や筆順誤りへの対処ということを掲げた。その単位として“へん”、“つくり”を基本とする部首を考えた。しかし、厳密に部首に従って漢字を分けると、細かく分割されてしまって、漢字の変形を吸収しうる単位ではなくなってしまうことがある。たとえば、“公”は部首に分けると“ハ”と“ム”に分割されるが、行書レベルでは“  ”のようにつながる場合が多い。しかし一方で、このようなつづけ字への対処を考慮して、これを含む大きな単位でだけサブ・パタンを設定すると、サブ・パタン間の共通部分の共有ができなくなり、サブ・パタンを用いた構造的な辞書の利点が生かされない(図２)。そこで、サブ・パタン自体に構造を持たせることで拡張にも対処できるようにした。漢字の部首といくつかのまとまった部分を最も基本的なサブ・パタン(以下これを基本サブ・パタンとする)とし、それらを組み合わせたものを、その上のレベルのサブ・パタン(以下これを中間サブ・パタンとする)とする。漢字の構造によっては何段階ものサブ・パタンから成る。基本サブ・パタンを段階的に組み合わせて、最終的にはそれが一つの文字となる。図３にその概念図を示す。

億 ::= イ + 意　　とすると、

意 ::= 意|意|竟|意

意 ::= GAGGAGIAAFKHH ——— それぞれ別のパタンとして

意 ::= G$GIAAFKHH ——— 持つことになる

竟 ::= GAGGAG#KHH

意 ::= G$GILFKHH ——— (英記号はストローク・コード)

図2　サブ・パタンのサブ・パタンを持たない場合

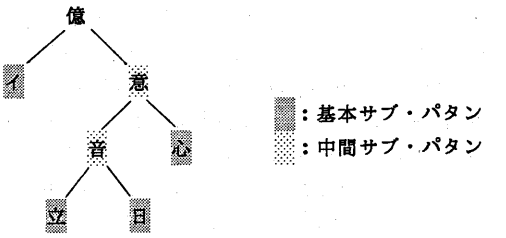


図3　サブ・パタン概念図

以上の方針に基づき、まず基本サブ・パタンの設定を行なった。その一部を図4に示す。

基本サブ・パタンNo.5
基本サブ・パタン名B 01005
出現確率　50
ストローク
No.4　画数 1　I

基本サブ・パタンNo.14
基本サブ・パタン名B 02009
出現確率　50
ストローク
No.10　画数 2　I V

基本サブ・パタンNo.7
基本サブ・パタン名B 02002
出現確率　50
ストローク
No.6　画数 2　Y K

基本サブ・パタンNo.16
基本サブ・パタン名B 02011
出現確率　50
ストローク
No.12　画数 2　Y G
No.5　画数 2　A G

図4　基本サブ・パタン(一部)

“へん”、“つくり”を基本として分けていくうえで問題になった部分がいくつかあった。その一つに、“くにがまえ”と“ぎょうがまえ”がある。“くにがまえ”を一つのサブ・パタンとして扱えると都合がよい。というのは、“くにがまえ”を持つ漢字は、“国”、“困”、“回”、“囲”、“団”、など10余りある。拡張性から考えてもサブ・パタンとして扱いたい。しかし、筆順を考えると、例えば“国”は“冂+玉+一”となり“くにがまえ”としてとりだすのは困難である。そこで“くにがまえ”を“冂”と“一”の２つに分離し、別々のパタンとして用いた。“ぎょうがまえ”については、左側の“ぎょうにんべん”が、それ自身サブ・パタンとして扱えるので“彳”と“亍”は分離した。

つづけ字は、世間一般的、習慣的に行なわれている方法による字を対象とした[9]。このようにして設定したサブ・パタンは全部で約460パタンほどになった。字種を常用漢字にまで拡張した時点でも、この基本サブ・パタンをそのまま用いることができる。また、これによる基本サブ・パタンの追加はほとんど行なわないですむ。

２．４　サブ・パタンの用い方

前節で述べたように設定したサブ・パタンをどのように認識処理に用いるか、について述べる。サブ・パタンは、次のようにして用いる。

(1) 手続きではなく、表現(データ)で持つ。

(2) トップ・ダウンのマッチングを行なう。

上記(1)の、“表現でサブ・パタンを持つ”ことにしたのは、追加、変更の容易さが挙げられる。JOLIS-1 においてもデータの形で持つことにより追加、変更が容易となった。実験システムとして、サブ・パタンの効果、問題点を明らかにし、さらに修正を加えていく、という観点から、辞書を容易に変更できることが必要となる。

上記(2)については、現時点でstochastic syntactic parsing を実現するのは困難である、という理由からである。手書き文字の不安定さから、その構文中にエラーが生じる可能性は充分に考えられる。これに対処するためにはstochastic syntactic parsing が必要となる。しかし、これを現時点で実現するには計算時間がかかりすぎるなど、実用的でない。さらに、認識する文字は範囲が限られており、既知である。そこで、候補文字の文法表現からストローク列に展開して入力ストローク列と相違度マッチングを行なうこととした。これは最も単純なstochastic top-down parsing とも考えられる。いずれにしても、相違度マッチングは実現が容易であり、今後、予期しないつづけ字、ストロークの欠落、分断に対処するためのDP-matching に発展させることも可能である。

上記理由によりトップ・ダウンの相違度マッチングを行なうことにしたが、その方法も、検索時間がかかりすぎるという問題点がある。辞書が大きいと手書きの速度に間に合わないということも十分に考えられる。しかし、それに対しては検索の前に大分類等で候補を絞る、といった方法で対処することが可能であると考える。本研究では、まず実験システムとしてサブ・パタンの有効性及び問題点を見極めること、そして、統計的相違度を用いることによる文字変形への対処の効果を明らかにすることが第一目的である。サブ・パタンの有効性が明らかとなれば、検索速度(手間)については並列サーチなどが考えられる。サブ・パタンの導入による構造解析的手法と、統計的相違度による確率的手法との融合を図ることとその効果を調査することが重要である、と考える。

3．文字辞書

3．1　辞書の構成

辞書はサブ・パタンを基本とし構造的に構成する。各漢字は基本サブ・パタンあるいは中間サブ・パタンを組み合わせて構成される。例を図５に示す。１つのサブ・パタンは２つのサブ・パタンから成るようにし、二分木の構成とする。左右下に延びているポインタはサブ・パタンのサブ・パタンを示す。横に延びているポインタは同レベルのサブ・パタンを示す。また、図５をBNFで記述すると、図６のようになる。図６の“|”は図５における横へのポインタに相当する。

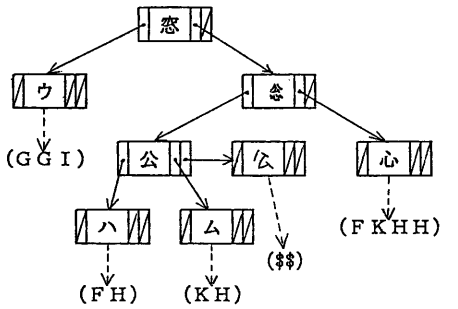


図５　辞書の構造

窓::=ウ 念
念::=公 心
公::=ハ ム | 仏

基本サブ・パタン
ウ::=GGI
ハ::=FH
ム::=KH
心::=FKHH

図６　BNFによるサブ・パタンの記述

3．2　文字表現

認識に用いる情報は、JOLIS-1 と同じストローク列表現(図７参照)と付加条件(図８参照)とする。文字辞書において、各文字コードはその文字を構成するサブ・パタン・コードの情報のみを持つ。サブ・パタンはそのサブ・パタンを構成する下位のサブ・パタン・コード及びストローク列表現を持ち、基本サブ・パタンが直接にストローク列表現を持つことになる。

付加条件は同一ストローク列文字の識別に必要であり、文字全体としての位相関係を調べるので、文字のレベルで持つことにする。

4．認識方法

4．1　統計的相違度

統計的相違度の扱いは、JOLIS-1 と同様に扱う。理論式

$$-\log P(T \mid S)=\sum_{i=1}^{n}-\log P(s_i|t_i)-\log P(T)+\log P(S)$$

に基づき、文字間相違度はストローク間相違度の和として表わす。ストローク間相違度は手書き文字データベースを用いて算出する。

Straight Strokes		Cursive Strokes			
A	→	I	つ つ つ	HD	し し
B	↗	J	へ	HE	ひ
C	↑	K	し く く	HF	ん
D	＼	L	て て	HG	○ ○
E	←	M	ろ ろ	HH	↱
F	↙	O	ら ら ら	HI	↴
G	↓	P	ろ	HJ	み み
H	＼	Q	ら	HK	て
		HA	の の	HL	ら
		HB	み	HM	ゆ
		HC	る る		

図７　JOLIS-1の基本ストローク

付加条件	文字		
	土	土	工
B(1)<yB(2)	Y	Y	N
B(1)<xB(3)	Y	N	N

Y：満足する，N：満足しない
B(n)：第nストロークの始点
<x,y：X(Y)座標の大小判定

図８　付加条件の例

4．2　検索方法

検索方法は、基本的にはJOLIS-1 と同じである。入力ストローク列と辞書内の標準ストローク列とを１対１でマッチングし、文字間相違度を算出する。辞書をシーケンシャルに検索して最も相違度の小さいものを認識結果とする。その際、標準ストローク列を得るためにサブ･パタンをストローク列に展開する必要がある。文字辞書内の、文字を構成する各サブ･パタンの中で最も相違度の小さいストローク列を選びだし、その和をとって文字間相違度とする。サブ･パタンの展開を行なうと、それだけ検索時間が増加し、手書きの速度に間に合わないことが考えられるが、構造を持った文字辞書で、統計的相違度を用いた認識処理を行なうことによる効果を明らかにすることが重要である。

サブ･パタンを直接辞書検索に用いることはしないが、表現としてもつことにより検索段階で次のような効果が期待できる。

よく行なわれるような一定の略字やつづけ字、筆順誤りを持つサブ･パタンにその表現を登録しておくと、そのサブ･パタンを展開したストローク列が、入力ストローク列のその部分と相違度最小になる。したがって、その字体変形を持つ表現の文字間相違度が最小となって認識できることになる(図９参照)。

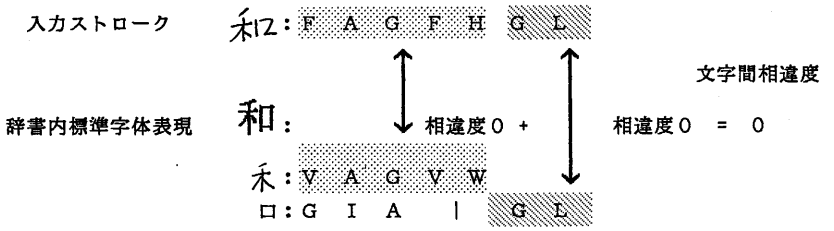


図９　文字変形の吸収

5．実験と解析

5．1　実験の目的

サブ･パタン導入の効果を調べるために、辞書登録と認識の２つの面から調査した。登録の面では、構造化によって登録表現数がどうなったかを調べ、登録手間について考察する。認識の面では、サブ･パタンを用いることによる文字

変形への対処ができているかを調べる。

５．２　文字辞書構造化の効果

文字辞書をサブ・パタンを用いて構造化にすることによって、どのような効果があったかを調査した。構造的にすることによって考えられる利点の一つに、登録の手間の減少ということがあった。このことを調べるために、基本サブ・パタンの使用頻度を調べた。登録文字は、教育漢字１～９画 519文字で、各漢字１表現ずつ登録した。その結果を図１０および表１に示す。平均使用頻度は２である。最も使用頻度の高いパタンは“イ”であり、48回も使用している。次いで使用頻度の高いのは、“口”であり、36回使用している。この“口”のつづけ字に対処するパタンをただ１回登録するだけで36文字分の“口”に対するつづけ字パタンを登録したことになる。１～９画までの教育漢字において、登録に必要なストローク数を算出した。サブ・パタンを用いないでJOLIS-1 のように１表現づつ登録した場合と、サブ・パタンを用いた場合のストローク数を表１に示す。この結果から、登録ストローク数は、サブ・パタンを導入しない場合と比べて、約1/2となっていることがわかる。つまり、登録の手間を登録ストローク数と考えると、 1/2の手間で済むことになる。さらに字種を追加し、登録することで、また、つづけ字等の変形パタンを登録することで、以上のような効果がさらに期待できる。

基本サブ・パタン	漢字例	基本サブ・パタン	漢字例
イ	位、伝	寸	寺、討
口	味、君	ウ	安、空
一	担、合	シ	泳、沿
力	加、努	木	村、困
ム	広、私	月	胃、明
ヒ	比、老	リ	利、則
日	明、暗	扌	持、指
土	社、径	冂	周、国
女	好、委	子	季、存

図１０　使用頻度の高い基本サブ・パタン
（10回以上）

表１　登録ストローク数

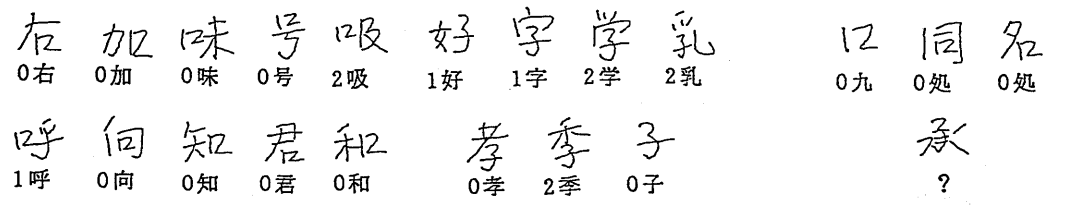
サブ・パタンを用いた場合	1466
サブ・パタンを用いない場合	3433
使用した基本サブ・パタン	319
平均サブ・パタン使用頻度	2
２回以上使用したサブ・パタン	160

５．３　認識におけるサブ・パタンの効果

使用頻度の高い“口”や“子”などに対して、つづけ字のパタンを14種類追加登録し、これらのサブ・パタンを含む、漢字に対して、つづけ字、筆順誤りを含む文字の認識結果を図１１に示す。

“口”のつづけ字“𠂉”を含むパタンや“子”のつづけ字“孑”を含むパタンは、比較的認識している(図１１-(a)参照)。しかし、低画数文字では他の文字のパタンと一致してしまう、という問題点もあげられる(図１１-(b)参照)。これらの文字に対しては、従来JOLIS-1 で行なったような付加条件を強化するという方向で対処できると考える。

次に、認識時間を測定したところ、平均で約0.5sec.、最悪の場合で1sec.程であった。最悪の場合では500以上の文字表現を検索していることになる。これは十分現実的な認識時間である。もちろん、認識対象文字を増やすことおよびマッチングを高度化することで、この時間は大きくなるが、それでも各種の方法によって高速化が可能なオーダーである。



(a) 正認識した文字　　(b) 誤認識、リジェクトした文字

図１１　つづけ字の認識結果

6．今後の課題

6．1　サブ・パタンに関する課題

つづけ字に対処した表現を登録したところ、その表現が他の文字の表現と一致してしまい、認識時に誤認識の原因となってしまう場合がある。特に低画数文字にこの傾向が見られた。これに対処するために、付加条件の設定が必要である。また、同一ストローク列表現となる基本サブ・パタンに対して、再設定の検討が必要であろう。同じストローク列表現となる基本サブ・パタンが複数あることは、パタンの識別に必ずしも有効な単位とは言えないであろう。表現が一致した基本サブ・パタンを１つにまとめるか、もしくは、それらを識別する情報を取り入れる等、何らかの対処を考える必要がある。

6．2　後継システムへの課題

後継システムへの課題として、次のことが挙げられる。

(1) 認識に用いる情報量を豊富にし、低画数文字の認識率を高める。ストロークとストロークの筆の流れを示す“裏のストローク”やその他の位置情報など、認識に必要な情報量を質、量ともに検討する必要がある。

(2) 認識手続きを構造化する。トップ・ダウンの構造解析を行なっているので辞書検索時間を減少させるためには大分類等の候補を絞る処理を行なう必要がある。また、予期しないつづけ字に対処するための認識手続きを検討する必要がある[８]。

(3) 文字辞書の半自動生成を行なうツールを開発する。文字辞書の作成は、単純作業が多いが、それに取られる時間は膨大である。サブ・パタンの導入により、登録の手間を多少減少化することができたが、それをさらに押し進めて、半自動生成を行なえるようにすべきである。

7．おわりに

漢字の構造を認識に反映させる場合、表現(representation)に反映させることがまず重要である、と考え、漢字の“へん”や“つくり”を基本としたサブ・パタンを設定し、文字辞書の表現に用いた。これにより、

(1) 文字辞書の構造化

(2) 一定のつづけ字、筆順誤りへの統一的対処

(3) 文字辞書の表現を約1/2に縮小化

(4) 辞書登録や管理の手間の減少

などの効果を確認した。また、サブ・パタンを反映させた辞書と、統計的相違度を用いた認識で、パタン認識の構造的手法と統計的手法との融合を図ることができた。認識に用いる情報量を豊富にすることでサブ・パタンの効果をさらに高めることができると考える。

謝辞

本研究を行なうにあたり、有益な御討論と御指導を頂いた阿刀田央一助教授、本多庸悟教授に深く感謝する。また、JOLIS-1 の移植に多大な貢献をしてくれた研究生の沢井良一氏、本学大学院生の相澤正氏、本学学部生の志村和英氏、平松徹氏に心から感謝する。さらに、辞書作成に協力してくれた本学大学院生の里山元章氏、牛山一也氏に心から御礼申しあげる。

参考文献

[1] 相澤正：“手書き文字の構造解析的認識における特徴抽出方式の研究”，東京農工大学 修士論文(1986)

[2] 池田裕治，他：“オンライン手書き文字認識システム(JOLIS)の研究開発支援ツール・セット”,情報処理学会日本文入力方式研究会資料16-2(1984)

[3] 菰田千冬，他：“オンライン手書き文字認識システムJOLIS-1 の定量的評価(統計的相違度の設定とその効果)”果)”,情報処理学会第30回全国大会予稿 4N-7(1985)

[4] 菰田千冬，他：“オンライン手書き文字認識システムJOLIS-2 のフィージビリティテスト(サブ・パタンの導入)”,情報処理学会第32回全国大会予稿 4P-10(1986)

[5] 中川正樹，他：“オンライン手書き文字認識システムJOLIS-1 の設計と試作”，情報処理学会日本文入力方式研究会資料3-1(1982)

[6] Nakagawa,M.,et al：“On-Line Recognition of Handwritten Japanese Characters in JOLIS-1”,Proc. of ICPR,Germany(1982)

[7] 中川正樹，他：“オンライン手書き文字認識システムJOLIS-1 の定量的評価”,情報処理学会日本文入力方式研究会資料16-1(1984)

[8] 沢井良一，他：“つづけ字によるストローク結合のオンライン手書き文字認識手法”，情報処理学会日本語文書処理研究会資料(1986)

[9] 藤原宏，他“模範 漢字くずし方字典”，第一法規(1985)

[10] 萬木正義，他：“階層分析法によるオンライン文字認識”，電子通信学会論文誌，Vol.J68-D,No.6,pp.1320-1327(1985)