IV－14

ＯＣＲシステムによる印刷情報の電算化について

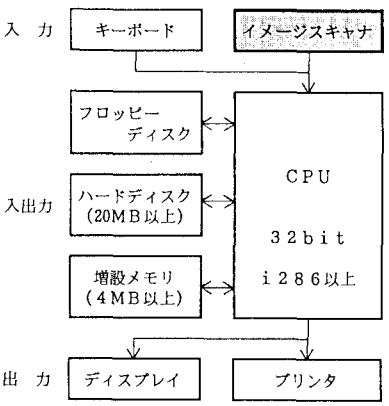
北見工業大学　正員　中岡良司
北見工業大学　正員　森　弘

１．はじめに

コンピュータとりわけパソコンの普及によって、大学研究者の情報処理能力は飛躍的に増大したが、今日においてもデータの初期入力に負担を感じる研究者は数多い。計画系の研究者は実験系の研究者と較べ統計資料や歴史的文献を取り扱う機会が多く、その負担は一層大きなものとなっている。そこで、本研究では、印刷された各種文字情報をＯＣＲシステムで自動入力する方法について、その有効性および可能性を検討した。ＯＣＲとは Optical Character Recognition（光学式文字認識）の略であり、印刷物あるいは手書き原稿から文字を認識してコンピュータで扱える形にしようとするものである。なお、本研究では市販のＯＣＲソフトを使用し、システムとしての総合的実用性を検証している。

２．ＯＣＲシステム

１）ハードウェア　ＯＣＲシステムのハードウェアの構成例を図－１に示す。標準的なパソコンシステムに過ぎないが、イメージスキャナは320DPI（Dot Per Inch）以上の性能が要求される。320DPIでは約13分の１ミリの線を読み取ることができる。本研究では、解像度320DPIの「PC-IN503G」（ＮＥＣ㈱）およびＣＰＵにPC-9801RXのi286を使用している。

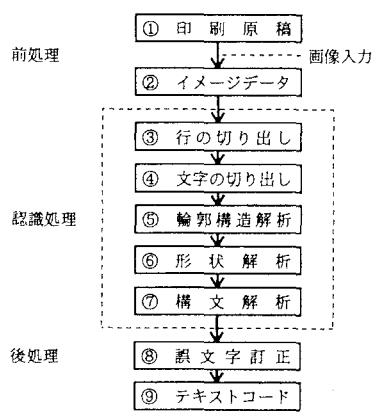


図－１　ハードウェア

２）ソフトウェア　上記ハードウェアをＯＣＲシステムとして利用するにはＯＣＲソフトが必要である。PC-9800シリーズ用のＯＣＲソフトとしては、1988年６月に英数字用ソフト（「PCR-SWAN Ver.1.1」（バーズ情報科学研究所））が、1990年１月に日本語ＯＣＲソフト（「採字帳　日本語版」（㈱テックメイト））が発売になっている。本研究で使用した「ワードアイ」（マイクロニクス㈱）は1991年９月に発売になったもので、文字認識にファジィ理論の応用による線分量子化方式を採用している点に最も特徴がある。最近では、ウインドウズの登場により新たなソフトが次々と開発され認識率も高まっており、今後一層の進歩が期待されている分野である。

３）ＯＣＲ処理のフロー　ＯＣＲソフトの利用を中心に、ＯＣＲ処理全体の流れを図－２に示す。実際の作業工程としてＯＣＲ処理は、前処理（準備作業）、認識処理（ソフト作動時間）、後処理（訂正作業）に分けられる。いずれのＯＣＲソフトも認識処理時間しか提示していないが総所要時間での検証が必要である。

①印刷原稿：多くの原稿は、最適な文字サイズにするためのコピー作業が必要となる。原稿を２分割、３分割以上しなければならない場合も多い。　②イメージデータ：原稿はスキャナに平



図－２　ＯＣＲ処理のフロー

A Study on Encoding of Printed Character to Computer by OCR System
by Ryoji NAKAOKA , Hiroshi MORI

行にセットする必要がある。ある程度の熟練を要する。また、イメージデータの転送にはＡ４版原稿で３分程度は要する。　③**行の切り出し**：最初にソフト側では白地部分と黒地部分から行を識別する。原稿が傾いていると切り出しが不調となり認識精度は著しく低下する。　④**文字の切り出し**：日本語文字は偏（へん）と旁（つくり）に分離するとともに合成して文字領域として切り出す。　⑤**輪郭構造解析**：切り出した文字を画素に細分し文字の輪郭部分のみを抽出し候補文字を探り出す。　⑥**形状解析**：文字画像を細線化し量子化数に置き換え認識辞書と照合する。照合にはファジィのあいまいさを応用する。　⑦**構文解析**：認識した文字を日本語辞書（約５万語）と照合し文法的なチェックを行う。　⑧**誤文字訂正**：変換後の文字列は必ず原稿と照合する必要がある。１字１字チェックしなければならないため意外に時間がかかる。

⑨テキストコード：訂正した文書内容をテキストコードファイルとして保存する。日本語ワープロ等に読み込んで合成する作業が必要である。　なお、上記③～⑦は使用ソフト特有の処理内容であるが、多くのソフトが程度の差は有れ同様の処理を行っていると考えて良い。

３．ＯＣＲシステムの所要時間

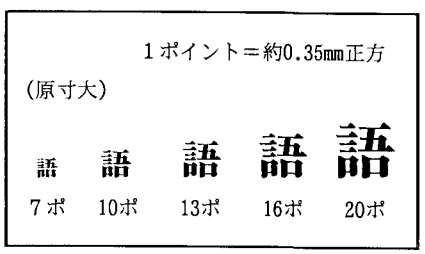
１）入力に関する諸条件

ＯＣＲシステムを利用するための標準的な諸条件を以下に示す。

①　文字情報　字体には、明朝体、ゴシック体、新聞書体、教科書体などがあるが、書体が混在する文書であっても、本ソフトの形状解析の効果により認識率に大きな相違はなかった。文字サイズは、印刷原稿の場合、ポイントという単位で表現する。図－３に例を示す。一定の大きさがなければ認識率が低下するのは当然である。本ソフトの場合、13ポイントが最良の文字サイズであった。印字品質は認識率には大きく影響しないが、新聞等の品質が悪い原稿にはノイズ（ゴミ）を消す作業が必要である。

②　ソフト設定　本ソフトには分析レベルによって簡易モード、通常モード、詳細モードがあるが、認識時間との相関からは通常モードが最適である。学習機能とは、誤認識した文字に対して正しい文字をあらかじめ認識辞書に登録しておくことであり、当然、認識率は上がるが自動処理の原則からはかけ離れていく。どの程度の相違があるかは、後の結果を参照していただきたい。

③　ハード設定　イメージスキャナの読み取り線密度は性能の上限に設定する。濃度調整は原稿に依存するため試行を繰り返すしかない。パソコンとイメージスキャナの接続には、転送速度が早い GP-IBを使用すべきである。また、認識辞書はＲＡＭディスクに置く。



図－３　文字サイズ（ポイント）

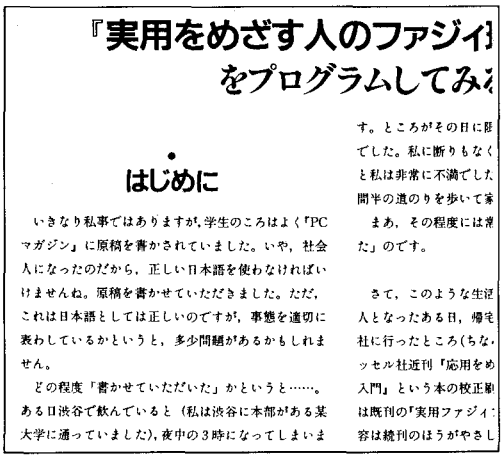
２）ＯＣＲ入力と手入力の比較

表－１は、標準的原稿（Ａ４判、明朝体、13ポイント文字、８００字）の処理時間をＯＣＲ入力と手入力の場合で比較した結果である。ＯＣＲ入力では学習機能のオンオフの２通りで計測した。その結果、①学習機能の有無は総所要時間の観点からは効果が認められない、②合計時間から見るとＯＣＲ入力は手入力の1.5倍から2.3倍の時間を要していることがわかる。ただし、この結果はあくまで本研究で使用したＯＣＲシステムによるものであり、ＣＰＵの性能の向上によりその差は確実に縮まっている（あるいは逆転している）。また、学習機能は大量の文書処理には有効であろう。可能性という観点からは、手入力の速度および作業量には限界があり、ＯＣＲシステムには性能の向上が大いに期待される。

表－１　ＯＣＲ入力と手入力の比較

入力 / 時間	ＯＣＲ入力		手入力
	学習機能無し	学習機能有り	
前処理時間	3～5分	3～5分	1～2分
認識時間	30～40	40～50	16～32
後処理時間	10～15	3～5	3～5
合計時間	43～60分	46～60分	20～39分
手入力比	1.5～2.2	1.5～2.3	1

注）入力原稿：Ａ４判、明朝体、１３ポイント、８００字



『実用をめざす人のファジィ
をプログラムしてみ

•

はじめに

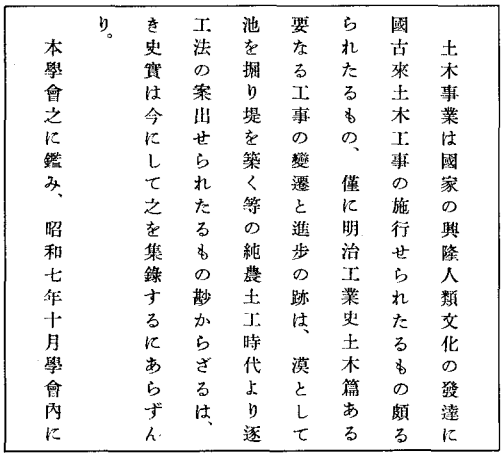
いきなり私事ではありますが，学生のころはよく「PC マガジン」に原稿を書かされていました。いや，社会人になったのだから，正しい日本語を使わなければいけませんね。原稿を書かせていただきました。ただ，これは日本語としては正しいのですが，事態を適切に表わしているかというと，多少問題があるかもしれません。

どの程度「書かせていただいた」かというと……。ある日渋谷で飲んでいると（私は渋谷に本部がある某大学に通っていました），夜中の3時になってしまいま

す。ところがその日に限 でした。私に断りもなく と私は非常に不満でした 間半の道のりを歩いて家 まあ，その程度には常 た」のです。

さて，このような生活 人となったある日，帰宅 社に行ったところ（ちな ッセル社近刊「応用をめ 入門」という本の校正刷 は既刊の「実用ファジィ 容は続刊のほうがやさし

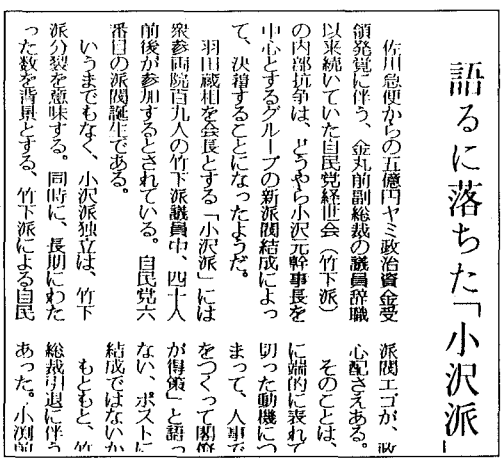
原稿A　（一般書籍）



土木事業は國家の興隆人類文化の發達に
國古來土木工事の施行せられたるもの頗る
られたるもの、僅に明治工業史土木篇ある
要なる工事の變遷と進歩の跡は、漠として
池を掘り堤を築く等の純農土工時代より逐
工法の案出せられたるもの尠からざるは、
き史實は今にして之を集錄するにあらずん
り。

本學會之に鑑み、昭和七年十月學會内に

原稿B　（旧漢字文献）



語るに落ちた「小沢派」

佐川急便からの五億円ヤミ政治資金受
領発覚に伴う、金丸前副総裁の議員辞職
以来続いていた自民党経世会（竹下派）
の内部抗争は、どちらも小沢元幹事長を
中心とするグループの新派閥結成によっ
て、決着することになったようだ。

羽田孜相を会長とする「小沢派」には
衆参両院百九人の竹下派議員中、四十人
前後が参加するとされている。自民党六
番目の派閥誕生である。

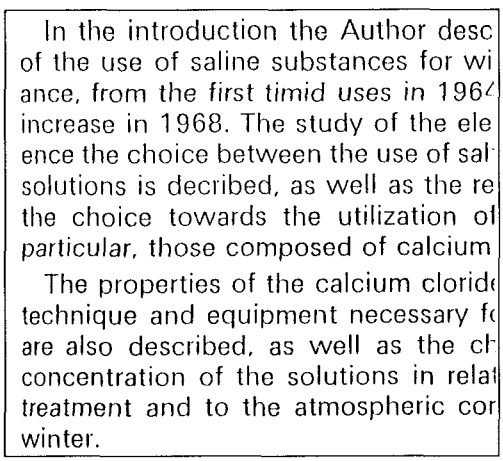
いうまでもなく、小沢派独立は、竹下
派分裂を意味する。同時に、長期にわた
った数を背景とする、竹下派による自民

派閥エゴが、政
心配さえある。

そのことは、
に端的に表れ
切った動機に
まって、人事
をつくって関係
が傍観」と語っ
ない、ポスト
結成ではないか

もともと、竹
総裁引退に伴う
あった。小渕

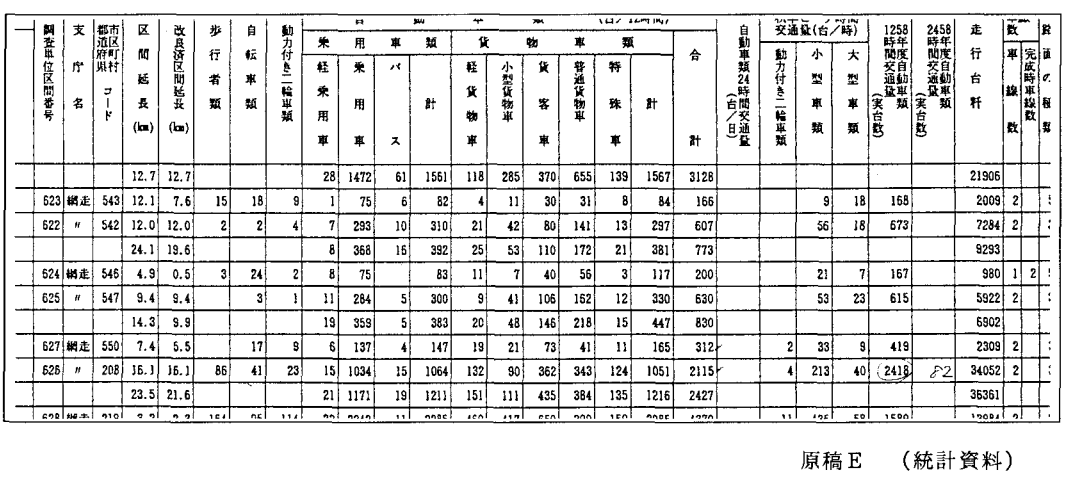
原稿C　（新聞記事）



In the introduction the Author desc
of the use of saline substances for wi
ance, from the first timid uses in 196
increase in 1968. The study of the ele
ence the choice between the use of sal
solutions is decribed, as well as the re
the choice towards the utilization of
particular, those composed of calcium

The properties of the calcium cloride
technique and equipment necessary fo
are also described, as well as the ch
concentration of the solutions in relat
treatment and to the atmospheric cor
winter.

原稿D　（英文）



調査単位区間番号	支庁名	市区町村道府県コード	区間延長(km)	改良済区間延長(km)	歩行者類	自転車類	動力付き二輪車類	自動車類				自動車類24時間交通量（台／日）	交通量(台／時)			1258時間自動車類交通量（実台数）	2458時間自動車類交通量（実台数）	走行台キロ	数		路線の概要
								乗用車類					動力付き二輪車類	小型車類	大型車類				車線数	完成時車線数	
								軽乗用車	乗用車	バス	計										
			12.7	12.7				28	1472	61	1561	118	285	370	655	139	1567	3128						21906			
623	網走	543	12.1	7.6	15	18	9	1	75	6	82	4	11	30	31	8	84	166		9	18	168		2009	2		
622	〃	542	12.0	12.0	2	2	4	7	293	10	310	21	42	80	141	13	297	607		56	18	673		7284	2		
			24.1	19.6				8	368	16	392	25	53	110	172	21	381	773						9293			
624	網走	546	4.9	0.5	3	24	2	8	75		83	11	7	40	56	3	117	200		21	7	167		980	1	2	
625	〃	547	9.4	9.4		3	1	11	284	5	300	9	41	106	152	12	330	630		53	23	615		5922	2		
			14.3	9.9				19	359	5	383	20	48	146	218	15	447	830						6902			
627	網走	550	7.4	5.5		17	9	6	137	4	147	19	21	73	41	11	165	312		2	33	9	419		2309	2		
626	〃	208	16.1	16.1	86	41	23	15	1034	15	1064	132	90	362	343	124	1051	2115		4	213	40	(2418)	62	34052	2		
			23.5	21.6				21	1171	19	1211	151	111	435	384	135	1216	2427						36361			
628	網走	210	2.2	2.2	154	95	114	20	2040	11	2085	459	417	650	900	150	2085	4270		11	425	59	1590		13084	2		

原稿E　（統計資料）

図－4　各種文献原稿例　（いずれも抜粋）

4．各種文献情報のＯＣＲ処理

本節では、様々な文献情報（原稿）を対象に、実際にＯＣＲでどの程度の処理が可能であるかを検証した結果を示す。使用原稿の例を図－４に、結果を表－２に示す。なお、認識率とは総文字数に対する正答文字数の割合であり、学習率とは総文字数に対する学習文字数の割合である。

１）一般書籍（原稿Ａ）

一般書籍においては、字体の混在、字サイズの混在、段組文書の認識を検討した。その結果、字体や字サイズの混在は特に認識率に影響無く、テスト原稿の使用漢字種類の影響が大きかった。また、段組文書は段毎にトリミングすることによって可能であった。学習前の認識率は平均して８割に至らず、約５％の漢字を学習させた場合、98.2％の認識率まで高めることが可能であった。

表－２　各種文献情報のＯＣＲ処理結果

原稿	種類	縦横	認識率		学習率	備考
			学習前	学習後		
原稿Ａ	一般書籍	横	76.3	98.2	5.2	字サイズ混在 ２段組原稿
原稿Ｂ	旧漢字	縦	77.5	82.6	7.5	２７字が旧漢字体 構文解析不能
原稿Ｃ	新聞	縦	90.9	95.3	1.3	１３ポに拡大 コピーノイズ消去
原稿Ｄ	英文	横	97.3	97.4	0.3	英文書体辞書利用
原稿Ｅ	統計資料	横	－	97.7	100.0	０～９の数値のみ 桁位置は不揃い

注）　認識率（％）＝ 正答字数 ／ 総文字数
　　学習率（％）＝ 学習字数 ／ 総文字数

２）旧漢字使用文献（原稿Ｂ）

旧漢字や旧仮名遣いの原稿に対する認識率を調べるため、原稿に「明治以前　日本土木史」（岩波書店）を用いた。全280字中、22種、27字が旧漢字である。これらを学習させたが、認識率は82.6％までしか向上しなかった。使用ソフトの構文解析辞書は現代文を対象としているためと考えられる。

３）新聞記事（原稿Ｃ）

新聞に使用される文字は、①少し平体のかかった特殊な書体、②使用漢字が限定、③印字品質が低いなどの特徴がある。そこで、新聞書体を選択し、文字サイズを13ポイントに拡大コピーするとともにノイズをあらかじめ消去した。その結果、学習前では90.9％の高い認識率を示したとともに、 1.3％の学習率の後では95.3％へと向上している。

４）英文原稿（原稿Ｄ）

英数字は文字数が限定されているため、一般に認識率は高く。認識辞書を英文書体辞書に切り替えて認識させた結果、学習前においても97.3％の高い認識率を示した。今後さらに改良が進めば、英文に関してのＯＣＲシステムは極めて有効と考えられる。

５）統計資料（原稿Ｅ）

統計資料の多くは、①同一形式でデータ量が膨大、②重要部分は数値のみ、③大判書式で文字は小さいなどの特徴がある。検討にあたっては、Ｂ４判の原稿を６分割し約10ポイントに拡大するとともに、０～９の10個の数字をあらかじめ学習させた。その結果、認識率は97.7％と高いものとなった。ただし、①結果は空白で区切られた桁が不揃いの状態となる、②６分割したため最終的に合成する手間がかかるなど問題も多い。データの桁が不揃いになる点に関しては、空白をデータの区切りとして読み込み可能なデータベースソフト（管理工学研究所の「桐」）で数表に復元することが可能であった。

5．おわりに

現在のＯＣＲシステムは実用と非実用の分水嶺にある。パソコンの性能向上と新たな認識アルゴリズの開発によって、実用性のある入力方法となる日も近いと考えられるが、その実用性はあくまですべての作業時間を考慮した総所要時間としてとらえなければ期待を裏切られるであろう。最後に、実際の作業を担当してくれた本学卒業生 久保俊明 君（北海道開発局）の労苦に謝意を表する。