JICST における日本語抄録の電算機処理*

——漢字処理の実施例——

大　場　賢　一**

1. システムのあらまし

1.1 文献速報自動作成システム

JICST は国内・国外で刊行されている科学技術文献を網羅的に収集し，有効と思われる情報を取捨選択して，日本語の抄録を作成し，科学技術の各専門分野別に抄録誌「科学技術文献速報」を発行している．昭和 43 年度の収集一次資料数は

①定期刊行物	6,800 種	国外	4,700 種
		国内	2,100 種
②会議資料	200 件		
③不定期資料	500 種		
④レポート類	10,500 件	PB レポート	3,000 件
		原子力関係	7,500 件
⑤特許	48,700 件		

であるが，定期刊行物は全世界で，約 40,000 種が発行されているといわれる．

収集された一次資料をもとに，内部の科学技術専門職員によって，一定規準のもとに取捨選択されたのち，全国の科学技術専門家に送られ，300 字以内で日本語の抄録にまとめられる．できあがった抄録原稿は再び JICST に集められ，各専門職員によって抄録内容のチェック，用語の統一などを行なったのち，主題内容の分析をして，文献速報用の分類表に従って対応する分類コードを付記すると同時に，キーワードを抽出，付記する．さらに，雑誌番号，巻，号，ページ，発行年，発行国，言語，原資料に掲載されている写真・図・表・参考文献の数などが記入される．

こうして完全になった抄録原稿は，従来は，15×12 cm 大のカード（抄録カード）にタイプされ，校正段階を経たのち，分類コードの指示している関連専門分野への重出処理がなされ，各専門分野別に仕分けしてから，分類別に配列し，抄録全部に一連番号（記事番号）を付番して印刷工程にまわされていたが，電子計算機および特殊漢字ラインプリンターの導入によって，抄録カードの作成から編集を経て，印刷の版下を作成するまでの過程を自動化したのが，文献速報自動作成システムである．

文献速報自動作成システムでは，抄録原稿を漢字テレタイプ（漢字けん盤さん孔機：新興製作所）とデータライタ(富士通)によって紙テープにパンチされる．年間抄録件数は約 40 万件で，総パンチ字数は約 2 億字となり，漢字テレタイプ 33 台，データライタ 12 台（英文 8 台，露文 4 台）を要している．

パンチされた紙テープは，電子計算機に読み込んで磁気テープファイルを作成し，2 段階の校正のため，漢字プリンタ (PT 700) にて校正用ゲラをアウトプットし，ゲラ校正を行なって，訂正を要するものがあれば，訂正データを紙テープでインプットし，磁気テープファイルの内容を正確にしていく．校正が終わって，内容が正確になった磁気テープの抄録は，版下作成のための編集処理にはいる．編集処理工程は，従来抄録カードで人間が行なっていた重出処理，記事番号付番，専門分野別仕分け，分類項目順の配列および印刷上のレイアウトのために文頭・文末の Justification と禁則処理，欧文単語の切れ目の Hyphenation そう入など，そして最後に文献速報の様式に合わせて，ページの割付け（枠組）まですべて自動的に行なわれ印刷版下作成にはいる．版下は漢字プリンタ (PT-200 フィルムレコーダ) にて，フィルム露光または印画紙印字され，これがそのまま印刷原版となるのである．さらに記事番号付番処理後の磁気テープは，年間を通して累積保存され，年間索引（日本語標題索引，著者名索引，雑誌名，レポート索引）が作られる（**第 1 図**参照）．

1.2 日本語文献情報の機械検索

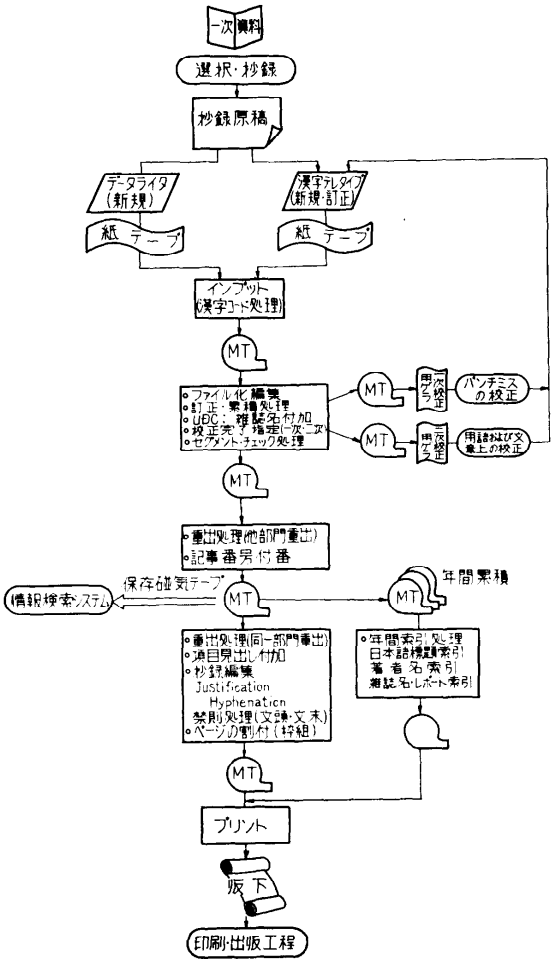
文献速報自動作成システムの効果は，さまざまの観点から考えることができるが，年間 40 万件と膨大な蓄積情報の中から，利用者が要求する特定の情報を迅速かつ正確に提供できる情報検索 (IR; Information Storage and Retrieval) システムの完成がまたれた．

JICST では富士通と協力して総合的な情報検索シス

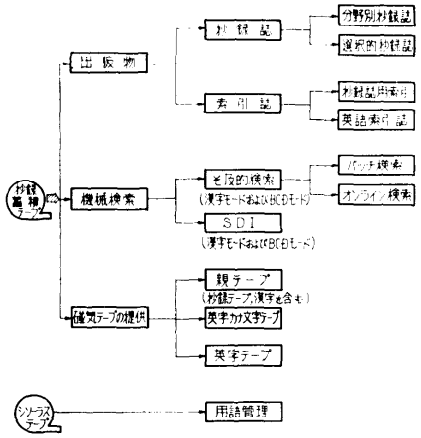
---

* Computerized compilation of abstructs journal at JICST, by Kenich Ohba (Computer Center, JICST)

** JICST 計算機室



第1図　文献速報自動作成システムフロー



第2図　総合情報検索システム図

テムを開発しており，その一部である漢字日本語文献の機械検索実験にはいった．漢字モードによる機械検索システムには，そ及的検索(バッチ検索)と情報選択提供 (SDI; Selective Dissemination of Information) があり，蓄積情報は文献速報自動化開始後の電気部門抄録を対象に行なっている．質問項目としては，キーワード，UDC, 著者名，言語，発行年などがあり，これらの論理関係の指定に従って探索する．探索結果の回答は，書誌事項，抄録，キーワードの任意の組合せで漢字プリンタ (PT 700) にてプリントアウトする (**第2図**参照)．検索精度を上げるためは，抄録の質的向上とシソーラスの完備が必要であるが，前者については，人間が読むための抄録から，その内容を適確に表現するキーワードの抽出が当面の課題となり，後者は抽出されたキーワードを大量に蓄積し，電子計算機で自動的に編成することも可能となったが，いずれも人間の介在が不可決であり難解である．この面でも用語管理の自動化が必要であり，総合システムの一環として企画中である．

一方，国外との情報交換および漢字プリンタをもたないユーザへの磁気テープの提供など，さらに，電子計算機と会話をしながら，ユーザ自身直接検索できるオンライン検索システムの開発を行なっている．

2. 機器構成

2.1 FACOM 230-50 機器構成図 (第3図参照)

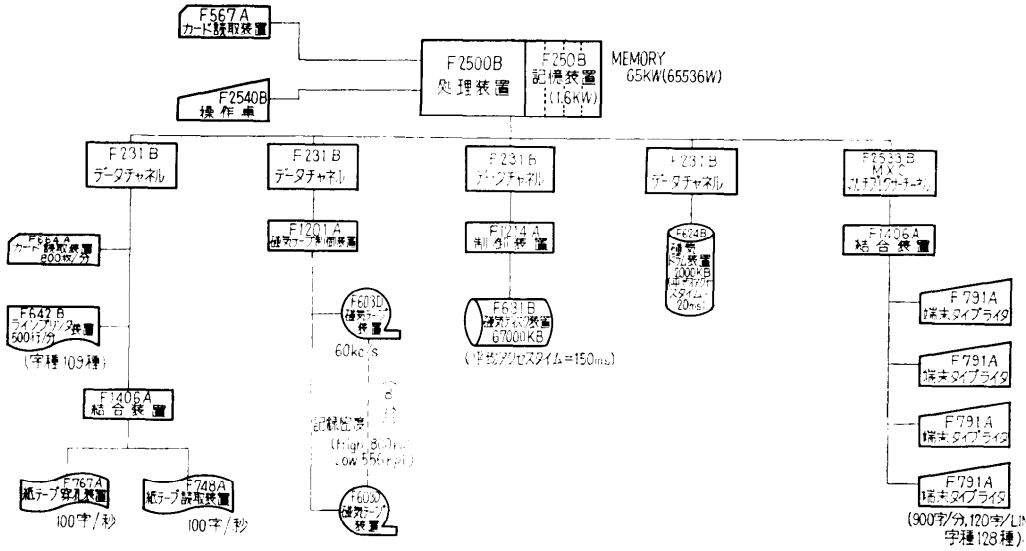
2.2 FACOM 270-20 機器構成図 (第4図参照)

3. 入力関係

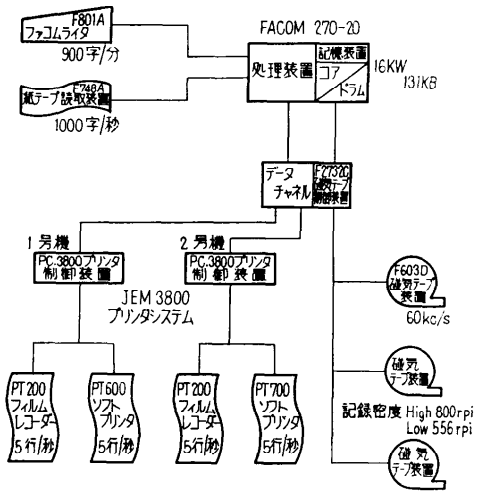
3.1 漢字テレタイプ (漢字鍵盤さん孔機)

文字けん盤 (**第5図**) は科学技術情報に出てくる文字頻度調査により選択されその配置も使用頻度を考慮して決定された．使用文字種は**第1表**のとおりである．

この他，科学技術情報の表現に多いイタリックやゴシック体文字，H_2O, 10^{-2} などのサフイックス文字，3などの合成文字がインプットできるようファンクションキーを用いて効率を高めた．漢テレの文字けん盤は4ブロックに分れ，1ブロックは 48 個のキーからなり，1個のキーには 13 文字収容されている (第5図)．漢字1字は $6u+6u$ の 12 bit で表わされるが，さん孔紙テープは8単位用を用いて，漢テレハードミスによるビット落ちを発見するため X, Y の判定ビッ



第3図 FACOM 230-50 機器構成図



第4図 FACOM 270-20 機器構成図

トをつけた.

シフトキーを押すと S_1 のコードが対応し，該当マスターキーを押すと，S_2 および MA のコードが対応して紙テープ上に 1〜6 bits と X，7〜12 bits と Y の順でさん孔される．第5図はシフトキー①とマスターキー④を押した場合の例である．

ファンクションキーとして，(合成)，(上つき，下つき，正常)，(イタリック，ゴシック，標準)，(字削)，(行削)，(オールマーク) などがある．たとえば

$$K_m=\int_{\lambda_0}^{\lambda_1} V_\lambda$$

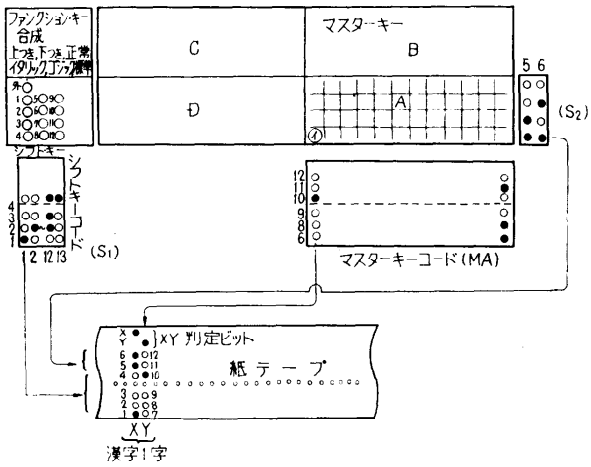
第1表

字種	インプット	アウトプット
漢字	1,861	1,861 (明朝)
カタカナ	81	162 (明朝，ゴシック)
ひらがな	77	154 (〃)
英欧文字	65	195 (ローマン，イタリック，ボールド)
ロシア文字	66	198 (〃)
ギリシャ文字	33	99 (〃)
アラビア文字	10	30 (〃)
ローマ数字	20	40 (ローマン，イタリック)
記号	199	248 (ローマン，ボールド，一部イタリック)
スペース	6	6
予備(空白)	78	78
合計	2,496	3,071

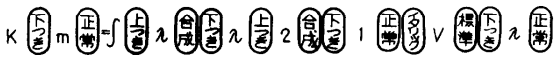
をパンチする場合**第6図**となる．

紙テープ上の漢字コードは，コンピュータ内部では，1ワードのビット構成を考慮して，紙テープコードをそのままキャラクターコードとして 12 bits で表わすが，上つき，下つき，合成，ピッチなどを表現する 4 bits を加え，18 bits (3 byte) で1字を構成するとともに，ソートの関係で漢テレコードの X と Y を入れ換えて YX としている．また，ファンクションコードは，紙テープ読み込みプログラムで指定のコードに変換するとともに，文字所有のピッチを付加している．コンピュータ内部でのコード表現を**第7図**に示した．

年間約 40 万件の抄録を消化するのに，漢テレパンチ字数は 12,000 万字と推測され，40 台が必要である．パンチ能率は 50 字/分であるといわれているが，



第5図



第6図

抄録内容の難易さにより，また，英文，カタカナタイプライタのように，目を手もとにおかずにタイプできるなどは不可能なので，仕事初めと終わりとでは，相当の差が生じるのではないかと思われる．

漢テレパンチを開始した当初，漢テレ不慣れによる操作ミスのため，また，漢テレ自身のハードミスによるビット落ちなどが発生し相当に悩まされた．操作ミスによる紙テープエラーはほとんどなくなったが，ハードミスエラーが時折発生するのは防ぐのがむずかしい．

3.2 データライタ

抄録原稿の中で漢字のない部分は，さん孔処理能率を高める意味で，データライタでパンチする．

データライタコードは，1字8単位であり，漢字コードと異なるため，紙テープインプットプログラムにて漢字コードに変換する．データライター分の年間パンチ数は約8,000万と推測され，英文用8台，露文用4台を要している．

4. 校正方法および訂正方法

4.1 校 正

漢字テレタイプは文字のキーをタッチすると，紙テープ上に漢字コードがさん孔されるのみで，目的の文字が正確であるかどうか，視覚文字で確認することができない．したがって，一度紙テープにパンチさん孔されたものが正確であるかどうかは，紙テープ上の文字コードを読まなければならないが，事実上不可能である．パンチャーは結局，感に頼って間違えたと思ったときは，一字分バックしてその文字を削除して，もう一度正確文字をタッチするようである．

文献速報自動作成システムにおける校正方法は，紙テープを読み込み磁気テープファイルにした後，校正作業（人間作業）および訂正データを作成しやすいように，レイアウトを整えて漢字プリンタ(PT 700)でモニタリスト（校正ゲラ）をアウトプットする．校正後のモニタリストは，そのまま訂正データ作成時のインプット原票となる．

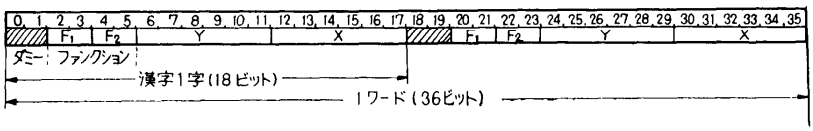
一次校正はパンチミスを発見するために行なうもので，校正作業はパンチャー側で行なう．二次校正は，各専門分野職員が文章上の修正などの必要性があるために行なわれるもので，校正作業は一次校正モニタリストと全く同じレイアウトである．一次校正モニタリストの例を**第8図**に示し，校正サイクルフローチャートを**第9図**に示す．

情報検索システムの漢字モードバッチ検索システムにおいても，利用者プロファイル，質問データを紙テープパンチしてインプットし，磁気テープファイルを作成するのであるが，やはり，文献速報システムと同様の方法をとっている．

4.2 訂正方法

校正の結果，訂正の要がおこったら，磁気テープファイルを修正していく．訂正データは校正ゲラを原票にして漢テレパンチを行ない，紙テープからインプットする．

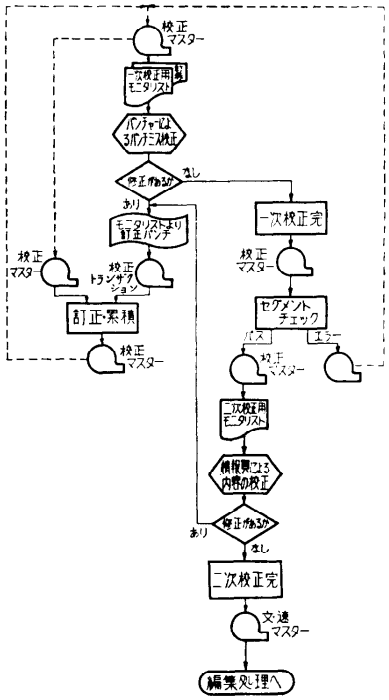
訂正機能としては



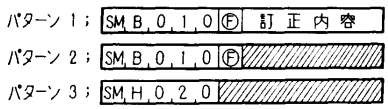
第7図

```
SMA Ⓐ03378006Ⓑ014726Ⓒ5  06Ⓒ150
SMB010  Ⓟ690318Ⓒ  Ⓚa1Ⓚ  ⓊE324  ⓋE
        Ⓝ  Ⓟ  Ⓚ68
SMB011  B  ▲Talanta㊢
SMB020  Ⓟ0000Ⓒ003Ⓚ002Ⓚ008Ⓤ15.12㊢
SMC010 Ⓟ  ⓒCA0404032▲化▲543.4/5 others ㊢
SMD010  Ⓟ1371-1376㊢
SME010  Ⓟ  ⓒMEALOR D㊢
SME020  Ⓟ  ⓒTOWNSHEND A㊢
SMF010  ⓅApplicatioⒸns of enzyⓂme-catalysⓀed reactio▲
SMF020  Ⓟns in tracⓒe analysisⓀ - III. DeⓀterminatio▲
SMF030  Ⓟn of silveⓒr and thioⓊurea by thⓀeir combin▲
SMF040  Ⓟed inhibitⒸion of invⓀertase. ㊢
SMG010  Ⓟ酵素触媒反応の微量分ⒸⅢ析への応用 Ⅲ  イⓀンベルターゼの阻害にⓀよる銀とチオ尿素の定▲
SMC020  Ⓟ益㊢
SMH010 Ⓟ  ⓒインベルターゼ(接触分析)㊢
SMH011  Ⓟインベルターゼ㊢
SMH020 Ⓟ  ⓒ銀㊢
SMH021  Ⓟギン㊢
SMH030 Ⓟ  ⓒチオ尿素(接触分析/インベルターゼ)㊢
SMH031  Ⓟチオニョウソ㊢
SMI010  ⓅインベルターゼはしょⒸ糖のぶどう糖および果Ⓚ糖への加水分解を促進Ⓚする。銀イオンはイン▲
SMI020  Ⓟベルターゼの触媒作用Ⓒを阻害するので、そのⓀ阻害の程度から逆に銀Ⓚイオンを定量できる。▲
SMI030  Ⓟ銀イオンと強く結合するⓒ陰イオンは、銀イオⓀンの阻害作用を減少さⓀせる傾向があるが、チ▲
SMI040  Ⓟオ尿素は逆に銀イオンⒸの阻害作用を高める。Ⓚこれらの効果を利用しⓀ、$1\sim5\times10^{-7}M$▲
SMI050  Ⓟ銀と、$10^{-7}\sim10Ⓒ^{-6}M$チオ尿素を定量Ⓚした。なお、チオ尿素Ⓚがどうして阻害作用を▲
SMI060  Ⓟ高めるのかの機構につⓒき考察した。分析は種Ⓚ々の場合の相対活性をⓀ調べて行なった㊢
```

第8図 校正用モニタリスト例



第9図 文速献報自動作成システム
校正サイクルフロー



第10図

パターン 1: アイテム単位のそう入，入換えおよびレコード単位のそう入．

パターン 2: アイテム単位の削除．

パターン 3: レコード単位の削除．

がある．訂正データの紙テープパンチ方法は**第10図**のようであるが，実際は SM A Ⓐ原稿 No Ⓑ補助 No ©校正コードをパンチしてから，訂正データをパンチする．©の校正コードは，校正フローの各段階で異動するデータを，一つのマスターファイルに集約して効率を高める意味での運用コードであり，**第2表**にそのコード変化と意味を示す．

第2表

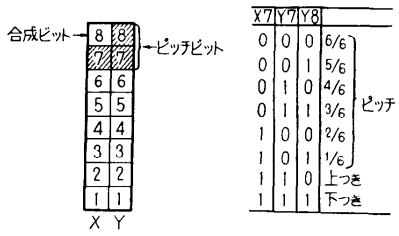
校正コード	機能	備考
0	新規データ	
1	一次校正用	0, 2 プリント →1
2	一次訂正データ	1 に対して訂正 →2
3	一次校正完了	一次校正完の指定 →3
4	セグメント・チェックパス	セグメント・チェックパス →4
5	二次校正用	4, 6 プリント →5
6	二次訂正データ	5 に対して訂正 →6
7	二次校正完了	二次校正完の指定 →7
8	テーブル関係	
9	訂正データ	

5. 出力関係

漢字まじり日本語の情報処理は，新聞社・通信社などで漢字テレタイプを用い，紙テープにてインプットし，モノタイプや写植機などの植字機にアウトプットしていたが，あまりにも植字速度が遅すぎた．したがって，電子計算機による日本語処理システムが要求され，その不可欠の要件として，高速漢字ラインプリンタの開発が待たれたのである．日本電子産業株式会社は，CRT 表示プリント方式を用いた高速全電子式写植機（JEM-3800）を開発し，ここに日本語の情報処理が，いよいよ計算機で扱えるようになった．

JEM-3800 で扱える文字種は，3,071 字種もあり，200～600 字/秒の速度で，また，6～10.5 ポイントの指定された大きさで，自動的に連続プリントができる．

このプリンタの中心部である文字発生部は，いわゆる，フライングスポット CRT 方式を用いているので，多種文字を発生するものとしては，他の方式に比べ安価であり，しかも，文字種の増加に対して，プリント速度は変わらないという特徴がある．さらに，文字発生部の文字マトリックス板（80 mm×80 mm：1,000 字）は容易に交換できるので，必要な文字書体のマトリックス板を用意しておけば，広範囲なプリントが可能である．このプリンタへの入力は計算機システムとのオンライン，あるいはオフラインの磁気テープ，紙テープリーダである．文字コードは2バイトで1字を表わし，また，英字，数字，露字，カタカナ，ひらがな，記号などのソートがそのままできる JJ-67 コードシステムを採用している．1字のビット構成は，12 ビットで文字種，3 ビットで文字幅（ピッチ）およびサフィックス位置指定，1 ビットで合成プリントを表わし，文字の大きさ（ポイント）はシフトファンクションコードを用いる（**第 11 図**）．



第 11 図

プリントの1行の長さは最大 30 *II*（パイカ）；126.51 mm が標準で，1行ごとにラインプリントされ，その行間隔は1ポイントピッチでコントロールされる．プリント文字種は 6，7.5，10.5 ポイントとも 3,071 字が可能であり，プリント速度は5～10行/秒である．また，日本字の文字幅は普通全角(6/6 ピッチ)であるが，英数字，記号などは 1/6 ピッチ単位で可変プリントされるし，化学式・数式などのサフィックス，文字と文字の合成などインプットデータに指定しておけば，自動的にプリントされる．プリント文字の解像力は，20 本/mm である．プリンタ・ユニットとして用途により，PT-200 フィルムレコーダと PT-700 ソフトプリンタの2種類がある．

PT-200 フィルムレコーダは，記録媒体がフィルムまたは印画紙で，CRT 上に表示された1行分の情報を，レンズ系を通して露光プリントするもので，解像力が高く，そのまま印刷版下に使用できる．

第 3 表

項　目	PT-200	PT-700
記録媒体	フィルムまたは印画紙	クイックコピー紙
媒体幅 (mm)	92	160
媒体長さ (m)	100	100
プリント文字の大きさ（ポイント）	6，7.5，10.5	10.5 のみ
一行の長さ (mm)	max 30 パイカ (126)	max 30 パイカ (126)
行間隔	1ポイント単位で可変	
印字速度 (行/秒)	5～10	5～10
現像	露光のみ	湿式内蔵
解像力 (本/mm)	20 以上	

PT-700 ソフトプリンタは，記録媒体がクイックコピー紙であり，CRT に高解像度ファイドオプティクス CRT を用いて，これに表示される1行分の情報を直接媒体に露光し，現像・定着工程（湿式）を経て高速にプリントするものである．

上記のプリンタの性能を**第 3 表**に示す．

文献速報自動作成システムおよび漢字モード情報検索システムにおいて，漢字プリンターを使用するデータは，すべてプリントイメージに編集されて，MTCP (FACOM 230-50 のビット構成を F 270-20 のビット構成に変換するプログラム名）により，モード変換してからプリントされる．

文献速報自動作成システムにおいて，daily に約 1,000 件の抄録データがインプットされるので，一次校正用モニタリストがプリントされるし，また，二次校正モニタリストもほぼ1,000件プリントされるので，PT-700 を2台必要とし，一方，版下は8シリーズが2回/月，1シリーズが3回/月で作成するので，1回/2日の割となり，1シリーズの抄録件数が 1,500～4,000 件であるので，PT-200 も2台を要している．PT-200，PT-700 とも FACOM 270-20 とオンライン接続して，2台とも同時にプリント操作できるように配慮している．

6. 消耗品関係

文献速報自動作成システムで消費される消耗品関係についてしるすと

（1） 紙テープ

daily に種々のデータが異動するので，データ別に色分けして，データの混同を防いでいる．紙テープは8単位用，300 m 巻きを使用しており，漢字テレタイプは1巻に 50,000 字分，データライタは1字が6ビット構成なので，100,000 字収容されることになる．しかし，紙テープの前後および各抄録別の区切りと，

セグメントの区切りにフィード部を置くようにしているので，使用効率は80%くらいである．年間400,000件の抄録情報に対し，漢字テレタイプ分12,000万字として2,400本，データライタ分8,000万字として600本，合計3,000本の紙テープが費やされる．

（2） クイックコピー紙

160 mm 幅，100 m 巻きの富士クイックコピー紙は，抄録1件分のプリントに平均25 cm を要するとし，1巻に400件分がプリントできるが，前後の送り部空白など考慮し，一次校正モニタリスト，二次校正モニタリスト，一次・二次訂正分リストなど，年間400,000件の処理で約3,000本を要する．

（3） 版下用ロールフィルム

92 mm 幅，100 m 巻きのロールフィルムは，文献速報誌の版下原版に使用されるもので，年間400本を要する．

7. 漢字処理の問題

文献速報自動作成システムが完成し，実ランにはいるまで，各段階でのソフトウェアは幾多の困難につきあたり，漢字の処理のむずかしさを痛感した．

まず，第1の問題点は，紙テープデータの処理である．漢字テレタイプは，初め文字コードに XY 判定ビットがなかったために，ハードミスによるビット落ちのため，桁ずれコードが発生する原因となった．XY 判定ビットの設定により，桁ずれの発生を最小にできたことは，大きな利得であった．紙テープそれ自体，カードに比べ非常に扱いづらいことはたしかであり，初めのうち，よく途中でちぎれたりして苦労した．第2はソフトウェア上の困難さである．自動作成システムの場合，紙テープ上の漢字コード（12ビット）が計算機処理では18ビットを1字とし，1ワードに2字つめた．実際のコードは紙テープ上のコードと同じであるが，その頭に4ビットのファンクションビットを設けたため，ソートなどに多少工夫をしなければならなかったし，COBOL が主体であったため，小まわりのきくテクニックが使えず，結果的にステップ数がふえて時間がかかってしまった．このことはアッセンブラ言語によるプログラムの変更で，チェッキングが詳細にでき，しかも，処理時間がアップできた．

現在，最大の欠陥となっていることは，校正作業および訂正データの作成に膨大な時間を費し，そのために情報の速報性が遅れてきつつあることである．この原因をさかのぼっていくと，結局パンチミスにたどりつく．1個のパンチミスを修正するための訂正パンチにまたミスをおこせば，それが二重のエラーを発生させることになり，かえって，エラーの数をふやしていくということになる．この欠陥は，磁気テープにインプットする前に，できる限りデータを完全にし，インプット後のデータ修正を極力なくす方法をとらなければ，除去できないことになる．したがって，新規分のパンチ段階で，紙テープ上のパンチミスを発見・修正してから，インプットする方法を講じる必要がある．そうすることにより，校正用モニタリストの一次分は省略でき，経費の節約になるとともに，時間の浪費もなくなるであろう．

8. 将来の見通し

文献速報自動作成システムを中心に，日本語の計算機処理について述べ，その問題点をあらってみた．漢字コードの計算機内部での処理については問題はないが，インプット側およびアウトプット側に大きな問題を含んでいる．

インプット側，すなわち，漢字テレタイプによる紙テープデータの作成が，一般のタイプライタのように打鍵印字がなされないので，複雑多種の鍵盤をもつ漢テレパンチの際のミスパンチ発生の原因になっている．そのミスを計算機インプット前に発見し，修正する方法として，漢字ディスプレイモニタの活用が考えられる．このディスプレイモニタシステム機器は，漢テレに紙テープリーダ，文字発生装置，CRT ブラウン管文字表示装置，紙テープパンチャおよび磁気テープ装置が接続されている．JICST では，このディスプレイ装置を導入し，校正フローの簡略をはかるべく検討中である．新規分の漢テレパンチをこのディスプレイ装置で行なえば，打鍵印字されたと同様なので，ミスタッチを視覚的にとらえることができ，それをすみやかに修正できる．ただ，33台の漢テレに，すべてこの装置をつけるとなると膨大な出費となるので，新規パンチ紙テープの修正に使用するとすれば，10台くらいあればよいことになる．修正方法としては，新規紙テープを読んで，CRT 上に文字表示し，原稿と対照する．ミスが発見されれば，その部分を漢テレで打ちなおし，CRT 上で校正を行なう．校正が終わったら，CRT 上の文字を新たな紙テープにパンチアウトし，再び新規テープを読み込んで校正をくり返す．そして修正ずみの紙テープを計算機にインプットする．

（昭和44年4月30日受付）