新聞漢字調査の機械処理システム

著者	野村 雅昭
雑誌名	電子計算機による国語研究
巻	3
ページ	146-164
発行年	1971-03
シリーズ	国立国語研究所報告 ; 39
URL	http://doi.org/10.15084/00001009

新聞漢字調査の機械処理システム

野村雅昭

1. はじめに

この調査は，昭和42年度より，第三資料研究室が行なっている，「新聞語彙調査に伴う漢字および表記の研究」の一部をなすものである。漢字に関する調査・研究は，42年度からはじめられ，44年度までに，ほぼ3分の1の量のデータについて，電子計算機による処理および漢字テレタイプ（略称　漢テレ）による印字を終えた。ついで，同年から，人手による作業をすすめて，本45年度には，中間報告を発表すべく，現在，まとめの作業が進行中である。集計の結果については，すでに，その一部を発表したもの(1)もあり，近く発表予定の報告（国研資料集8『現代新聞の漢字調査（中間報告）』）にすべてが収められているので，ここでは，主として，電子計算機による処理方法を中心にして述べることにする。中間集計までの処理システムは，43年1月に起案されたものをもとにして，その後，多少の修正を加えながら，現在にいたっている。その一部については，すでに述べたこともある(2)が，ここで，あらためて，全体の処理手順および個々の処理方法の概容について，解説をくわえることにする。

2. 全体の流れ

この調査の最初の入力データとなるのは，語彙調査の処理過程で作成される，磁気テープ・ファイルである。このファイルは，長単位(3)とよばれる語の

(1) 「新聞使用漢字の試行的分析」（『電子計算機による国語研究』国研報告34）
「新聞の漢字と雑誌の漢字」（『国研ＬＤＰ　6』）

(2) 「漢字調査の機械処理について」（『国研ＬＤＰ　1』）

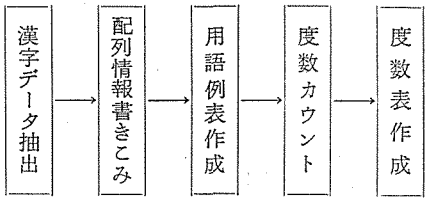
すべてを含み，それぞれに，出典と層別の情報を持っている。さらに，それぞれの長単位語の第1字目が漢字の場合は，それに読みがなとしての代表音がつけられており，変則的ではあるが，全レコードが五十音順に配列されている。（ただし，今回の中間集計では，別途の事情から，この配列情報を利用することはなかった。）そして，このファイルは，長単位語の度数カウントのひとつ手前のファイルであり，汎用性の強いものである。

長単位五十音順ファイル　フォーマット

代表音 (20)	出典情報 (12)	層別情報 (8)	長単位見出し語 (40)	E/I

この長単位ファイルから，漢字調査に必要なレコードを抜き出し，漢字配列に必要な情報を書きこみ，度数カウントをするとともに，見出し漢字，用語例のアウト・プットをするというのが，本システムの概略である。このシステムで作成したプログラムを機能によって分類すると，つぎの4種類になる。

図1　作業の流れ（略図）



1．漢字データ抽出プログラム

2．配列情報書きこみプログラム（部首コード・代表音コード）

3．度数カウントプログラム

4．印字処理プログラム（漢テレ・ラインプリンタ使用）

以上のほか，ソート・プログラムは，サービス・ルーチンを利用し，それにともなう，マージ・プログラムは，すべて，当方で作成した。このシステム全体の流れは，図2に示したとおりである。このほかに，人手による処理作業もあるわけだが，ここでは，省略する。

(3) 文節から付属語を除いたものと，ほぼ同じ。詳しくは，『電子計算機による新聞の語彙調査』（国研報告　37）を参照。

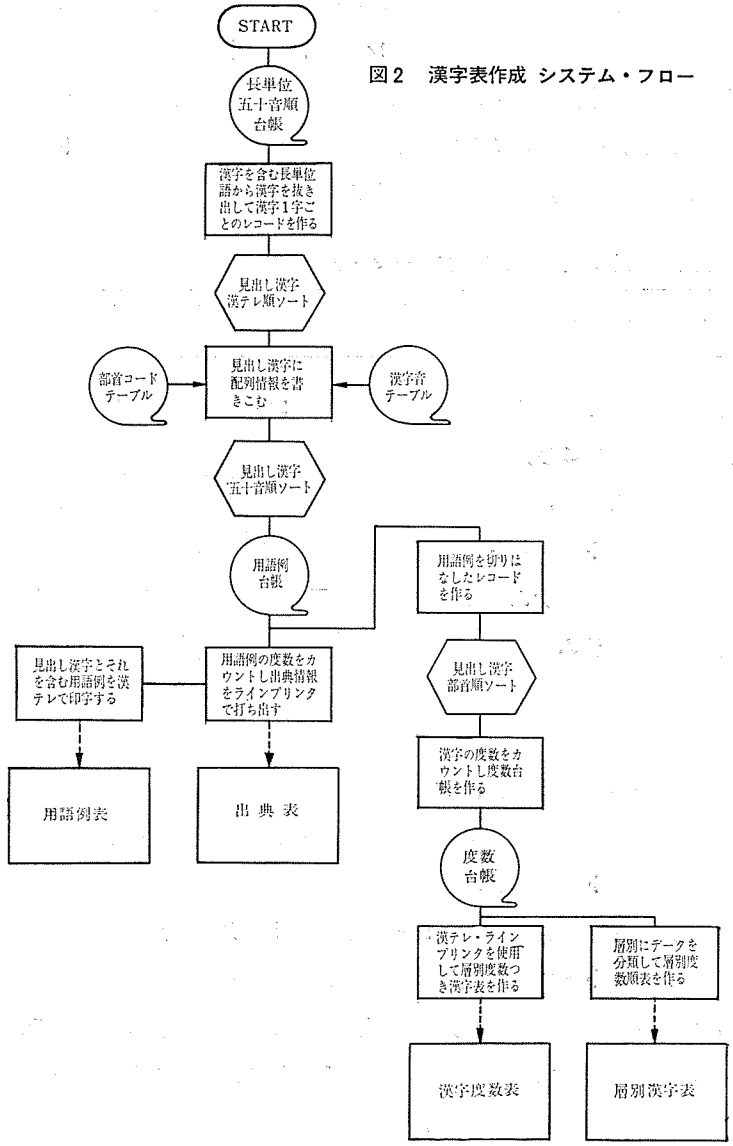


図2　漢字表作成 システム・フロー

3. 見出し漢字の抽出

このプログラムは，上述の長単位五十音順ファイルを，1レコードずつ読みこんで，漢字を含む語については，漢字1字ずつについて，それを見出し漢字として，1レコードを作成するものである。長単位語を1字ずつ読んで，それが漢字であるかいなかを判別するのは，漢テレコードでは漢字が1箇所に集まっていないため，多少の手つづきを要するが，漢字以外のすべての字種を判別するのにくらべれば，そうめんどうではない。ただ，なにをもって漢字と定義するかによって，多少，選択の手つづきは変わってくる。この調査では，漢字としてあつかうのは，便宜上，『大漢和辞典』（諸橋轍次著　約50,000字所収）に，字母としてのせられているものということにしてある。実際には，同辞典にない漢字もかなり出現しているが，おおむね字体の異同などによるもので，漢字かいなかの判断に迷うようなケースは，ほとんどなかった。

実際の新聞紙面で，漢字がどのように用いられているかという観点から，このプログラムで漢字として処理するものを分類すると，つぎの3類になる。

1．普通に使用される漢字

2．㊀・㊁など，記号的に使用される漢字

3．〇と々

1は，普通の記事や見出しなどに使われるものが大部分をしめる。各種の表や広告などの中には，記号的な使い方がされているものもあるが，それらもここに含めて，特に区別はしない。2は，テレビ・ラジオの番組欄などに現われるもので，記号としての機能が強いものであるが，1のグループと同様にあつかう。ただし，度数カウントの際には，記号的に使用されたものが，全体のうちどれだけあるかは，区別できるようにする。3は，普通には，漢字と考えられないし，前述の規則からもはずれるものであるが，漢字の用法をみる場合に参考となる可能性もあるので，この段階では，一応，漢字として処理して，用語例はアウトプットする。ただし，度数カウントの際には，対象から除く。なお，〇は一・二・三…などの漢数字とともに用いられた場合のみをとり，その

ほかの符号的な用法の場合は，とらない。

ところで，本研究所の漢テレは，盤面2,400字（600トップキー），そのうち漢字は，2,108字である。したがって，それ以外の漢字（盤外字）が出現した場合には，なんらかの処理法を考える必要がある。（それについては，あとでも多少ふれるが，詳しくは，下記の論文(4)を参照されたい。）そこで，漢テレによる入力という観点からいえば，上記の漢字として処理するものは，4つのタイプに分けて考えられる。

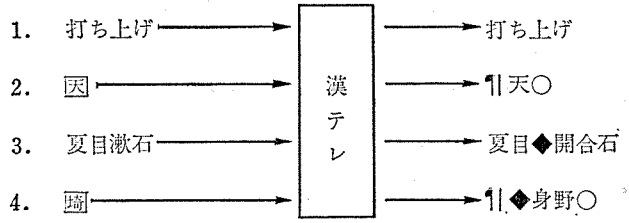
1．普通の用法の盤内字（〇・々を含む）

2．記号的用法の　〃

3．普通の用法の盤外字

4．記号的用法の　〃

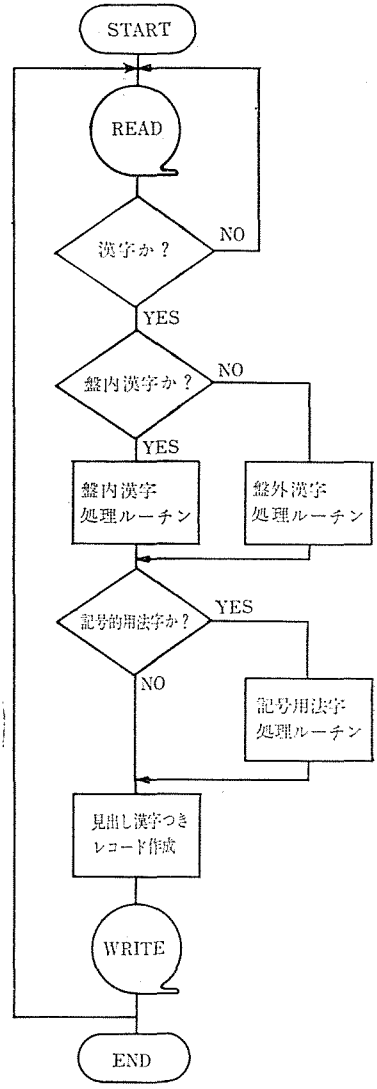
たとえば，実際の紙面に現われたものを，漢テレでインプットする場合にどうなるかを，上のタイプに従って示せば，つぎのようになる。



因は天気予報，埼は埼玉県のことで，番組欄で記号的に使われたものである。このように，盤内にない記号が現われた場合は，¶マークで示し，漢字の場合には，¶と〇の間に，その漢字をはさんで，3漢テレ字母で示す。したがって，¶マークを読みこんだ場合は，つぎの字が漢字であるかいなかを判定し，漢字の場合は，その漢字と〇の2字分を見出し漢字エリアにうつす。また，「漱」と「埼」は盤外字で，盤外字は，◆マークを打ったあと，盤内の2字の漢字の組み合わせで，これを表わすことになっている。（漱＝◆開合，埼＝◆身野）したがって，◆マークを判定した場合は，それに続く2字（◆を含めて

(4) 松本昭「国研用漢字テレタイプと同機利用の言語情報処理」（『電子計算機による国語研究』国研報告　31）

図3　漢字データ抽出 ブロック・チャート



3字）を見出し漢字エリアに移動することになる。

このようにして，長単位語を1字ずつ読んでは，漢字1字ごとに，長単位語を用語例としたレコードを作成する。つまり，「打ち上げ」からは，2レコード，「夏目◆開合石」からは，4レコードといった具合になる。この処理が終わったあとの，各レコードのフォーマットは，つぎのようになっている。

見出し漢字抽出データフォーマット

	見出し漢字 (8)	出典情報 (12)	層別情報 (8)	長単位用語例 (40)	E/I
1.	打			打ち上げ	E/I
2.	天○			¶天○	E/I
3.	◆開合			夏目◆開合石	E/I
4.	◆身野○			¶身野○	E/I

4. 配列情報の書きこみ

以上のような方法によって作成した，見出し漢字抽出ずみデータを，同じ見出し漢字ごとに集めるわけであるが，このまま計算機に配列をさせると，それぞれ一箇所には集まるが，漢テレコードによって配列されるため，全体としては，人間にとって意味のない並び方になってしまう。そこで，まず，それぞれの漢字に，部首理論コードをつけることにする。部首理論コードとは，盤外字の処理のところで少しふれたが，「大漢和辞典」の通し番号にしたがい，ノンシフトの盤内漢字2字の組み合わせによって，ほぼ部首順の配列になるように工夫されたものである(5)。盤外漢字の場合は，◆マークのあとに，その2字を打つわけであるが，盤内漢字の場合にも，この理論コードを書きこんでやればよい。そのためには，まず，全部のレコードを漢テレコード順にソートしておき，盤内字については，同じく漢テレコード順に配列してある，部首テーブル

(5) 前注（4）論文を参照。

を参照しながら，理論コードを書きこんでいく。このコードによって，ソートをすれば，ほぼ康熙字典順の配列とひとしくなるわけである。

部首テーブル　フォーマット

漢字 (2)	部首コード (4)	漢字 (2)	部首コード (4)	…	漢字 (2)	部首コード (4)	E/I

通し番号	漢	テレコード	部首コード
1	一	ＩＬ	計計　0101
2	丁	ＱＬ	〃形　0103
3	丐	—	〃型　0105
⋮	⋮	⋮	⋮　⋮
10	万	ＥＪ	計建　010G
⋮	⋮	⋮	⋮　⋮
73	中	8／	形奥　0341
⋮	⋮	⋮	⋮　⋮
80	串	—	形加　034C
⋮	⋮	⋮	⋮　⋮

ここで，ひとつめんどうな問題がある。それは，「大漢和辞典」には，正規の通し番号のほかに，′（ダッシュ）のついた番号の字があることと，字体の相違などのため，同辞典には収録していない字が少なからず出現することである。たとえば，「土」の部でいえば，「圧（4879′）」・「塁（5316′」・「増（5448′）」には，′がついており，「墨」・「堕」は，この字体では，同辞典に載っていない。このような字は，同じ部首の最後の字の通し番号のつぎの番号を順に当てることによって処理しているが，このため，普通の漢和辞典とことなる配列が生じる。たとえば，普通ならば，

土→圧→塁→堕→塩→増→墨

のように配列されるところが，このコードでは，

土→塩→圧→塁→増→墨→堕

のようになってしまう。実際には，同じ部首に属する字が，そう多く出現するものばかりではないから，それほどの不便は生じていないが，将来，なんらか

の解決をはかる必要がある。

上のような欠陥を補う意味と，部首順の配列がアルバイターなどの作業者には引きにくいことをも考慮して，さらに，五十音順の配列が可能なように，代表者コードを書きこむことにした。この代表音コードは，語彙調査で，長単位語をほぼ五十音順に配列するために，長単位語の第１字目が漢字である場合，その漢字の代表的な音を，実際の読み方とは関係なく，一義的に書きこむためのものである(6)。この中には，いわゆる音（訓に対する）だけでなく，音が一般的に用いられないものは，訓を代表音としているものもある。（たとえば，稲→いね，鹿→しか）そして，盤内漢字すべてに代表音テーブルが作成されている。

漢字音テーブル　フォーマット

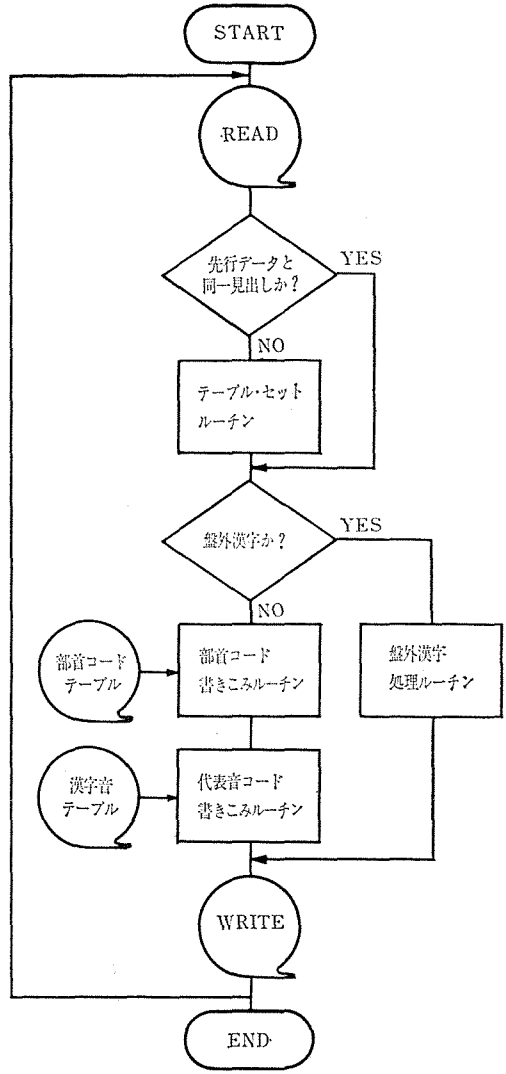
漢　字（2）	代表音（8）	E／I	漢　字（2）	代表音（8）	E／I	……

	漢テレコード	代表音コード
曲	00	きょくア
計	01	けいアア
巾	02	きんアア
⋮	⋮	⋮

この代表音を書きこむことによって，盤内漢字は，すべて，五十音順に配列される。ソートのときに，第１キーを代表音コードに，第二キーを部首コードにすれば，盤内漢字が五十音順に配列され（同音の場合は，部首順），そのあとに盤外漢字が部首順に並ぶということになっている。この盤外漢字が最後にくることと，語の配列のために作られたので，濁音が清音になっていることが，この処理の問題点であるが，それについては，最後の章で，ふたたびふれる。

(6) くわしくは，つぎの論文を参照。田中章夫「電子計算機によるワードリスト作成上の一問題」（『電子計算機による国語研究』国研報告　31）

図4　配列情報書きこみ ブロック・チャート



5. 用語例表の作成

以上のような配列情報を書きこんだファイルを，つぎのような優先順位によって，ソートする。

1．代表音コード

2．部首コード

3．用語例（漢テレコード）

4．出典情報

5．層別情報

これによって，同一漢字を用語例中に持つレコードが，一箇所に集まり，漢字および用語例の度数カウントが可能になる。ただし，漢字の度数カウントは，全体のもののみを行ない，層別の度数カウントは，後述の別途の処理による。このソートを終えたファイルを用語例台帳とよぶ。そのフォーマットは，つぎのとおりである。

用語例台帳　フォーマット

見出し漢字 (8)	代表音 (8)	出典情報 (12)	層別情報 (8)	長単位用語例 (40)	E/I
愛奥裁(SP)	あいアア	出典情報	層別情報	愛情物語(SP)…(SP)	E/I

この台帳をもとにして，各漢字と用語例の度数および出典・層別の情報がわかるような表を，ラインプリンタで打ち出す。これによって，各漢字の用法や疑問点を調べる必要がある場合には，原データにもどることが可能になる。この表を出典表とよぶ。また，見出し漢字と用語例は，紙テープにアウトプットして，漢テレで印字する。これを用語例表とよぶ。出典表と用語例表は，ページで対応するようになっているので，用語例表から出典表へさかのぼって，各種の情報を参照しうる。出典表および用語例表のフォーマットは，図6のとおりである。

図5　出典表作成 ブロック・チャート

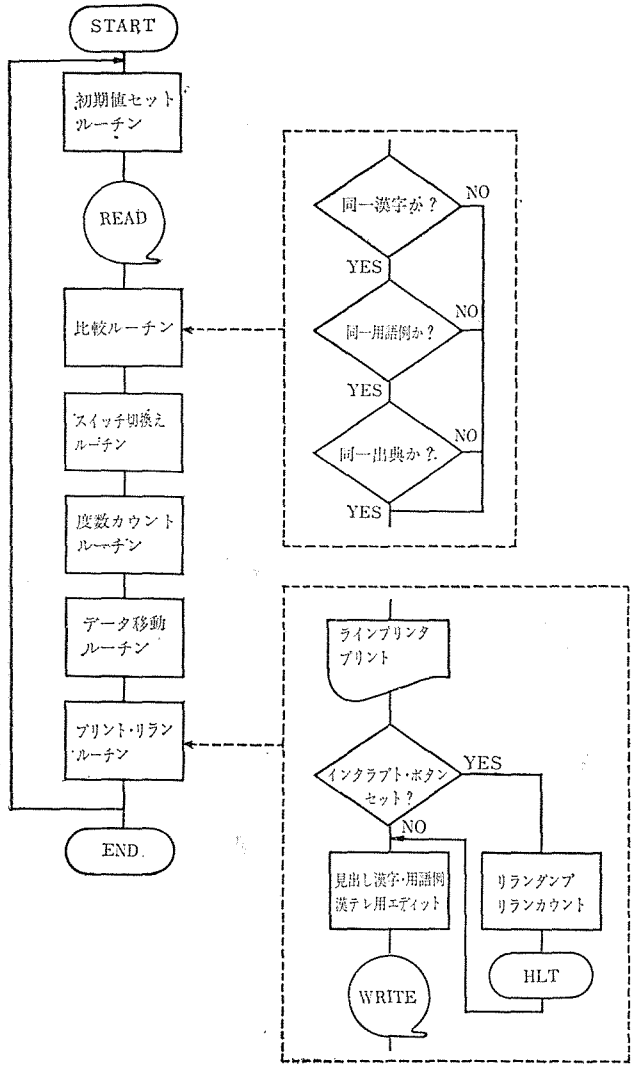


図6　用語例表・出典表フォーマット

8ページ

〔愛〕あい
愛
愛党精神
愛人

愛京子

10ページ

愛知
愛知県
愛知県体育館
愛知県知事選
愛知県内
愛知県本部
愛知化
愛知機

11ページ

愛知長官
愛知工場

PAGE 10

218L
C1 04518=0706 26327=0705 26327=0705 59787=1216 65124=0901
K0 40298=0310
YOREI—DOSU 6
218L0オ
A0 56625=0801
C1 06990=0910 18503=1216 48956=1216 83546=0506
J1 00665=0910 54559=0501 80677=01012
K0 04987=0501 46293=0505
YOREL—DOSU 11
218L0オ6, 2GO7
C1 01721=0910
YOREI—DOSU 1
218L0オ8LE54)
J1 80677=0101
YOREI—DOSU 1
218L0オ#V
J1 43488=0501
YOREI—DOSU 1
218L0オDZBZ
J1 80677=0101
YOREI—DOSU 1
218L49
A0 53449=0310
K0 06380=0310 18350=0310 37521=0310
YOREI—DOSU 4
218L7E
C1 02028=0310 62987=0310 64907=0310 77867=0310 82696=0310
J1 00135=0310 46215=0310 71175=0310
K0 33705=0301 57226=0310
YOREI—DOSU 10

6. 度数カウント

度数カウントのために，用語例台帳の各レコードから，用語例を切りおとしたものを取り出す。そして，見出し漢字によって，部首順にソートする。部首順にソートするのは，盤内字をあやまって盤外字として打鍵したものや異形同字などを発見しやすいからである。異形同字とは，「鷗—鴎，鑛—鉱，挿—插」などの類をいう。これらは，漢テレの盤内になくても，別字として打鍵することが可能であるが，今回の調査では，原データ→清書→パンチという作業の流れになっているため，正確を期しがたいので，集計の際には，合わせることにした。（用語例表では，別字として処理してある。）

度数カウントは，つぎのように行なう。1レコード読むごとに，全体の度数を1ずつ加える。それと同時に，層別の情報によって，計算機内のカウント・エリアの，該当するところにも1ずつ加える。（原則として，1レコードは，出現頻度1回に相当するが，同一文中に，同じ語が2回以上出現する場合も，1レコードになっている。その場合は，文内度数の情報によって加算する。）ここでいう層別とは，語彙調査のG種（文種別）層別とT種（話題別）層別の二つを組み合わせたもの(7)で，加算のときに，その指示を与えてやる。その結果，全体と12の層別の度数がカウントされるわけである。さらに，記号的用法のものは，全体の度数にも加えながら，別途に集計をする。こうして，出現した漢字の異なり数だけのレコードができあがることになる。これを度数台帳とよぶ。フォーマットは，つぎのようになっている。

度数台帳　フォーマット

見出し漢字（6）	代表音（8）	層別度数 (84)＝(7×12)	*	全体度数（7）	記号的用法度数（7）	E／I

今回の漢字調査では，用語例表の点検によって発見されたエラー（打鍵ミス・清書ミス・層別の誤判定など）による度数の異動は，すべて，この台帳で

(7) 前注（1）「新聞使用漢字の試行的分析」を参照。

修正することになっている。この台帳は，ノンバッチになっており，サービス・ルーチンなどによる修正もしやすくしてある。この台帳を，代表音コード，あるいは全体度数によってソートすれば，五十音順，出現度数順の度数台帳ができるわけである。

この出現度数順の度数台帳によって，層別累積度数表が作成される。（部首順あるいは五十音順台帳でもよいが，度数順のほうが，ソートの時間が短くてすむ。）まず，台帳の1字分のレコードを，各層ごとに分割して，層を識別するコードをつける。ある層の出現度数が0の場合は，レコードを作らない。全体度数についても，1レコードを作成するとすれば，1レコードから，（12−n）＋1のレコードができるわけで，総計は，各層ごとの異なり字数の総和に全体の異なり字数を加えたものと等しくなる。（nは，出現度数0の層の数。$0 \leqq n < 12$）これを度数順にソートしたのが層別度数台帳である。これを入力データとして，必要とする層の情報をパラメータで指示し，度数順に加算しながら，ラインプリンタでアウトプットすれば，層別累積度数表ができるわけである。また，見出し漢字を漢テレで印字することも可能である。

層別度数台帳　フォーマット

層別記号（2）	見出し漢字（8）	代表音（8）	層別度数（7）	全体度数（7）	E／I

7. 度数表の作成

各種順に配列された台帳を1レコードずつ読みこんで，ラインプリンタで打ち出す。フォーマットは，1レコード1行ですむ。見出し漢字の部分は，紙テープにアウトプットして，漢テレで印字する。プログラムの指示によって，ラインプリンタのフォーマットに合わせて，印字紙の紙幅・行数とも一致させられるので，あとで，人手で書きこむなり貼りつけるなりすることが可能である。ただし，盤外漢字は，コード・ブックによって，人間が翻訳しなければならない。出現異なり字数の約3分の1が盤外字なので，この手間は，かなりたいへんである。

図7　度数表作成 システム・フロー

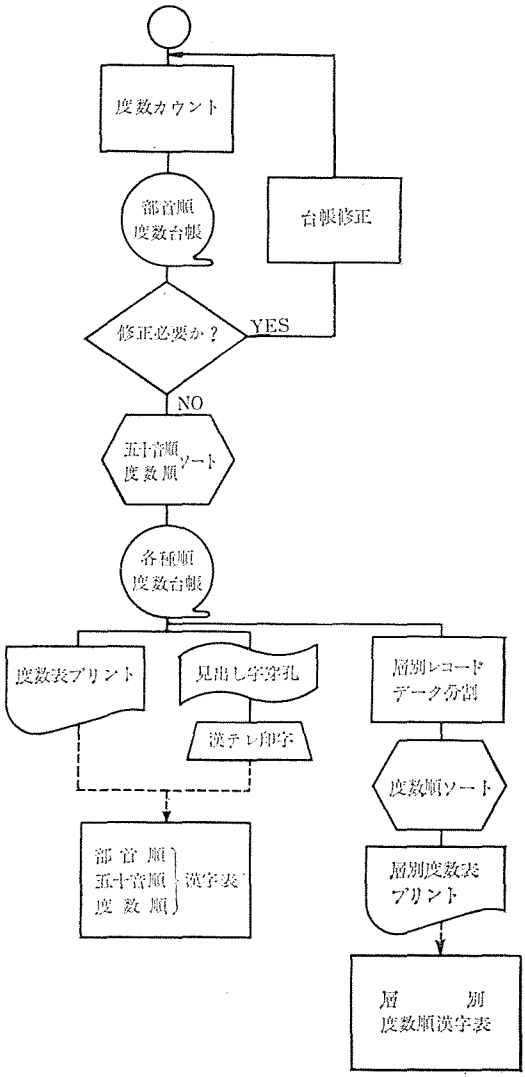
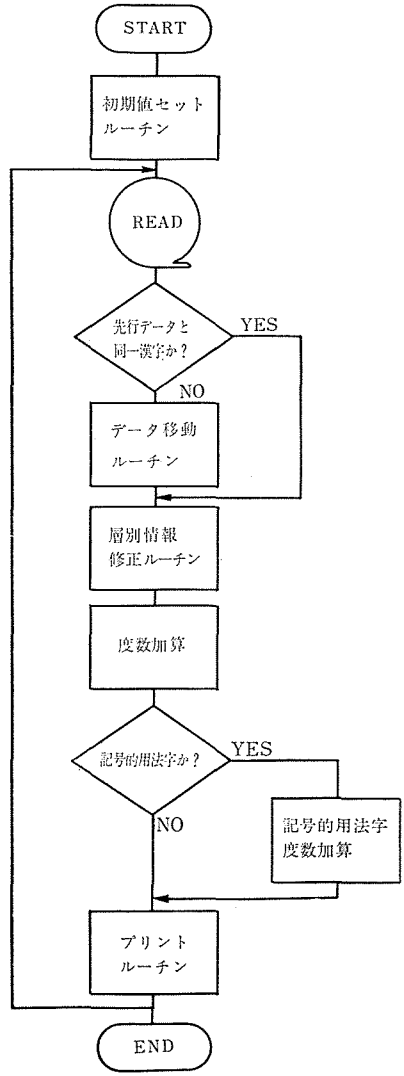


図8　度数カウント ブロック・チャート



8. おわりに

以上に述べた処理システムは，今回の中間集計のために作成したもので，数多くの問題点をかかえたまま，作業を強行したうらみがある。最終集計のための処理では，そうした不備をいくらかでも少なくすることが必要である。そこで，いくつかの問題点を指摘するとともに，その対策を考えてみることにする。

さきにも多少ふれたが，まず，見出し漢字の配列の問題がある。部首順配列における筆画数順の問題は，テーブルに情報を入れておけば，解決する。また，五十音順配列で，盤外字が最後に集まる問題は，全体の処理を行なう前に，盤外字だけをアウトプットして，テーブルに追加することができれば，解決する。清濁の問題は，読みがなをさしかえてやればよい。（ついでに一言するならば，代表音は，いわゆる音と訓の2種類あったほうがよい。）さらに，当用漢字・教育漢字などの情報も入れておけば，計算機内での集計が簡単にできるようになる。そうした要求を満たすためには，結局，漢字調査専用の総合漢字テーブルとでもいうべきものが必要となる。現在，その作成のための準備をすすめているが，詳細については，別の機会に述べることにする。

つぎは，データの修正の問題である。今回は，度数に関する修正は，すべて計算機内で処理できたが，用語例の修正は，すべて人手で行なわなければならなかった。これは，語彙調査のシステムと漢字調査のそれとがうまくかみ合っていないことにも原因がある。それを解決するためには，まず，漢字を用いて表記された長単位語をすべてアウトプットしたものを点検し，エラーを修正して再入力するしか方法がない。さもなければ，漢テレによる再出力を断念して，すべて人手で行なわなければならない。それには，所要時間，人手によって起こるミスなどの得失を十分に検討しなければならず，一概に論じるわけにはいかない。

また，今の問題とも関連するが，長単位語を含んだファイルのソートが，漢テレコード順と五十音順の2回あって，それに処理の大部分の時間を費さなけ

ればならないことがあげられる。それをいくらかでも短縮するため，1レコードの長さをできるだけ切りつめたが，たいした効果はあがらなかった。（長単位語の配列情報を落としたのは，そのためである。）これは，用語例表を漢テレでアウトプットしようとするかぎり避けられない問題である。上に述べたように，はじめに漢字表記語をアウトプットして，出典・層別の情報を修正したあと，用語例を切り離して，度数カウントの処理のみを行なうことにすれば，所要時間は，かなり短縮されることになる。

そのほか，細かい問題はかなりあるが，大部分は，実際の処理で改善することができる。残りのデータの処理を，いかに効率よく精度を高めることができるかは，今後の表記調査にも深いかかわりを持っている。現在の問題点の多くは，語彙調査のシステムがほぼ完成した段階から，漢字調査のシステム設計がはじめられたことに起因するものである。根本的には，語彙調査のシステムの中に，表記調査，漢字調査をどう位置づけるかということの検討が最後に残された問題であるといえよう。

（45. 10. 31）