A system of the word count program

電子計算機による語彙調査

主として長単位処理について

斎藤秀紀

○まえがき

国立国語研究所では，基礎語彙の選定や国語国字問題を解決する参考資料を得る目的を持って，数度の語彙調査を実施してきた。しかし流動的に変化する語彙の特徴を，短期間に大量に調査することは，従来の人手を用いた方法では不可能となり，機械による処理を考えねばならなくなった。そこで，昭和41年度より行なわれる，新聞の語彙調査に対し積極的な電子計算機の利用を計算しＨＩＴＡＣ－3010形電子計算機を導入した。

これによって，従来手作業に頼っていた，大部分の処理が機械によって置換えられ，語彙調査の速報性が生かされると共に言語情報処理に関する基礎資料を一般に提供できるものとなる。

本論文では，新聞の調査における長単位関係プログラムシステム及び短単位処理の進行状態を報告する。なお，現在までに作成されている長単位関係語彙表は次の三種であるが，この三種の語彙表によって基本的な数値は大体もうらできるものと考える。

出典語彙表
層別語彙表（度数順，配列順）
比　率　表（度数順）

1　システム構成

全体のフローチャートを図３に示す。ラン１で簡略50音順[1]（以下配列順と言う）情報の付加処理を行った後配列順にソートを行い，この調査全体のマス

ターファイルを作成する。配列順にソートされたデータは，必要に応じラン２で更新を行なうが，この調査のように，大量のデータを扱う作業では，一度に規定の順序に分類することは不可能であり，分割して処理を行い，分類後に再びこのデータの統合を行なわなければならない。また，このソートのためのキーは，漢字テレタイプ（以下漢テレと言う）でパンチされた見出し語の配列順位は，そのままソートしたのでは50音順にならず，付加情報を付け優先順位を変更しなければならない。そのため，見出し語の第一字目の漢字の代表音から理論コードを決め，簡略であるが50音順に近い形で並ぶよう考慮してある。

図１は配列情報付きマスターファイルの磁気テープフォーマットである

図１　配列順マスターファイル

	配列情報 (20)	出典情報 (12)	層別情報 (８)	見出し語 (40)	終記号	

ラン３では，新聞紙面の記事ごとに付けられた層列[(2)]の処理を行なう。

見出し語として立てられた語の各々は，新聞紙面の位置や，記事によって使われ方に相違のあることがある。そこでこの調査では，分析のさい各層間の比較から語の特徴と共通点を明らかにできるよう紙面を４層47項目に分類してある。図２は，このランで作成される磁気テープのファイル構成であるが，層内度数は左からG，P，S，Tの順である。

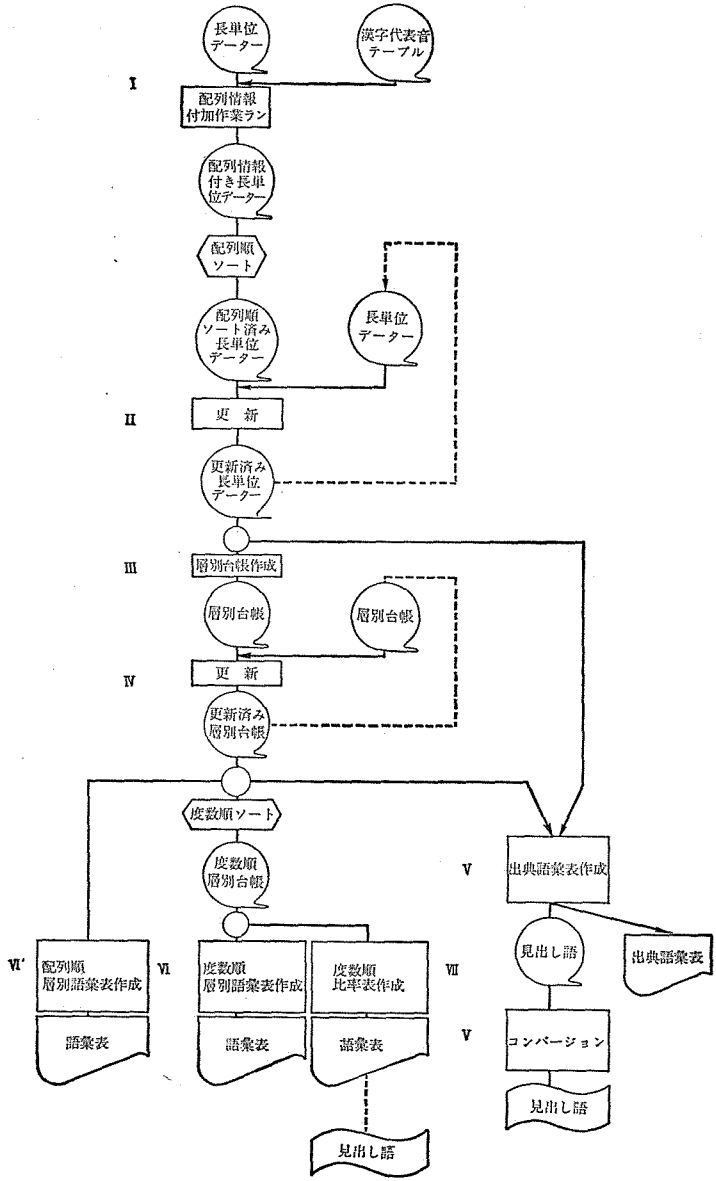
層別ファイルの作成は，配列順出典ファイルの見出し語順序と層別情報は，ほぼランダムに近いため，記憶装置をランダム装置と同様に扱い集計能率を上げるようはかってある。方法は，層内度数の集計用番地の最初の番地に層別の二桁の数を加算し，一致する記憶番地の内容をこの層の度数集計用作業番地として集計を行う。この方法は，集計すべき層の位置を個々に探索する必要はなく，任意に目的の位置に接近できる利点がある。

図２　層別マスターファイル

	層別内度数 (333)	総度数 (７)	配列情報 (20)	見出し語 (40)	終記号	

ラン４では，ラン３で作成された層別度数順及び配列順ファイルの更新を行なう。このファイルの更新は，配列の場合度数の更新を行っても配列順位は乱

図 3　用語調査用システムフローチャート



されないが，度数順にソートされた後では，第一優先順位である度数の変更はできない。この処理を行うさいは，データを異なり語に集約させた後にソートを行い更新のさい，度数情報に変更を加えない状態でなければならない。そのため，一度度数順に配列されたファイルは，順序を乱さずに任意の更新は不能であり，この部分のみ語彙表の必要に応じ，ラン３から重複して処理しなをさなければならない。

以上ラン４までは，データ更新に関するプログラムの説明であるが，各プログラムの接続にはデータの移動状態をチェックし，受けわたしによるデータの脱落を未然に防ぐようにしてある。また見出し語の長さは20字までを処理の対象とし，これを越えるものは最大40字までをレコードとして許した。

出典語彙表（図４）

この語彙表は，見出し語の出典を示すもので新聞名，ブロック番号，センテンス番号及び見出し語の出典度数からなり総索引の性格を持つものである。

語彙表の見出し語は，主に漢テレ印字したものを使用するが，処理の確認のためラインプリンタ（ＬＰ）用紙にも見出し語の機械コードを印字してある。このコードは，ハンドブックによって，漢テレ文字に各々対応させることもできるが，漢テレでデータを作成するさい，誤動作による脱さん孔，ビットの変化等のエラーデータの修正に有効な働きをする。

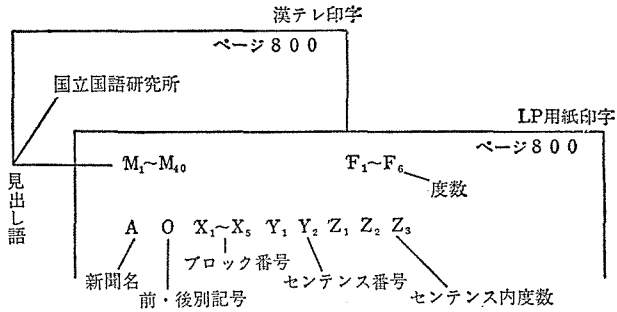
新聞名は，調査対象の三紙に対し表１に示す記号を与え，さらにサンプリングのとき分けられた各紙の一年分の前期（１月～６月）後期（７月～12月）の判定に数字の０と１で表わし，名称の簡略をはかってある。これはブロック番号と共に見出し語の在存する紙面の月，日，頁等を表わし，続く二桁のセンテンス番号と対になって，見出し語の正確な位置を示すのに必要な情報である。

表１

	朝　刊	夕　刊
朝　日	A	J
毎　日	B	K
続　売	C	L

これらは，単語の持つ意味が一義的に決定できない場合があるため，その語の持つ環境を無視できず，単語相互間の関係から語の用法を調べ，これを文脈から求めることが多いためである。また出典を示す五桁の数字は，簡

図 4　出典語彙表印字形式



単な計算によって，新聞の発刊された月，日，頁等に直接変換できる。

例　　A 0 3 2 3 2 5

アサヒ　前期　3月15日 8頁21ブロック

ブロック番号は，サンプリングのとき新聞紙面の面積によって分けられた区格番号で，一頁を30区格に分けてある。なお半年分，全ブロック数は次式で示される。

30（ブロック）×16（頁）×31（日）×６（月）＝89280

層別語彙表（図５）

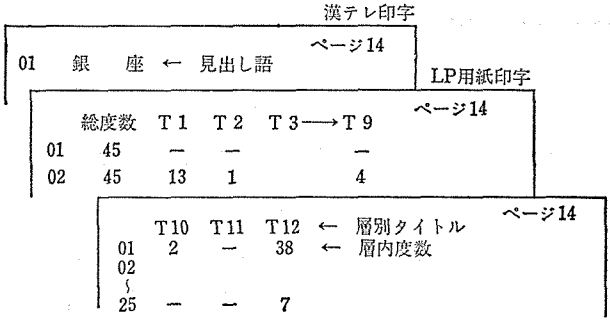
この語彙表は設定された四種の層を中心にして，見出し語の分布状態を調べるためのものである。層は新聞紙面の記事により

1　G　文種別（17）　　2　P　位置別　（7）

3　S　署名態度別（9）　　4　T　話題別　（12）

に分けられ，さらに項目ごとに（カッコ内の数字）細分し任意の組合せで分析が可能となっている。語彙表の印字形式は，四種の中から一層を任意に選択でき度数順，配列順と共に同一プログラムで処理できるようはかった。また印字が一頁におさまらない場合は，見開きの形で二頁に分け見出し語のみ紙テープ出力しオフラインで漢テレ印字する。紙テープは度数順の場合，他の度数順の語彙表にも共通して使用される。図５はこの層別語彙表の印字形式である。

図 5　層別語彙表印字形式



比率表（図 6）

比率語彙表は，各見出し語の出現度数順（下降順）に累積比率，順位等を計算したもので，印字形式は図 6 の通りである。出現頻度の等しい見出し語の順位は全体，記号外共に同順位とし表中のカッコ以後は記号類，エラーデータを省いた順位及び比率である。各比率の単位は全て0/00（パーミル）で示し，値は小数点以下四位で四捨五入を行ってある。また印字すべき度数の下限はパラメータにより任意に指定できる。

図 6.　比率語彙表印字形式

漢テレ印字

ページ　1
01　，
02　の
03　。
△
⋮

LP用紙印字

ページ　1

000000	ドスウ	ジュンイ	ヒリツ	ルイセキ	ジュンイ	ヒリツ	ルイセキ
01	28880	1	42.512	42.512			
02	28448	2	41.876	84.388	（ 1）	51.141	51.141
03	14790	3	21.771	106.159			
04							
△							
△							

2　情報の配列と転写

各プログラムの性格と語彙表の形式について説明したが，情報の付加を行なう場合人手によるものと辞書による方法があるが，一般に人手を用いた場合情報の付加作業は多人数にわたることから，付加された情報の不統一がおきやすく全体のデータの増加と共に清書、データパンチのさい誤りが入る欠点がある。しかし辞書式に比べ原文の細部にまでエデイトが可能となり，語の分析を主とした場合利点が多い。そこで全体の処理を短単位と長単位処理の二つに分け，短単位処理は長単位の機械処理の後に，その結果を利用してエデイトし，人間との作業が調査対象の異なり語についてのみ行なわれ，同一の見出し誤は重複して処理されないようはかった。このため作業の進行については，長単位の機械処理と平行して進められる短単位のエデイト作業のため，作業用の出典語彙表を順次作成して行なければならないが，ラン２で更新される配列順出典ファイルは最終的な語彙表の作成までファイルの更新のみでよく，周期の比較的長いマスターファイルとなる。そこでこの使用頻度の異った二つのファイルを能率よく使用するため，処理方法を二系統に分けてある。

ここで行なわれる処理は，一度出現した見出し語は全て機械で内部処理し作業の対象から省くこと，また配列順にソートするための情報を付加する，この二点である。

配列順のファイルの作成については，不規則に出現する見出し語の文字列の読みを個々に決定することは，磁気テープによる辞書の使用では処理時間が極めて長く問題が多い。そこで一度決定された読みを，辞書の見出し語との照合によって転写する場合と，新たに作成する場合とを分け，最初に入力されるデータを基礎に，情報転写のための辞書を作成し，二回目以後に入力されるデータのうち辞書に集容されている見出し語は全て配列情報と度数の転写を受けるようにした。

情報の転写を行なう場合，両ファイル共に磁気テープの配列順序を等しくしておくことが必要であるが，磁気テープを使用したバッチ処理方法では，処理

能力を上げるために配列順序の決め方は，その処理のつど目的本意に配列するのが通列である。図７は，短単位作業のための印字形式である。表中のカッコは作業用の仮名付けのための空欄である。

また作業台帳の作成と同次に見出し語を磁気テープに転写し，短単位作業用の見出し語管理ファイルを作成するが，これは短単位のデータの入力のさいデータの個数をチェックし長単位と短単位の見出し語の一致をはかるためのものである。機械処理と人間の作業の接続点では，特にデータ個数の管理を厳重に行なう必要がある。

図７．作業台帳印字形式

漢テレ印字

ページ　800

２年〔　　〕生〔　　〕

２年〔　　〕もの

２分〔　　〕10秒〔　　〕96

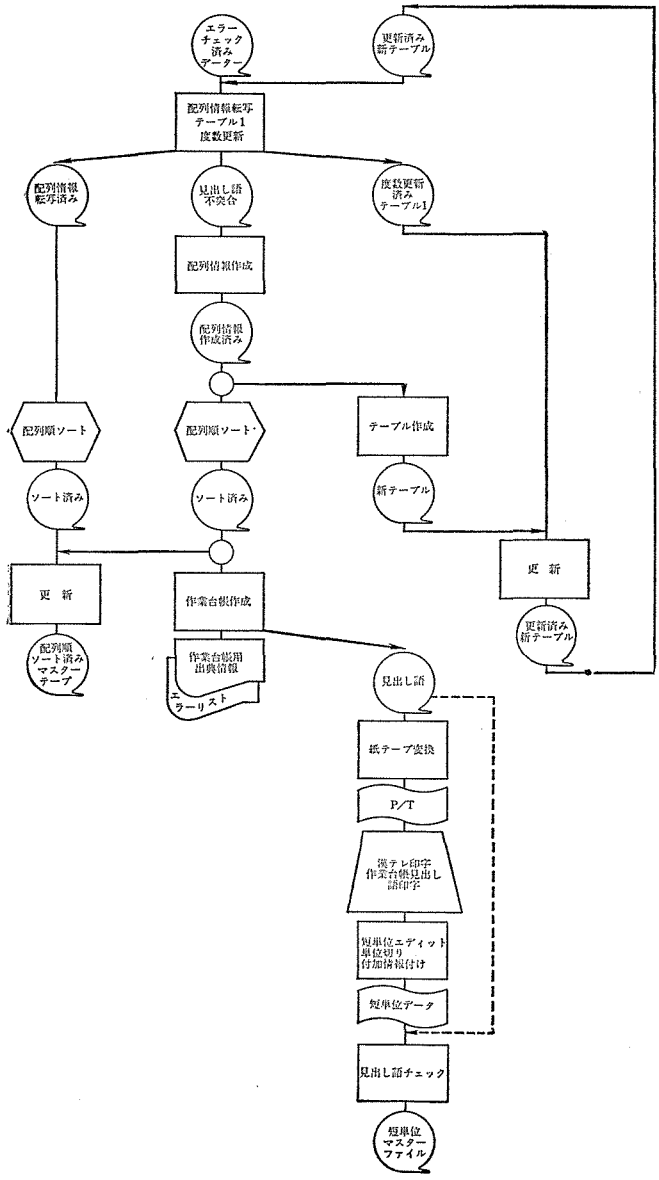
〜

以下略

３　結び

以上現在までに作成されている三種の語彙表とシステムの説明を終るが，電子計算機を使用しての大量のデータを扱う場合，語彙表や磁気テープ中の情報は全て検索の機能を持ち，分析のための資料として研究者に配布できなければならない。それは語の分析方法も従来の手作業の場合と異なり，定式化された部分は全て機械内部で処理され，語彙表としては，分析目的にそった必要な情報のみ選択して印字することが多いためである。これらは，情報検索の一種とも考えられが，データの検索と配布の方法は極めて重要であり，今後，語彙表その他情報の配布に関するシステムの充実をはかっていくことが必要であろう。終りに，このシステムの設計にあたっていろいろ検討していただいた，言語計量，第一資料室の方々，またプログラム作成については，研究補助員の花井夕起子氏に深く感謝いたします。

図 8　システムフローチャートII配列情報転写



参 考 文 献

1) 田中章夫　電子計算機によるワードリスト上の一問題（国立国語研究所報告31）

2) 林　四郎　新聞用語調査の概略と語彙分析法試案（同　上）

3) 石綿敏雄　語彙調査第一段階のプログラムの基本的な考え方（同　上）

付記　短単位処理の概略

Ⅰ　国立国語研究所で行なわれている語彙調査も現在第二段階を向かえ，短単位による調査が進められている。長単位による調査については，昭和41年度に起案され現在にいたっているが，調査の重点を速報性においた結果，付加情報は，最小限必要なものに限定されている。それは，長単位処理自体，短単位処理のプレエディットとしての性格を持ち，単位切りの能率と電子計算機を使用した上での大量のデータを扱う問題点をさぐる目的があったためである。そこでこの短単位作業においては，付加情報として，語種，品詞，活用形情報，その他漢字の仮名付けを行い，電子計算機による語の認定の自動化への方向付けを明確にした。また，異なる二つの語の単位（長単位，短単位）の接続をはかるため用例表を作成し原文の出典を容易に参照できるよう配慮してある。

これによって長単位の問題点であった，同形異語の判定を，この用例表を使用して分析できるはずである。

計算機による処理の概略は，図９のゼネラルフローチャートに示してあるが細部については，各担当者の論文を参照されたい。なお，この短単位処理のシステムの立案は主に斎藤，木村が行い，言語計量，第一資料室の全員によって検討された。

Ⅱ　単位システムの設計において，特に留意したことは次の二点である。

1・エラーデータ処理は，チエック点で判定記号を挿入し，ファイルの分離を行なわず他のチェック済みデータと同一ファイルにまとめた。

図9　短単位　システムフローチヤート

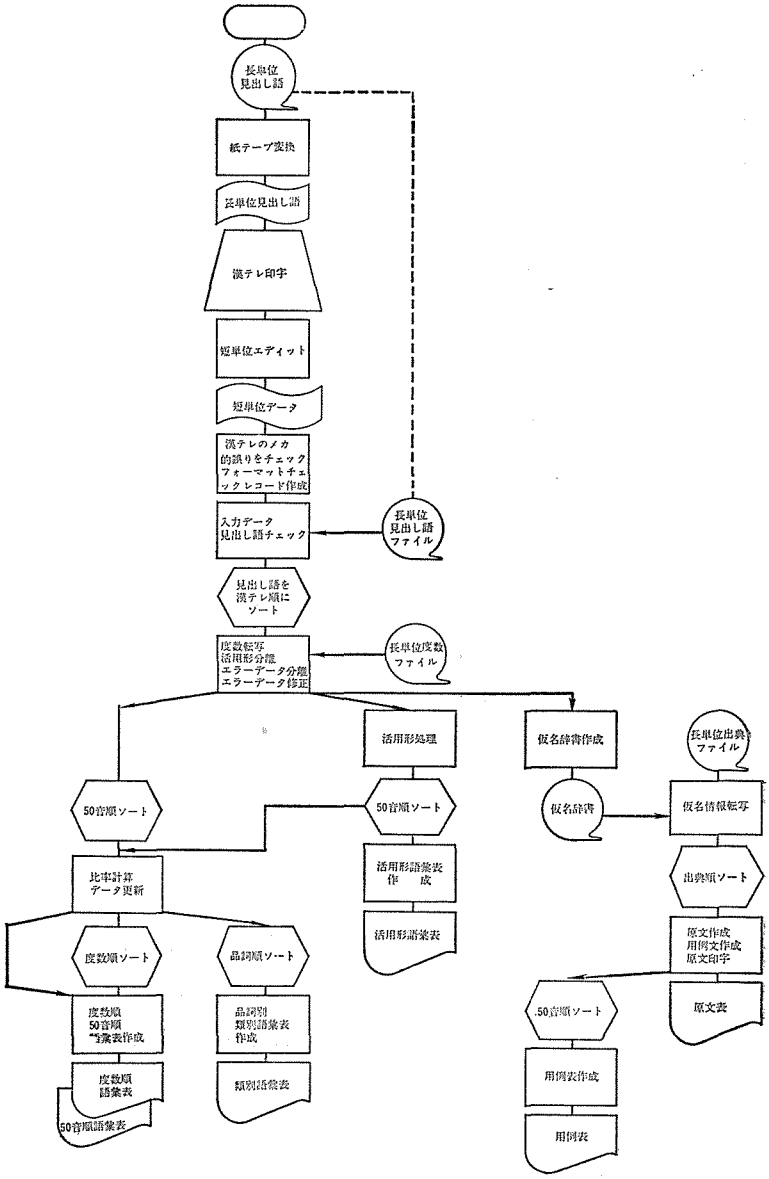
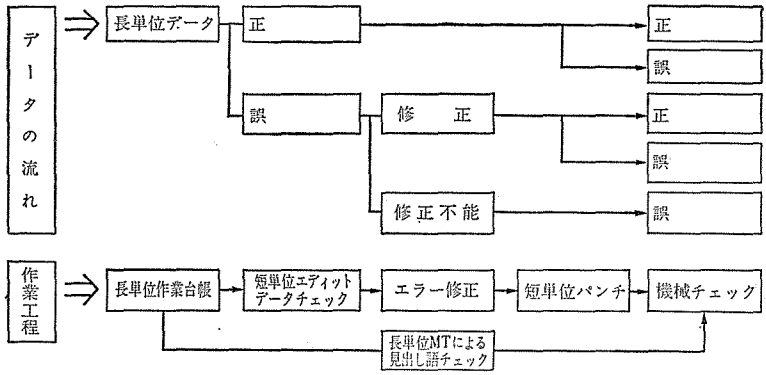


図10



2・磁気テープのフォーマットは，形式を規順化し，全体を印字処理とデータ処理関係の二種類に統一した。

バッチ処理形式の場合，エラーデータ処理は別ファイルに分離され，周期をずらして処理されることが多く，更新は比較的周期の長いものとなる。しかし，エラー処理は時間のかかる手作業の進行に合わせるため，任意の位置で修正を行なえることが望ましいが普通，データの修正は，見出し語を照合する情報として，エラーデータのビットの変化や脱落をそのまま再現して入力しなければならず，作業能率の向上はあまり期待できない。

1の方法で処理を行なった場合，エラーデータは全てファイルの中にあり，正しいデータと修正位置の指定によって置換，消去いずれも修正は容易であり，特にソート処理の後では，エラーデータは一ヵ所に集まり処理しやすい。またエラーの再投入によるデータの脱落を防ぎ，長単位と短単位の見出し語の一致をとりやすく，付加情報の転写を完全に行なえる等の利点がある。

2については前述のとおり，処理を長単位処理と短単位処理の二段階に分けた結果，エラーデータの種類が複雑になり個々のエラー別のファイルの作成は，無駄が多くなること，ファイルの追加，削除が1本のプログラムで任意の位置

で行なえ，処理の割りこみ等，早急に対処できるなどの理由による。
図10は予想されるエラーの位置と種類である。
この短単位処理のアウトプットとして予定しているものは，次の五種の語彙表である。

1　活用形語彙表

各活用語について代表形（終止形）と度数を示し，変化形別の度数カウントを行なう。

2　５０音順短単位表

見出し語別に語種，品詞，活用コード及び出現度数を５０音順に配列したもの。

3　度数順短単位表

５０音順の配列を度数順に再分類したもの。

4　語種品詞別語彙表（種別表）

各見出し語ごとに度数，類内順位，類内使用率を示す。

5　５０音順用例表

見出し語の用例を仮名文字で示したもの。

Ⅲ　以上五種の語彙表については各々必要に応じて紙テープによる見出し語のパンチを行ない，オフラインで漢テレ印字を行なう。また，見出し語を見やすくするため，ＬＰにも仮名文字で印字してある。

以上が語彙調査の短単位処理についての概要である。この調査では，付加情報としては日本語のスタティックな面だけにとどまり，意味的な面は調査の対象からはずされていた。しかし，今後人間の行動を含め人間の持つ価値感と意味の関係を明らかにし，モデルを構成する上に新しいウエイトの導入をはかり，言語のもつダイナミックな面の定式化への方向が必要となろう。また、同時に言語情報処理に関する基礎的研究として，文法，音声，情報理論等広い範囲を持った総合的な開発を行なっていかなければならない。この調査についての詳しい結果は，今後の論文に明らかにされると思うが，この調査の資料がこれら各方面の要求にたえられるものとなれば幸いである。