新聞語彙調査の類別語彙表について

新聞語彙調査の類別語彙表について

中野　洋

0．はじめに

現在，国立国語研究所，第一資料研究室・第三資料研究室・言語計量調査室でおこなわれている電子計算機による新聞語彙調査の短単位処理最終ＯＵＴＰＵＴの一つである類別語彙表の作製およびそれに付随する問題について報告する。

語彙調査に三つの方法――統計的方法，類型的方法，体系的方法――があり，それらを関連させながら調査を進めなければならないとは早くから言われていることである(注1)。

もちろん，調査目的によりその方法も異なろうが，類別語彙表はその類型的方法にあたるものといってよい。品詞による語分類は各説により異なるが，分類された語がある共通の意味・機能・形態をもち，多くの人が共通に理解できるものである以上，品詞による語分類をおこなうことは必要なことであり，常識的な方法でさえある。又，各語に品詞情報がつけられたということは，新聞語彙の品詞構成がわかるなど語彙論の研究に寄与するだけでなく，それぞれの品詞がもつ意味や機能や形態によって種々の研究(注2)を可能にすることにもなる。しかし，それらの研究は当面，類別語彙表の作製とは関係を持たない。

ここでいう類別語彙表とは新聞語彙調査の調査単位(注3)の一つである短単位に付けられた付加情報(注4)のうち語種・品詞・活用情報を利用し，各情報別に

注1）垣内松三「基本語彙学」上巻86ページ

注2）本報告50ページ参照

注3）田中章夫「国立国語研究所新聞語彙調査における言語単位」情報処理学会ＣＬ員会資料68―１参照

注4）本報告付記52ページ参照

つくられるいくつかの語彙表のことである。(注5)

短単位は，新聞からサンプリングされた文・見出し語などがこれも今回の調査単位である長単位に区切られ，長単位処理された後，人手によってさらに分割されたものである。これに，よみがな・付加情報をつけ短単位処理にまわす。

さて，短単位はすでに原文から切り離されており，原文中での短単位の意味，機能はわからない。そのため，付加情報は一語一情報を原則とするが，同表記異語についてはどちらの語をも認め，それぞれの付加情報をその一語につけることになる(注6)。たとえば，「と」は「戸」（付加情報は和語・純名詞・活用なし・動詞以外）の仮名表記、文頭の「と」（付加情報は和語・接続詞・活用なし・動詞以外），自立語などについた「と」（付加情報は和語・助詞・活用なし・動詞以外）の三種の「と」が考えられるから，それぞれの付加情報が「と」一語につく。又，「いき」は「息」（付加情報は和語・純名詞・活用なし・動詞以外）の仮名表記，「意気」（付加情報は漢語・純名詞・活用なし・動詞以外）の仮名表記，「この魚は生きが悪い」「東京行き」の「生き」「行き」（付加情報は和語・連用形転成名詞・活用なし・動詞以外）の仮名表記，「生きている」の「生き」（付加情報は和語・動詞・上一段活用・か行）の仮名表記，「行きます」の「行き」（付加情報は和語・動詞・四段，五段活用・か行）の仮名表記などが考えられるから，それぞれの付加情報が「いき」一語につく。

類別語彙表には，たとえば，上例の「と」は純名詞の表にも，接続詞の表にも，助詞の表にも出てこなければならない。したがって，類別語彙表の語単位は一語につき一付加情報の規則を守り，複数個の付加情報がついている語については，付加情報の数だけ見出し語をふやし，それぞれに一つ一つの付加情報をつける。したがって，類別語彙表の異なり語数，延べ語数は実際の数より大となる。

注5） 表１類別語彙表例47ページ参照

注6） 本報告付記53ページ参照

1．類別語彙表作製プログラム

これは二つのプログラムに分れる。一つ（RUN1）は複数個の付加情報がついている語を一語一付加情報になるようにすること，および各付加情報内での順位をつけ，付加情報に関する集計表をつくるプログラム。他の一つ（RUN2）はパラメータを解釈し，パラメータで指定された付加情報をもつ語を取り出し，取り出された語内での順位・出現率・累積比率などを計算し，指定された量だけを指定された順序に作表（ラインプリンタ・紙テープ）および磁気テープへの書きこみをしながらOUTPUTするものである。

RUN1により作られた磁気テープはそれ以後の類別語彙表の台帳的性格（類別語彙表の作製にはRUN1を通らず，この磁気テープを利用する）を持ち，集計表はRUN2実行前の，どのような付加情報をもつ語の類別語彙表を作るかの検討資料となる。RUN2は類別語彙表作製プログラムのメインプログラムであり，OUTPUTされるラインプリンタ紙および紙テープによる漢字テレタイプ印字紙は類別語彙表として用いられ，又，磁気テープは言語処理研究用として用いられる。

これらのプログラムのフローチャートは図1，パラメータ・フォーマットは図2のとおりである。

パラメータ　パラメータにはOUTPUTしたい語の付加情報およびOUTPUT量（これはOUTPUT数，度数，累積比率のどれか一つで決められる），OUTPUTの順序（度数順か50音順）を指定する。

例1. 和語・副詞・活用なし・動詞以外の語を30語だけ，50音順にしてOUTPUTしたい時，パラメータは図2のように作製する。このパラメータによるRUN2実行後の表は図3，見出し語・付加情報は表1のようになる。

図2　パラメータ・フォーマット（紙テープ）

(gap)　GOSHU=S,　HINSHI=C,　KATUYOUKEI=0,　KATUYOUGYOU=0.
OUTPUT-SUU=30.　50ONJUN.　E/B (gap)

S, C, 0, 0 は付加情報コードであり，和語，副詞，活用なし，動詞以外をあらわす。

図 1—1　類別語彙表作製プログラムフローチャート

RUN 1

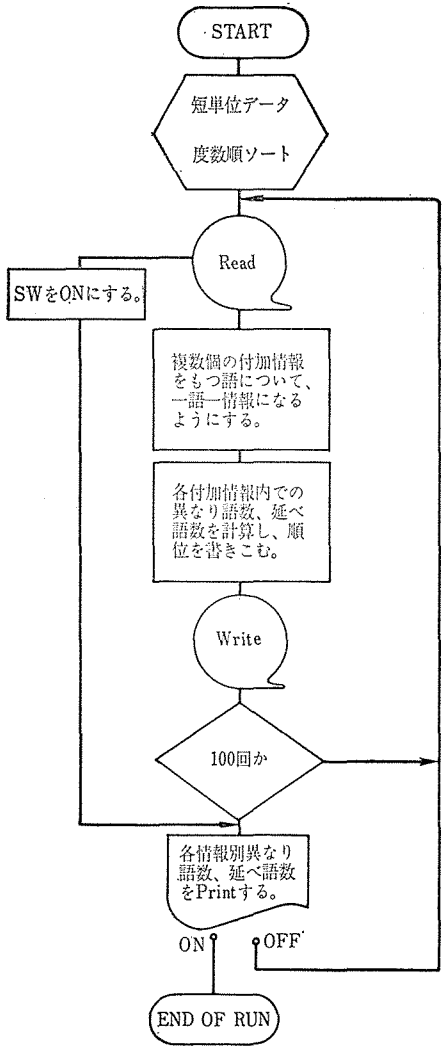
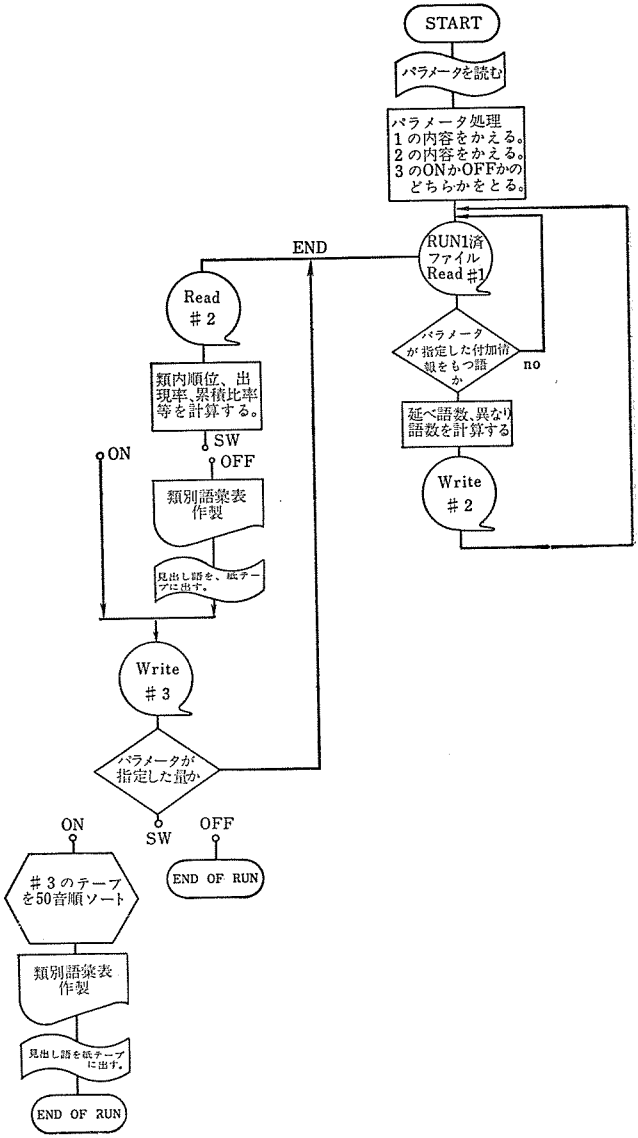


図 1—2　RUN 1


START
パラメータを読む
パラメータ処理
1の内容をかえる。
2の内容をかえる。
3のONかOFFかの
どちらかをとる。
RUN1済
ファイル
Read #1
END
パラメータ
が指定した付加情
報をもつ語
か
no
延べ語数、異なり
語数を計算する
Write
#2
Read
#2
類内順位、出
現率、累積比率
等を計算する。
SW
ON
OFF
類別語彙表
作製
見出し語を、紙テー
プに出す。
Write
#3
パラメータが
指定した量か
ON
SW
OFF
#3のテープ
を50音順ソート
END OF RUN
類別語彙表
作製
見出し語を紙テープ
に出す。
END OF RUN

2．何をOUTPUTするか

OUTPUTのキーは付加情報の一種，あるいは，語種・品詞・活用（2種類）情報の組みあわせである。語種情報9種，品詞情報23種，活用情報「A」12種，活用情報「B」17種のすべての組みあわせを計算すると42228組となり，それだけの数の類別語彙表ができる。しかし，活用情報が幾種類にもわかれるのは和語・混種語・語種不要，動詞・形容詞・動詞性接辞・形容詞性接辞・助動詞だけである。このように実在しえない組を除いて計算すると1628組になり，それだけの数の類別語彙表ができる。

もちろん，この1628組はいろいろな面からの分類であるから，一語がいくつかの組に含まれることになるが（たとえば，「書く」の付加情報はＳＥＦ３であるから15種――Ｓ，Ｅ，Ｆ，３，ＳＥ，ＥＦ，Ｆ３，ＳＦ，Ｅ３，Ｓ３，ＳＥＦ，ＥＦ３，ＳＥ３，ＳＦ３，ＳＥＦ３――の全ての語彙表に含まれる)。全組みあわせをOUTPUTするとなると尨大な量となる。そこで必要なものを，必要なだけOUTPUTしなければならない。

活用語（ただし，動詞，形容詞，動詞的接辞，形容詞的接辞の付加情報１個だけをもっているもの）に関する類別語彙表は，江川清氏の「終止形変換プログラム」中で，それぞれの活用形がその代表形（いわゆる終止形）にまとめられて作表される。したがって，「類別語彙表作製プログラム」では主に語種情報・品詞情報に関する類別語彙表を作製すればよい。

語種情報に関しては，和語，漢語，外来語，混種語の４種，品詞情報に関しては，算用・ローマ数字，記号・符号，品詞不明，情報無視を除いた19種の上位語について類別語彙表をつくれば，重要語のほとんどはどれかの表に含まれるだろう（この場合のプログラムは，付記でのべたように，語種，品詞コードをキーとしてソートする方が処理時間は短くてすむ)。しかし，語種情報と品詞情報とを組みあわせると興味ぶかい語彙表が作れる。

外来語は最初，純名詞又は固有名詞として日本語にとりいれられるが，つかいなれてくると他の品詞としても用いられる。だから外来語で純名詞，固有名

詞以外の品詞を取り出せば，それらは日本語にとけこんだ，外国語意識の薄い語といえる。その中でも最も多いのは「する」がついて動詞化される サ変語幹，「だ」「な」がついて形容動詞化される形動名である。サ変語幹となりえる語は動作などを表わす語であり，形動名となりえる語は状態をあらわす語である。たとえば，

サ変語幹：アルバイト・カアブ・サービス・スタート・チェンジ・デザイン

形動名　：オオバア・ショック・ロマンチック・スポオティ・シャン

その他　：アンチ・グラム・パアセント・ワン・ツウ

混種語についても同様のことが言えるだろうが，その外来語部分の要素は前者より日本語化されていると言える。例をみれば明らかである。

純名詞：ガラス戸・ゴム消し・やみドル・急ピッチ・ビール瓶・あんパン

動　詞：白模（つも）る・だぶる・サボる・ヤジる・アジる

漢語には外国語意識を感じないが，純名詞，サ変語幹，固有名詞以外の品詞は少ない。

付加情報は一語に複数個つくことがある。それらは付記52ページで記したように二つにわかれる。転成語ともとの語が同表記であるため区別されない語は，見方を変えれば，ある場合に限り別種の機能も持つことができる一つの語であるとも考えられる。おもなものをあげれば，

1)　名詞と名詞

(1)　純名詞と形動名　長単位分割の際，助詞・助動詞は自立語と分けられるが，形容動詞語幹とその語尾は分割されない。したがって，純名詞と形動名が同一語につくことはない。

i)　形動名だけのもの。「静か」「堂々」「静けさ」など。

ii)　同形見出しの別語として，純名詞も形動名もあるもの。「危険」「健康」「特別」など。これらは長単位の分割のしかたによって決まる。

(2)　純名詞と非用言的接辞

i)　非用言的接辞だけのもの。「お」「さん」「所（ただし，よみがなが〔じょ〕のもの）」など。

ii)　同形見出しの別語として，純名詞も非用言的接辞もあるもの。「様

（たとえば，『様になる』と『片岡様』などの場合）」「こと」「風」など。

(3) 純名詞とサ変語幹　「サ変語幹＋する」の形は長単位で分割されない。したがって，純名詞とサ変語幹が同一語につくことはない。（『昨日東京を出発，本日鹿児島に到着する』の「出発」は純名詞，「到着」はサ変語幹）

同形見出しの別語として，純名詞もサ変語幹もあるもの。　「勉強」「話」「スタート」など。

2) 名詞と動詞

純名詞と連用形転成名詞と動詞

i) 動詞からの転成が強く意識されるもの。　「泳ぎ」「走り」「出」「遊び」「泣き」など。付加情報は連用形転成名詞と動詞がつく。

ii) 動詞からの転成があまり意識されないもの。　「おび」「ひかり」「まつり」「はなし」など。付加情報は純名詞又は連用形転成名詞又は動詞，それらの組みあわさったものがついている。

iii) 動詞からの転成がほとんど意識されないもの。名詞から動詞へ転成した形があるもの。「相撲」「問答」など。付加情報は名詞がついている。

4) その他

(1) 同形見出しの別語として

i) 動詞性接辞と動詞　「だす」「とおす」「はじめる」など。

ii) 形容詞性接辞と形容詞　「くさい」「やすい」「よい」など。

iii) 形容詞性接辞と助動詞　「らしい」など。

(2) 一語に複数個の付加情報がつく

i) 名詞と助詞　「くらい」「ほど」「ところ」「こと」など。

ii) 名詞と副詞　「いちばん」「一切」「つゆ」「ゆめ」など。

iii) 接続詞と副詞　「なお」「また」など。

iv) 接続詞と動詞　「および」など。

v) 接続詞と助詞　「て」「が」「けれども」「と」など。

vi) 感動詞と副詞　「そう」など。

vii) 感動詞と助詞　「ね」など。

viii) 副詞と動詞「つまり」「たとえ」など。

ix) 副詞と形容詞　「いたく」「よく」など。

x) 副詞と助詞　「たって」など。

xi) 連体詞と動詞　「さる」「あくる」「きたる」など。

以上の語を集めた類別語彙表もつくれる。この表は，転成語をあつめたものであるとともに，どれかの品詞に決定することのむつかしい語をあつめたものであるともいえる。

各表は図3のように作製される。図3の例（和語であり，副詞である語の類別語彙表の一部を作ってみたもの）をもって説明する。「類内全体」は和語・副詞に含まれるものすべて（すなわち複数個の付加情報をもっているものを含む）を示し，「類内部分」は和語・副詞にだけしか含まれないもの（すなわち1個の付加情報を持っているものだけ）を示す。ラインプリンタ紙の番号は漢字テレタイプ印字紙の番号と一致し，見出し語は漢字テレタイプ印字紙にプリントされる。「語種・品詞・活用「A」・活用「B」における順位」とは，例えば番号1の「あまり」についてはその付加情報SC00（和語・副詞・活用なし・動詞以外）のS（和語）内では「あまり」の順位は101位，C（副詞）内では「あまり」の順位は4位……をあらわす。今，「雑誌90種の用語用字」の語彙表（表2）に新聞語彙調査の付加情報を付けて，類別語彙表を作ってみ

図3　RUN2 OUTPUT フォーマット

漢字テレタイプ印字紙

番号　見出し語（仮名）　見出し語（漢字・仮名まじり）

ラインプリンタ

番号	全体 度数	全体 順位	使用率	付加情報	類内全体 順位	類内全体 比率	類内全体 累積比率	類内部分 順位	類内部分 比率	累積 比率
記号				語種・品詞・活用A・活用Bにおける順位						

例　漢字テレタイプ印字紙

0001	あまり	あまり
0002	あるいは	或いは
0003	いか	如何

ラインプリンタ紙

```
1  700 167    .609  SC0 0 SI0 0 S+FE  1  7.543  7.543
*                101  14  580  589
2  534 413.5 .299  SCD 0 SA 0 0        2  4.752 12.295
*                237  24  621  621
3  473 479.5 .262  SC 0 0              3  4.213 16.508  1  10.485  10.485
                   267  29  672  672
```

表　1　類別語彙表　例

例1　外来語，サ変語幹（上位30語　度数順）

（見出し語）	（付加情報）	（全体順位）	（全体使用率）
デザイン	U300 U100	718	・185
スポオツ	U300 U100	1042.5	・130
サアビス	U300 U100	2512.5	・052
ヒット	U300 U100	2863	・046
プリント	U200 U100	2999.5	・044
ゲエム	U300 U100 U600	3150.5	・041
セット	U300 U100 U600	3297	・039
アルバイト	U300 U100	3469	・037
メモ	U300 U100	3469	・037
カアブ	U300 U100	3670	・034
コオチ	U300 U100	3888.5	・032
パイピング	U300 U100	3888.5	・032
プラス	U300 U100	3888.5	・032
タッチ	U300 U100	4425	・027
ハイキング	U300 U100	4425	・027
キャンプ	U300 U100	4757.5	・025
サイン	U300 U100	4757.5	・025
スタアト	U300 U100	4757.5	・025
スパイ	U300 U100	4757.5	・025
テスト	U300 U100	4757.5	・025
デビュウ	U300 U100	4757.5	・025
カット	U300 U100	5158.5	・023
カバア	U300 U100	5158.5	・023
スカウト	U300 U100	5158.5	・023
スキイ	U300 U100	5158.5	・023
トレエニング	U300 U100	5158.5	・023
ダンス	U300 U100	5610.5	・021
ノオト	U300 U100	6146.5	・081
ピクニック	U300 U100	6146.5	・081
マアク	U300 U100	6146.5	・081

例2　和語，副詞（上位30語　五十音順）

（見出し語）	（付加情報）	（全体順位）	（全体使用率）
アマリ	SC００ S100 S+FE	167	・669
アル(い)は	SC００ SA００	413.5	・299
イカ	SC００	479.5	・262
イロイロ	SC００	281	・425
オナジ	SC００ S100 SD００ SLN０	120	・906
カナラズ	SC００	382.5	・320
コウ	SC００	123	・891
サラニ	SC００	235	・500
スグ	SC００	241	・475
ススコシ	SC００	203	・583
ススデニ	SC００	326	・372
スベテ	SC００	362.5	・331
タダ	SC００ S100	182	・621
タトエバ	SC００	479.5	・262
タシカ	SC００ S400	560	・231
チョット	SC００	217	・545
トテモ	SC００	552.5	・233
ナオ	SC００	406	・304
ナカナカ	SC００	545.5	・235
ハジメ	SC００ S100 S200	366	・329
ハッキリ	SC００	419.5	・294
ホトンド	SC００	419.5	・294
マズ	SC００	248	・468
マダ	SC００	192	・599
マッタク	SC００	366	・329
モウ	SC００	111	・955
モシ	SC００	385.5	・317
モット	SC００	425	・292
モットモ	SC００ SA００	411	・301
ヤハリ	SC００	171	・650

（見出し語，見出し語の表記，全体順位全体使用率は「国立国語研究所報告21現代雑誌九十種の用語用字」第一分冊第二表によった。付加情報は新聞語彙調査 短単位付加情報つけの規則によってつけた。付加情報の記号は付加情報コードであり，52ページ参照）

ると，見出し語・付加情報・全体順位・全体使用率は表１のようにならぶ。

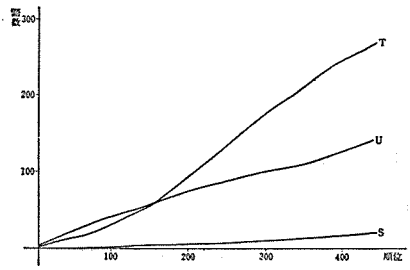
３．OUTPUT 量はどのようにして決めるか

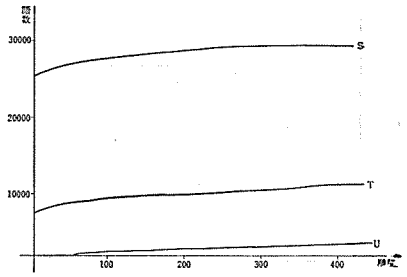
類別語彙表作製プログラムではパラメータの指定により各類別語彙表の OUTPUT 量を決めることができる。

付加情報の組みあわせによっては数えるほどの少量の語しか含まない類別語彙表や，全語彙の何割かをしめるような大量の語を含む類別語彙表などが出現する。たとえば，次のことは明らかに予想できる。純名詞，固有名詞，和語の動詞等は絶対量が多く，したがって，それらを含む付加情報の組みあわせに属する語は多い。又，接続詞，感動詞，連体詞，動詞性接辞，形容詞性接辞，助詞，助動詞，算用・ローマ数字，記号・符号などは絶対量が少なく，したがって，それらを含む付加情報の組みあわせに属する語は少ない。又，語種では，和語，漢語，語種不要などは多く，数字，記号などは少ない。もちろん，数多くの語を含む組みあわせでも OUTPUT したい語は少ない（たとえば固有名詞などは度数１や２の語を OUTPUT してみても，意味がない）場合や，数少ない語を含む組みあわせでも OUTPUT したい語は多く（新聞語彙にあらわれた語をすべて出したい時，たとえば，外来語の形動名や漢語の副詞や接続詞など）なる場合も少なくない。それらは，各組みあわせに対する語彙表利用者の用途，必要性，興味などによって決まる。しかし，それらを考える前に必要な事は，各語彙がどのように分布し，新聞語彙を構成しているのかを知る事だろう。

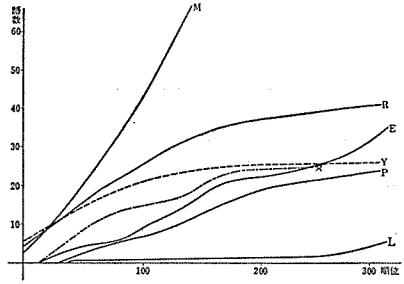
語彙の分布状態は延べ語数を縦軸に，異なり語数を横軸にとると，およそA図のようになる。ところで，類別語彙表作製プログラムはパラメータで OUTPUT 量を制御するが，それらは OUTPUT 数，各語の度数，累積比率のどれかで決定される。OUTPUT数，度数，累積比率と延べ語数，異なり語数の関係はA図によって大体知ることができる。実際には，OUTPUT 量は先に言った条件により決定されるが，それらを無視し，必要で最少限の語を OUTPUT しようとすれば，A図中，①，②，④より③の状態でOUTPUT するのが望ましい。

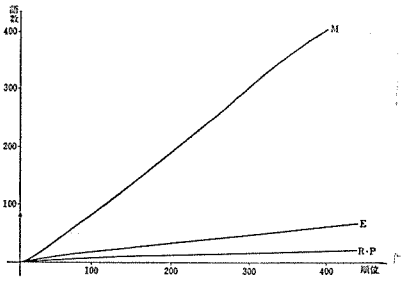
図4　1）語種別・累積異なり語数　　2）語種別・累積延べ語数



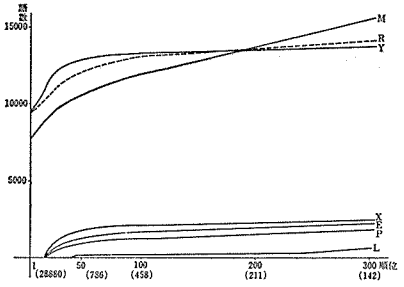


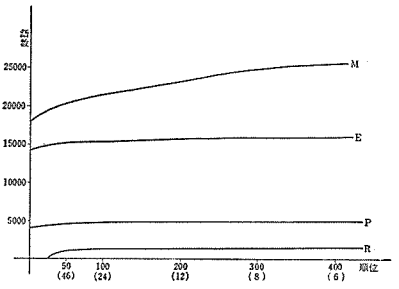
3）品詞別・累積異なり語数　　4）品詞別・累積異なり語数





5）品詞別・累積延べ語数　　6）品詞別・累積延べ語数



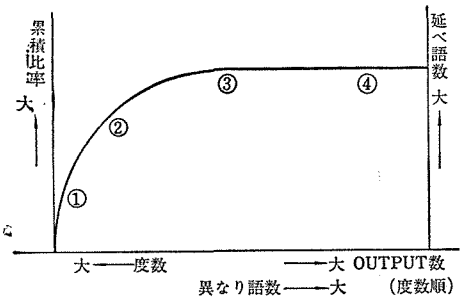


（注）グラフ3，5は長単位全体度数順語彙表から上位300語を抽出し作表した。グラフ1，2，4，6は長単位全体度数順語彙表から間隔30語で等間隔抽出法により度数6以上の語計440語を抽出し作表した。各グラフの横軸は標本内の順位である。グラフ5，6の横軸（　）内数字は実度数である。

グラフ5と6の内容が異なるのは，グラフ6が等間隔抽出法によったものであり，度数のきわめて高い上位も，度数の低い下位も同様に30語間隔で採集されたためと考えられる。

グラフ中の英文字は付加情報コードであり，付記52ページを参照。

A図



類別語彙表作製プログラム RUN 1には各語彙の分布状態がわかるような集計プログラム（度数順にならべた時の各情報別の異なり語数と延べ語数の増加の様子をあきらかにするもの。語種と品詞はその関係もわかる。）が，組みこまれているので，それも検討の上，OUTPUT 数は決定される。ここに，ごく簡単な予備調査(注1) の手集計の結果（図４）があるので参考にかかげておく。

この予備調査の目的はOUTPUT の方法を得ることであり，又，新聞語彙調査経過中の調査であるため，標本数が少なく，短単位処理後のものでない。それゆえ，この調査結果から母集団を推測することはきわめて危険であるが，先にのべた目的と，各語彙の大まかな様子を知るには有効だろう。

４．磁気テープの利用

類別語彙表作製プログラムの OUTPUT の一つである磁気テープは語彙表とは違った意味をもつ。磁気テープの最大の特徴はそのままの形で電子計算機に INPUT できることである。電子計算機にとっては単なる記号に過ぎなかった語は，磁気テープに書かれている付加情報を知ることによって，それがもつ意味・機能・形態をその範囲で得ることになる。又，大量の文字が書きこめること，処理時間が短かくてすむという特徴は，磁気テープの内容の消去，修正，加筆を迅速かつ大量に可能にする。したがって，類別表作製プログラムの磁気テープはすべての日本語とその意味・機能・形態・用例などあらゆる情報が書きこまれている日本語の総合辞書への基礎とならねばならない。それは又，あ

注1） 予備調査は新聞語彙調査の朝日新聞朝刊６ケ月分の長単位度数順語彙表から，度数６以上の語を，上位300 語までと，間隔30語の等間隔抽出法によって採集された語440語（どちらも長単位）を短単位に分割し，付加情報をつけて集計した。

らゆる言語処理研究用の辞書として活用できるものでなければならないからである。

現在の付加情報は新聞語彙調査のためのものであって，言語処理研究用としては限界がある。言語処理の研究に際し，語の形態だけによって処理をはかるとすれば，現有の付加情報活用コードがおおかた活用形処理(注1) のためのものであるように，機能面の充実をはからねばならない。これも又，その目的により情報のつけ方もわかれるが，総合辞書としての磁気テープにはそれらのすべてを満足させる情報をもっていることが望ましい。しかし，それは順次達成されるものである。今，ここに私が考える電子計算機による構文解析（語の形態・機能によってせまり，意味情報は極力つかわない）を例にとれば，各語には品詞・活用・係り結び・呼応表現などかかり・うけに関する豊富な情報が必要となる。助詞の分類（格助詞・副助詞の分類，接続助・終助詞の区別），副詞の分類（陳述副詞と述語の呼応・かかり方による分類），補助用言の情報とその分類，名詞の付属語をともなわない単独用法の有無に関する情報などがそれである。

この報告の最初に引用した体系的方法による語彙調査は新聞語彙調査においてはなされていない。しかし，「分類語彙表」によって，語を体系化する方法と電子計算機に入力可能な数値を得ることができる。言語処理への意味の導入が重要な問題になっている現在，我々はより充実した意味情報を得なければならない。

5．あとがき

以上が現在，作製ずみのプログラム，および実行計画である。類別語彙表およびその磁気テープは，調査単位が短単位であること，付加情報つけが原文から切り離された語になされていることなどおのずから限界もあるが，種々の研究資料として各方面に活用されることと信ずる。又，類別語彙表の分析およびそれを利用した研究は今後逐次おこなわれ，発表される。

注1） 本報告書55ページ　江川清「活用形処理」の自動化における一方式

付記　付加情報について

新聞語彙調査は長単位による第一次入力と，短単位による第二次入力によって，そのプロセスが大きく二分される。付加情報は第二次入力の際，短単位に分割された語一つ一つにつくものである。それには短単位処理中に消去される位置情報と，最後まで残る語種情報・品詞情報・活用情報（２種類）の４種がある。内容は下記のとおり。（　）内は付加情報コードである。

位置情報[注1]：単独（Ⓢ），前部分（ｎ），中部分（？），後部分（？），情報無視（％）の計５種。

語種情報　：和語（Ｓ），漢語（Ｔ），外来語（Ｕ），混種語（Ｖ），語種不要（Ｗ）[注2]，数字（Ｘ），記号（Ｙ），語種不要（Ｚ），情報無視（％）の計９種。

品詞情報　：純名詞（１），，連用形転成名詞（２），サ変語幹（３），形動名（４）[注3]，形容名（５）[注4]，非用言的接辞[注5]・助数詞（６），数詞（７），固有名詞（８），代名詞（９），接続詞（Ａ），感動詞（Ｂ），副詞（Ｃ），連体詞（Ｄ），動詞（Ｅ），動詞性接辞（＋）[注6]，形容詞性接辞（－）[注7]，形容詞（Ｌ），助動詞（Ｐ），助詞（Ｒ），算用・ローマ数字（Ｘ），記号・符号（Ｙ），品詞不明（Ｚ），情報無視（％）の計23種。

活用情報「Ａ」：活用なし（０），四段・五段活用（Ｆ），上一段活用（Ｇ），上二段活用（Ｈ），下一段活用（Ｉ），下二段活用（Ｊ），変格活

注1）長単位の中のどの部分であるかを表わす。

注2）品詞情報が固有名詞，助動詞，助詞，品詞不明のものである。

注3）形容動詞語幹および形容動詞おこりの名詞。形容動詞は短単位分割の際，語幹と語尾に分割され，語尾の品詞情報は助動詞とされる。「静か」「堂々」「しずけさ」など。

注4）形容詞おこりの名詞。「美くしさ」「深さ」「なつかしみ」など。

注5）接辞類で活用しないもの。「お」「さん」「殿」など

注6）接辞類で動詞型の活用をするもの。「めく」「じみる」「ぶる」など。

注7）接辞類で形容詞型の活用をするもの。「こい」「らしい」「ぽい」など。

用（K），口語形容詞（M），文語形容詞（N），助動詞（P），形容動詞語尾（Q），情報無視（%）の計12種。

活用情報「B」：動詞以外（0），わ・あ行（1），あ行（2），か行（3），が行（4），さ行（5）……わ行（F），情報無視（%）の計17種。

短単位入力例

η愛〔あい〕し（V200）
ʒ方〔かた〕（S600）
η沿〔えん〕岸〔がん〕（T100）
ϙ警〔けい〕備〔び〕（T100）
ʒ隊〔たい〕（T600）
Ⓢな（S100）（SB00）（WR00）（WPP0）

情報は一語につき一情報が原則だが，付加情報つけの作業は長単位処理後なので，原文中での意味・機能はわからない。したがって，同表記異語についてはその区別はつかず，一語に二情報以上の付加情報がつくことがある。それらは次の二種類である。

(1) 全くの別語が同表記のため区別がつかない語。
たとえば，「は」の付加情報はS100とWR00である。

(2) 転成語ともとの語が同表記であるため区別がつかない語。
たとえば，「つゆ」の付加情報はS100とSC00である。

付加情報コードは各情報内で下降順につけられている。語種・品詞・活用情報をそれぞれ第1，2，3，4ソートキーとしてソートすれば，各語はS100の語を先頭に，S200・S300……SEF1・SEF2……SEG1……%%%%，順にならぶ。類別語彙表プログラムではこの方法をとらない。理由は，各情報の組みあわせがどれかに決まっている時，この方法は有効だが，そうでない時，指定されうる情報の組みあわせは十数組にのぼり，そのたびごとに全体をソートしなければならないからである。

類別語彙表作製プログラムでは使わなかった付加情報のうちの一つ，位置情報は短単位が長単位のどの部分にあったかを示すものである。同じ短単位を含

む長単位をすべて集めて位置情報によって分けると，単独の用法・前部分として使われた例・中部分として使われた例・後部分として使われた例と並べることができる。表2の短単位用例表がその例である。

表 2　短単位用例表　例

お	625		大学		213
単独	0		単独	100	
前部分	お求め	34	前部分	大学講座	18
	お笑い	27		大学生	16
	お申込み	25		大学卒	14
	お問合せ	22		大学受験	13
	お送り	21		大学当局	8
	お手伝さん	21		大学院	7
	お知らせ	19		大学側	6
	おしゃれ	18	中部分		0
	おしらせ	17	後部分	短期大学	8
	お正月	15		早稲田大学	8
	お店	15		六大学	8
	お手伝	15		国立大学	7
	おけいこ	15			
	おなじみ	14	主義	48	
	お芝居	13			
	お答え	12	単独	0	
	おはなはん	12	前部分		0
	お茶	12	中部分	共産主義者	9
	・		後部分	社会主義	15
	・			帝国主義	11
中部分		0		米帝国主義	7
後部分		0		実力主義	6