A system analysis of the word count programs

語彙調査第一段階のプログラムの基本的な考え方

石綿敏雄

I

ここでいう語彙調査プログラムとは，語い調査を電子計算機によって行なうばあいの，機械処理プログラム全体をさしていうことにする。用語調査を電子計算機を用いて行なうばあいには，全体の行程を分析し，どの部分をどのように機械に乗せるのがよいかについて，まず考えなければならない。

プログラムのシステム・アナリシスを行なうばあい，原始データとしてどのような形のものがあり，計算機にかけるまでにどのように手を加えることができるか，入力としてどのような形にするのがよいか，出力としてどのようなものが必要であるのか，そのためにどのような機械処理が必要になるか，また可能であるか，また以上にあげたことが互にどのように関係させるか（前の行程から後ろの行程へというだけでなく，むしろ後ろの行程から前の行程への要請も含まれる），どのように配置すれば効率的であるか，結局全体としてどのような形を考えるのか，というような問題をとりあげなければならない。

語い調査のばあいには，原文，前処理，機械処理，後処理のあり方を検討することが必要である。ここでは，われわれが現在行なっている新聞用語調査のプログラムの，基本的な考え方について述べることにする。

このような処理についての分析は，従来各方面で実施された，または実行されつつある，電子計算機による用語調査の際にも行なわれているわけであり，本来，ここでもそれについてふれるべきであるが，いつさい省略する。[注]

いま従来の考え方や処理方法について言及する余裕がないので，筆者が見た文献の目録を示すにとどめる。

1) NSF : Current research and development in scientific documentation: 11 (1962)～ Washington
2) Академия Наук СССР : Автоматизация в лингвистике, сборник статей переведенных с английского, Французского и чешкого языков. Москва-Ленинград 1966.
3) cahier de lexicologie, Actes du colloque international sur la méchanisation des recherches lexicologiques, Besançon. 1961. Paris 1962.
3) A. Juilland ; Frequency dictionany of Spanish Words. Hague 1964.
4) 水谷静夫「フェランティマーキュリ計算機の言語的問題への応用」『こんぼ～と』no. 2 1963
5) 水谷静夫「電子計算機と古典の総索引作り」『国語と国文学』1964
6) 吉田昭「コンピュータによる『フランス語宝典』の編纂」『数理科学』1966
7) 菅野謙，石野博史「電子計算機によるラジオ・ニュース用語の分析」ほか『文研月報』1967

以上のうち2)は，たとえば A. J. Colin ; The automatic construction of glossary《Information and Control》1960 vol. 3 などを始めとする，西欧各国での業績を集めて翻訳したものである。

次に，筆者が今までに書いたものを並べておく。

i）「電子計算機による語い調査の一実験」『国語研究所論集2』1965
ii）「国立国語研究所における電子計算機の application」『HITAC ユーザ研究会第3回大回紀念論文集』1966
iii）「ことばと電子計算機」『数理科学』1966
iv）「スペイン語の語い調査」『スペイン図書』1966

語い調査全般のこと，語い調査を電子計算機で行なうばあいの全般的な問題，および この論文のなかで述べるわれわれの方法のための準備的研究などについては，以上(i～iv)にゆずって，ここではくりかえして述べない。なお，準備的研究である「65カンソ」のプログラムについては本報告書中の

斎藤秀紀「電子計算機と漢テレによる用語総索引の作成」(本報告書所収)を参照。

II

日本語の用語調査を考えるとき，どうしても文字のことを問題としてとりあげなければならない。電子計算機のような機械を用いるばあいにはことにそうであって，このことが全体の設計に大きな影響を与える。われわれが現在行なっている方法は，漢テレを利用し，漢字のままでIN,OUTできるように考えた。[注)]

前節において語い調査プログラムのばあいには，しごとの流れの分析として大きくみて，原データ，前処理，機械処理，後処理，結果表類の諸段階とその連絡について考える，といったが，これを取り扱うに当たって今回は次のように考えることにした。

プログラム全体のつくり方語い調査作業の電子計算機への乗せ方として，

① Pre-edit 作業で層別単位の切り方のすべて，漢字の読み方，同語異語などの情報のすべてを書きこんでしまい，あとは計算機による処理を行なう方法，と

② Pre-edit 作業では上記のうちのごく小部分に限定し，一度計算機による処理を行ない，あと再び人間が手を入れる方法

とが存在する。そして，②のなかでも，人間の後処理のあと再び計算機による処理を行なう方法も考えられる。さらに「ごく小部分」に何を含めるかによっても考え方が違ってくる。

結論から先にいうと，われわれのプログラム・アナリシスでは②の方法を採用し，Pre-edit としては，長単位語による単位切りと層別の情報を添えること，という二つの作業を行なうだけにした。

これは次のような理由による。

① 電子計算機への入力以前に各単語に情報を付けるとすると，出現する単語のすべてに出現順に情報をつけていかなければならない。このことは，単語の使用法の把握認定にとっては能率的で便利であるという利益もあるが，それによる損失もかなり大きい。すなわち，

漢テレを用いること，および漢テレ自体については，本報告書所収　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について」および48ページ所掲論文ii，iii参照。

1) 繰りかえし繰りかえしでてくるものにいちいち情報をつけなければならない。

2) そのようなばあいに，人間による作業では不統一なあつかいをしてしまうことが多く，これをさけることは困難である。

3) 人間作業による不統一な処理は，これをそのまま計算機によって data process すると，out put にそのままにあらわれてくる。

4) 同語異語の情報をあらかじめ記入するためには，結局ひとつの辞書のようなものをつくらなければならない。これはこれ自身大きなしごとになってしまう。

5) 長い単位は作業的にみて比較的切りやすい面があるが，短い単位は作業の等質性を保つのがむずかしい。[注1)]

② 長い単位に切って入力し，電子計算機で分類排列したあとで短い単位に切ったり種々の(同語異語など)情報を記入すれば，

1) 同じものに対してはまとめて処理することができるので，はじめ原文について情報を記入するばあいと比較すると，くりかえして記入する手間が省け，1回で行なえる。

2) 単位切り，同語異語の記入なども同様の理由でまとめて行なえるので，時間的な節約，記入の誤まりを防ぐことができる。

このような点から考えて，上記のように Pre-edit として長単位語による単位切りと層別の情報を添えるという方向をとることにしたのであった。

以上のことをたしかめるために，時間計算を行なってみた。長単二つの単位を用い，漢テレによって前後数語の用例を打ち出して，それによって同語異語の判別を行ない，最終的な語い表を作成するという段階まで考えてみた結果，やはり上記のと同様の結論が得られた。[注2)]

---

1. 今回の調査ではいわゆる文節にあたる長い単位と，短い単語，または造語成分にあたる短い単位，との二つの単位を共用することにしている。

2. このことについてもここでくわしくふれる余裕がない。次の文献を参照してほしい。山本武，小林さち子，石綿敏雄「語い調査プログラムの時間計算」『月報』(国立国語研究所第一資料研究室言語計量研究室発行) AUGUST 1965

III

国立国語研究所で現在行なっている新聞語い調査の第一段階プログラムは，その全体的なプランが昭和41年春に起案されたもので，このプログラムの意図はできるだけ早くある程度の段階のデータが得られることを目的としたものであって，「急行コース」と呼ばれる。ここではこれを全体的に展望することにする（細部にわたっての問題についてはそれぞれの関連論文をみてほしい）。「急行コース」のプログラムプランは NBC 山本武氏の協力を得て第一資料，言語計量研究室員の全員によって討議され，ほぼ同年中に実際の機械用プログラムが完成された。

語い調査急行コースは次のような目的をもって企画されたプログラムである。

1 簡単な pre-edit のままで入力し，それからできるだけ多くの情報が得られることを第一の目的とする。

2 同語異語のしわけはできないが，原文についての語形による索引ができること。また層別によるうちわけが示せること。

3 文字に関係する調査資料としてはかなり多くのことが得られること。すなわち文字の使用度数の集計とそれによる分類が可能なこと。また文字の使用された長単位語が用例として out put できること。

4 上記2の表があとの人間分析の作業に使えること。

計算機による処理を行なうためには，その前に原文について前処理を行ない，これをさん孔しなければならない。この急行コースでは原文に資料番号をそえ，原文の層別情報を与えておく。[注1)] 原文を長単位に切る。[注2)]

という前処理をほどこし，これをさん孔しやすいように特製の原稿用紙に清書し，これを漢テレによって紙テープにさん孔する。

計算機のなかでの操作について述べると次の通りである。

① その原文を計算機のなかによみこみ磁気テープに書く。

---

1. 本報告所収　林四郎「新聞語彙調査の概略と語彙分析法試案」について，参照。
2. 計算機で単位切りを行なうことも考えられる。この問題にも着手しているが，その実現はやや遠い将来の問題であろう。

② 原文の一つ一つの単語についてレコードを作成する。[注1)]

③ これを分類する。

④ 語毎に各種の集計を行なってこれを out put し，各種の表を作成する。(後述 out put 1～3)

⑤ 各字について一つ一つのレコードをつくる。

⑥ これを分類し集計して漢字分類のための表を作成する。(後述 out put 4～5)

上のような操作をほどこした。

部分的にさらにくわしくいえば，

1 計算機のなかに原文を入れるにあたって，pre-edit としての人間作業のあやまりや，紙テープにさん孔した漢テレ上のあやまりなどがあるかもしれない。漢テレのマシンチェックが必要になる。[注2)]

3 分類する，については，分類排列のとき何も手を加えないままで行なえば，漢テレコードの HITAC 3010 によって所持しているコードの大小順に並ぶことになる。

これだけでも索引を作れば使えないことはないが使いにくいので，あらかじめ漢字の代表音をきめた表を用意しておき，それを参照してデータの各レコードの第一字目の漢字に代表音をそえ，これによって分類排列して引きやすい語い表を作る方法をとった。[注3)]

急行コースの out put として予定しているものは，次の5種である。

① 50音順語い表

各語の総使用度数とそのひとつひとつの出典の表。この語い表は総索引の性格をもつ。

② 度数順語い表

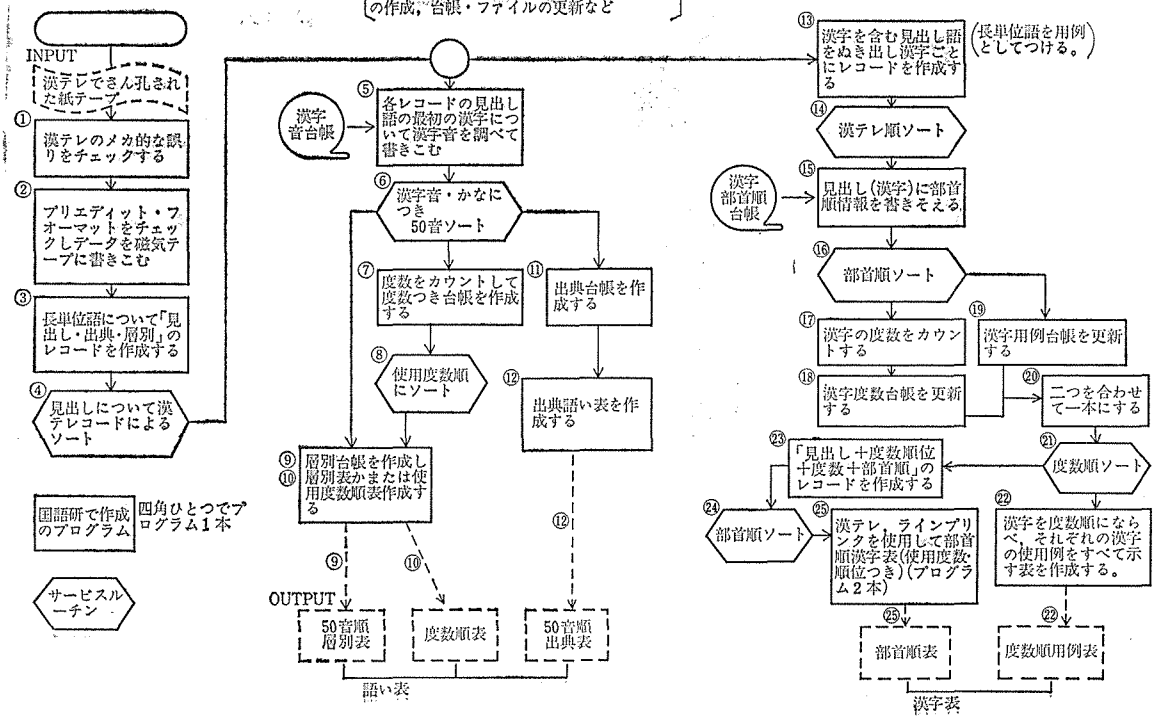
①の50音順語い表を，出典を示さず，度数順にならべかえたもの。

---

1. レコードを作成するというのは人間の手作業でいえば単語カードをつくるのに相当する。これを計算機の中で行なう。
2. これに関しては本報告所収の木村繁「漢テレ入力データのチェック」を参照
3. 本報告所収田中章夫「電子計算機によるワードリスト作成上の一問題」および田中章夫「用語調査におけるワードリスト作成上の一問題」HITAC ユーザ研究会第4回大会記念論文集 1967 参照。

語い調査　　プログラム・ブロックチャート。

この図は大体を示したものでこまかくいえばそれぞれのプログラムにいくつかのプログラムが付属していることがある。たとえば台帳の作成，台帳・ファイルの更新など



（後にこれに層別の数字が付加された。）

③ 50音順層例語い表。

50音順語い表に出典を示さず層別度数表を付記したもの。

④ 度数順用例付漢字表

漢字を使用度数順にならべて度数と順位を示し，その漢字がどんな用語に用いられたかを示す表

⑤ 部首順漢字表

部首順に排列した漢字表，度数と順位付き。

度数順用例付漢字表のための索引として使用することができる。

out put の表作成にあたっては，どうしても漢字を用いなければならない部分だけ漢テレを用い，数字のみで足りる部分はラインプリンタで打ち出すようにし，前者と後者が連絡できるように考えた。これは out put の時間をできるだけ少なくするためである。したがって漢テレの印字フォーマットも工夫し，むだなスペース動作を行なわないように考えた。

計算機の中でのデータ処理の手順は，前ページのゼネラルフローに示す通りである。（当初の計画）

Ⅳ

この急行コースの結果のアウトプットでは，次のような点に問題が存する。

1 ここで得られる語い表は，もとの語表記をそのままとったもので，このままでは不完全である。すなわち漢字にその完全な読みがきめられていないし，語い調査の過程上，同語異語についての操作（同形異語の弁別と異形同語の集合）をへていない。

2 上に示したような操作をどこまで計算機で行なうことができるか。完全な自動化を急に行なうことができないが，人間の操作を加えるとしてこれをどの点まで行なうことにするか。[注]

この問題に関しては田中章夫「電子計算機による漢字の自動解読とその問題点」『計量国語学』および48ページ文献i参照。文献iで述べた方法はわれわれの段階ではいまだまったくの実験的なこころみでしかないが，48ページ文献6によれば，フランスでは大規模な設査で実際に使用するという。

現在のところ多分に人間の作業に頼らなければならないが，そうだとすれば人間の作業をどのような形で行なって最終的な語表作成にいたるか。分析をどのように行なうか。

以上のような問題および計画遂行中に行なわれた変更などについては，別の機会に発表されることになろう。