高校教科書用語調査システムとその設計思想

高校教科書用語調査システムとその設計思想

土屋信一・中野　洋

1. はじめに

高校教科書の用語調査は電子計算機による大規模な語彙調査の二弾目である。現在，その進行も終了段階で，データの修正がほぼ終わり，集計整理の段階に達している。この調査を遂行している間，たびたび語彙調査論，語彙調査システム論について意見をたたかわせた。それは企画時に最も盛んであって，言語計量研究部の研究員全員での会議が何回となく開かれた。その第一回は，電子計算機による新聞の語彙調査終了時の昭和47年12月に始まり，昭和49年12月ころ実質上の企画会議は終わっている。それ以降も全体の打ち合わせはたびたび持たれたし，各人はそれぞれの分担において，調査全体について考えることが多かった。

土屋は高校教科書全体を総轄する立場に立ち，中野は主に電子計算機使用の立場に立ってこれらの問題を考えた。この調査の終了段階を迎えて，これまでの討議内容を頭においた上で，筆者らの考えを加え，語彙調査システム論の立場から，高校教科書の用語調査を見なおしてみようと考える。

まず，今回の調査における作業システム，プログラムシステムの特徴をあげ，それらが語彙調査論とどう関わるかを考える。次に具体的な作業の流れを概観する。

2. プログラムシステム・作業システムの特徴

2-1. プログラムシステムの特徴

1. KWIC を使ったこと …同語異語判別を完全にやったこと
　　…エラー率をかなり低くすることが出来たこと

…スピードが落ちたこと

2. データ管理システムに組みこんだこと

3. オペレータ管理をシステムに組みこんだこと

4. 検査出力に漢字プリンタを使ったこと

5. 自動処理は採用しなかったこと

6. 処理時間に余裕をつけたこと…失敗を恐れないこと
…成果を捨てる勇気をもつこと

7. 修正ルーチンを充実させたこと（修正こそが肝要である）

2-2. 作業システムの特徴

1. 二つの調査単位を同時に採用し調査を行ったこと

2. サンプリング調査にしないで全数調査にしたこと

3. 分析までを計画にくみこんだこと（これが果されるかどうかは今後）

4. エラーのおこりやすい箇所を調べ，その対策を講じたこと

5. 1/20のサンプルデータの調査を先行させたこと

6. 修正はポストエディットが主体であったこと

7. 同語異語判別を完全に行ったこと

8. すべてのデータについて研究員が検査をしたこと

3. 語彙調査論その思想

ここでは，われわれが，また国立国語研究所が，なぜ語彙調査を行うのかというところまで戻って語彙調査論を展開するつもりはない。報告13・94ページ「語彙調査の成立根拠と基本的諸概念の定義」に示された「語彙調査」を行うものとする。ただし，報告13の当時と比べて，語彙調査に求められるものが多様化してきていると言うことはできる。ここでは，電子計算機を用いた語彙調査が，どのような形で進められるべきかについて考える。

3-1. 電子計算機を用いた語彙調査のあるべき姿

1. 電子計算機を使った調査の特色は，大量のデータを短時間で処理することであるかのように言われ，思われていた。しかし本当はそうではない。それが真実であるのは「それが単純作業であれば」という条件がつく。語

彙調査の総体は，決して単純作業の連続ではない。電子計算機を用いた調査の特色は，「単純作業は機械で，複雑な作業は人間がやること」にある。これにより，より複雑で，かつより精密な調査が可能になる。

2. 語彙調査において，見出し語の配列や使用率の計算，語彙表の作成は単純な作業に属する。これらは機械に任せてよい。しかし同語異語の判別や，単位切り，情報つけ，読みがなつけ，エラーデータの発見とその修正などは最も複雑な作業である。同語異語の判別は語の意味，文法的性格，形態，表記など言語学的な知識がなければ処理しきれない。また，何がエラーで何がエラーではないかはデータそのものを知っていなければ出来ない。そのデータに，どのような分析がほどこされるのかを知っていなければわからないのである。

3. このような特徴を発揮するためには，配列，計算はもちろんのこと，複雑な人間の作業を容易にするために，いつ，どんなところででも，修正や情報付加が計算機で出来るようなシステムをつくりあげた。その機能は修正システム判別システムなどのサブシステムに生かされた。しかし，最終的には漢字ディスプレイの採用，ＴＳＳの利用，データベースの利用などが実現してはじめて完成するだろう。今回はその第一歩を印したものである。

4. それでは，現段階では手作業で行うのと変わりがない，または，手作業の方が精度が高いのではないかという疑問が起こる。しかし，そう判断するのは早計である。なぜなら，データの配列，情報の付加（修正）は，その方法が決まっていれば，計算機は人間より処理が格段に早く，正確である。実例をあげると，同じ種類のエラーをたちどころに集め（ひらがなの「へ」とカタカナの「ヘ」の違いを集めるなど），その修正を一挙に済ませることができる。

5. 電子計算機を使うもうひとつの利点は，一度作ったファイルを何度も，加工しなおして使えるということである。

コピー機械の発達で，同じ書類を何枚もコピーするのはそれ程苦にはならなくなった。しかし，書類のある部分だけ取り出し，その順序を変える

などの機能は電子計算機でなくてはできない。たとえば，教科書の調査では，M単位とW単位の二つの調査単位を同時に調査している。これが可能になるのは今言った機能があるからである。すなわち，

W国立
M国語
M研究
M所　　先頭の文字MWは単位の区切りを示す。

のデータにおいて，M単位のデータを作るにはこれをすべてコピーし，単位をMにする。すなわち，

M国立
M国語
M研究
M所

とすればよい。また，W単位のデータを作るには文字データだけをつなげてコピーし，その全体をW単位とすればよい。

W国立国語研究所

これは，コピー機能と，任意の部分を取り出す機能をかねそなえていてはじめて可能になる作業である。

6. この機能は，いったんデータを作ればあらゆる観点から同じ規模の調査が何回も可能になることを示している。もちろん，新しい分析にはその観点から見て有用な情報がデータ内になくてはならない。言語データはいわば多面的な情報をもったデータであるため，いろいろ分析が可能である。しかし，データはこれに耐え得る質を維持していなくてはならないことは当然である。

7. 電子計算機を使った語彙調査は以上の理由により，大量を処理するだけ，短時間で処理するだけで満足していたのでは何にもならない。質のよいデータを作ること，利用価値の高いデータを作ることが最も重要である。

8. われわれが作った語彙調査データは，質のよいものである。電子計算機

の特長を十分に生かすためにも語彙調査システムは，多面的な分析が得られるものでなくてはならない。したがって，われわれのシステムではただ語彙表を作るだけではなく，各種の分析ができるように設計されている。各種の分析とは16ページに示すとおりである。

9. いろいろな分析が可能なデータとはどのようなものであろう。あらかじめ分析項目が決っていれば，それ用の最も有効なデータを作ることができる。たとえば語彙調査のために設計されたサンプリングなどはその典型であろう。大量の語彙を推定するに少量のデータですませるのである。では，あらかじめ分析項目が決っていないものについてデータはどのような形がよいか。答えは，原データをできるだけそのまま入力することである。この理由により，サンプリング調査は行わず全数調査にした。サンプリング調査では出来なくて，今回の調査で可能になった分析は，語彙調査におけるサンプリングの研究，文章の分析である。逆に，こうしたことによって生じた欠点はサンプリングに比べ，述べる範囲が小さいことである。これが良いか悪いかはこの調査の分析結果による。

3-2. プログラムシステムの設計思想

大量の語彙調査を計算機で行った例は，昭和41年の新聞三紙の調査がある。われわれはこの調査のプログラムシステムを基礎とし，これを上回る性能を持たせることを目標とした。

1. 新聞の語彙調査のプログラムシステムの特徴は，αとβの二種類の単位の調査を，はじめαを，次にαの結果を使ってβに切りβの調査を，というように直列に結びついて調査を行ったことである。

   教科書では，MとWの二単位を調査したのは同じだが，二つを同時に調査した（前部分を共通に，のち並列的に処理した）。

2. 新聞での修正はプレエディット（入力データの校正検査）が中心で，のち集計段階で少量の修正を行った。

   これに対し，教科書ではポストエディットが中心であった。つまり，集計後は検査校正するシステム，具体的には KWIC を使っての，および語

彙表をいったん出力してからの検査校正に力を入れ，入力データの検査校正は簡単に済ませた。この方法の方がエラーデータの発見効率がよいと考えたからである(参考文献参照)。エラーは分析してはじめて発見できるものだという考えによる。しかし，この方法は前の調査では不可能であった，すなわち，この方法は漢字の大量の出力が不可欠だが，今回の調査では高速漢字プリンタを使用することによってこれが可能になったのである。

3. 新聞では，表記別調査であって，完全な同語異語判別は行っていない。ただし活用語はその表記内で終止形にまとめられている。

教科書ではまず仮名表記をつけることにより，異表記語をまとめ，その後，仮名表記で同語になった語の判別を行った。これによって，同語異語の判別は完全になった。

4. 修正の方法は紙テープによった。これは新聞も教科書も同じである。
5. 新聞では語種，品詞の情報を付加した。教科書ではこれらの情報はつけていない。

3-3. 電子計算機を使った語彙調査の特徴を生かすためにはプログラムシステムはどうであらねばならないか。

1. まず単純作業(語彙調査では配列，計算，フォーマットをかえての出力)は計算機で行う。いろいろな段階で自由に使えるようにする。
2. エラーを発見するために， KWIC や語彙表などのある意味での分析プログラムを早い段階で通す。こうするとエラーを見つけやすく，かつ直しやすい。
3. 修正の方法を簡単にする。また，どんな所でも修正できるようにする。
4. 一度作ったデータを何度も使えるようにする。MとWの作業はできるだけ共通にして，同じことを二つの単位で二度行うというようなことはしない。Mで行った作業はそれがそのままWに生かされるようにする。
5. いろいろな分析に耐える質のよいデータをつくる。これは修正機能を充実させることで実現する。また，原データ，入力データはすべてマスターファイルで生かす。このようにすれば，マスターファイルを使うことによって，語彙調査だけではなく他の分析にも使うことができる。マスターフ

ァイルをデータベースと考える。

6. プログラム管理，ジョブ管理，データ管理の機能をつける。具体的には次のとおりである。

システムチャートを全員に配布し，各プログラムの関連を示した。各プログラムの仕様を同様に内部資料「季報」に報告した。各プログラムの詳細仕様書（プログラムの概要，ファイル定義書，ファイルレイアウト，フローチャート，ランブック）とプログラム・リストを完備した。これはプロジェクトメンバーの各人がプログラムの関連と内容を知ることにより，各分担の位置，作業の進行状況，重要度を知るとともに，後の分析に資するためである。これをプログラム管理とよぶ。

プログラムのランを行う際，オペレータの名前，データの名前を入力し，ラン終了時には，入力した名前とともに，処理量の項目別集計値を出力した。また，磁気テープに貼るラベルの型式を統一し，磁気テープ内に記録されるファイルＩＤのチェックを厳密に行った。これは，他教科ファイルの混入や別のオペレータが同じジョブを二回行うことを防ぐ，磁気テープボリュームの脱落や重複を避けることをねらいとする。このようなエラーが起こった場合，最終出力後，その分析後でなければエラーがみつからないのが常である。少なくとも機械的なエラーは起こらないからである。このようなまちがいは起こすはずがないと考えるのはあやまりであって，大量調査においてはこのようなまちがいは必ず起こると考えるべきである。これらをジョブ管理，データ管理とよぶ。

3-4. このプログラムシステムでは実現したくても種々の事情で実現できなかったこと。

1. 修正システムの効率化

現在は紙テープベースである。データを紙テープで作り，これとマッチングしたものを正しいデータに置き換える。この方法ではキーとなる番号，またはデータを新たにパンチしなければならないが，それにエラーが起こる可能性があり，このための検査校正が必要になる。能率的ではない。

データを直接見ることができ，かつ，その場で修正し，修正した結果を確認できる方法があればよい。これは，磁気テープなり磁気ディスクなりを直

接漢字ディスプレイに表示することで実現できる。最近各社で漢字ディスプレイが開発され，このような修正方法も十分に実現可能である。教科書システムのシステム設計当時ではこの方法は考えられなかった。

中間段階としてフロッピーディスクによる修正がある。しかし，これは中間にしかすぎない。データそのものを直すのではなく，新しくマッチングするためのキーを作らなくて済むこと，及び，データの中の誤り箇所だけを，修正すればよいことなどの利点があるが，マッチング修正作業はやはり必要である。

これに対し，磁気ディスクの中のデータ（データベース）を，直接，エラーの箇所だけ修正し，その確認がとれるシステムは最も良い。もちろん，ＴＳＳ制御のもとに動かす。修正作業が人間の速度に落ち，ＴＴＳを使わなければ計算機に無駄が生じるためである。TSS データベースを直接修正するシステムはデータベースから派生する各種の分析データについても一挙に修正することになり効率的である。

上に述べた修正システムについての比較を次にかかげる。

（例）修正システムの比較

紙テープベース	フロッピーベース	データベース
マスターファイル	マスターファイル	マスターファイル
印字	印字	印字
検査校正	検査校正	検査校正
修正データの作成	データの抜き出し	ディスプレイ修正
｛キー	校正データにより修正	
｛エラーを含むデータ	マッチング修正	
｛修正指示		
修正データのパンチ←		
修正データの印字		
修正データの検査校正		
エラーがあれば——		
マッチング修正・確認		

(例) 校正のしかたの比較

	エラーデータ	修正データ		修正結果
紙テープベース	0010 デーた	0010 データ	(置き換え) →0010	データ
フロッピーベース	0010 デーた	た→タ	(マッチング) 0010	データ
データベース	0010 デーた	た→タ	0010	データ

2. 自動処理

教科書システムの設計当時，できるだけ自動化するのか，または，人間作業の補助機械として機械を使うのかという方針決定に論議が分かれた。結局，われわれは自動化を採用しなかった。当時，われわれはいくらかの自動化プログラムを持っていた。自動単位切り，自動読みがなつけ（辞書方式とプログラム方式），自動品詞認定などがそれである。

それらのプログラムを採用しなかった理由は以下のとおりである。

(1) 自動処理の精度が80%台である。

(2) 有効な修正方法が開発されていない。(1)の結果，相当量の修正が予想される。

(3) 新聞の語彙調査が機械処理中心であったのに対し，教科書は手作業中心でデータの質を高めようとした。これら二つの調査により機械を使った調査の最適システムを探ろうとした。

(4) 大量のデータを自動処理した経験がなかった。したがって，起こりえるエラーの種類がつかめなかった。

3. データベースシステム

筆者（中野）はデータベースの意義を現在のようにはっきり意識はしていなかったものの，大量のデータをかかえ，その有効な利用法を探る過程で，データベース的な方法を思考していた。

部内資料ＬＤＰ11 (1973.3) に表わした中野洋「言語データのデータベースについて」は「データベース」という言葉を用いた最初である。

計算機を使うひとつの利点は，先に述べたように各種の分析研究に語彙調査データを共通に用いることが出来るという点にある。しかし，現状ではマスターファイルから派生する各種の加工データの氾濫に呆然自失するのがおち

である。また，各種の分析をするということは別の面からデータを眺めるということであり，それだけ新しい検査をし，新しいエラーが見つかる。しかし，われわれは，そのエラーをマスターファイルにもどって修正するということは，現実には行っていない。なぜなら，ただ，マスターファイルを直せばいいのではなく，そこから生じた各種の分析用データすべてを修正しなければならない。現実問題としてこれは可能であろうか。答えはノーである。たとえば，新聞の語彙調査の最終段階に調査に用いた磁気テープを数えたら，2000巻を越えていた。数本のマスターテープのエラーの修正から，少なく見積って，1000巻のデータすべてを修正することができるか。コピーするだけでも数十時間はかかるであろう。これは現実には不可能といえる。

データベースの考え方は，マスターファイルの管理とそこから発生するファイルの管理である。データベースを直せば自動的にすべてのファイルを直していることになるようなシステムである。

すべての分析は，データベースを使うことによって行われる。もちろん，なんらかの加工は行われるが，すべてがデータベースを通じて行われるようにする。このようなシステムでは処理時間が膨大になるが，これは計算機自身の処理スピードの向上とオペレートの自動化によって苦にはならない。

データベースを実現するためには，計算機の内部メモリーを大きくすること，処理速度を早くすること，オペレーティングシステムを充実すること，ＴＳＳ制御を導入すること，有用なデータベース用ソフトを作ること，端末を多くすることが必要である。しかし，これらはほとんどが，ハード上の問題であり，技術的には解決されていると考えられる。それより，データの作成，加工の方が人も金もくうのである。

しかしながら，このシステムは教科書調査では採用されなかった。機械も金も理解者もなかったからである。

4. 高校教科書語彙調査システムの概略

4-1 システム概略

システムの詳細は語彙表とともに詳しく報告される。ここではその考え方と概略について述べる。

システム設計の土台となったのは，当時 HITAC-3010を使い，KWIC によるエラーの発見，修正を行っていた索引作成システムであった。このシステムによって調査を行った場合，起こるであろうエラーとその原因を洗い出したのが表1「教科書調査システム（案）におけるエラーの種類と原因」である。これは研究第1部第二研究室「語彙調査に生ずる狂いの種類・原因・対策」（年報5）の方法によったものである。この結果，管理システムの確立とともに，発見が困難なケアレスミスを防ぐための士気の高揚の必要が痛感せられた。以下に表1から得られたエラー対策を列挙する。

a．チェックシステムの確立。

b．オペレートのジョブ・ランブックを充実する。記録をとる。

c．作業の整理表・うけわたし表を充実する。

d．分担者の連絡を緊密にとる。

e．作業者の負担を軽減する。……一度に二種類以上の仕事をさせない。

f．士気の高揚をはかる。……アルバイタに至るすべての人々が，この調査の

〔表1〕

教科書調査システム(案)におけるエラーの種類と原因（ミニクイックまで）
「語彙調査に生ずる狂いの種類・原因・対策」年報5の方法による
原因　1…作業者の不注意　2…管理不十分　3…システム　4…作業の指導　5…その他

作業段階	番号	狂いの種類	発見の困難さ	発生頻度	原因	対策
基本方針の設定	010	妥当性・信頼性・客観性・適応性・再現性を満足していない	絶望的		立案者の無能	
テキストの決定・購入	020	調査目的に合った資料でない	〃		〃	
入力対象とする部分の指示	031	不適当な対象を指示した。	困難		企画時	
	032	計画通りの指示をしなかった。	容易		①，4	検査
	033	計画通りの指示をしたのに，対象としなかった。	〃		①，2	検査
	034	計画通りの指示をしたのに，対象外のものを入れた。	〃		①，2	検査
単位切り	041	単位切りミス（三種類）	やや困難	○	④，1	検査，ミニクイック
単位切り検査・修正	051	検査もれ（数ページ全部・検査のこし）	困難	○	①	整理表チェック

作業段階	番号	狂いの種類	発見の困難さ	発生頻度	原因	対策
	052	修正もれ	〃	×	①, 3	士気の高揚
	053	修正エラー	〃	△	①, ④	規則の整備・徹底・教育・校正
清書	0611	清書ミス	容易	○	①, 4	校正
	0612	清書の重複	〃	△	①	校正
	0613	清書の脱落	〃	△	①	校正・整理表のチェック
よみがな	0621	よみがなをつけるべきなのに，つけなかった	容易	△	①, 4	検査
	0622	よみがなをつけなくてよいのにつけた	困難	○	①, 4	代表形と まぎらう 検査
	0623	ひらがなでいれるべきなのに，カタカナにした	容易	×	4	検査 Pro. チェック
	0624	よみがな記号〔 〕を間違って，( )や〈 〉にした。	やや困難	△	1, 4	検査
	0625	よみがなのつけまちがい	困難	○	④, 1	検査
代表形	0631	代表形をつけるべきなのに，つけなかった	やや困難	○	1, 4	検査
	0632	代表形をつけなくてよいのに，つけた	やや困難	△	1, 4	検査
	0633	代表形をカタカナで入れた	容易	×	1, 4	検査 Pro. チェック
	0634	代表形のつけまちがい	困難	○	1, 4	検査 Pro. チェック ミニクイック
ふりがな情報	0641	ふりがな情報のつけ落し	やや困難	○	1	検査
	0642	ふりがな情報はいらないのに，つけた	〃	△	1	検査
単位情報	0711	単位情報をつけ誤った	〃	○	1, 4	検査
	0712	単位情報をつけなかった	容易	△	1	検査 Pro. チェック
	0713	単位情報の重複	やや困難	×	1	検査
出典情報	0721	本の名前を入れなかった	容易	○	1	Pro. チェック
	0722	本の名前を入れまちがった	容易（注意すれば	△	1	Pro. チェック 検査
	0731	ページ情報を入れなかった	容易(〃)	△	1	Pro. チェック 検査
	0732	ページ情報を入れまちがった	容易(〃)	○	1	Pro. チェック 検査
	0741	行情報を入れなかった	容易(〃)	○	1	〃
	0742	行情報を入れまちがった	〃	○	1	〃
	0751	段落情報を入れなかった	〃	○	1	検査 Pro. チェック不可能
	0752	段落情報を入れまちがった	〃	△	1	検査 〃
	0761	見出し情報を入れなかった	〃	○	1	〃
	0762	見出し情報を入れまちがった	〃	△	1	〃

作業段階	番号	狂いの種類	発見の困難さ	発生頻度	原因	対策
校正	0811	校正もれ	困難	△	1，2	整理表，士気の高揚
	0812	校正をまちがった	困難	×	1，2，4	ミニクイック
	0821	検査もれ	困難	△	1，2	整理表，士気の高揚
	0822	修正もれ	困難	×	1，3	士気の高揚
	0823	修正エラー	困難	×	1，4	〃　クイック
入力パンチ	091	パンチミス	やや困難	○	1，3	ミニクイック
	092	ファンクションコードの打鍵もれ	〃	△	1，3	Pro. チェック
	093	漢テレの機械エラー	〃	○	整備	Pro. チェック，クイック
	094	漢テレの切れが悪い	〃	◎	〃	〃
原文よみこみ	101	P/T の粉失	容易（注意すれば	×	2	うけわたし表
	102	P/T よみこみの脱落	〃	△	オペ	Pro. チェック
	103	〃　重複	〃	△	オペ	〃
	104	エリアオーバー	容易	×	ページおとしプログラム	ページを入れる。（1ページ1巻の場合簡単）
修正データの作成	121	エラーリストの粉失	困難	○	2	エラーリストのページングうけわたし表
	122	修正ミス	困難	△	1，④	クイック
	123	修正もれ	困難	△	1，④	士気の高揚
	124	語番号の書きまちがい	困難	○	1	〃
修正データのパンチ	125	パンチミス	やや困難	○	1，3	印字，校正
	126	ファンクションコードの打鍵もれ	〃	△	1，3	〃
	127	漢テレの機械エラー	〃	○	整備	〃
	128	漢テレの切れが悪い	〃	◎	〃	〃
修正	131	修正データの粉失	困難	○	2，3，1	
	132	修正データの重複	困難	△	オペ，2	Pro. チェック
	133	修正データの脱落	困難	△	オペ2，2	対策なし
ソート・マージ	161	リールのかけまちがい	困難	△	オペ，2	JOBランブックの充実　士気の高揚
	162	マスター・テープの破損	容易	×	オペ，2	コピー，前段階のファイル保存
	163	ファィルのかけまちがい	容易	△	オペ，2	ラベルチェックの重複　JOBランブックの充実

作業段階	番号	狂いの種類	発見の困難さ	発生頻度	原因	対策
校正（ミニクイック）	171	校正もれ	困難	○	1，4，2	士気の高揚
	172	校正エラー	困難	○	4，1	
検査（ミニクイック）	173	検査もれ	困難	○	4，1	
	174	修正エラー	困難	○	4，1	規則の整備，検討会の開催
	175	修正もれ	困難	○	1	士気の高揚
エラー語番号の書き出し	181	エラー語番号の書きまちがい	困難	○	1	〃
	191	パンチ・ミス	やや困難	○	1	印字・校正
	192	機械エラー	〃	△	整備	
校正（原文）	201	校正もれ	困難	○	1	士気の高揚
	202	校正エラー	〃	○	1	〃
エラー語番号の書き出し	211	エラー語番号の取り出しミス	〃	○	1，3	
	212	〃　書きまちがい	〃	○	1	
パンチ	221	パンチ・ミス	やや困難	○	1	印字・校正
	222	機械エラー	〃	△	整備	

◉ かなり多い　○ 多い　△ 少しある　× ほとんどない
原因の欄で○囲いの数字はそれが主要な原因であることを示す
Pro. はプログラムの略。

意義を確認し，調査全体のどの部分を担っているかを知ることが必要。

g．オペレータは入力データ・出力結果の内容を知っていなければならない。

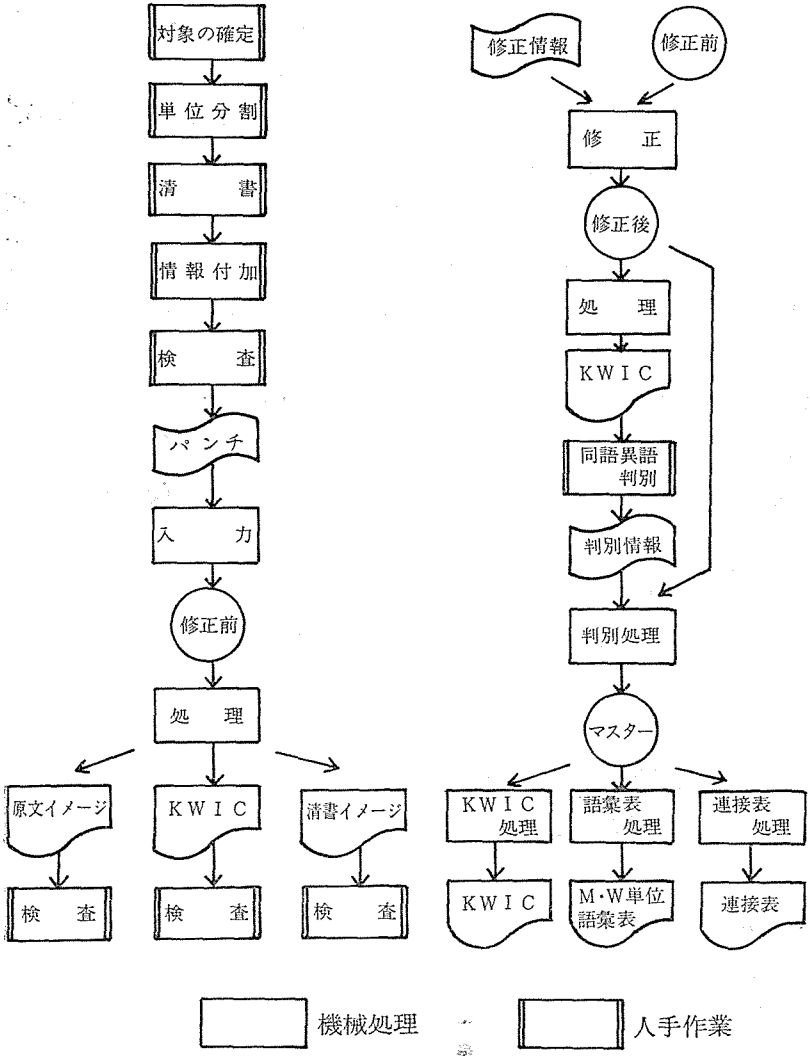
h．作業者の教育

このような検討のもとに，データ管理オペレート管理システムを含んだシステムが作られた。図1は現在おこなわれている調査システムの概要である。このシステムの特徴は検査・校正・修正が処理後であること，同語異語の判別作業が入っていることである。機械処理は複数のプログラムによって動いている。その総数は73本である。

4-2　入力情報と各種の出力結果およびその分析計画

どのような結果を出力するかは，何を入力するかによって決定される。今回の調査は用語用字調査ということばで示されるように，語彙表，漢字表の出力

図 1　語彙調査システムフロー



をメインにおき，その他各種の調査が行えるように設計した。

その結果，入力データは次のように決った。

原文　　ただし，脚注，図，表およびその注記等は入力しない。

単位切り情報　W単位とM単位の二種類を採用。本報告鷲岡論文参照。

よみがな　　主に，結果を50音順に並べるために用いる。

代表形　　異形同語をまとめるために用いる。

判別情報　　同じ代表形を持つ異語（動詞の「ある」と連体詞の「ある」など）を分離するために用いる。

助辞情報　　助辞であることを示す。

ルビ情報　　教科書原文にルビがついていることを示す。

見出し語情報　タイトルに用いられた語であることを示す。

出典情報　　教科書名，ページ情報，段落情報

このデータにより，次の表が出力される。例を17・18ページに示す。

M単位／W単位 ┬ 度数順語彙表 ┬ 全体表（記号類，助辞類を含む）（全体）
　　　　　　　└ 50音順語彙表 └ 部分表（語彙分類別）　　（教科別）

連接表

用例表（KWIC）

その他，以下の分析が計画されていた(73. 12. 19野村雅昭まとめによる)。

教科書基本語彙の抽出

教科別特徴語彙の抽出

対象記述用語と方法記述用語の分析

文章における語彙構造の分析

述語文の構造の分析

語構成の分析

基礎概念語と応用概念語の分析

語基概念表示性による量的把握

説明文的特徴語基の抽出

説明文的文末表現の抽出

これらの計画は，それ以後の人事移動，入力方式の変更等のために実現でき

20／1全教科　度数順語彙表

見出し	J	判	度数	活	(GA) 比率	累比	順位	(GB) 比率	累比	順位	(GC) 比率	累比	順位	(GD) 比率	累比	順位	(GE) 比率	累比	順位
の	J	の	1844	1	5.866'	5.866'	1.0	6.618	6.618	1.0							22.825'	20.824	4.0
,		,	1791	1	5.697'	11.562	2.0										50.084'	50.084'	1.0
に	J	に	1262	1	4.014	15.577'	3.0	4.529	11.148'	2.0							15.621'	36.445	5.0
を	J	を	994	1	3.162'	18.738	4.0	3.568'	14.715	3.0							12.304'	62.553'	7.0
は	J	は	906	1	2.882'	21.620	5.0	3.252'	17.967	4.0							11.214	73.767	8.0
.		.	870	1	2.767	24.388'	6.0										24.329'	74.413'	2.0
て		て	853	2	2.713	27.101'	7.0	3.062'	21.029'	5.0	4.312'	4.312'	1.0	4.522	4.522	1.0	4.522	78.264'	22.0
する		し	828	7	2.634'	29.735'	8.0	2.972'	24.000	6.0	4.195	8.497	2.0	4.390'	8.912'	2.0	4.390'	87.110'	24.0
が	J	が	792	1	2.519	32.254'	9.0	2.843'	26.843'	7.0							9.803	94.657	10.0

20／1全教科　50音順語彙表

見出し	J	判	度数	活	(GA) 比率	累比	順位	(GB) 比率	累比	順位	(GC) 比率	累比	順位	(GD) 比率	累比	順位	(GE) 比率	累比	順位
あいんしゅたいん		ア	1	1	0.003	99.645'	2886.0	0.004'	94.215'	2867.0	0.005	99.447	2833.5	0.005	91.604	2817.5	0.005	99.522	2892.5
アインシュタイン			1																
あう		合	20	8	0.064'	65.468'	163.5	0.072'	61.298'	153.5	0.101	45.872	136.5	0.106	43.349'	126.5	0.106	42.550'	180.5
あい			2																
合う			2																
あう			3																
合っ			3																
合っ			1																
あっ			2																
会わ			1																
合わ			6																

20万の1	全体	普通名 (−)	接辞 (F)	助辞 (J)	数字 (N)	記号 (S)
のべ	31438	18863	0	8079	920	3576
ことなり	3762	3687	0	37	18	20

GA ＝ (−)＋(F)＋(J)＋(N)＋(S)
GB ＝ (−)＋(F)＋(J)＋(N)
GC ＝ (−)＋(F)　　＋(N)
GD ＝ (−)
GE ＝ その語のグループ内

連接表

うんどう　45

	運動	13	代表形
		(13)	あ
	運動　して	4	
	運動　神経単位	1	
	運動　できる	1	
	運動　方程式	4	
		(10)	
イギリス商品ボイコット	運動	1	
インド民族	運動	1	
改革	運動	1	
原水爆禁止	運動	1	
国民精神総動員	運動	1	
署名	運動	1	
等加速度直線	運動	1	あーる
東遊	運動	1	あい
日本留学	運動	1	
年周	運動	2	
反占領	運動	1	
反仏独立	運動	1	
平和	運動	2	
放物	運動	1	
民権	運動	1	あい
民族	運動	3	
民族主義	運動	1	
民族独立	運動	1	
		(22)	

（同語異語判別用ミニKWIC）　　M単位

判	科	通し番号	頁	段落	出現形		
亜	化	00236700	16602	塩化水素HClと次	亜	塩素酸HClOとができている
亜	化	00238900	16603	・97)，▽O式次	亜	塩素酸は弱酸で，次のように電
亜	化	00241100	16604	に達する．▽O式次	亜	塩素酸とその塩とは，酸化剤と
亜	化	00243300	16604	用する．それは，次	亜	塩素酸が他の物質に酸素を与え
亜	化	00247900	16606	▽O式塩素水中の次	亜	塩素酸がなくなれば，はじめの
亜	化	00251300	16606	用して，ふたたび次	亜	塩素酸となる．このようにして
亜	化	00256400	16607	れる．サラシ粉は次	亜	塩素酸イオンを含んでいるので
亜	学	00200800	19101	・光沢（金属光沢・	亜	金属光沢・非金属光沢の区別）
亜	学	00216700	19104	状2）目がら状3）	亜	目がら状4）多片状5）土状6
ア	物	00049500	02705	面積の単位として1	アール	というような独自の単位も
相	政	00260500	08102	品物とが互いに有無	相	通ずる結果になるから，双方に
相	地	00186800	13001	てある．おもな貿易	相	手国は，アメリカ合衆国とEC
相	日	00311200	22301	イギリス船の来航が	あい	つぎ，日本に開国を求めてき
相	日	00231400	15001	についたが，その後	あい	ついて来日した宣教師は，布
相	物	00247200	16504	じ固有振動数を持つ	相	手にはこの力による仕事が有効
合	生	00197320	16202	発生過程をたどる場	合	に，胚胚の各部が何になるかを
合	生	00260020	16902	面が頭端にちかい場	合	には，後方のきり口からも頭を
合	生	00260370	16902	を，尾端にちかい場	合	には，前方のきり口からも尾を
合	政	00205300	07902	人に依存している度	合	は，その国の経済の発達や資源
合	政	00347500	10401	いので安定しない場	合	が多い．けれども，この中小企
合	政	00055950	02101	色は，絶対王政の場	合	と異なり，君主が無制限な支配

ないものもある。実現可能なものでも，そのためにいかに時間と費用をかけるかによるものもある。今後の課題である。当初，教科書システムはこれらが実現できるように計画されたものである。

同語異語の判別については，報告61所収土屋論文に詳しいのでここでは述べない。

5. おわりに

電子計算機を用いた調査は非人間的だと言われてきた。この種の調査を紹介すると，いつもその処理スピード，その処理量に感心されても，処理の精度については，機械のすることだからと大目に見てこられた。電子計算機を本当に調査の道具として使うなら，それではいけない。処理のスピードや量をおとさず精度を上げるにはどうすればよいか，機械の欠点を人間がいかにカバーするか。人間と機械の調和のとれたシステムとはどんなものか。そのようなことを求めて，高校教科書の用語調査システムは設計されたのである。

参 考 文 献

1 斎藤秀紀・靏岡昭夫・中野洋・米田正人.「大量言語処理におけるエラーと対策」情報処理学会・計算言語研究会資料ＣＬ4-1. 1975. 12. 5.

2 斎賀秀夫・土屋信一・靏岡昭夫・野村雅昭・佐竹秀雄・斎藤秀紀・田中卓史.「高校教科書語彙調査の概要」(情報処理学会・計算言語研究会資料ＣＬ10-2. 1977.6. 24

3 中野洋・堀江久美子・米田純子.「高校教科書用語調査におけるエラーデータ」国研内部資料. 季報1977冬

4 中野洋.「言語処理における一貫処理の研究」電子計算機による国語研究Ⅸ.1978.3.

5 言語計量研究部第一研究室.「高校教科書の用語用字調査―データ修正の記録―」国研内部資料季報1979秋