高校教科書の同語異語判別システム

著者	土屋 信一
雑誌名	電子計算機による国語研究
巻	9
ページ	1-16
発行年	1978-03
シリーズ	国立国語研究所報告 ; 61
URL	http://doi.org/10.15084/00001052

高校教科書の同語異語判別システム

土屋信一

1. はじめに

高校教科書用語調査は，電子計算機を使った用語調査の第２回目のものである。第１回目の新聞の用語用字調査とは違った，いくつかの目標を持っている。同語異語判別作業を，全てはできなくとも，一部分のデータについて，行ってみようというのも，この目標のひとつである。このように考えたのは，前回の新聞調査が，まず機械（電子計算機）を使いこなすこと，すなわち，言語データを計算機に扱わせることが，大きな目標であったのに対して，その目標は，多くの方々の努力によって，すでに達せられており，次は，調査内容の向上にあると考えたためである。調査の質の向上は，対象の選択・目標の定め方・分析方法の検討等，さまざまなところで十分に吟味をし，実行に移してゆくことであるが，機械（電子計算機）と言語データとの関係する部分では，正確なデータを作ることであり，そのためには，誤データの修正と情報の付加作業を正確・迅速に行うことである。同語異語判別作業は，この部分における主要な作業段階であり，この作業なしでは語彙調査の性格まで変わってしまうほど重要なものであるが，前回の新聞調査では，作業の困難性のゆえに，調査システムに組み入れられなかったものである。今回の調査でも，機械装置が進歩したとは言え，まだまだ困難な作業である。しかし，幸いというか，今回の教科書調査は，新聞調査に比べて語彙量が少なく，また，用語・文体・表記なども比較的まとまりのあるものなので，実現は比較的容易である。そこで，電子計算機を使った同語異語判別システムの作成を試み，この方面での調査の質の向上に役立てたいと考えた。

なお，この同語異語判別システムは，教科書調査の実行グループの中で，幾

度か意見を交換し，検討した結果生まれたものであり，土屋はそれらをまとめて報告するという立場に立っていることを付記する。

2. これまでの語彙調査と同語異語判別作業

ここに同語異語判別作業と呼ぶものは，次のように規定される。

> 語彙調査において，調査対象となった語集団の個々の語を類別し，同一語ごとに一まとめにする作業。具体的には，集計単位を定め，同語とする語の範囲を規定してから，異形態・異表記の中で同語のものを統合し，同形態・同表記の語の中で異語のものを分離する作業。この作業は，調査対象を，表記形や語形のレベルでなく，語のレベルでとらえるものである。

次に，これまで国語研究所で進めてきた語彙調査において，同語異語判別作業がどのように位置づけられ，進められてきたを概観する。

カードを使った手集計の語彙調査では，この作業は，カードを採る段階で，作業者の頭の中で大部分が行われてしまい，大仕事だとは意識されていない。そのため，国立国語研究所資料集2「語彙調査—現代新聞用語の一例—」(1952) にも，国立国語研究所報告4「婦人雑誌の用語」(1953) にも，ほとんど記述はない。わずかに，見出しの掲げ方として，前者の22ページ「見出しの書き方は，次のとおりである。」以下に少々，後者の54ページ「3.21見出しの掲げ方」に少々記されているのみである。

しかし，総合雑誌・雑誌九十種調査では，この作業は明確に規定されるようになった。すなわち，「調査単位」のほかに「集計単位」を定め，同語の範囲を厳密に示している。

総合雑誌…国立国語研究所報告13・19ページ～24ページ「2.4集計単位の定め方」参照。

雑誌九十種…国立国語研究所報告21・14ページ～20ページ「3.3集計単位の定め方」参照。

(このほかに，見出しの掲げ方について記したところがあるが，省略する。)

さらに，この二つの語彙調査には，次の研究論文がある。

総合雑誌…国立国語研究所報告13・108ページ～115ページ

「同じ語か異なる語かの線型判別函数による決定」

雑誌九十種…国立国語研究所報告25・294ページ～330ページ

「同じ語か異なる語かの判別」

このうち，判別函数については，その概略が，見坊豪紀「辞書をつくる」(1976年）の「辞書の姿」64ページ～65ページに紹介されている。これは「言語生活」114号昭和36年3月号の「辞書の姿」の再掲載であるが，要を得ているので，ここに引用する。(原文はたて書き)

国立国語研究所の書きことば研究室では，現在（昭和36年）調査中の四三万枚にのぼる採集カード（付属語をふくまない）を五十音順に配列したさい，一枚一枚のカードを同語か別語かという観点から分けてきている。このために，同語別語の判定規準を作ることが要求され，経験的に次のような公式を考えた。

$$Z=0.066+x_1+0.555\,x_2+0.423\,x_3+0.199\,x_4$$

x_1 から x_4 までは，判定のために使う四種類の規準（後述）の得点を示す。0.066などの値は経験的に得られた常数である。判定者（複数）の与えた得点を項目ごとに集計して計算した結果，Zが0より小さければ問題の二つの語は別なことばである。0または0より大きいときは同じことばである。

この判別式に利用した規準と各規準への配点は次のとおりである。

(1) **触れ合い** (i)意味し方の共通観点の有無，(ii)指された概念・事物のカテゴリーの異同。(＋2から－2まで)

(2) **漢字表記** その意味を表わすのに使っている漢字が同じかどうか。(＋2から－2まで)

(3) **アクセント**の型が同じかどうか。(＋2から－2まで)

(4) それぞれの**類語・対語**(ついご)が同じかどうか。(＋1から－1まで)

この公式は，意味の重なり方の判定をもっとも重視し，漢字とアクセントをそれぞれ約半分の重さと評定し，類語・対語の事情を多少考慮した，という構成である。この重みづけも規準の取り上げ方も，当時の所員たちの判定を土台にして，経験的に作りあげたものであるから，広く深い学識

をそなえた天才的個人の出現をこばむものではない。

国立国語研究所報告13には，上記の事がらの詳細な説明があり，この判別式を作るため，問題になる20の語形に対して５人の所員が11の規準から点数を与え，式を作り，判別の結果と比べたところ，上記の４規準を使ったものが成功率が94パーセントで「語彙調査の実施上実用になる」こと，さらにその後有効と思われる規準を追加して新たな判別函数を作ったが，成功率はかえって低くなったこと，このような「客観的操作が出来なければ，すなわち同じ語か否かを主観だけによって決めるのでは，語彙調査の基礎はくずれてしまう」ことなどが，述べられている。

この判別函数は，国立国語研究所報告 25・301 ページ「判別の実例一覧表・前書き」によると，雑誌九十種調査でも使われたようだが，高校教科書調査でもこのまま使ってよいかは，疑問である。それは調査単位が異なり，対象が異なるため，当然，定数も違ってくるだろうと思われるからである。この判別函数を作るために使った20の語形について，現所員数人に対して，単位のみ高校教科書調査の調査単位（同語異語判別作業のためにはM単位のみ）で調査し，新たな判別函数を作るとか，それに先立ちβ単位で切ったものを現所員が判別してかつての判別函数と一致するものかどうか確認するとか，あるいは，この判別函数作成に当った所員に，M単位で切ったものを判別してもらって函数を作るとか，あるいは，全く別に幾語かを選んで函数を作ってみるとか，いろいろなことが考えられる。ただ，楽観的な見方をするならば，判別函数作成のために調べた20語のうち19語はM単位と同じであり，異なるのは「シラナミ」（白い波の意と盗賊の意）が「シラ」と「ナミ」に切ったに過ぎないので，この判別函数をそのまま使ってもそれほど危険はないと言うことができる。なお，調査語等については，国語研究所図書館蔵の「総合雑誌の調査（プリント集）」によって知ることができる。

判別函数についての紹介が長くなってしまったが，要するに，総合雑誌・雑誌九十種調査は，同語異語判別作業に非常な努力をはらって，厳密に遂行した。それは，おそらく，水谷静夫氏によって計量語彙論の立場から語彙調査が定義づけられ，その中で，単位切り（単位語認定）と並んで，同語異語判別

（見出し語認定）が語彙調査の基盤として明確に位置づけられたためと思われる。

これについては，次の論文が発表されている。

○国語研報告13（1958）94ページ～96ページ「語彙調査の成立根拠と基本諸概念の定義」

○「国語学」40輯（1960）水谷静夫「計量語彙論の基礎づけ」

○「国語学」62集（1965）水谷静夫「語彙論の術語をめぐって」

また，岩波講座「日本語」9（1977）の水谷静夫「語彙の量的構造」の第一章「計量語彙論の前提」にも大筋が述べられている。

3. 計算機を使った語彙調査における同語異語判別作業

国語研究所では，手集計による語彙調査は，雑誌九十種調査で終わり，昭和41年の新聞を対象とした，新聞語彙調査からは，電子計算機を利用した語彙調査が始まった。この新聞語彙調査では，同語異語判別作業はシステムの中に組み込まれず，すなわち，調査単位で切ったままの表記形を集計して長単位表を作成し，さらに長単位を切って読み仮名を付し短単位表を作成するという方式を採った。

この点について，国立国語研究所報告37「電子計算機による新聞の語彙調査」では，次のように述べられている。

○ 今回の用語調査は，電子計算機を用いて行なう第一回のものであるので，技術的に必ずしもすべての見通しがつけられていたわけではない。解決できなかった問題として，大きなものは漢字の読み（「通った」のカヨッタ，トオッタなど）も含めて，同語異語の処理のそれがある。すなわち同形異語の判別，異形同語の処理がなされていないのである。ただ，漢テレを用いて漢字をそのまま入力しているので，かなで書くときの同音語がそのまま全部区別されていないわけではない。かなり救われている。その反面，かな表記と漢字表記の語の identification ができなくなっている。いずれにせよ，同語異語の処理はできなかったのである。この報告書の語彙表は，すなわち同表記同形語の使用度数表であって，従来国語研究所で

作成してきたようないわゆる語彙表とは，やや性格を異にするものである。　（同書3ページ　2.1.　企画条件と調査内容）

○　調査単位のまとめ方。

この調査ではいわゆる同語異語の判別を行なっていない。得られた単位は同表記同形の語について度数をカウントした表である。表記形で同じである「いき」（「粋」の意）と「いき」（「行き」の意）とは区別されず同じ語とカウントされ，表記形が異なる「いき」（「行き」の意）と「行き」とは別な語として別にカウントされ，整理された度数表である。すなわち異なった語でも表記が同じであれば区別されず，同じ語でも表記や語形が異なれば別の語として処理されている。しかしすべての語についてこのようなことがあるのではないから，この表はそのようなことを考慮に入れた上で使用すれば十分役に立つと思う。

（同書12ページ　3.3.　調査単位と度数の数え方）

このように，手作業による語彙調査と比べると，調査単位レベルの集計しか行われていない。集計単位レベルの集計を行わなかったかわりに，同表記形の集計は文字列の集計として役立つと考えて実施したと考えるべきであろう。同表記形の集計が同語異語判別の済んだ集計単位の集計の代用とはなり得ないことは調査当時すでに自明のことであったからである。

なぜ新聞語彙調査で同語異語判別作業がなされなかったか。それは「大量・迅速」・「人手がかからない」を特色としてスタートした，計算機による語彙調査の第一回目で，その特色が失われてしまうためであると思われる。機械処理に入る前の人手の作業（プレ・エディット）の段階で，単位切りや清書などの作業とともに同語異語判別作業をしてしまうことは，不可能に近い。正確な見出し語（集計単位の表記形）を付するためには，全調査単位を見渡さなければならず，そのためにはカード等を使って語彙調査をほぼやってしまうことになる。その結果を入力したのでは，語彙調査に計算機を使う意味がほとんどない。

また，機械処理後の人手の作業（ポスト・エディット）の段階で同語異語判別作業を行うことも，新聞語彙調査の時点では，次のような技術上の難点があ

った。

1. 同語異語判別をするためには，その調査単位を取りまく文脈が必要であるが，そのためにデータを入れた磁気テープが膨大な量になり，扱いきれなくなる。新聞データは簡易五十音順長単位ファイルで，当時21巻あった。その中に収められている約200万長単位語に，それぞれ数十字分の文脈をつけると，データ・テープは50巻を越す。修正・ソート・マージ等の作業の段階では，その何倍かのテープを使用するから，機械処理の時間・人手も，テープの保管場所も，われわれの態勢では実現不可能となる。(また，文脈を添えず，出典情報のみとすれば磁気テープの増量は防げるが，作業が，いわゆる人海作戦となるため，膨大な人件費が必要となる。これは技術上の難点ではないが，実現不可能なことには変わりない。)

2. 大量の文脈を印字する方法がなかった。調査当時，漢字プリンタはなく，片仮名・アルファベット・数字等を打ち出すラインプリンタと，1分間120字を印字する漢テレしかなかった。前者では同語異語判別用の文脈の印字には不十分であり，後者では時間がかかり過ぎる。石綿敏雄氏は前者の方法に漢字を数字・アルファベット・記号で表わす手順を加え，三分の一のデータについてKWIC(文脈付き用例表)を作成し，土屋も後者の方法で，約2万語の長単位仮名書き語を文脈付きで印字したが，いずれも大仕事であった。

この二点の機械処理上の制約のほかに，同語異語判別作業自体が，大量の人件費と期間を要するという難点があり，ポスト・エディットの段階での同語異語判別作業も行われなかった。

以上，新聞語彙調査で，同語異語判別作業が行われなかった事情について述べたが，石綿敏雄氏「電子計算による用語調査と同語異語の処理」(国立国語研究所報告49「電子計算機による国語研究V」1～21ページ)によれば，計算機を使った語彙調査では，同じような問題が英語・ドイツ語でもあり，未解決のままになっているという。

4. 教科書調査における同語異語判別作業

上に挙げた石綿氏の論文では，「現段階での解決案」として，中間段階でカ

ードで出力し，同語異語判別作業をしたあとで，再入力する方式が提案され，「将来の解決案」として自動同語異語判別の可能性が述べられている。今回高校教科書調査で採った方式は，まだ未確定の部分はあるが，石綿氏の「現段階での解決案」に近く，それよりも人力に頼る部分を多く含むシステムである。また，「将来の解決案」である，同語異語判別の自動化ということも，この調査の語彙表作成のためには全く考えていない。これは，データ自体にミスがあり，それを修正する作業でまた新たなミスを生ずるであろうと考えてデータ修正システムを整備したので，それと同様な方式で同語異語判別作業も進めていこうと考えたためである。同語異語判別作業も，やはりミスを重ねながら，何度か修正してゆくものである。また，データのミスは全て修正されること，同語異語の判別された結果も正しく各単位ごとに付けられること，そしてそれらデータが磁気テープまたは磁気ディスクに納められることを目標とした。

語彙表を作成するだけなら，ポスト・エディットで人力を投入すればできる。新聞語彙調査の時点と異なり，KWIC システムは完成しており，高速漢字プリンタも実用化したので，先に述べたポスト・エディット方式の難点はほぼ克服された。しかし，語彙表を作っただけでは語彙調査は終わらない。つぎにさまざまの分析・記述が行われねばならず，そのためには電子計算機を利用することが絶対的に優位である。手作業で作られた語彙表とそれを使ったカード利用の分析・記述では，計量的な分析は簡単なものしか行いがたく，結局，浅い分析・記述内容になってしまうであろう。また，同語異語判別作業の自動化の研究を含めて，さまざまの言語情報処理（各種語彙表・用例表の作成といった低い段階から自動単位切り・読み仮名つけ・構文解析など高度なものまでを含む）の前進のためにも，確かな語彙データが，コンピュータに納まっていなければならない。以上のように考え，同語異語判別作業は人間ができるだけ丁寧に行い，判別の結果は磁気テープ（または磁気ドラム）に納められることとした。今回の同語異語判別作業は，大きく次のⅠ・Ⅱ・Ⅲの三段階に分かれている。

Ⅰ．プレ・エディットの段階で「代表形」を付ける。

これは入力データの段階で，活用語（助辞を除く）には終止形を，語形に二

種以上のゆれのあるものは，その中の最も一般的な語形を代表形として添える作業である。代表形付けに関する作業規則およびその説明として，以下のものを作成し，作業者に配布した。

○清書規則・付　代表形について（1975.3.8　第3版）

○代表形のつけ方（言語計量研究部「季報」1975年秋・第1分冊）

○代表形の問題点　その1「集計単位に注目しよう」（同「季報」1977年夏）

これまでにも述べたように，集計単位は語彙全体を見渡して初めて定まるものなので，この段階で完全な代表形が付けられるわけではない。しかし，異形態同語・異表記同語は，この段階で大部分同一箇所またはその付近に集められる。作業者が代表形をつけ忘れたり，ゆれていると意識していなかったり，あるいは全く同語と意識しなかったりして代表形を付けず，たまたま二種以上の形が調査する対象の中に出てくる場合などには，ばらばらになる。

なお，実際の作業の段階で，あまり多く代表形を付けることは作業を遅らせることになるという配慮のもとに，代表形を付けるのを，一部あるいは全データにわたって，差し控えたものもある。

例　れる・られる⟶れる　　せる・させる⟶せる

　　う・よう⟶う　　ない・ぬ・ん⟶ない

　　大きい・大きな⟶おおき　　細かい・細かな⟶こまか

　　読む・読める⟶よむ

これらは，もっと先の段階で，労力少なく，代表形を付けることが可能であろうと考えている。

II　KWIC 作成ののち同語異語判別作業を行う。

これは検査・修正用の KWIC を利用して同語異語判別作業を行うものである。すなわち，データの検査・修正が終了したのち，同じシステムでKWICを作成し，その KWIC を見ながら，情報を付加・削除し，新しいデータ・ファイルを作成する作業である。Ⅰの作業で異形態同語はほぼ集められているので，ここではその確認をする。代表形が必要なのに付いていない場合，あるいは同語に二種以上の代表形が付いている場合などには，代表形の付加あるいは差し換えを行う。また，同形態異語・同表記異語を分離する。同表記異語とい

っても，漢字表記語は入力段階で読み仮名を付しているので「工夫（こうふ・くふう）」とか「生物（せいぶつ・なまもの）」といった同表記語は問題とならず，もっぱら「水をかける（掛）」・「大地をかける（駆）」・「人員がかける（欠）」の「かける」のような，仮名表記の同語の分離が中心となる。これによって，同語異語判別作業はほぼ終了する。

Ⅱの段階でも，いろいろな方法が考えられる。

（1） **代表形に手を加え，変形させて，1集計単位は1代表形とする。** これは同音語が多数ある場合，代表形に手を加えすぎることになり，不自然な代表形を付けることは，非常な技術を要するので，好ましくない。例えば，「書く」「掻く」「欠く」を区別するために，それぞれの代表形を「かくしょ」「かくそう」「かくけつ」と音を付加する方法や，「かく1」「かく2」「かく3」，あるいは「かくA」「かくB」「かくC」と記号を付加する方法などがこれである。代表形をこのように変形しても，配列用の読み仮名（普通の読み仮名から濁点・半濁点を除き拗音・長音などに手を加えて平仮名で表わしたもの）を元のままにしておけば，配列上は問題がないであろうが，検索のため，さらには仮名漢字変換の実験等のためには，代表形は見出し語そのものであったほうがよいので，この方法は採らない。

（2） **代表形のほかに漢字1字を添えて同形異語を分離する。** 異形同語が一つの代表形のもとに集まっているかどうかを，KWIC を検査し確認する一方で，同一代表形の中で分離しなければならない語に出会ったら，代表形のほかに，何らかの情報を付加して分離する。この結果，1.代表形，2.付加情報（すでに入力段階で付いている助辞の情報を含む）が同じなら同語，同じでなければ別語となり，同語異語判別作業は完了する。

この場合，付加する情報としては，上にしるした漢字1字だけでなく，次のような，いろいろな方式が考えられる。

a．代表的な漢字表記を1字だけでなく全ての長さにおいて添える。

例　でんき＜伝記＞・でんき＜電気＞・でんき＜電器＞

b．意味情報，例えば「分類語彙表」の番号のようなものを添える。

c．意味を表わす語，例えば，言い換え程度の語釈，漢語なら訓読み，その

他，使用上の注意すべき点で語の識別に役立ち得るものを添える。

d．数字・記号などを添える。

例　でんき1，でんき2，……

このうち，bは非常に興味深い方式であり，同語異語判別作業の機械化のためには，最も試みるべき方式と思われる。現在「分類語彙表」（国立国語研究所資料集6）には，意味上関連のある語をグループ化して，そのグループに番号を付しているが，これを各異なり語ごとに番号を付し，新出語にも番号を付するようにすれば，これをテーブルにすることによって，各種語彙調査の語彙表の統合やデータファイルの統合，各種語彙調査の語彙の対比研究，さらには，同語異語判別作業をしていないデータを計算機によって判別することも可能になってくる。同語異語判別作業を完全に自動化することは不可能かも知れないが，テーブルを充実することによって，大量語彙調査の場合は，かなり精度を上げられると思う。あるいは誤差の範囲内として許容される程度まで同語異語判別の精度が上がるのではないかと思う。そうなれば，「自動化」は実現したことになる。また，そうならないまでも，計算機によって同語異語判別を行い，処理できなかった部分あるいは誤った判別をした部分について手作業で同語異語判別作業を進めるようにすれば，能率が上がると思われる。

しかし，今回は実用化を急ぐため，漢字1字方式を採った。漢字1字におさえたのは，高速漢字プリンターによる判別作業台帳（データ修正の済んだデータで作成した修正用ミニ KWIC M単位表）のスペースを考慮したためである。しかし，堅実な方式ではあるが，後述するように，自動化のための基礎実験も兼ねているものである。なお，漢字1字とはいうものの，国語研究所の漢テレで情報を付加・入力するため，盤内字2,110字に制限される。実際に作業を進めてみると，例えば，「汗をかく」の「かく」に付ける漢字がない（「掻」は盤外字）など，不便を感ずる場合もあるが，作業担当者の工夫にまつことにした。

以下に，今回の作業の方針とⅡの段階における作業手順をかかげる。

作業方針

1．　同語異語判別作業はM単位についておこなう。

1. 1　M単位は文字列単位と語（形態素）単位が混在しているが，同語異語判別作業は，語の単位について行う。句読点・記号・数字は語として認めがたいので除く。

1. 2　「語」の範囲の中でも，助辞は同語異語判別作業から除く。

2.　同語と認める範囲は，雑誌九十種調査の集計単位に準ずる。

作業手順（Ⅱ段階のみ）

Ⅱの段階で，すなわち，修正処理の終了後，最終出力までのあいだに，代表形の同じなかで，同語のものに漢字を１字判別情報として添える。

(1)　代表形と判別情報・助辞情報の有無が一致するものは同語であり，一致しないものは別語である。

(2)　作業としては，出現形の先頭１字を機械処理で判別情報として取り入れ，それを変更するものについてのみ，別字を指定するものとする。

(3)　仮名書きの語にも盤内字の中から１字を適宜添える。

例　いま―今　ニワトリ―鶏　イモリ―い（(4)参照。）

(4)　判別情報は，仮名であってもよい。出現形が大部分仮名表記で，漢字を添えることが大変な労力となる場合は，別語のまぎれこむことに注意しつつ，この規定を適用する。

例　ある・いる

(5)　判別情報に，盤外字は，漢字１字という制限のため使えないので，盤内字で用をなさない場合に限り，表記に関係のない，品詞・意味・用法等に関する情報を１字（盤内字）添えることが許される。

(6)　前項の場合，先頭から２字目の漢字で用をなす場合は，それでもよい。

例　電気／電器　→前者に「気」，後者に「器」

確率／確立　→　〃　「率」　〃　「立」

精細／精彩　→　〃　「細」　〃　「彩」

国際／国債　→　〃　「際」　〃　「債」

Ⅲ　語彙表作成の段階で補正作業を行う。

同語異語判別作業は，Ⅰ・Ⅱの段階でほぼ終了するはずであるが，一部には作業を残すものもある。不注意で残ったものはⅠ・Ⅱを繰り返すことによって

処理できるが，ここにⅠ・Ⅱの作業段階では労力がかかりすぎ，あえて残されるものが出てくるだろうという推測がある。同語異語判別作業は，この稿執筆の時では，まだ試作検討の段階であるため，十分な見通しを立てることができない。しかし，すでにⅠの段階で代表形を付けるのを差し控えた「れる・られる」・「ない・ぬ・ん」など，一群の高頻度の語については，Ⅲの段階を設け補正する必要があることは明白である。これらの語に大量の代表形を添えることは労力の点で得策ではないため，あと回しにしたものである。また，代表形をⅠの段階で添えたとしても，代表形を付け忘れ，あるいは別の代表形を付けて修正を必要とすることも十分起り得ることで，その場合，入力データの語番号をもとに代表形を付け換える現在の修正システムでは，十分な修正を行うには労力がかかり過ぎると思われる。

そこで，語彙表作成の段階で，手直しをするのが適当と考えられる。この場合，語彙表を人手によって書き改めるというのも一方法であるが，磁気テープに納まっている状態の語彙表データに，ディスプレイなどを使って，一括して代表形を差し換える（例えば全ての「られる」の代表形を「れる」とするなど）の方式が望ましい。なぜなら，こうすることによって，同語異語判別の済んだ磁気テープ・ファイルが作成されるからであり，この磁気テープ・ファイルは先に述べたように，さまざまの分析・研究を可能にするものだからである。

ただ，Ⅲの段階の機械処理システムは，まだ，全く作成されていない現状なので，確実な見通しを述べることはできない。

5. 同語異語判別作業と機械処理システム

以上，高校教科書の同語異語判別作業の考え方・作業の進めかたのあらましを，人手が中心であることから，人手による作業の側から述べてきたが，機械処理システムの中でどう位置づけるか，各装置をどう利用するかなどについて，箇条書きにしてみる。同語異語判別作業の機械処理の部分は，まだ動き出していないので，以下の事がらの多くは，検討中のものである。

(1) 語彙調査システムの流れの中での位置づけ

検査用 KWIC による検査・修正の終わったデータに対して，同語異語判別作業を行う。修正作業の段階で同語異語判別を行うとか，修正の手の入った作業台帳を使って同語異語判別作業を進めることも考えられるが，ミスを生ずること，また，完全な判別作業を期しがたいことから，好ましくない。語彙調査において，検査と修正は大きな作業量であるから，判別作業は，調査の流れの中では，最終的な位置に置かれることになる。

(2) 判別後のデータ・ファイルの扱い

同語異語判別作業の結果，KWIC データのファイルには，当然判別結果が付加されるが，入力原文の形式のデータ・ファイルにも，判別の情報が付いていることが望ましい。これは，判別情報の付加処理を入力原文形式のデータで行えば，直ちに実現することである。KWIC データに直接，情報を添えるシステムを採る場合は，後に原文形式のデータに情報を移す手間が加わるわけであるが，そのデータでさまざまの実験・分析を試みるためには，手間を惜しまず，情報を転写する処理をしておくべきだと思う。

(3) 判別作業と機械装置との関係

人手による判別作業の結果を，電子計算機に再入力して，次の作業である語彙表作成等の機械処理に備えてはじめて，同語異語判別の段階は終ったことになる。この再入力の機械処理について，現在，二つの方法が考えられている。

(1) データ修正システムと同じく，付加情報を紙テープで入力し，情報付きの磁気テープ・ファイルを作成する。

(2) データを磁気ディスクに納め，中央処理装置のコンソール・ディスプレイによって，情報を付け，情報付きの磁気ディスク・ファイルを作成する。

(1)の方法は，大量・迅速に機械処理をすることができるが，完璧を期しがたく，また，間違えた場合，そのミスを発見するために幾つかの機械処理を経た後の印字結果を待たねばならない。(2)の方法は，正確に情報を付けることはできるが，ひとつひとつキーを押して付けてゆくため，非常に時間がかかる。結局，両方法を併用し，(1)の処理ののち，処理し残したものについてのみ，(2)の処理を施すことになろうか。

このほかに，漢字ディスプレイ装置による方法，国語研報告49で石綿敏雄氏の提案されたＩＢＭカードによる方法等も考えられるが，いずれも現在の機械装置とソフト・ウェアでは難点があり，実現しにくい。

6. 今後のシステムの拡大

この同語異語判別作業が，人手と機械によって順調に動き出してのち，次のように，作業の内容を広げてゆくことが可能であり，また，当然そのようになっていかなければならない。この高校教科書語彙調査というプロジェクトは，昭和53年度までに終了するものなので，この中では，同語異語判別作業の実現の可能性を探り，おそらく一部データについて実施するにとどまると思われるが，今後の方向として挙げておく。

(1) W単位の同語異語判別システム

現在のところ同語異語判別作業は，M単位についてだけ考えられている。しかし，W単位語彙表を作成し，W単位の見出し語レベルの分析・記述を進めるためには，W単位の同語異語判別作業を行わなければならない。現在のところ，W単位の集計単位はまだ考えられていないし，それを考えるために十分なだけのW単位に対する理論づけもできていない。また，調査システムの流れの中では，入力段階でM単位に対して代表形を付しているだけなので，M単位の同語異語判別と同じく，W単位の同語異語判別でも代表形を付けるとすれば，はたしてM単位の代表形からW単位のそれを機械的に合成すればよいのかどうかも，まだ十分に考えられていない。これらは今後引き続き検討していかなければならない問題である。

(2) 情報を付加するシステム

同語異語判別作業のシステムを利用して，品詞・語種その他の情報を付加していくことは容易に考えられることである。付加した結果を分析に使うだけでなく，情報を蓄積し，あるいは更新していくように，ファイルの運用管理システムを考えねばならない。ただ，判別結果を計算機中のデータに正確に付加する方式がまだ十分に考えられていないので，確かなデータ修正システムの設計とともに，今後，まず検討していかなければならない。

(3) 同語異語判別の自動化

同語異語判別作業で，人手に依らなければならないことは，今後も同じであろう。しかし，同語異語判別済みのデータが今後大量に蓄積されるならば，それを利用して，ある程度機械に同語異語判別作業をさせ，機械でできなかったものだけ人手でするという方式が考えられてよい。さらに進んで，蓄積されたデータとソフト・ウェアによって，機械による同語異語判別が，十分利用できるだけの精度をあげ得るようになれば，それは同語異語判別の自動化が完成したといってよいであろう。実際に，新しい調査でそのような精度を上げることは不可能としても，同語異語判別済みの大量のファイルがあり，それに類似して，しかもばらつきの少ないデータを対象とするならば，十分考えうることである。