Application of the Kana-Kanji conversion process to the identification of homonyms

国立国語研究所報告85　研究報告集7（1986）

同形異語判別への仮名・漢字変換処理の応用

斎藤秀紀

1. はじめに

国立国語研究所（以下国研）では，国語政策を決定するうえで参考になる資料を収集・作成するため，各種の用語用字調査を行ってきた。用語調査は，昭和27年に実施された『現代新聞の用語の一例』〔文献7〕が最初である。これは，引き続き行われる，婦人雑誌〔文献8〕・総合雑誌〔文献9〕・雑誌九十種〔文献11〕調査の試行をかねたものである。また，昭和41年には，調査規模の拡大をはかるためコンピュータを導入，同年から開始される新聞3紙(朝日・読売・毎日朝夕刊3紙1年分)〔文献13〕の大量処理に使用された。以後コンピュータは，小・中・高教科書〔文献14〕，日独・大都市における言語の社会調査，文献調査，KWIC用例集の作成など，資料収集と言語研究に利用されてきた。しかし，コンピュータを使用した初期の用語調査は，人手を中心とした作業の延長上にあり，機能を十分に活用しているとはいいがたかった。特に，新聞調査では終了まで約9年を要しており，この点からも基本処理に重点がおかれていたといえる。これは，漢字処理に対する技術力の未成熟さと，大量データ処理に対応可能な単語分割方法の確定に問題があったためと思われる。そのほか，調査の中間に人の判断を挿入したことも原因の一つにあげられる。

本稿では，昭和59，60年度文部省科学研究費「国定読本の用語の研究」における50万長単位語調査（第3期～第6期）用に開発したシステム機能を中心に述べ，人間—機械間の相補処理によって，調査期間と経費の削減が可能

になることを示す。なお，OCR 方式の利用上の問題点について，4以下に OCR 方式に対する改善試案を示した。

システムは「言語処理におけるターンアラウンド・システム」〔文献1〕の考え方を基本に次の二点を拡張した。第一は，同語異語判別処理にかんする部分であり，第二はデータの統計的性質をシステムの運用に応用したことである。

第一の機能によって，同語異語判別処理を中心に，仮名表記の辞書による意味の確定，見出し語と五十音配列用理論コードの併用など，付加情報の統一処理への道をひらいた。第二の機能からは，作業期間の短縮である。これは，付加情報の分離一括処理，OCR (Optical Character Reader: 光学式文字読み取り装置) 用紙上の情報に対する複写機能，同一語形の類型化による作業データの疑似的削減である。また，本稿では，第一と第二の結合によって将来拡張すべき機能，すなわち辞書の総合化，データ分類上の動的キー指示，仮名・漢字変換処理の同語異語判別処理への応用，について方向を明確にした。ここで，同語異語判別処理とは，国定読本に現れる表記形で，同じ語か異なる語かを判別する操作をいう。基本形は，音・形・義の三要素を組み合わせ，8種の類型から同音か異音かの判別をはぶいた4種について処理を行う。

同形	同語	行く (いく)／(ゆく)，夜 (よ)／(よる)
	異語	いま (名詞)／(副詞)，か (終助詞)／(副助詞)／(並立助詞)
異形	同語	あまり／あんまり，みな／みんな，木／き／キ
	異語	麻 (あさ)／朝，入る (いる)／居る／射る

2. OCR 方式による処理の概要

コンピュータを使用した用語用字調査は，新聞を対象とし昭和41年から本格的に行われた。調査システムは，それまでの人手による調査の経験をもとにコンピュータ化した。人手による方法の利点は，コンピュータ導入以前の

用語調査法，組織運用に人手による作業形態を生かせることにある。また，調査内容を熟知している作業者の確保，要員教育から業務への段階的移行，コンピュータ導入の初期抵抗を柔らげる効果が得られる。しかし一方では，人手を中心とした作業を，そのままコンピュータ処理へ移行させることは，システム設計，人員配分の適正化など，効率面で満足できない点も少なくない。原因は，調査の重要部分をコンピュータ入力前にすべて人手によって処理するため，データ修正，入力データ量の増加，入力原稿作成にともなう清書・転記作業が全体の効率をさげるためである。当然，効率の低下は調査の長期化をまねき，運用経費の増加，要員の確保と異動にともなう業務の引き継ぎ，コンピュータの切り替えで問題が発生する。OCR 方式は，このような問題に対し，調査期間の短縮と費用の削減をはかる目的で計画された。要求した機能は次の4点である。

1) 入力原稿に対する事前編集事項を少なくすること。
2) 一次入力の対象となるデータ数の削減により経費軽減が可能なこと。
3) 入出力媒体と作業用帳票の共通化をはかり，中間で発生するデータとの照合が容易なこと。また転記・清書などの中間作業を省略できること。
4) 入力データ，作業内容が直視でき，調査者にとって作業内容が確認しやすいこと。

4項目の要求事項の採用によって，新聞調査で障害となっていた点は解決される。国定読本調査の実用システムは，調査量50万長単位語に対し約800万円，期間2年間で対応しなければならない。このことからも，従来の方式と比べ費用，期間，必要人員など二分の一以下におさえることが必要になる。本システムで採用した OCR 方式は，用紙価格，保存媒体の耐久性，手書き文字の作業者への教育，濁音，半濁音，拗音，撥音，促音などに特殊表現が必要になる欠点がある。しかし，以下に示した利点を有しており，前述の条件を満たすためには有効な方式である。

OCR 方式の利点

1) 作業台帳，入出力媒体をOCR用紙によって共用化できるため，データ転記のさいの誤り防止と清書など中間作業の省略が可能。
2) OCR用紙へデータを直接記入することによって，データ入力件数と入力経費の削減が可能。
3) 作業対象になる任意の用語が，事前に分類・配列できるため，作業目的別の帳票編集により作業の効率化が可能。
4) OCR用紙によるデータ記入処理の分散と機械処理の簡素化によりシステムの非専門家への解放が可能。
5) 人間—機械系間の相補処理による有機的システムへの拡張と運用の最適化が可能。
6) OCR用紙上のデータおよび作業手続きの直接確認による機能変更，システム管理の容易性，プログラム作成本数の削減が可能。

システムの作業工程を図1に示した。システムは大きく前処理部分と後処理部分に分けられる。前部分は，データ作成・修正，単位切り処理を行う。後部分では，単位切り済み用語に対する見出し語づけ，同語異語判別，品詞・注記（表3）の各処理である。ここで，二種の作業用紙は，単位切り処理用OCR紙を「01」，後処理部分で使用する用紙を「02」と名づける。また「03」は，KWIC用例出力用紙を，「04」用紙は，各種の語彙表を一括表現する場合を想定している。処理の対象となるのは，01から03の範囲の処理である。以下にOCR用紙と図番号，作業内容を示す。

用紙名	図番号	作業内容
01	2	単位切り処理
02	3	見出し語，品詞・注記，同語異語判別処理
03	4	KWIC用例集
04	—	各種語彙表

01～02の各作業は，用例作成に必要な文長の決定と単語分割（単位切り）への準備が必要である。文と単語の分割符号は，それぞれ「#」「／」記号を使用する。指定記号は「#」を入力原稿の事前編集で，「／」は01用紙の手書

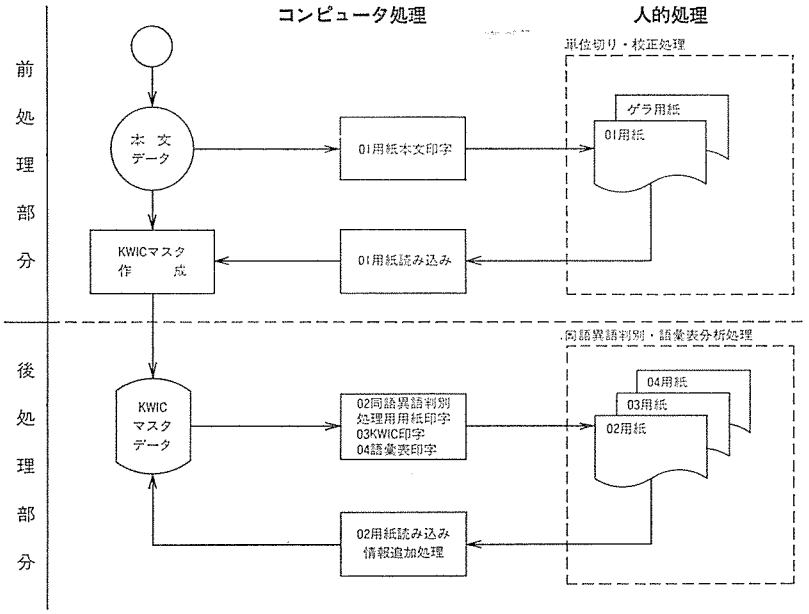

図 1　国定読本の用語調査作業工程

き部分で対応させる。「／」記号は，OCR 用紙を読み込んだのち，コンピュータ処理によって本文データへ挿入，単位切りと KWIC 形式へのデータ・フォーマット変換を行う。一用例単位は「＃」記号で示した文長である。KWIC 形式に変換された用例は，単位切り用校正台帳，02作業帳票の二業務で共用されるが，校正は同一語形群にまとめられた KWIC の見出し語に着目し，目視による検査を主体とする。修正は，出典情報を手がかりに，該当する01用紙上で行い，誤りがなくなるまで修正・再入力をくりかえす。

一方，02用紙上の用例は，同語異語判別の付加情報の作業用に使用し，出現形に対する見出し語・品詞・注記，仮名表記データへの意味コードを KWIC を参照しながら記入する。同形異語の判別には，出現形が漢字の場合は読み仮名，仮名表記では漢字を付加する。本システムでは，漢字の直接付加に替えて，しかるべき辞書または判別用リストを基準に，見出し語につけられているコード番号を用いた。見出し語と追記された意味コードは，

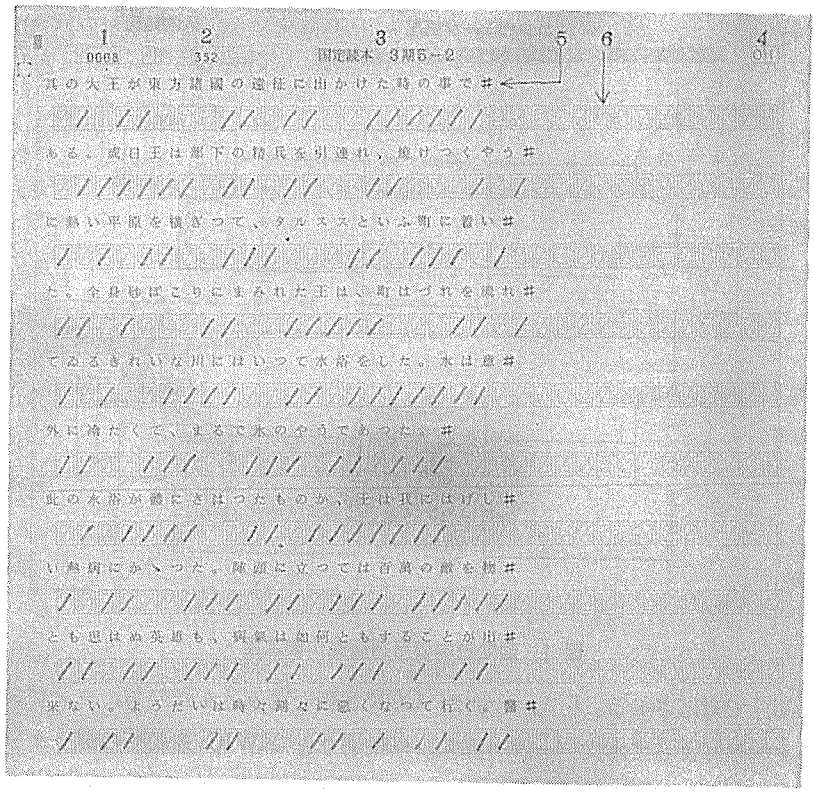

図 2　単位切り処理用紙（01用紙）

01，02用紙共通情報（桁数．枝番号）

1. 帳票シーケンス番号（4.1）
   OCR 用紙の通し番号。帳票脱落，追加，入力順序の確認用。
2. 識別番号（3.0）
   教科書の期・学年・巻数。
3. 出典タイトル（20.0）
   識別番号および表題用
4. 帳票 ID（2.0）
   01，02の二種の数値で示される OCR 用紙の識別番号。

01用紙（455字／頁）

5. 本文データ（35字／1行，13行／頁）
   単位切りの対象になる本文データ。
6. マーク記入枠
   印字された本文データの真下にある単位切り処理記号欄。

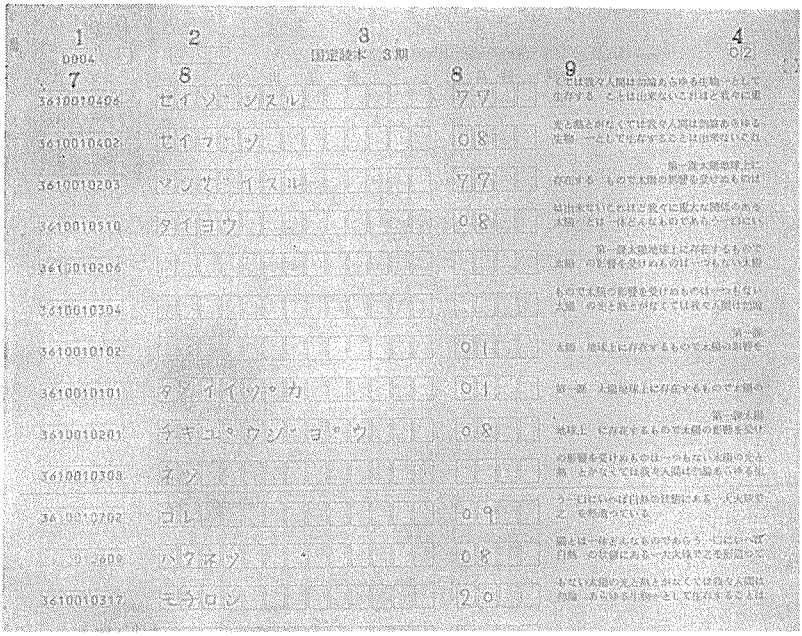

図 3　見出し語・品詞・注記用紙（02用紙）

02用紙（21行／頁）

7.　出典情報（10.0）
　　教科書の出典情報，期(1)・学年(1)・巻数(1)・頁(3)・行(2)・単語番号(2)。

8.　見出し語記入枠（13.0），品詞・注記記入枠（4.0）
　　手書き用片仮名文字記入枠。
　　品詞・注記，同音異語判別用識別番号。

9.　用例データ（40字）
　　付加情報・用例印字欄。

02用紙で使用する特殊機能

＋　：用例中の出現形を見出し語欄へ複写。

＃　：読み仮名が13字以上の場合13字目に挿入。12字までを見出し語以下省略。

空白：直前に記入された情報の複写。

相互に参照できるため，将来機械辞書の総合化をはかるうえで重要である。02用紙における同語異語判別処理は，調査の中核であり効率化をいかに進めるかが調査進行のかぎになる。特に，作業量を減らすためには，同形異語判別語数を全データ数から異なり語数に近づける方法が必要になる。その方法として，本システムでは次の処理を行った。第一は，見出し語・品詞・注記

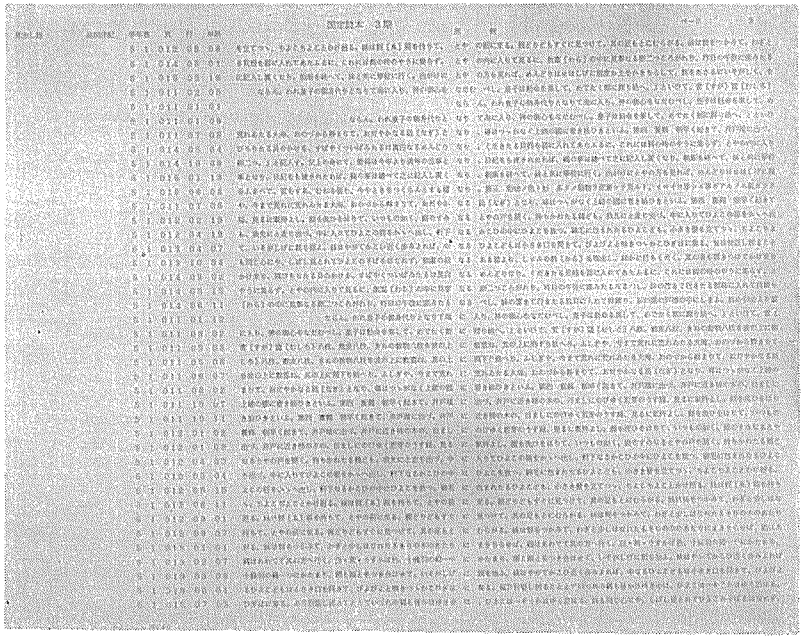

図 4　KWIC 用例（03用紙）

の記入量の削減と手書き文字の誤読率をさげるため，各記入欄に複写機能を持たせた。第二は，KWIC 用例を02用紙に印字するさい，作業に合った帳票編集と出現形が同一語群にまとめられるようはかった。第三は，作業台帳・最終 KWIC 処理を同一プログラムで対応させたことである。

複写機能は，記入項目数の削減が，また帳票編集は，作業効率と精度を向上させ，同一語形群の集約は，実質的に作業を早める効果が得られる。作業台帳および最終出力用 KWIC の共用方式は，プログラム本数の削減，操作，運用管理を容易にする。さらに KWIC 上のデータ修正は，キー語のみとすることによって，作業に支障が生じない限り用例修正は省略できる。最終の誤り修正は，すべての付加作業が終了した時点で各情報を KWIC マスタファイルに追記，最終ファイルの一括再編成によって対応させる。最終 KWIC 出力と修正は，同一プログラムによる反復処理をとるため，両処理の運用は

スパイラル形式になり，データは反復処理のなかで収束する。

以上が，OCR方式によるターンアラウンド処理を利用した用語調査システムの概要である。開発したプログラムは，入出力関係7本，データの併合など3本，計10本である。入出力中心のシステム構成になったが，プログラムの設計・作成・運用の簡素化を目的とした当初の計画は達成されたと思われる。

システム設計にあたって参加者全員による討議を行った。この作業が可能になったのも，OCR 用紙上にシステムの全体を明示させ，入力データ，調査過程，出力処理を調査者に見せた結果と考えられる。このため，調査の最終結果，入力データ，中間作業のおのおのが机上で検討でき，調査者の各作業分担，システム内の位置，変更にともなう確認が明確になり，把握が容易になった。システム全体の効率化を進める方法として重要な手順であったと思われる。

3. 分離処理による処理の効率化

付加情報づけを行うには，事前に対象データを同一語形群に分類，環境を均一化したうえで作業を進めることが，作業効果をあげるうえで重要である。しかし，同音異語判別情報の付加は，出現した用語のすべてを対象にしており，判別処理が調査全体の進行を決定していた。これは，同語異語判別情報の作業効率の向上が，調査期間の短縮に不可欠であることを意味している。本システムでは，入力データの削減，中間作業の短縮，プログラムのモジュール化による作成本数を削減させるため，OCRによるターンアラウンド方式を導入した。

省力化の二次対応は，OCR 用紙に記入された情報の複写機能と手書き文字の記入作業の削減，特定用語の分離による品詞処理の半自動化である。判別処理の効率からいえば，本来完全な自動処理にあるが，本システムでは，人間—機械系の相互補足方式を基本にした。対応法は，出現した単語の度数上位語を選択的に分離すること。品詞情報は，特定品詞をかりに付加し（表

1，表4）他の品詞については，修正・補正処理で対応させたことである。分離処理は，上位出現語の特定の数語によって全データの相当数が占有される特性を利用するものである。

たとえば，第1期の国定読本（延べ31619語，助詞10663語，表1およびこの数値は集計初期の値を使用しており最終結果と異なる場合がある）では，上位5語で全データの約29.9%を占めている。この数値をもとにすると，第3期以降で処理すべき50万語については，上位5語で約15万語，10語で16.4万語，20語で20.9万語が対象になると推定される。すなわち分離処理によれば02用紙で直接処理すべきものは，それぞれ35万語，33.6万語，29万語になり，実質的作業量の減少と調査期間の短縮に結びつく。同様に，同表記異語の「散らばり度」も特定の品詞，用語に集中することが知られているため，仮定的処理として付加情報を一義的に決定し，曖昧さをともなう語は，データの校正・修正段階で補正することにする。

以上の点をまとめると，分離処理すべき語を抽出する条件は，既知であるデータの語彙表から推定できること。抽出する上位語は，一義的に付加情報が決定できることを前提にする。逆に，同一語形であるが同語異語判別処理で曖昧性が高く，複数の品詞，または意味に分散される語は除外する。実際に抽出した5語の分離対象語は，いずれも，語形変化のない語，読みが一義的に決定できる語，これに近い条件を基準に選択した。しかし，データ配列に，より詳細な条件が設定できれば，語の類型化から，さらに自動化を指向した処理への道をひらくことになる。現行は，複数個の品詞を持つ語に対し，出現形の直前・直後の一単語をそれぞれ第二，第三キーとした。このキー語によって，同じ環境の語を集め品詞決定の補助処理としている。そのほか，分離処理で得られる効果には，作業台帳でOCR用紙の代替紙として低価格用紙の利用，一括処理との併用による人的労力と人件費の削減も含まれる。また，本調査で記号系を不要語としたのも省力化の一つになる。

表1に見出し語として候補にあげた5語は，第1期の出現上位の同音同表記（表4：出現形では片仮名，平仮名の10語分）の出現頻度の高い語から恣

表 1　分離処理対象語の候補（国定読本第1期調査表から選択）

No.	見　出　し　語	出現数(出現率)%	偏り度%	出現予想数（異語数）
1	は（係助詞）	1552 (4.91)	97.4	24550 (639)
2	て（接続助詞）	1495 (4.73)	98.5	23650 (355)
3	を（格助詞）	1092 (3.45)	99.6	17250 (69)
4	た（助動詞）	977 (3.09)	97.5	15450 (387)
5	も（係助詞）	389 (1.23)	99.7	6150 (19)
計		5505(17.41)		87050(1469)

意的に選んだ。見出し語欄の情報は，付加された品詞の例である。出現度数は，第1期の延べ31619語に現れた各語の出現度数と百分率である。偏り度，出現予想数は，第3期以降の予想される調査量50万語に対する作業量の予測で，第1期出現度数の百分率から推定した。出現予想度数のカッコ内の数値は，同表記異語の予想数で偏り度の百分値の差から求めた。この出現度数のなかには，それぞれ見出し語につけて，指定した品詞にあたるもの以外の同表記の語を含むため，同表記異語を省いた指定の語の百分率を「偏り度」として示した。

表1の「は」の場合，第1期の出現度数は，1552語で，全体の4.91%を占めている。特定の意味・品詞への集約，すなわち係助詞として97.4%，そのほか2.6%が他の品詞である。2.6%の内容は，仮名文字練習用（ひらがなドリル）1例，「葉」の仮名表記34例，漢字表記6例である（表4）。出現予想数のカッコ内の数値は，見出し語で出現が予想される総語数中，異なる意味，または品詞・注記である。「見出し語」の例では，第1期，第2期以外の50万語では，「は」については24550例，そのうち639語が助詞以外の語であり，39語に1語の割合で修正処理を必要とすることを意味する。分離処理では，とりあえず合併してある情報から，別語として修正する部分に埋め込まれている誤りデータの見過ごしが問題になる。一義的に付加した情報から修正部分を抽出するさいの実用化の可否はこの部分にあり，さらにコンピュータとの相補処理を充実させる必要がある。

4. OCR 方式に対する改善試案

4.1 仮名・漢字変換処理の拡張と応用

試案は，データ入力に仮名・漢字変換処理を使うことによって，OCR 方式の問題点を一部改善できることを示す。改善案の第一は，仮名・漢字変換処理にともなう，データ入力と単位切り処理，読み仮名の同語異語判別情報としての利用法，配列用理論コード機能の三点である。第二は，変換用辞書のあり方である。ここで，読み情報を一次情報とし，二次情報を変換対象語，双方を対応させる関係表を辞書とする。

変換方式を使って漢字入力を行う場合，漢字タブレットによる直接入力方式に比べ，原稿表記を仮名またはローマ字で表現可能な「読み」に変換する間接的な入力法をとらなければならない。変換方式による漢字入力は，一次情報の「読み」を入力，辞書による変換をとおし「原稿表記」へ再変換する二重処理を行っている。さらに，読みと原稿表記への二度の変換過程は，データを正しく入力するために，それぞれの段階において人間の判断による同語異語判別，変換された用語・原稿表記間の照合，入力データの検査を暗黙に処理している。これは，一次入力情報の指定と同時に，「単位切り」，「漢字の読み仮名付加」，「仮名表記の漢字付加」を変換処理で対応させていることになる。言い替えると，これらの一連の処理は，OCR 方式の01～02用紙作業を，仮名・漢字変換方式による入力処理で，吸収できる可能性があることを示している。

また，入力部分で単位切り，読み仮名などの情報づけは，OCR 方式における作業上の問題点を改善することができる。たとえば，02用紙上に印字された KWIC の配列がコード順である場合，異表記同語は，異なった位置に並び，用語の検索，異形同語処理ともに，02作業に対する作業を難しくする。これに対し，漢字の五十音順配列では，02用紙で使用する作業用 KWIC 用例の配列に，異なった語形が集められ02用紙の手書き文字の誤読問題，文字記入量，システムに必要な作業日数を減らし，全調査に必要な経費を削減

させる。しかし，これらの問題をメーカから提供されている仮名・漢字変換プログラムで対応させるためには，仮名・ローマ字入力機能に，一次情報である読みと，変換された漢字ともに情報の保存が必要になる。保存および同語異語判別情報を得るための出力は，以下に示した形式になる。なお，斜線は単位切り符号である。

変換処理	原稿内容	出力形式	意　味
1)変　換	①漢　字	仮名（漢字）／	読み仮名づけ処理
	②仮　名	漢字（仮名）／	同語異語判別処理
	③漢字（ルビ付き）	仮名〔漢字〕／	ルビによる処理
2)無変換	①仮　名	仮名（仮名）／	仮名処理
	②英数記号(ANK)	ANK(ANK)／	英数字記号処理

「出力形式」で示した表現は，辞書を使用した入力データの変換後の形である。見出し語に相当する部分は，指標形式で現し，カッコ内の表示は原稿上の出現形である。カッコ類の各記号の意味は，出現形および読み仮名，同語異語判別情報，仮名表現とルビの分離情報を示している。ここで，入力データを出力形式に変換するためには，仮名表記についても漢字変換が必要になる。また，後処理方式では不要であった漢字の読み仮名，単位切り符号の原稿上への事前記入・編集など，人手を中心にした前処理方式に近い作業が要求される。単位切りについても，現行方式の仮名・漢字変換方式は，複数の単位についての認定方法が確立されていないため，同一辞書で混在させるか，調査単位ごとに語変換用辞書を用意しなければならない。この単語の二重性は，入力処理と並行した，辞書の保守とシステム管理など，二重の対応が必要になることを意味する。さらに，拡張機能を仮名・漢字変換処理で実用化させるには，次の二つの問題を解決しなければならない。第一は，辞書に登録されている用語と読み仮名が一対一に対応しているため，同音異語判別への選択回数が増加することである。第二は，入力打鍵数の削減のためにとられる，代表音訓による漢字選択では，正しい読みが指定できない場合が生じることである。この二つの問題は，変換処理と付加情報の生成が漢字を

「正しく読む」ことを前提にしているため，入力用読みと出力される語についてN対1の基準化が必要になる。この点については，それぞれ別語として辞書へ登録しなければならない。同様に，漢字変換処理には，入力情報の保存・併記出力機能は含まれないため，辞書機能の変更，項目の自由登録，諸機能の利用者への開放，支援プログラムの拡張が必須となる。メーカから提供されているプログラムの変更と，利用者に対する辞書へのアクセス法の開放には，プログラムの部品化による機能分割の適正化とユーティリティへの基本的な考え方を変えることが必要である。

最後に，OCR 方式と改良案との相違を図5に示した。この図からも，現行処理の相当部分は省略できることが明らかであり，新システムへの移行には，改良案の妥当性の確認が必要になる。日本語に対する同語異語判別を容易にするためにも，入力データと変換用辞書の在り方（仮名・漢字併用出力機能も含む），漢字の多義性への対応，学習効果の研究が今後の課題である。

4.2 ソートキーの動的指示

データの分類処理（SORT）は，データをグループ化，または最終印字のために配列する場合と，未知のデータの特性を抽出する試行的手段として使用する二つが考えられる。コンピュータによる分類処理は，メーカから提供されるプログラムの使用が多いが，利用にあたって次の条件を満たしていなければならない。

1) データ中に分類基準となるキー項目が存在すること。

2) キー項目の属性・位置・桁数が明確に与えられていること。

現在使用しているソートプログラムのパラメータは，対象になる全データに一様に適用される。しかし，ソート処理によってデータの構造を抽出するには，キー項目はデータ構造を間接的に表現していることが必要である。構造の抽出は，データ構造をキーに対応させるさい，結果をある程度予測してキー項目を設定し，最適配列が得られるまでキーの変更と多数回の試行を行う。この点で，データ特性を無視したキー指定は，基本的に成立しないこと

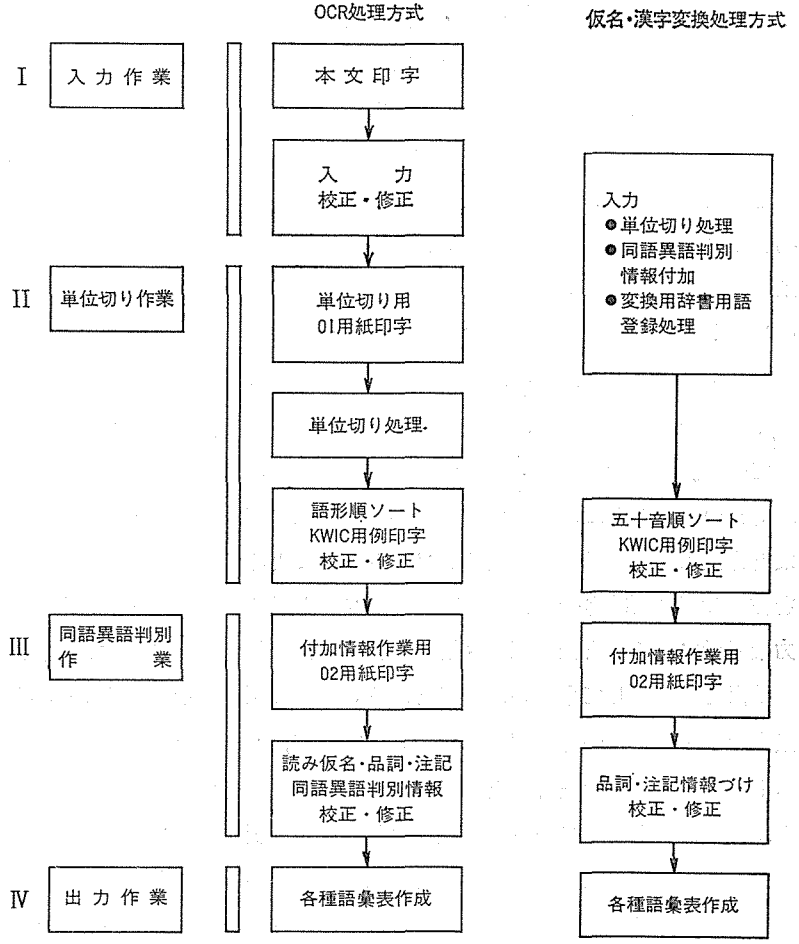

図 5　OCR方式と仮名・漢字変換方式の比較

になる。また，ソート処理において現在行われている文字列への一括処理は特殊であるといえる。

以上の各点を言語処理用ソートプログラムに応用した場合，文脈のなかで変化する語の意味・品詞ごとに，キー設定条件は異なることになり，従来のキー項目の設定方法に比べ，より自由度の高い指示機能が要求される。たとえば，レコード中の任意の項目に対するキー設定と，語頭・語尾・昇順・降

順，キー項目の優先順位の変更・解除などの機能である（キー機能については，データと辞書間の対応規則としての見方ができ，再検討の心要があるが第一次検証ではこのまま使用する）。この処理は，ソートプログラムの固定化されたキー項目の指定に，動的指示プログラム開発の有効性を示唆しているものと思われる。ソートキーの動的指示は，データ自身が均一特性を持つ場合をはぶき，データ階層または特徴群単位にキーを指定し，編成効果を作業に反映させる処理になる。しかし現行処理で，これらの機能をメーカ提供のソートプログラムで対応させるには，データ配列単位にキー指示と配列後の併合が必要になり，多数回のデータ処理が入ることになる。

キー項目の動的指示は，実務段階で OCR 02用紙で使用することを前提にしている。02用紙処理における出現形の読み仮名，品詞情報づけ作業は，それぞれの単語ごとに，前後の語環境からの意味の確定，語頭・語尾配列順序など作業に与える効果が異なる。ソート処理によるデータ配列順序の有効性を作業効率に反映させるためには，事前ソートがデータ特性・構造を知る強力な手段になる。また，データ構造の把握は，大量データに対する必要部分の抽出処理に対しても重要な機能となる。動的分類基準を与えるキー項目の設定は，辞書形式からの引用を想定しているが，これによって分離処理用リスト，仮名・漢字変換辞書との三種のリストの総合化の道がひらかれる。さらに，レコード中のソートキー項目の辞書からの引用は，ソート対象レコード長の短縮，物理的レコードを規定したキー指示から独立させる。動的キー指示の利用法は，基本実験の終了を待たなければならないが，ホスト・コンピュータ側での並行処理によるソート時間の短縮，データベースへのデータ構造別記録，データベースから磁気テープへのデータ排出処理に有効と思われる。

並列処理はホスト・コンピュータの記憶域管理方式の特性に制限を受ける。しかし，記憶管理が動的である場合は，キー単位ごとに実行領域を割り当てることができ，疑似的並列処理によるソート時間が短縮される。同様に，データベース処理も，入力データの特性にそった事前ソートが必要であ

るのに対し，複数のソートキー指定による単一処理が可能になる。また，データベース中の構造化されたデータを標準形式で磁気テープに出力するさい，データ特性の逆指定による分類処理によって，併合処理を省略できることになる。動的キー指示によって得られる，これらの機能は，従来のソート処理に比べ，多様化を進めるためにも不可欠である。

4.3 辞書の統合とデータ処理の基準化

国研における機械処理用辞書は，漢字に対する外字表現と解読用として，単語用は五十音配列用の理論コード生成用に開発された。基本配列は，漢字用が部首順・代表音訓順の二種の配列法を，単語は読み仮名による五十音配列，または漢字の代表音訓順を採用している。その後漢字用辞書は，高速漢字プリンタの導入，JIS コードの採用によって，複数コードの変換用に拡張されてきた。このコード変換用辞書は，データを長期的に利用するためデータ内容を保証する重要な意味を持つことになる。また，データをバックアップするため，辞書には国研外字表現形式をメタコードとして位置づけ，検字用辞書への対応用中間コードの役割りを与えてある。ここで検字用辞書は，市販されている漢和辞書中，最大のものと，最小と思われるものを選び，ともに辞書で使用されている検字番号を利用できるようはかっている。一方，コンピュータメーカから提供される辞書類は，国研で開発されたものと同類の辞書が用意されているが，日本語入力用に使用される仮名・漢字変換用辞書は，五十音配列用理論コード生成辞書と共用できることが多い。これら辞書類は，今後のコンピュータ処理の中心的役割を果たしていくものと思われるが，単語の読みを利用した五十音配列用辞書は，漢字の代表音訓に比べ，より自然な語配列が可能になる。本システムにおける基本処理と省力化もこれら辞書に負うところが多い。なお，辞書を使用する場合の効果は，次の二点が考えられる。

1) データ処理の基準化。

2) 辞書の使用による自動処理化。

国研で使用してきた機械辞書は，コード変換用，意味分類用を主な目的に

していた。しかし，頻度表では最終結果を対象にするため，調査過程で得られる中間情報は無視されてきた。二次情報は，もとになる一次情報から作成されるとすると，精度は，一次情報の性質・特性・調査目的によって大きく変化する。中間情報を無視したデータ利用は，結果の精度を保証する場合問題があることになる。度数をはじめとする中間情報は，操作のための手続き情報として，データ精度の確定に主導的支援を与える。その点で，データ処理効率をあげるためには，既知のデータをもとにした統計値の引用が有効であり，既知データから任意の段階の統計値を得られることが望ましいことになる。このことから，情報の辞書からの引用には，辞書項目は既知データの調査経過を示す詳細値と，調査結果を含んでいることが必要である。辞書を中心とした調査は，結果を辞書に還元する操作を挿入することによって，辞書と入力データの整合性を高め，辞書の並行保守を容易にする。

以上の点から，辞書に対する基本事項は，外部利用者に対する資料としての性格，調査結果・経過の統計処理可能な環境の維持，人間—機械系のインタフェースとしての役割を明確にすることが必要である。さらに，一次資料から二次資料を作成する過程を利用者に正確に伝える手段と効果を検討しなければならない。一次資料の短絡的使用を避けるためにも留意すべき点である。そのほか，仮名・漢字変換と同語異語判別情報は，ともに仮名表記と漢字表記の相互参照によって処理させるため，入力データと辞書項目は，常に一対一に対応していなければならない。また，五十音配列用に使用する読み仮名は，変換のさいの一次入力情報を利用するため，辞書項目のすべてに読み仮名が付加されていることが必要になる。入力データと辞書項目が一致しない場合，辞書に登録されているすべてのデータへの対応が不能になり，その後の処理に問題を残すことになる。辞書は，利用者に処理中のデータ内容を知らせる重要な手掛かりを与える。

現在，新コンピュータへの移行処置の中で，漢字辞書類の総合化を行っている。統合は，コード変換用，表記関係辞書，メーカ提供の辞書との項目の照合，市販辞書の検字番号の登録の各作業である。最終目標は，漢字・単語

辞書の結合にある〔文献4〕が，一次作業で収容を予定している項目は以下の通りである。

1) 漢字テレタイプ用盤内・盤外（外字）コード
2) JIS コード・区点情報
3) 日本電気・日立コード
4) 大漢和辞典・新字源・大字典検字番号
5) 部首情報・総画情報
6) 当用漢字・常用漢字・教育漢字・人名漢字識別情報
7) 読み仮名
8) 雑誌九十種・新聞用字・教科書調査出現度数

以上の8項目は，コード変換に対するもの，漢字属性，調査結果の三種に分けられる。コード変換用辞書は，蓄積されている各種データの新コンピュータへの移行と継続利用のための手段を提供し，属性はデータ配列・検字用に使用される。度数は，国研で調査された漢字出現度数の結果である。特に度数は，国定読本の用語調査における分離処理導入のもとになった。その点で，用語・漢字辞書についても見出し語に付加された度数は，出現形ごとの度数とともに「偏り度」を示す基本的な情報となる。出現形の度数は，集計過程で得られる補助情報として統計処理のさい重要である。

4.4 データ・プログラムの仮想結合

最近のインテリジェント端末による分散処理の普及によって，OA (Office Automation) 用プログラムの利用が盛んになっている。しかし，OA 用に開発されたプログラムは，それぞれ独立しており，本来同一思想のもとで設計作成されていたにもかかわらず，ファイル間の互換性，付加のフロッピーディスクの容量不足による機能低下など，パッケージ間の整合性に問題が多い。また，オンライン化されたシステムでは，ホスト側と端末側プログラムが重複していることも少なくない。プログラムの二重性は，データ・プログラムの互換性，操作法の継続性，マニュアル類の利用，辞書，コードにかんし別々に運用管理しなければならない。この点でOA用プログラムは，非

専門家の利用を前提に開発されていたにもかかわらず，暗に専門家による保守・管理体制を要求していることになる。利用者は，開発段階にあるOA用プログラムを使用するうえで不安定な状態に置かれているといってよい。利用者への問題を解決するためには，多様化し発展段階にあるOA用プログラムを，ホスト・コンピュータとの統一思想のもとで再編成する必要があると思われる。

再編成は，端末装置の機能を中心としたホスト・コンピュータのブラックボックス化であり，当面の改善方法として有力な手段である。この対応は，データ，ファイル，機械処理用辞書，プログラム，オペレーションの統一的管理を可能にし，ホスト・端末プログラムの双方に相補的かつ仮想化による一元化を進める。データおよびプログラムにかんする仮想結合の与え方は，端末側で作成されたデータ・プログラムの分散処理〔文献4〕の一形態となるが，コンピュータ利用者の負担を軽くする。しかし，ホスト側プログラムは，端末装置に比べ，より大きな機能を持っており，差を埋める手段として端末側で使用するコマンド機能の補足が必要である。ホストと端末側機能の複合化による疑似的拡張〔文献3〕は，端末側でも独立して保有すべきと思われるが，コマンドの疑似的統合機能を含め，ホスト系のコマンドとの一体化をはかるべきである。

プログラムの疑似的結合は，プログラム間の回線使用の自動切り替えをともなうが，データ転送のさい端末の利用状況とスケジューリング調整の最適化によって回線使用の効率化が期待できる。ローカルエリアネットワークの利用拡大とともに回線の問題は，システム全体の効率化に影響する。プログラム間の疑似的結合は，プログラムがディスプレイ上の各情報を通して密結合を持つと同様，回線を通したプログラム間の結合も，データを介した疎結合状態を持つのと同様の効果がある。これは，両処理の疑似的結合が相似形式であり，ともに同一効果が得られることを示している。プログラムの結合用パラメータは，各プログラム機能を利用目的別に再調整しディスプレイ上に表示する。パラメータは，プログラム機能の抽出，応答形式の取りまとめ

の過程で最適化される。

そのほか，光ディスクについては，関係形式データベースのバックアップ用装置として，データベースは，端末用簡易表操作プログラムの仮想ファイル化を計画している。ファイル，データベース，光ディスク間の仮想化は，端末用簡易表操作プログラムと関係形式のデータベースの間で一部実現しており，新コンピュータ導入の一環としてメーカから提供される予定である。OA 化における仮想化の対象は，利用者の多い日本語ワードプロセッサについても同様の対応が必要であり，端末側の代表的プログラムの一元化の方向を明確にすべきと思われる。

5. おわりに

国定読本の用語の調査について，基本的な考え方と問題点に対する改良案を述べた。限られた期間・経費で調査を進めるためには，費用対効果比，システム効率を最大にすることが急務である。また，このことが調査完了のかぎとなっていた。システムは，これらの条件を満足させることを前提に設計された。設計時の条件の一つは，データ・プログラムともに外注によって達成できる見通しがついた。条件の第二は，ターンアラウンド処理と，後処理方式の導入によって，読み仮名づけ，単位切り，品詞情報づけの効率化で対応可能になった。これらの処理は，特定用語の分離一括処理，総合辞書による処理の基準化に負うところが多い。現在，システムの改善を行うため仮名・漢字変換処理による単位切り，同語異語判別処理への効果，ソートキーの動的指示，仮想ファイルの拡張など，新方式導入の妥当性について検討を進めている。実用化には，まだ解決しなければならない問題がある。

一つはソフトウェア機能に利用者の目的にそった拡張・変更への対応がなされていないことである。この問題は，提供されるソフトウェアが単一処理を目的に開発されており，端末・ホスト双方ともにプログラムは，個々に独立して使用される前提で開発されてきたことが原因となっている。インテリジェント端末・分散処理下でのプログラムとしては，見直しの時期にあると

思われる。統一をはかるためには，プログラムの部品化と部品の結合による拡大機能を持つプログラムを作成し，疑似的結合の標準インタフェースの設定，処理結果と統計値の開放が必要になる。

最後に，OCR 方式のシステム設計にあたって，対象データ量を把握しておくことは，外注費用の算出など予算配分上重要である。しかし，第1期の調査は，出版のための作業が行われており，事前調査による詳細データが得られにくい状態にある。そこで，第1期作業の経験とサンプリングによる数値をもとに，第3期以降のデータ量の予測を行った。

表2に示した，従来の用語調査方式で必要とするデータ入力量は，次のようになる。漢字を含んだ表記は，40%約20万語と予想されるため，入力原稿上の文字数16字に対し 320 万字となる。以下同様に計算し記号系をはぶいた総計では約 604 万字である（ただし同語異語判別対象語と仮名表記語は重複している)。ターンアラウンド方式では，入力データを約150万字と見積ったため，両システム間の差は 454 万字になる。必要とする金額は，単価によって変動するが，差と同程度のひらきは出るものと思われる（OCR 用紙価格が高いため 4 万枚で約80万円の増額になる)。

各作業の必要日数は，単位切り処理に 94～135 人日（1人1日35～50枚処理すると仮定）程度と推定される。また02用紙では，同様の算定規準で 480～685 人日かかると予想した。総日数は，574～820 人日である。作業は，修正処理のための諸作業を含んでいないが，29 カ月～41カ月（20 日/月として計算）を要することになる。これは，2人の作業者で15カ月から21カ月で対応できることを示している。OCR 方式は，従来方式による調査期間・費用に対し，推定50%～70%が短縮でき，本システムの設計目標は達成されたと思われる。また，試案 4.1「仮名・漢字変換処理の拡張と応用」で述べた方法による場合，品詞・注記処理は残るが，漢字を含んだ語への見出し語づけ処理の40%，20万語（表2）にあたるものを02用紙からはぶくことができる。これは，分離処理と合わせて全体で約60%，30万語が対象になる。試案の実用化は，用紙・期間の短縮ともに漢字部分の対応で十分効果は得られる

ことになる。

以上の各数値は，算出基準が経験によっているため，正確な推定値を示していない可能性がある。システム設計に必要な値は，入力データ総数の５％程度の誤差は許されるため，あえてこの数値を基本に算出した。

〔謝辞〕システム設計にあたり，メンバの高梨信博氏には第１期からの各種情報をまとめていただいた。記して謝意を表する。なお本稿は，文部省科学研究費助成「国定読本の用語の研究」（一般研究Ａ研究代表者飛田良文）の一部である。

(1985. 6. 27)

参考文献

1) 斎藤秀紀 (1976)「言語処理におけるターンアラウンド・システム」『電子計算機による国語研究Ⅷ』（国研報告59）63-111。
2) …… (1980)「分散処理システムへの試み」『電子計算機による国語研究Ⅹ』（国研報吾67）73-88。
3) …… (1983)『分散処理による大量日本語処理の効率化に関する研究』（昭和57年度科学研究費補助金　一般研究Ｃ研究成果報告書）。
4) …… (1984)「会話処理によるファイル管理情報の生成」『研究報告集５』（国研報告79）145-162。
5) ……他 (1984)「日本語とパーソナルコンピュータ」『電子通信学会誌』Vol. 67, No. 4, 57-103。
6) …… (1985)「漢字コードの拡張法に関する試案」『研究報告集６』（国研報告83）57-103。
7) 国立国語研究所 (1952)『語彙調査―現代新聞の用語の一例』（国研報告２）。
8) …… (1953)『婦人雑誌の用語』（国研報告４）。
9) …… (1957)『総合雑誌の用語（前編）』（国研報告12）。
10) …… (1958)『総合雑誌の用語（後編）』『国研報告13）。
11) …… (1962)『現代雑誌九十種の用語用字（第一分冊総記・語彙表）』（国研報告21）。
12) …… (1963)『現代雑誌九十種の用語用字（第二分冊漢字表）』（国研報告22）。
13) …… (1973)『電子計算機による新聞の語彙語査（Ⅳ）』（国研報告48）。

14) …… (1983)『高校数科書の語彙調査』(国研報告76)。
15) …… (1976)『現代新聞の漢字』(国研報告56)。
16) …… (1983)『電子計算機と国語研究』。
17) 稲永紘之他 (1982)「日本語処理のための機械辞書」『情報処理』Vol. 23, No. 2, 140-146。
18) 豊島正之 (1982)「文献学的研究の為の索引を電子計算機で作る上での諸問題に就て」『言語研究の中の計算機』(計算機利用言語学研究会編東大) 41-52。
19) 林四郎他編 (1984)『例解新国語辞典』第1刷 (三省堂)。
20) 小川環樹他編 (1985)『新字源』230版 (角川書店)。
21) 諸橋轍次編 (1971)『大漢和辞典』第3印刷 (大修館書店)。
22) 上田万年他編 (1971)『大字典』第56版 (講談社)。
23)『OCR ターンアラウンド処理基本設計書』(国研-FS-001)。
24)『OCR ターンアラウンド処理 OCR 帳票案』(国研-EM-010)。
25) 石綿敏雄 (1984)「情報処理における最適化表現」『正書法・造語法の資料と研究法—日本語の正書法及び造語法とそのあり方 (中間報告集)』(昭和58年文部省科学研究費補助金特定研究(1)研究課題番号58107016研究代表者林大) 90-102。

表 2　第1期からの文字出現予想（延べ語数）

第1期データ 推　定　値	漢字表記 40% (20万語)	仮名表記 50% (25万語)	記　号 2% (1万語)	同語異語判別対象語 8% (4万語)
平均単語長	3 字	3 字	1 字	3 字
読み仮名長	6	—	—	—
品詞・注記	4	4	—	4
編集記号	3	3	1	1
計 推定全文字数	16 字 320万字	10 字 250万字	2 字 2万字	8 字 32万字

表 3　品詞・注記の略号と番号

品　詞	略　号	番　号	品　詞	略　号	番　号
〔名詞〕			接続助詞	接助	64
課名	課名	01	並立助詞	並助	65
話手名	話手	02	準体助詞	準助	66
人名	人名	03	終助詞	終助	67
地名	地名	04	間投助詞	間投	68
〔予備〕	0 5	05	〔動詞〕		
	0 6	06	四段	四	70
	0 7	07	五段	五	71
名詞	名	08	上二段	上二	72
代名詞	代名	09	上一段	上一	73
形状詞	形状	10	下二段	下二	74
副詞	副	20	下一段	下一	75
連体詞	連体	30	カ行変格	カ変	76
接続詞	接	40	サ行変格	サ変	77
感動詞	感	50	ナ行変格	ナ変	78
〔助詞〕			ラ行変格	ラ変	79
格助詞	格助	61	形容詞	形	80
副助詞	副助	62	助動詞	助動	90
係助詞	係助	63			

表 4　出現上位の同表記語

No.	語	使用度数	同表記異語	かな／漢字表記数
1	は（助）	1552 (1593)	は*	1
			葉	34/6
2	て（助）	1495 (1518)	て*	1
			て＋	1
			手	5/16
3	の（格助）	1438 (1646)	の*	1
			野	0/4
			の（準助）	203
4	に（格助）	1336 (1604)	に*	1
			に＋	1
			二（課名）	0/25
			二	0/7
			荷に（接助）	2
			似（似ルのー）	6
			に（ナリのー）	11
			に（ダのー）	55
				157
5	を（助）	1092 (1096)	を*	1
			尾	3
6	た（助動）	1003 (1028)	た*	1
			た＋	1
			他	0/8
			田	1/14
7	が（格助）	788 (861)	が*	1
			蛾	3
			が（接助）	69
8	ます（助動）	641		
	ましす			
	ますれ			
	ませ			
9	と（格助）	559 (840)	と*	1
			と＋	1
			と（接助）	101
			と（並助）	176
			戸	0/1
			斗	0/1
10	だ（助動）	441		
	だっら			
	だでなら			
	なに			
その他	も（助）	389 (390)	も*	1

*：ひらがなドリル　＋：文字　－：活用形

カッコ内の数字は仮名・漢字表記数を加えたもの。