国立国語研究所報告94　研究報告集 9 (1988)

漢字情報データベース

斎 藤 秀 紀

SAITO Hidenori : Kanji Information Data-Base

要旨：本稿は，国立国語研究所における機械辞書の歴史的な背景，各種漢字調査情報と市販の漢和辞書情報の結合によって期待できる利用上の相乗効果，機械辞書のデータベース化と項目内容（見出し漢字：9731字，付加情報：40項目）の検索方法について述べた。また，データベース化された漢字情報は，調査情報の履歴管理，蓄積データに対する索引機能，共通インタフェースの多様化と情報接点の拡張，コンピュータ処理費用の軽減にも有効であることを示した。

その他，JIS 2バイト系の拡張計画に対し，現在すでに拡張漢字として使用している漢字コードとの間に問題が生じる可能性を指摘した。同様に，市販漢和辞書の CD-ROM (Compact Disc-Read Only Memory) 化は，日本語の外字処理の軽減が期待される反面，字形の相違が情報交換上の問題を広げることについてもふれた。
キーワード：機械辞書，漢字辞書，CD-ROM，漢字情報データベース。

Abstract: This paper discusses the Kanji information data-base developed in The National Language Research Institute. It begins with a brief history of its development and covers in detail the process of machine dictionary data-base compilation, the effect of combining the research information with the published Chinese Japanese character dictionary information, and the search methed for the contents (9731 Kanji with 40 additional items).

It is pointed out that the kanji information data-base helps control the research information, enhances the index function, expands the mutual interface as well as the information contact, and reduces the cost of computer processing.

In the course of discussion, some problematic aspects are pointed out. First there could be some discrepancy between the JIS (2 byte) codes and the expanded modes when the former is to be expanded.

Similarly, though the processing of extra symbols in Japanese maybe eased greatly, the difference in the shape of Kanji may create problems in exchange of information.
Key words: machine dictionary, kanji dictionary, CD-ROM, Kanji information data-base.

1. はじめに

国立国語研究所（以下国語研）で作成されたコンピュータ処理用の機械辞書には，三つの流れがある。第一は，装置に係わるコード・外字登録を管理するコードブックから派生したもの。他の二つは，漢和辞書情報と漢字調査及び用語調査の結果を統合したものである。国語研における辞書の作成目的は，これらの辞書の一部または複数を使い，次に示す利用を想定したものであった。

1) 漢字データに対する付加情報の標準化。

2) 漢字辞書・調査データに対するデータ管理。

3) 漢和辞書情報及び調査データのデータベース化と索引化。

4) 漢字処理の自動化を進めるための基本情報の収集。

漢字データベースを作成するに当たり，統合化の対象となったファイルは，(1)コードブック：漢字テレタイプライタの盤外字コード，高速漢字プリンタ，日本語入力装置，コンピュータ内部コード，JIS コードの五種を統合したもの，(2)表記テーブル：JIS 漢字に対し，大漢和辞典・大字典の検字番号，新字源の検字番号・読み・画数・部首の各情報を基本に，雑誌九十種，現代新聞の漢字調査の順位・度数情報を付加したもの，(3)調査データ：中学教科書・高校教科書の調査結果から漢字の度数を再集計し付加したものの三種である。ここで，総合辞書に登録した見出し漢字は，日本電気から提供された印字可能な漢字パターン9731字である。

基本ファイルは，JIS・各装置コード・旧漢字テレタイプコードに対応する外字コード（言語計量研究部・第三研究室作成），漢字表記テーブル及び漢字調査データ（第二研究室作成），中学・高等学校教科書調査データ（第一研究室作成）を使用した。また，辞書情報の総画数・部首・読みのチェックは，日本電気・日立製作所で提供しているコンピュータ処理用の機械辞書を利用した。

2. 漢字情報の総合化の背景

2. 1 国語研の漢字辞書の歴史

機械処理を目的にした漢字辞書の考え方は，昭和41年（1966）にコンピュータと漢字入力用テレタイプライタを導入したときに始まる。最初の辞書は，漢字を部首順または五十音順に配列するための対応表である〔文献1, 2〕。これと同時に，漢字テレタイプライタの盤外字に対するコード化・解読用ハンドブックの作成があった〔文献3〕。外字コードは，漢字テレタイプライタの盤内字２文字を組み合わせ，大漢和辞典の検字番号に対応させたものである。外字コードを大漢和の検字番号と対応させたことは，大漢和辞典をそのままコードブックとして使用できることを意味する。また，外字コードは，部首順配列のための理論コードとしても使用できることになる。

以後，昭和50・55年（1975, 1980）に高速漢字プリンタ，昭和55年（1980）の日本語入力装置の導入にあわせ対応表の拡張を行ってきた。使用できる漢字数も当初の2110字からJIS対応の6353字に拡張され，ACOS-S550の導入とともに9731字が使用可能になった。この間，コンピュータの切り替えが２回行われている。その後，外字表現用の理論コードは，JISコードの採用後に補助的なものになったが，新出漢字の追加作業の中でインタフェース用情報として整備されてきた。以上が第一期の漢字辞書の流れである。

第二の流れは，表記テーブルを中心とする辞書である。表記テーブルは，大量用語用字調査の効率化のために，標準化された情報付加用辞書として作成された。対象漢字は，JISに限られているが雑誌九十種・現代新聞の漢字調査で得られた度数〔文献4, 5〕，市販辞書から大漢和・大字典・新字源の検字番号〔文献8, 9, 10〕，新字源から読み・総画数・部首情報，国の政策で規定した当用・常用・教育・人名漢字の識別情報を含んでいる。第三の流れは，中学及び高等学校の教科書調査〔文献７，６〕から漢字の出現度数を整理したものである。

これらの流れをまとめると，漢和辞書は，市販の漢字辞書の定性的な面に重点をおいた資料からのもの，調査データなど定量的な情報を対象に収集したものの二つに分類できる。前者は，静的であり後者は動的特性をもつデータであると言える。以下，静的情報と動的情報を総合した場合の相乗効果，

将来の外字処理の方向，共用データの索引，外部データとの接続インタフェースの役割について述べる。

2．2　漢字情報の総合化に対する基本的な考え方

コンピュータを利用した大量調査を行って20年が経過した。調査で収集された用例・用語用字データは，冊子体・マイクロフィルム・磁気テープなどに保存されてきた。一方，個別に研究され蓄積されたデータは，媒体・記録形式・データ構造，コードに配当された字形の違いがデータの結合を妨げ，経年調査で得られたデータの比較に問題があった。従来，メーカから提供されている漢字パターンは，JIS で規格化されているものが多く，単位の揺らぎがないことから結合には問題は少ないとされていた。しかし，JIS の改訂，個別追加などが結果として字形の時系列的な管理を利用者に課していた。さらに，配列と分類基準が統一されていないため，データ間の照合に必要なキーと付属情報の共通化にも障害となっていた。データに付加する情報の多様化は，蓄積されるデータの重複部分を増加させ，磁気テープ，ディスクの利用効率を下げることになる〔文献12, 13, 14〕。

これら照合用の情報は，インタフェースを通してキーの形で具体化されるが，大量データの共有化とデータ交換がすすむにつれ，キーとそれを支える付属情報の管理が重要な課題になる。不特定の利用者に対するデータ提供に伴う管理は，字形の揺らぎの許容範囲をどの程度付加情報で補うことができるか，データを送る側・受け取る側の双方でどの程度標準化できるかにかかっている。各項目間の関係は，対象データから未登録項目へフィードバックさせ細分化する操作，細分化された情報を対象データへ還元させ精密化していく循環過程が背景にあり相補的であることが要求される。

個別データを結合し蓄積していくためには，対象データ間の整合性を調整できる多様なインタフェースを設定することが管理の上で必要である。調整機能とは，インタフェースを通して行われる二次情報の一次情報への還元操作である。二次情報とは，辞書項目に対する属性情報と，索引化された辞書に接続される外部データの二つである。二次情報が一次キーを補正・拡張す

る操作の中でキーの精密化に使用できることは，キーは時系列的な履歴管理とデータ間の整合性を補正する，動的インタフェースとしての機能が暗黙に与えられていることになる。これは，個別データの総合化，データの長期保存の場合のいずれにも当てはまり，理論コード設定のさいの基本事項として重要である。

一方，山田は，システム開発の効率化にデータ中心のシステム設計の導入が有効であることを述べている〔文献11〕。システム設計の目標には，(1)データの部品化と情報生産の確立，(2)システム構造の確立，(3)システム内部統制の確立，が重要であることを指摘している。さらに，データを資源として統制する利点に，次の三点をあげている。

1) データは，プログラムより安定しておりコンピュータ処理から独立して設計できる。

2) データの重複を排除できデータの標準化をすすめやすい。

3) データの評価と価値の測定が容易である。

本稿で述べる辞書の総合化の目標は，(1)大量データの長期保存とデータの統合管理，(2)データの統合化による重複部分の排除，(3)共用データの辞書化と辞書の共通項目の分離管理である。システム開発の効率化とデータ保存問題は，データの部品化と情報生産方法の標準化，資源の統制などと目的は同じになる。

この類似点は，データの総合管理がシステムの大型化に伴う開発費用の削減，大量データの保守・管理の効率化にも対応できる可能性を示している。総合管理とは，調査データからは漢字の利用実態の把握を，辞書からは事前に整理された漢字情報の引用など，異なる情報源を統合した漢字の総合的なデータベース化の方向づけである。

2. 3 インタフェースとしての外字コード

外字コードを理論コードとして使用する利点は，内部コードよりコード化できる対象が広いことである。外字コードは，設定のさい処理できるすべての漢字に，特定の目的を前提にした統一的な見方を反映できる。統一的な見

方のもとでの処理は，次のような利点がある。

1) 理論コードは，漢字の物理的コードと独立に設定でき，データの入力・保存処理をメーカの漢字パターンの提供能力から自立できる。また，入力・出力・保存コードの分離は，コンピュータ処理に対するコード変更の影響を軽減させる。

2) 物理的コードと理論コードの分離は，入力に対する人間・機械間の最適化を進め，コードに対する統一的な見方を反映したモデルを通して情報交換用インタフェースを設定できる。

3) 理論コードには，二次情報と一次情報の間に変換機能を埋め込むことができ，辞書の再編成の範囲を広げる。

データを長期間保存する場合，データに対して常に検索・加工・印字手段が保証されていなければならない。しかし，使用するコードは，JIS 漢字においても5年ごとの見直しがあり，利用者による追加のほかメーカ提供の文字セットも装置によって異なることがある。JIS 規格が情報交換を目的に設定されたとはいえ，使用する装置との間で事前調整が必要になる。

辞書の利用は，装置の更新に伴う内部コードの変項，JIS 規格の改訂・プログラムの変更など，コンピュータの利用環境の変化からデータを独立させる。一方，メーカから提供されている漢字パターンは，出力処理の範囲を決める。これは，出力処理の制限が入力データに影響を与え，入力段階で調査者のデータに対する情報の一部を損なうことを意味する。対応には，入力・保存コードを出力コードから分離させる方法がある。物理データに対するメタコードの設定である。理論コードの使用は，疑似的に取り扱う文字セットを拡大させ，目的に応じた統一的な配列・分類の基準化をすすめる。さらに，物理コード及びコードに配当されたJIS漢字は，標準的なインタフェースとしての役割が強いのに対し，疑似コードはそれを細分し理論化したモデルの結果として位置づけできる。

物理コードと理論コードの分離は，人間・機械の双方に適した疑似的コードを設定できる。双方の系における最適化とは，その系のモデル化と解釈の

具体化である。ここで，疑似コードの役割は二つある。一つは，最適化されたモデルを通して見る，辞書の拡張基準と分類の精密化であり，他の一つは複数データを接続するためのインタフェース機能である。モデルで仮定したコードは，具体化される過程で明確に基準化されインタフェースになる。その点でモデルの精度は，見出し漢字を説明する属性情報の数と質に依存しているとも言える。

また，属性情報は，キーの意味の広がりを表すが，検索時には意味の絞り込み条件に使用できる。例えば，漢字データベースを使い字形の類似度を調べるとき，どの属性情報が利用できるかと言うことである。利用者が辞書を使ってデータを検索するとき，検索用キーは利用者に対して多様な接点をもつこと，また検索の結果が成功しなかったとき，二次情報から類似したキーを探すため逆引きできることが必要になる。この操作は，外部データまたは説明項目から得られた情報は，見出し語を細分化するための情報として，キーの補充と接続条件の拡張に利用できることを意味する。

二次キーから一次キーへの情報の還元は，辞書自身の中で対応できることが前提にあるが，利用者から見た辞書は，機械処理用，人間・機械系で使用するコードブックのいずれも，二つの情報の橋渡しを行う変換機構としての役割をもつことになる。人間と機械との間のインタフェースに漢和辞書が利用できることは次の利点がある。

1) 辞書は，普及度が高くコードブックとしての利用に抵抗が少ない。

2) コードブックの標準化と作成労力の省力化ができる。

3) 市販漢和辞書の検字番号は，外字入力方法の標準化をすすめる。

4) 辞書の索引は，多様な検字手段と外部情報との接点を多様化できる。

5) 将来，漢和辞書の CD-ROM 化により外字処理を減少させる。

2. 4 CD-ROM 化された漢字辞書との結合

CD-ROM は，長時間の音楽再生用として開発されたが，ディジタル・コードを記録できることから，コンピュータの補助記憶装置としての利用が注目されている。CD-ROM は，小型・軽量であり540メガバイトと大容量のデータが

記録できる。また,図形・音声・コードなどの情報を大量に安く出版でき,パーソナル・コンピュータを利用した情報検索・加工が容易であるといった利点がある。現在,特許情報の提供,電話帳,辞書など主に出版関係で利用が計画されており,磁気テープにかわるコンピュータ可読媒体として注目されている。

CD-ROMのコンピュータ利用には,漢字パターンが指定の装置上で表示・出力できることが前提にある。この前提に立てば,コンピュータ処理におけるデータ入力・出力処理をJIS規格以外の世界に拡大させる。外字処理とコードブックの修正作業の事実上の解消である。CD-ROM情報の共通利用は,標準化の方向にあるが,記録するデータのインタフェース,データ間の整合性を保証する物理的コードと字形の標準化が必要である〔文献15〕。

特に,字形については,辞書・JISともに揺らぎがあり,漢字辞書のデータベース化を行う前に目視による確認が必要になる。字形の同一性を基準化する方法は,事前に字形を整理しシソーラス化する方法がある。CD-ROM化された漢和辞書の利用は,疑似的に表示可能な標準文字セットを増加させ,パーソナル・コンピュータとの結合は,検索・加工の容易性とともに柔軟な人間・機械間のインタフェースを確保できる。この二つは,シソーラス作成の有効な道具になりうる。

CD-ROMを使用したデータ提供システムは,処理の分散が基本にある。処理の分散は,データの分散化をすすめデータの一機関への過度の集中を防ぐ。さらに,(1)データ破壊への保安,(2)研究者に対するデータ利用機会の保証,(3)国内外の研究者への同時サービス,(4)関係する資料・文献を収集している機関との資料収集の調整,(5)原資料の収集量の物理的限界への対応,(6)関係資料の把握とメタ情報による二次資料化など,情報の作成・運用上の問題を軽減する。

しかし,CD-ROMを使用した大量データの交換は,大量データを収容できるが故に,一機関で処理できる文献・資料には量・対象とともに限界がある。入手できる情報と媒体の多様化は,情報加工の手段としてのコンピュータ化に時間・費用・人材が確保できないことが予想されるためである。その点で,

今後のデータ利用の形は，複数の組織で作成された二次データの有機的な結合が不可欠である。

また，データの信頼性については，それぞれ専門とする研究者または研究機関ごとの処理の分担が，結果として情報の質を高める。CD-ROMの大量データの記録能力は，データの過度の集中化を部分的に解決し，小規模のデータ保存媒体としても，データ交換媒体としても，費用・時間の点で要求に十分答えられる。以下に，CD-ROMの特徴とされる内容を示す〔文献15〕。

1) 記憶容量が大きく(540MB)，他の媒体に比べ蓄積費用が安い。

2) 傷・ほこりに強く常温での保存と，データの長期保存に優れている。

3) 読み取り専用機能は，データの改ざんを防止できる。

4) コード・イメージ・音声の併用記録ができ大量複製が容易である。

5) 媒体は，小型・軽量であり保存場所をとらない。

6) パソコンを使用したデータ検索・加工・データ交換が容易である。

7) データ作成と利用の非対象性から処理別に最適化が可能である。

3. 漢字辞書項目の概要

本節では，漢字総合辞書に収容した情報の内容を説明する。総合漢字辞書は，ACOS-S550上での利用を前提に作成されている〔文献16〕。辞書項目の項番1「見出し漢字」は，辞書の見出しとキーの役目をもつためACOS-S550の内部コードを使った。それ以外の項目も，日本電気の2バイト系漢字コード（JIPS(E)）で一文字を表している。2バイト系漢字コードは，数値情報も漢字扱いとなり，直接演算処理には使用できない。演算には，1バイト系の内部コードへの変換が必要になる。

コード体系が異なるコンピュータでの辞書の使用は，見出し漢字を目標とするコンピュータの内部コードに変換しなければならない。コードが辞書に登録されていない場合は，対応するコードを辞書に追加した後，その辞書をデータとしたコード変換を行う。この処理は，辞書の複数コードに対する管理機能を使うことになる。

各項目の先頭につけた番号は，表1（45ページ）で示したデータ項目との

照合用である。カッコ内の数字は，表1のデータ項目の始めと終わりのカラムを示し，コロンの後はその項目のバイト数である。なお，コードは，16進4桁の数値を漢字1文字2バイト系モードで表現した。その他は10進表現である。各項目の長さは，バイト長で表しているため，漢字モードでは表1で示した長さの2分の1の字数になる。検索結果の画面は，図1に示した。番号は，データ項目の内容で説明した番号と一致させた。

漢字辞書データ項目の内容

1) 見出し漢字・データ結合用キー(1-2:2) ホスト・コンピュータで処理するため，該当する漢字をJIPS(E) コードで表現したもの。JIPS(E) コードは，日本電気で規定した漢字コードでJISコードの1バイト系符号を対応するEBCDICコードで表現したもの。記号類を省いた漢字数は，G0領域（JIS：6349字），G1領域（拡張：3382字）の9731字である。非漢字を含めた字数は，基本文字7461字（漢字6349，その他の文字345，特殊文字108，罫線など659字），拡張文字4064字（漢字3382，記号530，変体仮名152字）の11525字である。

2) 区・点番号(3-10：8)「区」：JIS-C6226-1978で規定した2バイトコードで先頭のバイトで表される10進2桁の数字。「点」：2バイトコードで後のバイトで表される10進2桁の数字。区点は1から94まであり，第一水準は16-01から47-51，第二水準は48-01から83-94が割り当てられている。

3) 改訂情報(11-12：2) JIS-C6226-1978に対し1983年に改訂された漢字の識別情報に＊印を表示。

4) JIPS(E)コード(13-20：8) ACOS-S550で使用している漢字コード。見出し漢字に配当されたJIPS(E)コードを16進4桁で表現したもの。コードの範囲は，G0領域（2121から7E7E）及びG1領域（A1A1からFEFE）までの17672字分である。

5) JIPS(J) コード(21-28：8) ホスト・コンピュータで使用している漢字コードを外部記憶媒体へ出力するときに使用する外部表現コード。JIS

コードにメーカで追加した拡張漢字（G1 領域）を加えたもの。

6) 端末外部コード(29-36：8) 端末系（N6300-55N）から出力媒体へ記録する外部表現コード。端末系内部コードを EBCDIC 表現したもの。使用できる漢字数は，ホスト・コンピュータと同数である(項目1参照)。

7) 端末内部コード(37-44：8) 端末系（N6300-55N）コンピュータで使用している内部コード。1バイト系と2バイト系の切り替えに重みづけによる識別を行ったもの。

8) 漢テレ盤内字コード(45-52：8) 旧漢字テレタイプで使用している8進数4桁コード。

9) 漢テレ盤外字コード(53-74 22) 漢字テレタイプの外字表現用コード。外字表示記号◇と盤内字2文字の組み合わせで外字1字を表現したもの。データ接続用のインタフェースと統一配列用理論コードを兼ねている。コードは8進数を16進表現している。例）「愛」一文字を表現した項目内容〔04011215◇奥裁〕。

10) 日立コード(75-82：8) HITAC-M150 コンピュータで使用した漢字コード。JIS コード（8ビット中7ビット使用）の未定義ビットの先頭に16進（8080）を加えたもの。

11) 旧日電コード(83-90：8) NEAC-N7370高速漢字プリンタの漢字コード。JIS コードを基本に JIS の未定義ビットを0表現したもの。

12) 見出し部首(91-92：2) JIS-C6226 の字形索引で示された部首情報。ただし，定義されていない字は，字形索引で示されている見出し部首の次の漢字で代用した。

No.	部首と番号	代替漢字と区点番号
1)	丨（002）	个（48-04）
2)	疒（104）	疔（65-43）
3)	内（114）	禹（67-27）
4)	癶（162）	辶（77-72）

13) 部首コード(93-98：6) JIS の字形索引で使われている通し番号。康熙字典の部首番号。「一：001」から「龠：214」まで数字情報。

14) 画数(99-102：4) 見出し漢字に対する総画情報。新字源の親字につ

けられた総画情報。

15) 部首内画数(103-106：4) 見出し漢字から部首部分の画数を省いた画数。新字源から引用。

16) 新字源番号(107-116：10) 新字源の親字につけられた検字番号。見出し漢字が新字源にない場合，5桁の数字0を表示。下1桁は追加用の枝番号として使用。

17) 大漢和番号(117-128：12) 大漢和辞典につけられた検字番号。見出し漢字にない場合は，6桁の0を表示。下1桁は追加用の枝番号として使用。「ダッシュ」つきの検字番号は，枝番号「5」を記入。

18) 大字典番号(129-138：10) 大字典につけられた5桁の検字番号。見出し漢字が辞書にない場合は，5桁の0を表示。

19) 教育漢字1(139-140：2) 昭和33年(1958)に小学校学習指導要領で示された学年別漢字配当表(881字)に対する学年情報(1-6学年)。いわゆる教育漢字。

20) 教育漢字2(141-142：2) 昭和52年(1977)に小学校学習指導要領で示された漢字996字に関する学年配当漢字(1-6学年)。図1では学習漢字として表示。教育漢字に備考漢字115字を加えたもの。

21) 当用漢字(143-144：2) 昭和21年(1946)内閣告示の訓令。当用漢字表として示された1850字の識別記号（表内字は1，表外字は0)。

22) 当用漢字補正1(145-147：2) 昭和29年(1954)に国語審議会から出された当用漢字表補正資料による，当用漢字表から削る28字の候補の識別情報（該当漢字は1，非該当漢字は0)。

23) 当用漢字補正2(148-150：2) 当用漢字表に加える28字の候補の識別情報(該当漢字は1，非該当漢字は0)。この項目は項目22と重複しているが将来は統一する。

24) 常用漢字(151-152：2) 昭和56年(1981)国語審議会答申による常用漢字表の本表で示された1945字の識別情報(該当漢字は1，非該当漢字は0)。

25) 人名漢字1(153-154：2) 昭和26年(1951)内閣告示・訓令の人名用漢

字別表92字に関する識別情報（該当漢字は１，非該当漢字は０）。

26) 人名漢字２(155-156：2) 昭和51年(1976)内閣告示・訓令の人名用漢字追加表による28字に関する識別情報(該当漢字は１，非該当漢字は０)。

27) 人名漢字３(157-158：2) 昭和56年(1981)に新たに追加された54字に関する識別情報（該当漢字は１，非該当漢字は０)。

28) 新聞順位(159-166：8) 昭和41年(1966)発行朝日・毎日・読売三紙に関する漢字調査結果の漢字出現度数をもとにした順位情報（漢字数，延べ99.1万，異なり3213字)。

29) 雑誌順位(167-174：8) 昭和31年(1956)発行の現代雑誌九十種調査に関する漢字調査結果の漢字出現度数をもとにした順位情報（漢字延べ28万，異なり3328字)。

30) 新聞度数(175-182：10) 新聞三紙の調査から得られた漢字の出現度数。

31) 雑誌度数(183-192：10) 雑誌九十種の調査から得られた漢字の出現度数。

32) 新聞人名度数(193-202：10) 新聞三紙の調査で人名に使われた度数。

33) 雑誌人名度数(203-212：10) 雑誌九十種の調査で人名に使われた度数。

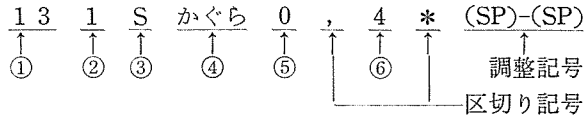
34) 新聞地名度数(213-222：10) 新聞三紙の調査で地名に使われた度数。

35) 雑誌地名度数(223-232：10) 雑誌九十種の調査で地名に使われた度数。

36) 読み別度数(233-988：756) 常用漢字表で規定された読み方度数。項目は，一つの読み単位に雑誌・新聞調査の出現度数を付加したもの。一データ項目長は36バイト，最大21項目。項目の詳細は以下の通り。

①1-4 バイト
項目番号
②5-6 バイト
0：常用漢字表内音訓。
1：特別な表内音訓
（雨に対する”あま”など)。
2：表外音訓。
8：表内音訓で出現度数が 0のもの。
③7-8 バイト
S：熟字訓・あて字。
空白：上記に該当しない読み。
9-36バイト（可変長)
④読み，⑤雑誌，⑥新聞の度数。

使用例



37) 高校教科書度数(989-998：10)　昭和49年(1974)度の高等学校で使用した教科書（理科・社会科など９教科）調査結果の漢字出現度数（漢字数延べ推定35万，異なり未集計)。

38)　中学教科書度数(999-1008：10)　昭和56年(1981)度の中学校で使用した教科書（理科・社会科など７教科）調査結果の漢字出現度数（漢字数延べ推定14万，異なり1770字)。

39)　余白(1009-1038：30)　予備の空白欄。スペース記号(4F4F)を挿入。

40)　音読み(1039-1058：20)　漢和辞書「新字源」にもとづく漢字の「音」読み情報。

41)　訓読み(1059-1118：60)　漢和辞書「新字源」にもとづく漢字の「訓」読み情報。

4.　おわりに

総合漢字辞書の作成によって，プロトタイプではあるが漢字データを統一的に利用できる環境を作った。その作業によって，データの分散と集中，双方の特徴を併せた相補的な利用法も明確になった。基本的な考え方は，漢字から単語辞書へ拡張できる見通しもついた。しかし，総合辞書の作成・整備は，データ入力方式に仮名・漢字変換方式をとる限り，保守・管理の一環として続けていく必要がある。辞書の管理は，結果の集約されたものとして実務的な作業の中で成果と結びついていく。これは，データの累積と蓄積過程の履歴管理にほかならない。また，総合漢字辞書は，多様なデータへの整合性を取るインタフェースとしての重要な役割をもつことになる。

漢和辞書とJIS漢字との対応は，辞書を直接コードブックとして利用できるかどうかを確認するためのものであった。確認の過程で得られたJISにあり辞書にない94字の漢字は（表2)，JIS 規格漢字に問題があるのか，国の規格で定めた漢字は当然辞書に載せるべきなのか，JIS 漢字の利用者にとって対象漢字の意味をどのように調べるのか問題が多い。

総合漢字辞書に関する基本的な考え方を述べたが，JIS コードは1988年を目標に二回目の改訂作業が行われつつある。この作業と並行して２バイト系

データの拡張法も検討されている。2バイト系データの拡張とは，JIS 規格の8836字を最大26508字にまで広げる案である。見直しは，追加漢字の候補を選定する作業も含めたものになるはずである。しかし，この拡張作業には，次のような問題を解決しなければならない。

第一は，各メーカともに現行のJIS以外に，拡張漢字を独自の領域に配当していることである。JIS 規格の拡張は，配当後のコード順序，拡張領域にある漢字コードを使った既存の辞書・データのすべてに影響を与える。これは，JIS 83年改訂版における第一，第二水準間の漢字の一部入れ替えによって，旧版との間に情報交換上の互換性を崩すことになったことと同じ問題を生じさせる。

第二は，CD-ROMを使用した漢字辞書の出版である。CD-ROM化された漢字辞書は，コンピュータで処理できる漢字を飛躍的に増加させるが，使用する文字セット，コードともに閉じた世界を作る。辞書を閉じた世界におくことは，JIS 規格以外の領域でコード，文字セットから独立した文章作成が行われることになる。これは，情報交換の基本的な精神に逆行し，辞書間の字形の相違が増加した場合，情報交換上の不整合問題を広げることになる。

これらの点からも，JIS 規格の検討に当たっては，新しい流通媒体を使用した情報配布の効果と，それによって受ける影響も考慮しておくべきであろう。辞書は，データ交換・結合のさい標準化した接続点を利用者に見せるが，接続点の揺らぎは利用者に無用の混乱を課すことになるからである。なお，今後の漢字総合辞書に関する作業として次の五項目を予定している。

1) 二次情報の一次情報に対する関係を調べる。

2) 辞書項目の最適管理を図るため，項目の適性分割法を検討する。

3) キーの多様化を図り，任意の字形・偏・旁・冠，また現在中国で試みられている。各種の検字法を取り込んだシステムを作成する。

4) 検索システムのデータベース・システムへの拡張と，追記型光ディスクまたはCD-ROMによるデータ配布方式の確立を図る。

5) 総合漢字辞書と新聞 KWIC 用語索引との連結を図り，用例との接続

効果を調べる。

最後に，データベースを作成するさい使用した新字源には，JIS 漢字と一致しない漢字が現在確認した段階で496字ある。照合できなかった漢字情報は，大字典・大漢和の双方から補填したが，引用した辞書の識別は行っていない。使用のさいは，注意が必要である。

本稿では，漢字辞書の総合化の基本理念と辞書項目の内容について述べてきた。しかし，辞書は完全にデータ修正が終了しているわけではない。修正には，まだ多数の時間と人的労力を必要とし，利用する研究者の協力が必要となる。本稿で述べた機械処理用の辞書は，共有情報として利用していく過程で評価が定まっていく。その点で，プロトタイプとしての機械辞書が，実務に耐えきれるかどうかを検討するためには，広く利用されその問題点が辞書にフィードバックされるシステムを作成することが重要である。

〔付記〕 システムの開発に当たって，プログラムは研究補助員の米田純子が担当した。漢字の情報付加と見直しは，アルバイタの太田幸代さんの協力による。漢字の点検には，言語計量研究部長・野村雅昭の助言を受けた。また，日立・日本電気両社の辞書を利用させていただいた。記して謝意を表する。 (1987.8.31)

5. 参 考 文 献

1) 松本 昭 (1968)「国研用漢字テレタイプと同機利用の言語情報処理」『電子計算機による国語研究』(報告31)57-90。
2) 田中 章夫 (1968)「電子計算機によるワードリスト作成上の一問題」『電子計算機による国語研究』(報告31)115-132。
3) 国立国語研究所(1967)『漢字コードブック』。
4) —— (1962)『現代雑誌九十種の用語用字（第二分冊漢字表)』(報告22)。
5) —— (1976)『現代新聞の漢字』(報告56)。
6) —— (1983)『高校教科書の語彙調査』(報告76)。
7) —— (1986)『中学校教科書の語彙調査』(報告87)。
8) 諸橋轍次・編 (1971)『大漢和辞典』第3刷（大修館書店)。
9) 上田万年・他編(1974)『大字典』第21刷（講談社)。
10) 小川環樹・他編(1984)『新字源』第230版（角川書店)。

11) 山田　進・他　(1985)「データ中心システム設計技法」
『日経コンピュータ』5月7日号，167-183。
12) 斎藤　秀紀　(1985)「漢字コードの拡張法に対する試案」
『研究報告集(6)』(報告83) 57-103。
13) ——　(1986)「電子計算機による用語調査法の開発」
『国定読本第1期「尋常小学校読本」の用語』
(昭和59・60年度文部省科学研究費補助金一般研究
A，国定読本の用語の研究，研究課題番号59410011
研究代表者飛田良文) 139-147。
14) ——　(1986)「同形異語判別への仮名・漢字変換処理の応用」
『研究報告集(7)』(報告85) 109-134。
15) ——　(1987)「光ディスクを使用した大量日本語データの蓄積」
『研究報告集(8)』(報告90) 95-123。
16) 米田　純子　(1987)「漢字総合辞書」『CL通信第8号』38-47。
17) 玉井　鉄夫　(1966)「初級講座情報科学と情報技術第四回分類法」
『情報管理』Vol. 9, No. 4, 172-182。
18) 田島　一夫　(1979)「JIS漢字表の利用上の問題 ―漢字処理システムにおける漢字のデザインと管理」
『情報管理』Vol. 21, No. 10, 753-761。
19) 林　　大　(1984)「字体・字形・書体をめぐって」
『日本語学』Vol. 3, 10-15。
20)「日本語処理・カナ漢字変換，コード体系の不統一が深刻に」
『日経コンピュータ』(柳田俊彦) 1987. 3. 2, 77-85。
21) 武部　良明　(1981)『日本語表記法の課題』(三省堂)。
22) 野村　雅昭　(1984)「JISC6226情報交換用漢字符号系の改正」
『標準化ジャーナル』(日本規格協会)
Vol. 14, No. 3, 4-9。

表1　辞書データ項目の内容

項番	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
項目名	見出し漢字	区・点番号	改訂情報	JIPSEコード	JIPSJコード	端末外部コード	端末内部コード	漢テレ盤内字コード	漢テレ盤外字コード	日立コード	旧日電コード	見出し部首	部首コード	画数	部首内画数
長さ	02	08	02	08	08	08	08	08	22	08	08	02	06	04	04

16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
新字源番号	大漢和番号	大字典番号	教育漢字1	教育漢字2	当用漢字	当用漢字補正1	当用漢字補正2	常用漢字	人名漢字1	人名漢字2	人名漢字3	新聞順位	雑誌順位	新聞度数	雑誌度数
10	12	10	02	02	02	02	02	02	02	02	02	08	08	10	10

32	33	34	35	36	37	38	39	40	41
新聞人名度数	雑誌人名度数	新聞地名度数	雑誌地名度数	読み別度数	高校教科書度数	中学教科書度数	余白	音読み	訓読み
10	10	10	10	756	10	10	30	20	60

ファイル名：DCL3.SOUGOU，ファイル形式：索引順，レコード長：1118Byte

表2　大漢和・新字源・大字典にない JIS 漢字一覧（1978年版）

漢字	区番	JISコード
侭	48-54	5056
劔	49-91	517B
昕	51-06	5326
噛	51-85	5375
坤	52-11	542B
垉	52-18	5432
垣	52-21	5435
垳	52-24	5438
垜	52-27	543B
垉	52-34	5442
塗	52-43	544B
塀	52-49	5451
塔	52-55	5457
塩	52-60	545C
壜	52-63	545F
嵜	54-12	562C
岬	54-16	5630
峅	54-18	5632
峠	54-19	5633
峙	54-31	563F
峺	54-46	564E
幣	54-82	5672
弼	55-27	573B
悠	55-78	576E
捌	57-43	594B
捻	57-62	595E
撹	57-88	5978
攝	58-58	5A5A
晁	58-83	5A73
昡	59-06	5B26
杁	59-21	5B35
朶	59-32	5B40
杤	59-37	5B45
椁	59-67	5B63
椿	59-77	5B6D
椢	59-90	5B7A
椁	59-91	5B7B
椀	60-09	5C29
椙	60-13	5C2D
椄	60-14	5C2E
椧	60-16	5C30
椢	60-17	5C31
椙	60-47	5C4F
椶	60-51	5C53
椪	60-57	5C59
椙	60-81	5C71
汢	61-73	5D69

漢字	区番	JISコード
涛	62-25	5E39
涮	62-67	5E63
琺	64-94	607E
畢	65-22	6136
畝	65-30	613E
畴	65-39	6147
畠	65-42	614A
砿	66-72	6268
砒	66-77	626D
碵	66-83	6273
稭	67-46	634E
筑	67-83	6373
篋	68-01	6421
箏	68-24	6438
籘	68-35	6443
籌	68-44	644C
籏	68-57	6459
粁	68-68	6464
粭	68-70	6466
糎	68-72	6468
糠	68-84	6474
紵	69-78	656E
膤	71-19	6733
莫	72-20	6834
莵	72-25	6839
菴	72-45	684D
蜷	74-12	6A2C
祢	74-57	6A59
蓑	74-62	6A5E
蟶	77-32	6D40
輌	77-50	6D52
轌	77-58	6D5A
溢	77-90	6D7A
鈑	78-63	6F5F
鈬	78-69	6F65
錦	78-93	6F7D
鑁	79-39	6F47
鑪	79-47	6F4F
閧	79-64	6F60
隣	80-03	7023
騙	81-50	7152
鰍	82-32	7240
鯵	82-45	724D
鴬	82-84	7274
鵈	82-94	727E
鷆	83-23	7337
麩	83-48	7350

JIS　JISE　盤内　盤外　内部E　内部J　修正

①愛 ⑤3026 ④F050 ⑧0201 ⑨◆奥裁 ⑥7450 ⑦6F26 ③

②区番	16-06	⑰大漢和	109470
⑫部首	心	⑯新字源	25740
⑬部首コード	061	⑱大字典	03405
⑭総画数	13		
⑮部首内画数	09		

⑩おんよみ　アイ

⑪くんよみ　メ(デル)/オ(シム)/イト(シイ)/カナ(シイ)/

⑳学習漢字	4年	⑲教育漢字	4年
㉓常用漢字	◎	㉑当用漢字	◎
㉕人名漢字			

愛　各種語彙調査の集計結果

	雑誌	新聞	高校教科書	中学教科書
㉙使用順位	0232	㉘ 0432		
㉛使用度数	00286	㉚ 00598	㊲ 00093	㊳ 00017
㉝人名度数	00007	㉜ 00042		
㉟地名度数	00015	㉞ 00099		

愛 ㊴ よみ別度数表

			雑誌	新聞
0		アイ	259	443
2		いとしい	1	1
2		かなしい	2	1
2		まな	2	2
2		いとおしい	0	1
2	S	かわいい	0	9

0／表内音訓，1／特別な表内音訓，2／表外音訓，S／熟字訓・あて字

図1　漢字情報の検索画面