Unification of a new Japanese data base which uses key words of varying lengths

キーの階層性を利用した異なる日本語データベースの統合

齋藤秀紀

SAITŌ Hidenori: Unification of a New Japanese Data Base Which Uses Key Words of Varying Lengths

要旨：東アジア諸国（中国・日本・韓国）との間で科学技術の交流が盛んになり，日本語教育に対する重要性が増している。しかし，日本語教育に利用できる資料は，十分であるとはいいがたい。一方，国立国語研究所には，用語用字調査で得た現代日本語に関する資料が500万KWIC用例，漢字データベースなどがあり，日本語研究教材作成に利用できる環境にある。

本稿では，これらの資料を総合的にコンピュータで管理する方法と，日本語研究者にデータ提供を円滑に行うためのシステムの試案を述べる。また，蓄積されている漢字・単語・用例などのキー長の異なるデータを統合する方法として，疎結合方式が有効であることを示す。さらに，中国・日本・韓国の相互のデータ交換を想定した統一漢字コードを提案する。

キーワード：日本語データベース，キーの階層性，日本語教育教材，東アジア漢字圏．統一漢字コード

Abstract: Amidst the active exchange in science and technology among the East Asian countries (China, Japan, Korea), the importance of teaching Japanese as a foreign language has increased. However, one can hardly say that the data presently available for teaching Japanese as a foreign language is sufficient. On the other hand, the data on modern Japanese, which has been collected by the National Language Research Institute in surveys of vocabulary and writing forms, include 5,000,000 KWIC examples (words in context) and a large KANZI data base and are in a form which can be used for research on the Japanese language and preparation of teaching materials.

In this paper we described a tentative plan for methods of integrated computer management control of and easy access to these data by Japanese researchers. We demonstrated the effectiveness of the loose coupling method as a method for unifying these data of varying key lengths which include previously collected characters, words and examples. We also proposed a KANZI unification code which allows for mutual data exchange among China, Japan and Korea.

Key words: Japanese data base, key words of varying lengths, materials for teaching Japanese as a foreign language, East Asian KANZI zone, KANZI unification code

1. はじめに

国立国語研究所（以下国語研）では，用語用字の使用実態を知る目的で各種の調査を行ってきた。初期の調査は，主に冊子体で報告されてきたが昭和41年のコンピュータ導入後は，磁気テープ，マイクロフィルムなど機械可読媒体でも保存されるようになった。コンピュータ可読資料は，蓄積データの再利用が容易になると同時に，データ加工・相互参照への自由度を広げた。しかし，再利用には，媒体・記録形式・利用手引書・字形・コード管理などデータと資料の二つの面で管理が必要になった。

一方，言語計量研究部第三研究室では，蓄積データの共用を図るため，基礎実験といくつかの提案を行ってきた。その一部は，「漢字コードの拡張法に対する試案」（文献5），「漢字情報データベース」（文献9）で報告している。また，大量のコードデータを集中管理するため追記型光ディスクを使ったKWIC用例データ検索システムを開発した（文献8）。光ディスクの採用は，大規模データベースを作成・運用していくうえで長期・安定型媒体利用への道を開いたことになる。現在，これをコンパクト・ディスク(CD-ROM)へ移植し，大量データの出版媒体としての利用を実験中である。

その他，処理方法では，国定読本の用語調査（第3期）に光学式手書き文字入力（Optical Character Reader : OCR）システムを作成し（文献6, 7），同語異語判別を含む約91,200語を300人日程度で処理できる方法を開発した。

以上が，これまでのシステム開発の概要である。しかし，データベース化の対象になる用例・語彙表は，目的によって「調査単位」が異なり，資料の比較は，専門家の見るところでも困難であることが多い(文献4)。異なる調査間のデータをあつかうためには，漢字および漢字の基本性をもとに，漢字・単語・用例などの各種資料を横断的に参照できるシステムが必要になる。この考え方は，荒木（文献2）が行っているが，本稿では，これを漢字から元データ間の双方向の結合を許し，異なる調査の相互参照とデータ管理への

対応に拡張する。

以下，本稿では国語研内に蓄積されている KWIC 用例集 500 万件，各種の語彙表，漢字辞書などコンピュータ可読データを中心に，総合的な日本語研究のためのデータベース作成の試案を示す。日本語データベースとは，国語研および外部で作成された漢字および用例などの研究成果を統合し，データの共有化と個別データを相互に参照できる環境を作ることである。最後に，中国・韓国・日本語間のデータ交換を行うため，統一コードについて試案を示す。

2. 機械処理用辞書とキーの管理

これまで，国語研における機械処理用の辞書は，漢字と用語の二つを利用してきた。漢字辞書は，コンピュータで印字可能な漢字9731字に，漢和辞典情報のほか漢字調査度数（雑誌九十種，現代新聞の漢字，中学・高等学校教科書，雑誌用語の変遷），外字，旧コンピュータ・入出力装置コードなど 41 項目の情報を統合したものである（文献12-17）。

漢字辞書作成の第一の目的は，20年間に使用したデータコードの管理であった。第二は，データに対する付加情報づけの標準化を進めるコンピュータ処理用辞書として。第三は，漢字調査データの総合管理である。本システムの基本的な考え方は，この漢字情報データベース作成の延長上にあるが，第三番目の機能を発展させ，国語研における全データの集中管理のため用例に対する索引機能の導入とその問題点を探ることである。

用語辞書は，一部データを五十音順に配列する理論コードとして利用されてきたが，日本語入力における仮名・漢字変換処理の利用が主なものである。簡易型日本語入力の方法に，仮名またはローマ字入力・漢字変換方式が定着したことは，機械処理用辞書の重要性が増したことになる。

辞書を利用した変換処理は，入力するすべてのデータと辞書項目が一対一に対応していることが前提にある。仮名・漢字変換処理を通した辞書とは，辞書項目を指標としたデータ管理機能の直接的利用である。

データ利用の活性化には，データ交換を基本にしたシステム開発が重要になる。また，開放的なデータ利用への道を開くには，分析結果を元データへ還元させる機能が不可欠である。処理結果のデータへの還元は，利用者自身が辞書項目の拡張と細分化に参加すると同時に，システムが必要とする方向を探る指標情報に利用できる。データの試行検索の履歴を通して見るシステムの利用実態の把握とデータ特性の習得・再確認である。

これらの還元データは，元データに対して重層構造を作り，システム開発環境をプログラム優先の方法からデータ中心の考え方へと移行させる（文献7，8）。さらに，漢字辞書は東アジア漢字圏とのデータ交換用インタフェースに拡張できる。

以上をまとめると，機械辞書は辞書そのものとしての利用，異なるデータ間の接続機能，変換処理を通したデータ管理，データに対する索引，また変換時の未登録処理から辞書自体の学習機能を併せもつことになる。これらの機能を疎結合用キーに使用する利点は，次の三点である。

1) 漢字は，物理的な単位と理論的な単位が近く，かつ日本語における最小単位である。
2) 見出し語は，調査によって「長い単位」と「短い単位」の二つの単位があり，異なる単位の見出し語の共通キーに使用できる。
3) 標準化（例えばJIS）された字形を通して，市販辞書との併用ができる。

3. 蓄積データの共有化の問題

機械可読なデータとして大量に保存されているKWIC用例集は，新聞・文芸作品・教科書データなどがある。用語調査は，中・高等学校教科書である。これらの調査は，調査目的によって単語の単位が異なるが，基本的には，「長い単位」と「短い単位」の二つが使用されてきた。また，新聞のKWIC用例集は，短単位の一つである「ベータ」を，教科書調査では長・短に相当する「W・M」単位である。

データを共有のものとするためには，データの単位，作成された背景を利用者に明確にする必要がある。宮島は，「総索引への注文」（文献３）のなかで索引を使用する立場から，資料を比較するうえでたてられた見出し語の揺らぎが問題になることを指摘している（以下の三項目は文献３から引用）。

1) おなじことばが一つの索引のなかでまちまちにあつかわれているもの。

2) おなじ種類のことばが一つの索引のなかでまちまちにあつかわれているもの。

3) おなじ種類のことばが索引によってまちまちにあつかわれているもの。

第一の型は，不注意によるものであり，第二は作業方針の不統一によるもの，第三は編集者の意見の相違にあるとしている。さらに「雑誌用語の変遷」では，国語研で過去に行われた用語調査から，「長い単位」と「短い単位」を認める経過と単位認定後の見出し語をまとめるさいの揺れの「はば」の問題を説明している(文献16)。この二つの論文は，いずれも異なる資料間，同一資料内において研究目的の相違や単語認定の曖昧さが，資料間の照合を困難にすることを述べたものである。

用語表・度数表は，見出し語をたて調査目的にそって集計された出現回数が付加される。しかし，見出し語は，研究目的または編集目的にそって集められた単語の集団指標として使われ，文脈から「単語」が切りはなされた時点で意味の一義性が失われる。

異なる調査では，指標の形が同じでも含まれる内容・個数が異なる。指標で示される集団は，集団の等質性が保障されていなければならない。資料間の見出し語のたて方の曖昧さは，研究者の研究目的が異なれば，当然なものと言える。しかし，集団の内容が「異なる」か「同じ」であるかを確認するには，元データ・用例の参照による以外に方法がないことになる。コンピュータ処理では，比較・照合に指標をキーとして使うため集団の等質性とキーの長さが問題になる。

索引の見出し語の形は，利用者にとって固定されたものとして見える。これは，見出し語のたて方に揺らぎがあっても見出し語が固定されたものなら

ば，揺らぎの問題はコンピュータの検索キーの指定方法によって調整できることになる。以上述べた事柄は，蓄積データを利用するさい，利用者はデータ特性が不明のまま手探り状態で検索せざるを得ないことを示している。複数のデータベース，用例集間のデータ検索・参照は，キーが一対一に対応できないことを前提にシステム化しなければならないことを意味する。

4. 疎結合に対する基本的な考え方

4.1 疎結合による検索処理

大量データを対象にした検索処理は，いかに早く目的のデータへたどり着くかが重要である。試行検索は，キーの繰り返し指定が行われるが，対象が大量でありキーの曖昧さが固有のものであるならば，検索による絞り込みは目的のデータへ近づく最適手段となりうる。調査結果は，それ自体で完結している。完結しているものは，不特定多数の利用者が目的別にデータを共有するとき元データの個別管理が不可欠になる。蓄積データの再利用には，対象が他のデータから独立していることが重要である。

疎結合の基本は，データおよび検索キー特性が不明なとき個別に管理されたデータ間の疑似的結合の手段として，既知の検索用キーを使用することにある。疑似的な結合とは，複数の単位で構成されている見出し語を検索するさい，漢字を一次キーに使い検索されたデータの一部を次に検索するデータのキーに使用する方法である。データのキーと指定するキーの包含関係は，キーの曖昧さと長さに対する整合性を情報の絞り込み操作のなかで検索者の知識を利用する。その点で，この方法によるシステムは，利用者の知識と経験に依存する。

その他，疎結合は，データを検索する操作を通してデータの確認も同時に行っている。この操作は，例えば外部で作成されたデータを利用するさい，利用と並行してデータ内容を確認できる。データ利用時の事前整理にかかる時間を短縮できることになる。以上が疎結合方式の特徴である。この特徴は，次の三点にまとめられる。

1) 独立して調査されたデータは個別管理を基本にできる。

2) キーの指定は，利用者の知識を接続インタフェースの一部としてシステムに一体化できる。

3) データ間の参照と連結は，漢字を一次キーとして使い，以後検索されたデータの一部を「橋渡し」情報に使用できる。

疎結合方式によるデータ検索の方法は，二つある。第一の方法は，漢字データベースの付属情報を直接使用する場合である。漢字および漢字属性のキー化は，漢字が長さに依存しないため，データ照合に必要な最少限の情報を指定できる。第二の方法は，市販の漢和辞書（新字源・大漢和・大字典・漢英辞典など）の索引から検字番号を使い間接的に機械辞書につなぐ方法である。さらに，二次以降に検索されたキーの役割は二つある。

第一は，漢字を基本キーとして見出し語・用例と元情報へもどる場合である。漢字情報は，本文から最終的に集計された結果であるとすると，集計のさい失われた情報へさかのぼる利用法である(図1)。第二は，二次検索された情報を別データを探すキーとして使用する場合である。ここで，検索過程で二次以降に検索された見出し語または出現形は，次の検索用キーとして使用できれば，二次以降に検索されたキーは中間コードあるいはメタコードとして位置づけできる。漢字情報データベースの各項目は，データ間の「橋渡し情報」として選択的に指定できることになる(文献9)。

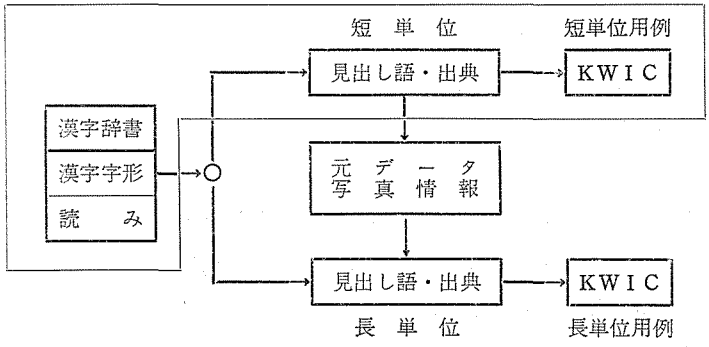

図1 漢字辞書と単位の異なるKWICの関係

第三は，データを抽出した資料などコード化しにくいデータの管理である。漢字・単語・用例は，いずれも加工後のデータである。これに対して，文・段落の参照，新聞であれば抽出データの環境情報，コード化できない古い資料，データ特性を説明する資料でコード化の必要のないものなど，コード情報と対になってコンピュータ管理しなければならない処理に利用できる。この関係を表すと図１のようになる。細罫線内の処理が今回の実験の範囲である。

見出し漢字	見出し部首	画数	教育1	教育2	常用	音読み
国	口	08	1	1	1	コク／

訓読み
クニ／

図２　一次検索された漢字辞書の一部

読み	出現形	出典
こっか	国家	024370010022J0
こっかい	国会	019310020047B0
こっかい	国会	019560030700B0
こっかい	ルタ〕五日発＝ロイター※M○米，	国家　安全保障会議開く※M○ワシントン
こっかい	力説，両国間の通商航海条約の比国	国会　における批准を反対者を説得しても
こっかい	くも二十六日までには印刷を終え，	国会　に提出できるだろう」と語った．元
こっかい	はそのなかで同国の憲法を停止し，	国会　を解散させたと述べた．また同中佐
こっかい	十分，パレスホテルで，経済気象台	国会　審議のもつれから新年度予算編成は
こっかい	力者をお願いしたい．一，予算案の	国会　提出は二十六，七日ごろとなろう．
しこく	板を繰出すのがねらい．憲法を停止	国会　も解散オートポルタ軍事政権※M○
しこく	結れ〃と予想予算，２６日までには	国会　提出田中幹事長談自民党の田中幹事
しょこく	，二，三回の総選挙により社会党が	国会　の過半数を占め，政権を握ると予想
しょこく	祝寿の会をつくり，徳島の同人誌『	四国　文学』が古希祝寿記念号を出した．
じこく	和※M○「元日から５日まで浅草と	四国　の舞台あいさつ．忙しいのは大いに
じこく	．ドゴール仏大統領も一日，「関係	諸国　の組織的な接触，外部からの武力干
ぜんこく	力均衡を意味しない．そこにアジア	諸国　の不在があるからである．それは中
ぜんこく	の性格からいって，核兵器の発動は	自国　の死亡にかかわるぎりぎりの時しか
ぜんこく	の核のカサなどはあてにならない．	自国　を焼け野原にする危険をおかしてま
ぜんこく	社の地方組織を動員して二百十社の	全国　主要企業に質問状を送り百八十九社
ぜんこく	にも非常に重視されるからである．	全国　どこでもできる副収入書道塾を開き
ぜんこく	間で一世帯一住宅を実現する※６○	全国　一律最低賃金制を実施し，当面一万
ぜんこく	中央魚相互貿大沢商築地魚日マタイ	全国　蕎大都魚黒糖山善機東菱自東トヨタ

図３　見出し・出現形・出典・用例の検索例

4.2 漢字の単位性とキー特性

一方，この疎結合による検索は，林が述べている単語と漢字に関する教育上の原理と類似している。林は，漢字と単語の関係に「形式上」と「意味上」の二つの原理を立て，漢字学習の要素は少ない数の原理を理解することにある，としている（以下の四条件は文献18から引用)。形式上の原理は，1）漢字が独立して一単位を表していること。2）漢字一字は独立しないが接続語となって他の単語に付着すること。3）造語要素となって他の造語要素と結合し，主に二字で単語を作ることがあること。4）漢字自身に要素性がなく，二字または三字が合体して一単語を作ること，の四項目を挙げている。

また，意味上の原理に，漢字一字の意味の支えがその字の「音・訓」に現れることがあること，「音・訓」のように定まった形には示されていないが，おのずから明らかで他のいくつかの単語でそれを示すことができるもの，などである。

このような造語成分が他の造語成分と結合していく過程は，利用者が漢字から用例へ検索する過程と表面的には同質であると仮定できる。造語成分である漢字を単語検索の基本キーに使用し，読みは意味上の原理の関係に対応させる。学習者が漢字を習得していく過程は，データ検索における試行の過程と同じとする考え方である。

ここで言う試行とは，データ利用者の目的とするデータにたどりつく過程に必要な知識を，漢字の造語過程と同一であるとし，検索のさい「橋渡し」情報として使用することである。検索で必要とする暗黙の知識・情報を単語間の「橋渡し」情報として使用できるならば，索引の見出し語の曖昧さは利用者の使用目的と専門家としての知識をキーの生成過程に埋め込むことができる。この方法は，資料間の異なる見出しの長さへの対応が漢字の最小単位を操作することによって部分的に解決できることになる。

5. 中国・日本・韓国の統一コードの試案

5.1 統一コード設定の背景・要求事項

コンピュータ利用の多様化とともに，中国・日本・韓国(Chinese Japanese

Korean: CJK) における科学技術文献の交換，日本語教材開発や対照研究の重要性が増した。このためには，アジアの漢字使用国とのデータ交換を前提にした漢字コードの設定が必要になっている。本節では，国語研の外字コードを拡張し CJK で使用している漢字データの相互交換を容易にする統一コードを提案する。統一コードは，漢字データベースとの字形を通したデータ交換へ発展させるためのものである。

国語研コードは，新聞調査のために昭和41年に導入した漢字テレタイプライタの盤外字管理のために設けた(文献1)。大漢和辞典をコードブックに使い，検字番号は盤内字2文字を組み合わせたものである。国語研コードは，JIS 制定後に主たる役割はなくなったが，JIS 外漢字の管理と統一配列コードとして使用されてきた。

これに対して，国外における JIS 相当の国家標準コードは，中国・韓国でも規格化されている。物理コードは，ISO 7ビットコードを基本に2バイト14ビットを一漢字に対応させている。各国の標準文字セットコードが国際標準に従ったことは，基本コードに同一操作を与えることができる。ここで提案するコードは，CJK コードの各々に「重み」をつけ，2バイトの統一体系を設定することである。

以下，CJK 国家標準字形を混在できる統一コードの第三機能として追加する試案を示す。基本的な考え方は，「漢字コードの拡張法に関する試案」(文献5)で述べた方法による。統一コードを設定する上での要求事項は，以下の三項目である。

1) 中国・日本・韓国など東アジア諸国の漢字情報を統一的に扱うことができること。
2) コードは保存用とし，また各国で定めた国家標準コードへの変換と JIS-1988 年版で追加される文字も吸収できること。
3) 物理コードの構造は2バイト16ビットとし，使用領域の指定は演算処理で選択可能なこと。

5.2 統一コード設定上の問題

CJK 各国で標準化されている漢字の 字形や字数は，その国の 政策・使用環境によって独自に規格化されている。また，漢字を使用している国々でも，中国の簡体字，韓国の漢字ハングル併用，台湾で使用している繁体字など，字形と意味の双方で違いがある。さらに，文字セットの変更もその国独自の考え方で実施されている。

統一コードの設定は，漢字を使用していること，物理コードが一致していること，の二つの条件をもとにいかに異なる部分をまとめるかにかかっている。例えば，漢字の単位性を基本に文字セットを字形によって再構成するか，国家標準を尊重するかなどである。

「字形」で整理した場合は，改定で生じる文字セットの入れ替え・削除・追加を各々改定版ごとに 対応させなければならない。 新規格への対応は，JIS制定以前の漢字利用と同様の問題を抱えることになり，変換テーブルによる一括処理が基本操作になる（実務的な運用は文献5参照)。一括変換方式は，データの処理量によって限界が生じ改定が不規則にかつ頻繁に行われるならば，蓄積されている旧データとの互換性が損なわれる。

しかし，変換方式は，テーブルそのものは過去のデータを間接的に管理しているため，新・旧文字セット・コード対応表を使うかぎり新・旧データ間の互換性は維持できる。さらに，変換テーブルは，管理機能を索引に拡張できる。本稿で述べるキーの階層性を利用した疎結合方式は，この特性を利用している。

字形および文字セットの改定は，各水準間の文字の入れ替え，字形を他の規格と調整すること，文字数を追加すること，などが行われる。しかし，JIS 1983年の改定では，同様の変更を行ったため旧データとの互換性を維持できなくなったことがある。この結果，1985年版の普及を遅らせることになった。このような背景は，標準文字セットの改定には，旧版との互換性を維持できることが不可欠であることを意味する。これは，国際標準においても同様である。

これに対して，各国で定めた標準文字セットを尊重した場合，国家標準字形のなかでは整合性は維持され，少なくとも国家間の文字セットの改定にともなう調整は不要になる。また，科学用語の新漢字の追加登録など最新情報の修得と，最新データとの互換性も保証される。今後とも，すべてを漢字表現する国では漢字の追加登録が多くなることが予想される。将来，国際規格として漢字コードが統一されるならば，新コードへの移行は重要な課題になる。この点からも新コードの設計は，現在使用しているコード体系から新コードへの変更を容易にする配慮が必要になる。

5.3 標準文字セットの拡張

次に，改定作業によって予想される問題は，漢字数の拡張である。JIS で決められている漢字は6349字（1978年版）であるが，1988年の第二次の改定では6000字前後が追加される可能性が強い。6000字と仮定すれば，改定後の総数は約 12000 である。統一漢字の領域として必要な単位は，従来の8000から16000程度に拡張しなければならないことになる。（台湾：基本4808字，拡張17077字，その他11660を３バイト表現している）。

提案するコード体系を２バイト単位とした理由は，２バイトで表現できる字数は65536あること。現在，日本を含めた３カ国の漢字総数は，24000字以下であり，台湾を省く中国・韓国は94区・点8836字を標準化していること。CJK 各国のコードも２バイト系を主に使用していることが挙げられる。これには汎用コンピュータの処理単位が４バイトであることも含まれる。

さらに，CJK で使用する漢字の総数が 24000 字以下であることは，65536を上限に，計算で４ないし８の領域がシフトコードなしに識別できること，元コードへの逆変換の容易性とともに，単位領域を 16384 字に設定できること，３カ国に絞れば文字セットの拡張と，JIS ２バイトと同様に処理できることが利点としてある。これを一般化すると次式になる。

コード領域＝コード値 DIV 8836

コード位置＝(コード値－重み) MOD 8836

DIV：整数の除算関数　　MOD：剰余関数

表1　領域に対する重み配分

重み		領域	重み		領域
16進数	10進数		16進数	10進数	
0	0	記号領域	75EF	30191	第三領域
0E63	3683	第零領域	9873	39027	第四領域
30E7	12519	第一領域	BAF7	47863	第五領域
536B	21355	第二領域	D07B	56699	第六領域

データとプログラムコードの混在への対応は，表現方法が二つある。計算方式では，記号・第0領域コードは，ともに商は“0”となり領域の識別ができない。3683を境界とする判別が必要である。ただし，記号・アルファベット領域として指定すればこの制限はなくなる。その他，1バイト系非漢字の混在は，“OX”形式で2バイト表現になる。第1バイト目が“ゼロ”，第2バイト目がアルファベット・数字・記号系である（2バイト目が“00”になるコードは，“XO”表現になる。この範囲では，非漢字コードが隣接している場合桁落ちに注意する必要がある)。EBCDIC は，現行の混在処理と同じである。

入力データの保存コードへの変換は，各領域に決められている「重み」を加える。JIS または ISO コード2バイトへの変換および出力・各国漢字への解読は，上記の逆演算を行う。変換処理は，表1に示した重みを縦軸方向のみ与える。表を二次元と見るならば，横軸は1から94までの繰り返しであり，縦軸は94進表示の項目番号になる。以下計算の手続きを示す。

1. 統一コードの作成方法
   1) JIS 16進コードを区点別に00から93の10進相対番号に変換する。
   2) 二次元区点表を（94×94）8836の一次元表に変換し重みを加える。
2. JIS コードへの逆変換
   1) コードの所属する領域番号を計算し，統一コードから重みを引いた後，二次元表へ戻す（区・点番号の計算)。

2) 区・点番号を結合し16進 JIS コードに変換する。

計算は，3カ国のコードを混在させることを想定した。一領域の大きさは，現在の規格をそのまま対応させるため94区・点の8836字である。2バイトコード体系では，7領域・1記号領域を定義域として確保できることになる。記号領域は，区・点で表される 8836 と，16ビット65536との間が整数関係にないため余り3682字（0000，FFFF は省く）をあてた。しかし，記号領域の設定は，非漢字を特定領域に移動できれば，現行の記号系の補塡と同時に漢字の追加も期待できる。記号系の見直しは，技術情報の交流のためにも不可欠であり，CJK の非漢字領域を含めた互換性の維持のためにも行うべき時期にある。

ソート処理は，保存コード・逆算後のコードいずれも目的別に使用できる。配列とソート順序は，日本では部首，総画，五十音順のなかで読み順配列の利用が多い。分類・配列の多様性の要求に答えるためには，並行して漢字データベースの拡張が必要になるが，漢字情報データベースとの併用は不可欠の関係になる。

以上，簡単に試案を述べた。この案は，国語研の外字コードの浮動表現方法を拡張したものである。基本的な機能は，すでに外字処理のなかで実用化されているが，統一コードの妥当性は，今後コンピュータ実験を通して確認する予定である。その他，試案では，物理コードの8ビットを JIS 7ビットコードに一種の論理コードとして対応させた。コードは，保留領域を使用することを前提にしているが，JIS 形式によって情報交換とコードの互換性の二つが確保できるため，内部コードと情報交換用コードを別の体系と考えるならば問題はなくなる。この処理は，表1の重みを2121またはA1A1におけば，領域を介して処理できる。

しかし，各国で定めた標準コード形式は，シフトコードの標準化，漢字パターンとの対応ができれば実務上の最低機能は整うことになる。また，各国が日本のように標準文字セットを増加させることを考えるならば，一領域に16000 は必要になる。情報交換用の国際標準は，国家標準のサブセットを集

めたものとすれば国家標準との互換性も保証される。

6. おわりに

内部で作成された用例・漢字データの統一管理の試案を述べた。試案は，両データの統轄を通して外部に開かれた情報交換の組織を作る。データ交換は，内部で必要とするデータの補塡とともに内部データに対する客観的評価への道を開く。また，キーとデータ間の階層性を仮名・漢字変換辞書項目の階層性に対応させることは，仮名・漢字変換辞書に用例検索用キーの管理と同時に漢字変換の二つの機能をもたせることになる。辞書による両処理の一体化である。

しかし，ここで提案したシステムは，方法論的な実験を一部開始したにすぎない。また，統一コードは，実用化が急がれるが必要とするメーカ・利用者ともに一部にかぎられている。実用化を進めるにあたって提案事項は，いくつかの検討事項が残されている。代表的なものは次の三点である。

1) 日本語のコードデータ以外の例えばイメージ情報の接続をどう考えるか。

新聞 KWIC 用例集は，本文と完全に照合していない。これは，照合には調査に要した時間程度の修正時間がかかるからである。修正時間をかけても完全な索引にするのは当然である。しかし，できるだけ早く利用できるよう公開することもまた重要であると考えたためである。本システムの次の拡張では，KWIC 用例の欠落部分は，利用者が KWIC の出典情報を手掛かりに直接新聞紙面を呼び出す機能を設ける予定である。この機能は，利用者にとって当面研究用データとして対応できるはずである。また，将来欠落データを補塡する場合も紙面が簡単に指定できれば，修正用の道具としても利用できる。

2) 蓄積データの保存期間および漢字・単語辞書の索引化と DD 化の問題。

関係形式データベース (Relational Data Base: RDB) の採用は，1) データの共有，2) データの完全性，3) データの独立性，をデータ運用の道具に利用するためのものであった。また，複数のデータベースにあるデータ内容

を一貫して管理するためには，データディクショナリ・ディレクトリ (Data Dictionary/Directory System: DD/DS) が必要になる。DD/DS によるデータ管理の実務処理は，まだ実験的な水準にあり実用化されている例は少ない。しかし，本稿で述べた漢字の索引機能は，将来この DD/DS 機能のなかにどの程度含めることができるかが，分散データの運用効率を進める要点になる。

また，データの共有は，異なる調査データを集約・統一管理するためにＲＤＢシステムを使い，データ利用と検索手続きの標準化を図るためのものである。データの完全性は，利用するデータの安定・正確の点で更新の多いデータとは異なるが，国語研のデータ利用形態が調査結果または最終的に安定した用例データを対象にするためデータの変更・修正以外に誤りはない。

この問題は，第三のデータの独立性とも関係する。個別の調査データは，個々に独立して管理すること，異なるデータ間の統合は，利用者が必要部分を疑似的に結合することによって，データの共有が保証される。

3)　１バイトコードの保留領域の使用と記号・表音文字の拡張。

第１領域を基本に現行の１バイト系の文字セットを入れることを前提に説明した。しかし，今後プログラムで使用するアルファベットは，従来の英米系の文字を主体にしていくのか，利用者独自のアルファベットの利用も許すのか再検討されなければならない。現行の JIS 非漢字領域に登録されている記号系・アルファベット・数字は，各専門領域で使用できる十分な数は用意されていない。また，国によって利用する記号が異なる場合がある。漢字圏諸国の国家標準コードを含め，学術的な領域と印刷で必要とする一般的な記号系の整理の時期にあるといえる。

一方，情報処理の普及は，情報交換が主体にある。この形態の一つに，コンパクトディスクによる出版が多くなると予想される。大量データのコンパクト化は，情報処理の形をかえ個人別のデータベースを普及させる足掛かりとなる。これらの条件を満足させるためには，記号・アルファベットなどの「数」と「登録領域」の再検討が必要である。ここで，独立した領域にこれ

らの記号系をまとめることができれば，記号・英数字系の充実と非漢字領域を段階的に拡張することができる。特に，最近の科学情報の細分化と範囲の広がりは，ますます記号系の拡張を必要としており，統一コードの設定では留意すべき点である。

次に，国語研外字コード化の方式は，使用するコードの定義域とコードの組合わせの方法が重要である。JIS 定義域の "8836" 字から2個字を組合わせる総数は約3900万ある。この漢字で表される4バイトコードの漢字世界は，2バイト16ビットの実領域で表現できる最大値をこえて定義できる。今後2バイト空間と4バイト空間の対応の方法がコード管理・運用上重要になろう。また，この領域は，コード情報のみならず直接イメージをあつかう2値情報への拡張も考えられる。

最後に，統一漢字コードの提案ではJIS 1バイトコードの保留コード領域を使用したが，将来この領域に制御機能をもったコードが配当されるならば，プログラム作成のうえで問題がおきる可能性がある。新コードを採用するうえで空間上の漢字配列とともに今後の重要な研究課題になろう。

[付記] 統一コードのシステムの開発に当たって漢字辞書プログラムの作成は，研究補助員の米田純子が担当した。用語の単位については，言語体系研究部・宮島達夫部長から助言と資料の提供を受けた。また，RDBの利用と実験には，日本電気株式会社官庁システムの方々の協力を受けた。ここに記して謝意を表す。(1988.6.8)

7. 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) 荒木卓也・他 (1986)「多面的検索機能を備えた機械式辞書の試作」『計量国語学』Vol. 15, No. 6, 210-220

3) 宮島 達夫 (1969)「総索引への注文」『国語学』(武蔵野書院) 110-120

4) —— (1974)「単語の認定が語い調査の結果にどうひびくか」(国語研内部資料)

5) 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

6) —— (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

7) —— (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

8) —— (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

9) —— (1988)「漢字情報データベース」『研究報告集(9)』(報告94) 27-47

10) 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』(日経マグロウヒル社) 167-183

11) 堀内 一 (1987)「データ中心システム設計技法の可能性と問題点」『日経コンピュータ別冊』(日経マグロウヒル社) 266-277

12) 国立国語研究所(1962)『現代雑誌九十種の用語用字(第二分冊漢字表)』(報告22)

13) —— (1976)『現代新聞の漢字』(報告56)

14) —— (1983)『高校教科書の語彙調査』(報告76)

15) —— (1986)『中学教科書の語彙調査』(報告87)

16) —— (1987)『雑誌用語の変遷』(報告89)

17) —— (1985)『電子計算機と国語研究』

18) 林 四郎 (1987)「日本語の漢字」『漢字・語彙・文章への研究へ』(明治書院) 105-111

19) B. W. Leong-Hong・他(1986)『データディクショナリ／ディレクトリシステム』(穂鷹良介監訳，成田光彰訳) オーム社

20) Eiiti Wada (1987)「Three Byte Code Considered Harmful and Standardization of the Two Octet Character Sets」(First International Conference on Scholarly Information Network East Asian Applications & International Cooperation) Tokyo, December 8-11

21) B.W. Lee et al.(1987)「International Exchange of Bibliographic Information From an Asian Perspective」(First International Conference Scholary Information Network East Asian Applications & International Cooperation) Tokyo, December 8-11

22) 田島 一夫 (1983)「日本語情報処理における文字セットコントロールシ

ステム」『情報管理』Vol. 26, No. 7, 554-567
23) 『情報交換用漢字符号系 JIS C-6226-1983』日本規格協会，1984
24) 『中華人民共和国標準信息交換用漢字編碼字符集 GB2312-80』技術標準出版社，中国北京，1981
25) 「漢字コードの問題点を探る」『日経バイト』（根本　勝）102-107，1987. 11
26) 米田　純子　(1988)「漢字総合辞書」『CL 研究第 2 号』
（国立国語研究所・言語計量研究部）60-69