人文科学とコンピュータ 30－7
（1996. 5. 17）

# 漢字情報と文例情報を結合した日本語データベースの構築

斎藤 秀紀
国立国語研究所

現代日本語に関する電子化資料は、日本語研究や日本語教育を行うために重要である。しかし、日本語データベースを構築するためには、文字集合の拡張や多国語化法、文字と符号の規範性の確立や追加情報への対応など解決すべき点が多い。本稿では、昭和41年から国立国語研究所で行った「現代新聞の調査」（朝日・毎日・読売、朝夕刊1年分）データをもとに、各種の漢字調査の結果、漢和辞典情報、政令で規定する情報を統合した日本語データベースの概要を説明し併せて漢字符号に必要な機能を検討する。日本語データベースで扱った情報は、漢字6,349 字を含む45項目の付加情報および約200 万長単位語相当の新聞記事の本文である。

# A Database of the Japanese Language Combining Data on Sino-Japanese Characters with Examples of Usage

Hidenori Saito
The National Language Research Institute

Electronic materials on the contemporary Japanese language have come to play an integral part in language research and teaching. There are a great many problems that need to be overcome, however, such as how to provide for expansionof character sets and functionality in a multilingual environment. There is also much work yet to be done on establishing standards for charactersand their codes, and inproviding supplementary data. This article presents an outline of a Japanese language database drawn from a National Language Research Institute study which examined usage of Sino-Japanese characters (Kanji) in three major newspapers (Asahi, Mainichi and Yomiuri) in the year 1966. The database combines the results of the analyses undertaken in this study with data drawn from Kanji dictionaries, as well as official standards for Kanji usage.

## 1．はじめに

国立国語研究所では、昭和24年から国民各階層の言語実態を知るために各種の用語用字調査を行なってきた。これまでに行われた代表的な調査には、「婦人雑誌2 誌の調査」（昭和25年1 年分）や「総合雑誌13誌」（昭和28年7 月から昭和29年6 月分）、「雑誌九十種の用語用字調査」（昭和31年1 年分）がある。調査は、順位100 位で100 の用例を得るため、段階的に調査量を拡大してきた。しかし、調査は、人手を中心に処理していたため規模の拡大と共に、調査期間の長期化やデータ管理に限界が生じていた。

これらの問題を解決するため昭和41年から行われた「現代新聞の調査」では、コンピュータによる調査に切り替えられた（「現代新聞の用語用字調査」以下「新聞調査」、昭和41年度版、朝日・毎日・読売、朝夕刊3 紙1 年分、1/60の面サンプリング推定母集団1.2 億長単位語）。新聞調査は、我が国における最初の大規模調査となった。調査は、約10年をかけて終了した。本稿で述べる日本語データベースは、新聞調査で得られた各種の漢字や語彙表をもとに、新聞記事のデータ、漢和辞典で規定した各種の情報、政令で規定した各種の情報を統合したものである。漢字情報は、JIS C6226-1978版（漢字6,349字）に見出しを含む45種の情報を付加した。新聞記事データは、新聞調査のサンプリングで抽出した約200 万長単位語相当の本文である。

各種の用字調査の出現度数、漢和辞典の検字番号、政令で規定した情報は、第4研究部・第二研究室で作成した資料と一部データを使った。新聞記事の切り抜き資料は、第一研究室が用語調査のデータ入力用の基本台帳として作成したものを実験的に使用した。新聞記事データは、用語表の語形に付けられた出典情報を手掛かりに当時の第4研究部・言語計量調査室で再構成した。各種の漢字符号の統合とデータベースの取りまとめは、旧言語計量調査室（現電子計算機システム開発研究室）からの室員が引き続き行った。

## 2．漢字データと文例データの概要

新聞調査では、盤面から2110字の漢字を入力できる漢字テレタイプライタを使用した。漢字の入力は、盤面から入力できる2110字のほか、漢字2 文字を組合せて表外字とした。表外字は、漢字テレタイプライタの盤面に収容されている漢字を拡張するために、諸橋轍次編「大漢和辞典」で規定した検字番号を符号化した。大漢和辞典は、コードブックとして拡張され、表外字の管理や漢字テレタイプライタや高速漢字プリンタで作成されたデータの管理に使用された。表外字の利用は、JIS C6226-1978の制定と共に主な役割は終わったとされた。

しかし、JIS X0208 は、1983年に改正され1978年版との間で交換性を崩したことから、引き続き符号管理に利用された。JIS X0208-1983に対する符号管理は、理論符号としての表外字の位置を明確にした。同時に、理論符号としての表外字は、文字の拡張手段と装置から独立して文字情報を規定する利用法を確立させた。コードブックは、これらの情報を総合的に管理する手段として利用されるようになった。その後、符号管理に使用したコードブックは、各種の語彙調査や漢字調査で使用する読みや漢字の画数などを規定する基準書に発展した。

コードブックに記載された定形的な情報は、常用漢字や当用漢字の識別符号など政令で規定する情報、大漢和辞典・新字源・大字源・大字典の検字番号や新字源から読み・画数・部首の各情報である。そのほか、雑誌九十誌や婦人雑誌・中学・高等学校の教科書などの各調査で得た漢字の出現度数を付加した。コードブックで規定する見出し字形、符号、属性情報は、各種情報の規範を示すものであり、規定された情報は不変であることが要求される。

一方、漢字符号は、規定された漢字情報が利用実態の変化にあわせて追加や変更できる機能をもつことが必要になる。さらに、データを長期間保存し、任意に再現するためには、漢字符号と対応する文字集合の関係を安定的なものにすることが不可欠である。そためには、漢字符号が変化の履歴を記録できる機能をもつことが必要である。

コードブックは、これらの要求に対応するため電子媒体への変換が計画された。コードブックに記録した属性情報の種類は、(1) これまで使用した漢字テレタイプライタや漢字プリンタ装置など装置の漢字符号10種類、(2) 各種の語彙調査や漢字調査（雑誌九十種・新聞の調査・中高等学校教科書・文学作品など）に出現した漢字の度数や人名・地名で使われた漢字の識別情報など13項目、(3) 大漢和・新字源、大字源・大字典などの漢和辞書の検字番号、新字源の読み、画数などの10項目、(4) 政令で規定された常用漢字、当用漢字、教育漢字、学年別配当など11項目の識別情報である[1]。

## 3. 日本語データベースの概要

日本語データベースの構築は、コードブックを電子化した漢字データベース、新聞記事の本文データベースを中心に、新聞記事の切り抜きデータベース（来年度作成予定）を疑似的に結合する方法をとった。本文データは、大量の新聞記事データの単位切り作業と修正時間を省くため、単位情報をデータに埋め込むことを避けた。また、データの検索には、使用する研究者の知識を利用できることを前提に、任意の文字列を指定する方法をとった。これは、研究対象や研究者によって要求する単位が異なることから、要求の全てをデータでもつことが不可能と判断されたことによる。そのほか、漢字データベースの検索結果から新聞記事データベースの文例を検索するさいのキー長を調整する操作を省く効果をねらったものである。文字列検索の導入は、キーの語長からの独立と、異なるキー長をもつデータベースを疑似的に結合することができる。

各データベースの構成は、各データベースが独立に作成されたためCD-ROM上に併記させ、キー情報を受け渡すことによって疑似的に結合させた。新聞調査では、1/60で面サンプリングしていることから、一文がサンプリング区画を外れた場合に文脈が切れることがある。本文は、校正処理で修正できる範囲で補填した（切り抜き記事は、用語調査の1 サンプリング面に相当する）。切り抜き記事データベースは、画像処理に対する先行実験の性格をもつものとして扱った。

データベースは、独立に検索する場合と、漢字の属性情報（漢字の読み、画数、部首などの情報）で検索した結果を、二次情報として新聞記事のデータベースのキーに渡す方法に分けた。見出し漢字を変更した場合には、検索結果の「属性情報の一覧」、「詳細情報」、「用例」を関連付けて見られるように連動して変化させた。各情報はブラウジングが可能である。

検索条件に該当した一次検索結果は、見出し字形、区点番号、改正情報、画数、当用漢字、常用漢字、学習漢字を表示し、カーソルによる二次選択を許した。選択された情報は、詳細情報として装置の漢字符号、辞典類の検字番号、部首や画数、音訓の読み、常用漢字、教育漢字、学習漢字、人名で使われた識別記号を表示した。また、詳細情報と連動させて用例は、表示できる文長に制限があるため、全文を表示する選択機能を設けた。

検索は、検索条件の少ない属性情報については、Visual Basicのコンボ・ボックス機能を使い入力操作の軽減をはかった。7 種の検索条件の選択は、ボックスに表示された情報をクリックする。キーで複合条件を指定した場合には、論理積(and）が指定される。指定した条件に該当する情報が見つからない場合には、メッセージを表示した。

そのほか、これまでの用例集の作成では、キーとなる単語の長さが事前に決められ、単位規則にしたがって作業が進められていた。複数の単位に対応した用例集を作成するためには、一文に複数の単位を埋め込むことが必要になる。この処理は、大量データの用例を作成するためには、単位やデータに関する修正と校正を困難なものにした。一方、用例集の単位は、それぞれの研究者が希望するものであることが必要である。本プログラムでは、これらの条件に対応するため、文字列による検索を基本とした。

属性情報として指定できるキーの一覧

属性情報による条件: 漢字, 区点番号、改正情報、JIPS、総画数、部首内画数

部首属性による条件設定：部首、部首コード

辞書による条件設定：新字源、大字源、大漢和辞典、大字典の各検字番号

順位度数による条件設定：雑誌順位、雑誌度数、新聞度数、雑誌人名、新聞人名、雑誌地名、新聞地名、高校順位、中学順位、中学度数、文学順位、文学度数

読みによる条件設定：音読み、訓読み（新字源）

その他の条件：教育漢字1 、教育2 、学習漢字、当用漢字、当用（加）、当用（減）、常用漢字、人名1 、人名2 、人名3 、人名4 の識別符号

文字列による条件：検索文字列

雑誌：雑誌九十誌調査

新聞：昭和41年度発行, 朝日・毎日・読売朝夕刊1 年分の調査

中学：昭和56年度の中学校教科書調査（理科、社会科など7 教科の調査）

高校：昭和49年度の高等学校教科書（理科、社会科など 9教科の調査）

文学：直木賞受賞作品4 本

教育漢字1 ： 昭和52年小学校学習指導改訂後の漢字996 字に関する学年配当情報（1 から6）

教育漢字2 ： 平成6 年小学校学習指導改訂後の漢

字1006字に関する学年配当情報（1 から6）
人名1 ： 昭和26年内閣告示・訓令の人名用漢字別表92字に関する識別情報（該当1 非該当0）
人名2 ： 昭和51年内閣告示・訓令の人名用漢字別表120 字に関する識別情報（該当1 非該当0）
人名3 ： 昭和56年に新たに追加された166 字に関する識別情報（該当1 非該当0）
人名4 ： 平成2 年に新たに追加された284 字に関する識別情報（該当1 非該当0）
JIPS ：日本電気株式会社で規定したJIS X0208相当の漢字符号と文字集合
部首内画数：総画数から部首部分を省いた画数

「部首・画数・読みを複合条件で検索する例」

最初に図1の初期画面から「部首属性による条件」をクリックし、コンボボックスから部首を入力する（事例では部首「土」を指定した） （図1)。キー情報は、直接字形として入力する場合と別に設けた部首一覧から指定する場合の二つを選択できる。入力条件を確認した後、「OK」ボタンをクリックする。次に、「画数」による条件を選択し、「9 」を指定する（図2)。「漢字検索」ボタンを選択すると部首「土」で画数が9 の漢字の論理積に該当する情報が表示される（図3)。

さらに、この一覧を絞り込む場合には、「読みによる条件」を選択し、1 バイト系の文字で漢字の読みを入力する（入力事例では、「ｹｲ」を指定した）。正しく入力されていることを確認し「OK」をクリックする（図4)。検索結果に対して該当する漢字の用例が必要な場合には、「文例検索」ボタンをクリックする（図5)。その漢字の属する全文を必要とする場合には、用例が表示された画面の「原文表示」を選択する（図6)。

「文字列のよる検索例」

文例の検索は、最初に「文字列による条件」を選択する。表示された画面から文字列を入力し、「文例検索」ボタンをクリックする（キーは「国語」を指定)(図7)。漢字の検索結果をキーとして使う場合には、漢字に関する情報が検索された結果を確認し、画面の「文例検索」をクリックする。該当する字形が文例にある場合は、KWIC形式で文例が表示される。

さらに長い文例を必要とする場合には、用例表示画面の「原文表示」ボックスをクリックする。原文は、現在表示されている画面の上に表示される。検索された漢字の字形が含まれる本文情報を見たい場合の検索条件は、文字列を優先した検索を行う。該当する文字列が存在しない場合には、誤りのメッセージを表示して再入力を要求する。なお、本文データは、出典情報として新聞名、日付、ページ、サンプリング・ブロックの番号にセンテンスを付けたものである。情報の表示形式は、図6 に示したものと同じである。

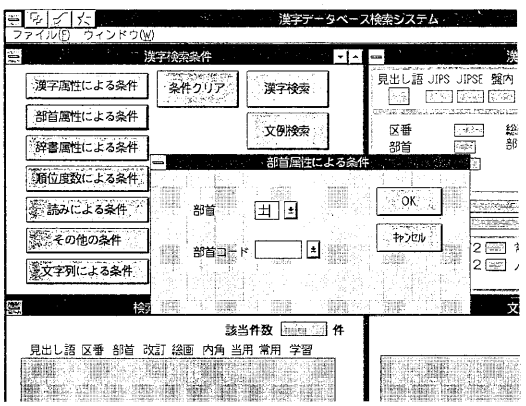

図1　部首情報の指定

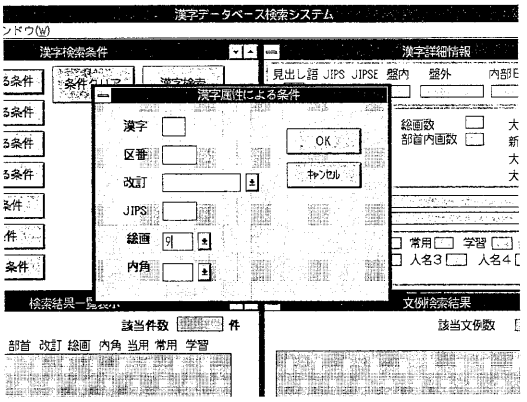

図2　画数情報の指定

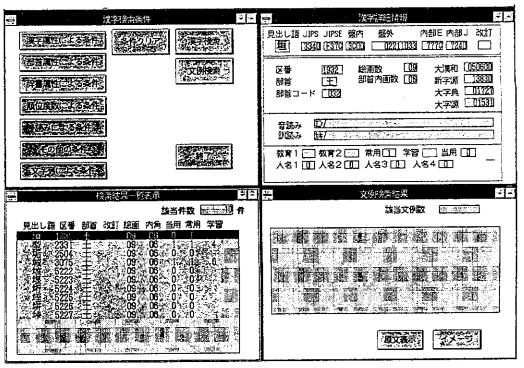
図3　部首と画数による一次検索結果

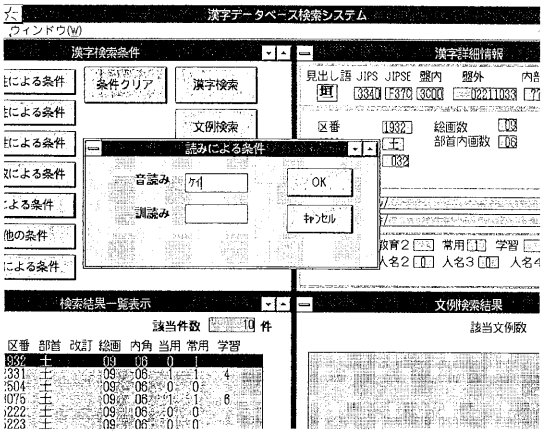


図4　二次検索用の読みの指定

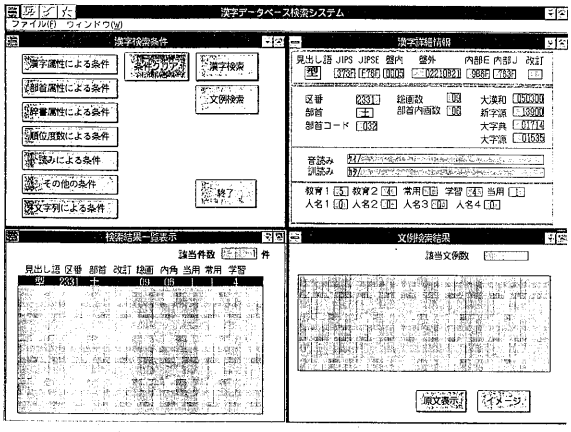


図5　部首・画数・読みによる検索結果

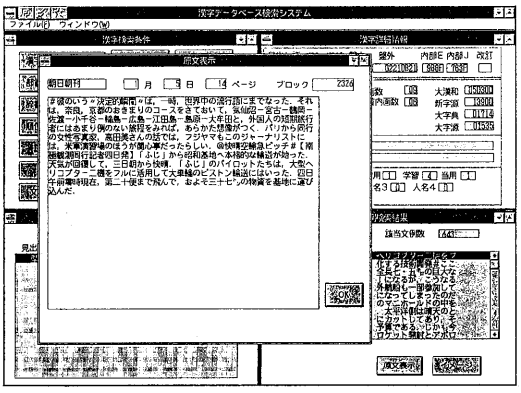


図6　原文検索結果

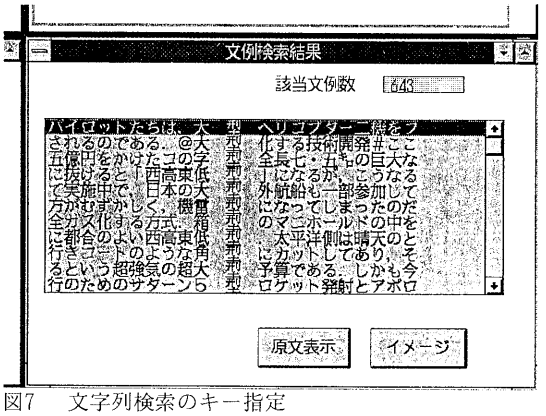


図7　文字列検索のキー指定

## 4．ファイル構造

漢字データを統合した漢字ファイルは、6,349 字の漢字に見出し漢字に44個の属性情報を付加したものである。属性情報は、検索のさいのキーになるものである。漢字の属性情報をキーとして検索を行なう場合には、多様な検索条件が指定される。処理時間の高速化に対応するためには、索引構成を個々の使用目的に合わせて配列することが必要になる。しかし、本体となる漢字ファイルは、詳細情報の全てが記録されており、検索の条件にあわせて再配列することは処理時間の点から実務的ではない。漢字ファイルは、この二つの条件を満たすために複数の配列条件に対応できる索引部と、基本配列が固定される本体部に分けて設定した。

索引部と本体部の分離は、データベースに使用する漢字と文例データが共に閉じた状態におけるため事前編集による最適構造を決定することができる。データ構成方法の最適化は、検索処理の高速化手段に対応できる。索引としてまとめた属性情報は、漢字符号系を省いた漢字字形、部首情報、本文などの40種類である。索引部と本体部の結合は、中間に「対応表」をおいた。対応表による二つの情報群の結合は、異なる配列をもった複数の索引を設定でき、本体情報の引用を検索条件にあった部分に絞ることができる。

この処理は、本体部分の全ての情報を読む必要がなくなることから、記憶装置におく情報を索引に限定できる。これは、記憶装置に読み込むデータ量を減少させ、検索処理に必要な時間を減少させる効果を期待できる。

特に、読み情報に関する索引は、同一漢字が複数

の読みをもつことが多いことから、実用化には検索速度の効率化が重要になる。この条件を満たすため、読みの索引は、構造を二段とした。第一次索引は、第二次索引に対する全ての読みに対する索引である。一次索引は、'ア'から'ワ'の読みの先頭1文字で構成する情報と二次索引への連結情報である。第二索引は、6,349字に付けた全ての読み情報を50音順に配列したものと漢字ファイルへの連結情報である。読みを指定した検索では、二つの索引と対応表を使い、キーとして入力した条件と一致する属性情報を漢字ファイルから読み出す。

部首情報の索引例は、部首「二」は二(ニ), 于(ウ), 井(イ・セイ), 云(イウ・ウン), 五(ゴ), 互(ゴ), 弍(ニ), 亙(コウ), 亘(セン・カン・ワタリ), 亜(ア), 些(サ・イササカ), 竺(ジク), 亞(ア), 亟(キョク・スミヤカ)の14字がある。検索処理では、14個の情報を漢字ファイルから読み込むことになる。部首に関する情報は、記憶装置に部首、部首コード、該当する部首にある漢字数、対応表への連結情報で構成されている。キーとなる部首から漢字数と連結番号を求め対応表を引く。次に、対応表に記録されている漢字ファイルへの連結情報から漢字ファイルへの連結処理を行う。ここで同一部首に属する漢字の全てを対応表に記録されている一覧を見て、漢字ファイルから読み込み表示する。

索引と対応表の連結情報は、漢字ファイルにある属性情報の先頭位置を開始点とし、各属性情報の位置を相対番号で表現したものを使った。対応表と漢字ファイル間の結合情報は、漢字ファイルの見出し漢字に1から6,349の連番をふり連結情報とした。対応表には、索引で指定される可能性のある検索条件を全て漢字ファイルとの連結情報として登録することになる（例えば部首の場合には214個の連結情報を設定する）。また、対応表には、連結情報のほか漢字ファイル情報が可変長の場合があるため、引用する情報の番号と長さの二つの情報を付加した。

## 5．今後の方針

漢字と属性情報、文例を含む日本データベースは、海外の日本語研究者や日本語教育関係者から教材として提供されることを期待されている。しかし、JIS X0208やJIS X0212には、大規模の漢和辞典や古典・漢籍を符号化できる充分な文字種がなく、インターネットを通して安定した状態で正確に伝送する方法も確立していない。

今後この研究に関しては、新聞記事データの校正を引き続き行うと共に、約12,000枚の新聞の切り抜き記事をデータベース化し、日本語データベースとの結合をはかる。また、大規模な文字種を表現するための漢字符号の伝送実験を日本語データベースをテストデータに使い、インターネット上で行う。日本語データベースは、これらの基礎実験の後、日本語教育の現場で教材としての日本語データベースの有効性をオーストラリアのモナシュ大学との間で確認する。

漢字データベースは、JIS C6226-1978で規定した文字集合を中心に構築した。ここで作成したデータベースは、一つの国で使用される漢字と本文をまとめたものであり、多国語の混在を想定したものではない。しかし、多国語化が進むにつれ同じ字形がどの国で使用されたものであるかを識別できなければ、電子化されたデータからの辞書情報の引用ができなくなる。従来、多国語処理は、入力時に付属情報としてデータに埋め込む形で処理されてきたが、事前にデータの編集が必要になると共に、扱うデータの長さが冗長になる問題がある。漢字処理の多国語化への要求は、避けられない課題であり、漢字の国名などを漢字符号に埋め込む実験は今後の新しい漢字符号のあり方を検討する手段としてデータベースを拡張するなかで検証する。

［付記］本研究は、平成7年度文部省科学研究費・重点領域研究(1)として「インターネットにおける学術漢字の符号化に関する基礎的研究」（代表者：斎藤秀紀，課題番号07207129）の交付を受けて行ったものの一部である。

## 参考文献

1)斎藤秀紀，漢字情報データベース，研究報告集(9)，国語研報告94, pp. 27-47(1988).
2)斎藤秀紀，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，研究報告集(10)，国語研報告96，pp. 173-192(1989).
3)斎藤秀紀，大漢和辞典の検字番号に基づく構造化4バイトコードの提案，情報処理学会論文誌，Vol. 35，No. 6, pp. 1119-1126(1994).
4)斎藤秀紀・柳沢好昭・横山詔一，インターネットにおける学術漢字の符号化に関する研究，情報処理学会「人文科学とコンピュータ」研究会資料28-1，pp. 19-24(1995).
5)斎藤秀紀，JISに無い字をどう扱うか，「人文学と情報処理」勉誠社，No. 10, pp. 22-25(1996).