公開シンポジウム

人文科学とデータベース

「データ」を読む・観る・解く

1995年12月25.26日

◆◆◆◆◆◆◆◆◆◆◆◆ 目　　次 ◆◆◆◆◆◆◆◆◆◆◆◆

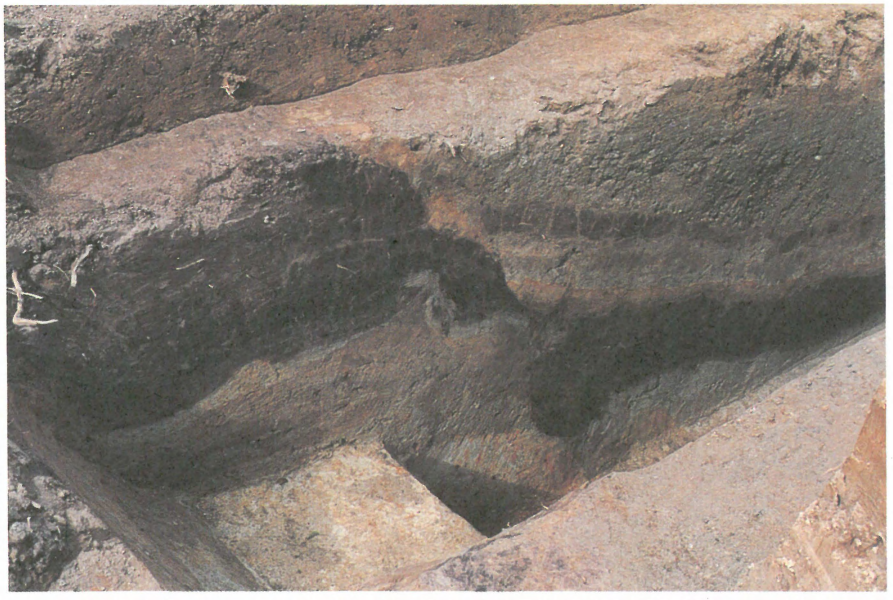
（特別講演）

（一般講演）

古地震データと活断層

寒川　旭　（通産省地質調査所大阪地域地質センター）

大阪市中央区大手前4－1－67　大阪合同庁舎2号館別館

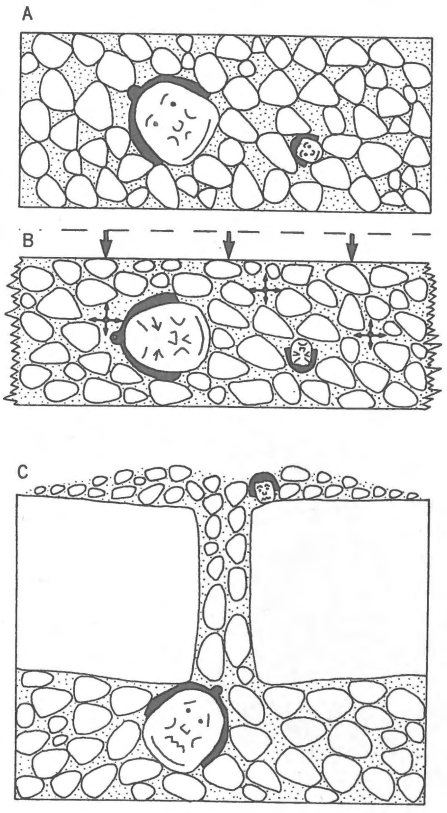


久留米市山川前田遺跡の活断層跡

（678年筑紫地震を発生させたもの）

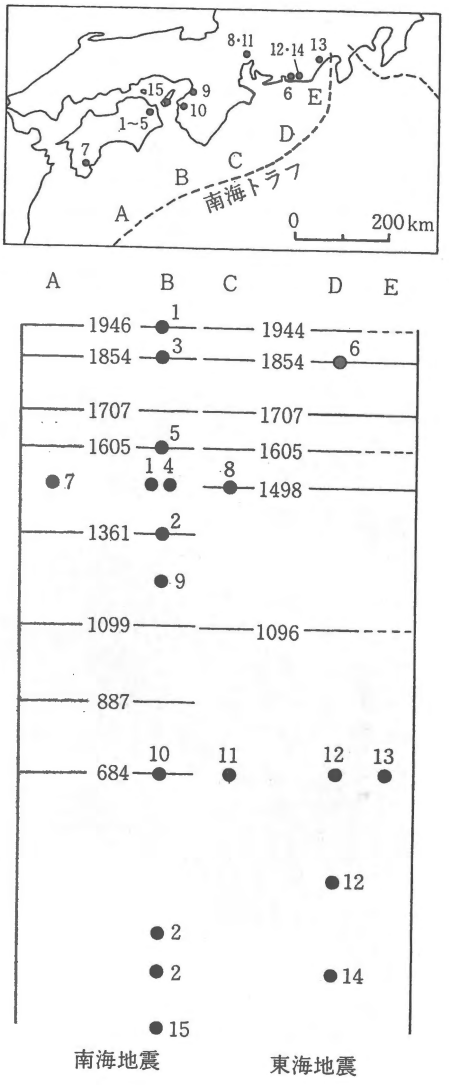
門真・守口両市の西三荘・八雲東遺跡の液状化跡

（1596年伏見地震による）



液状化現象のメカニズム

A　通常の状態
B　激しい地震動によって液状化現象が発生
C　上位の地層を引き裂いて噴砂(礫)が噴出

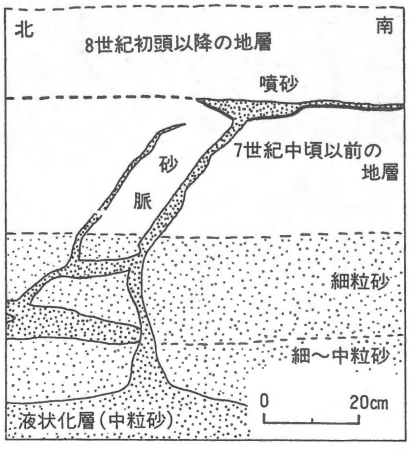


南海地震と東海地震の発生時期

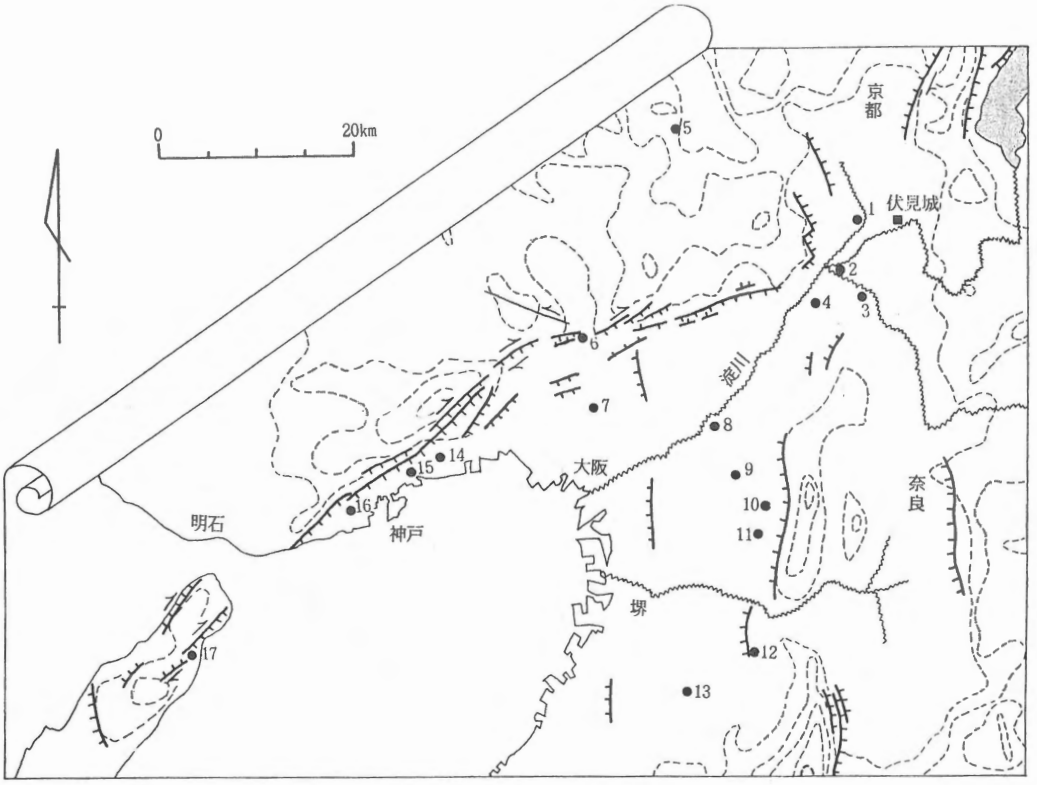
(黒丸印は遺跡で地震跡が検出されたもの)　1 宮ノ前遺跡, 2 黒谷川宮ノ前遺跡, 3 神宅遺跡, 4 古城遺跡, 5 黒谷川古城遺跡, 6 御殿二之宮遺跡, 7 アゾノ遺跡, 8 尾張国府跡, 9 石津太神社遺跡, 10 川辺遺跡, 11 田所遺跡, 12 坂尻遺跡, 13 川合遺跡, 14 鶴松遺跡, 15 下内膳遺跡.

階級	説　明
0	無　感　人体に感じないで地震計に記録される程度
I	微　震　静止している人や、特に地震に注意深い人だけに感ずる程度の地震
II	軽　震　大ぜいの人に感ずる程度のもので、戸障子がわずかに動くのがわかる程度の地震
III	弱　震　家屋がゆれ、戸障子がガタガタと鳴動し、電灯のようなつり下げ物は相当ゆれ、器内の水面の動くのがわかる程度の地震
IV	中　震　家屋の動揺が激しく、すわりの悪い花びんなどは倒れ、器内の水はあふれ出る。また、歩いている人にも感じられ、多くの人々は戸外に飛び出す程度の地震
V	強　震　壁に割れ目が入り、墓石・石どうろうが倒れたり、煙突・石垣などが破損する程度の地震
VI	烈　震　家屋の倒壊は30%以下で、山くずれが起き、地割れを生じ、多くの人々が立っていることができない程度の地震
VII	激　震　家屋の倒壊が30%以上に及び、山くずれ、地割れ、断層などを生じる

気象庁の震度階級

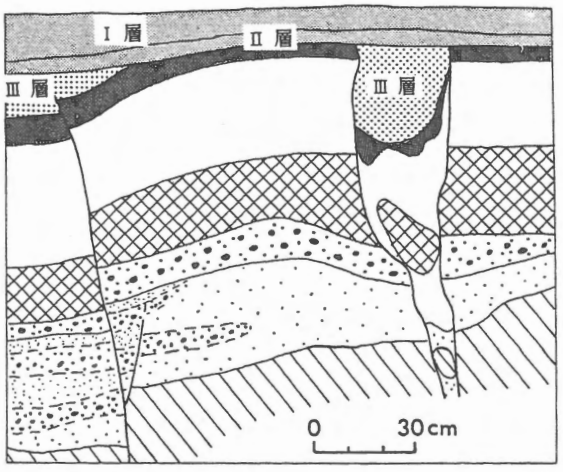


袋井市坂尻遺跡の液状化跡
(684年の東海地震による)



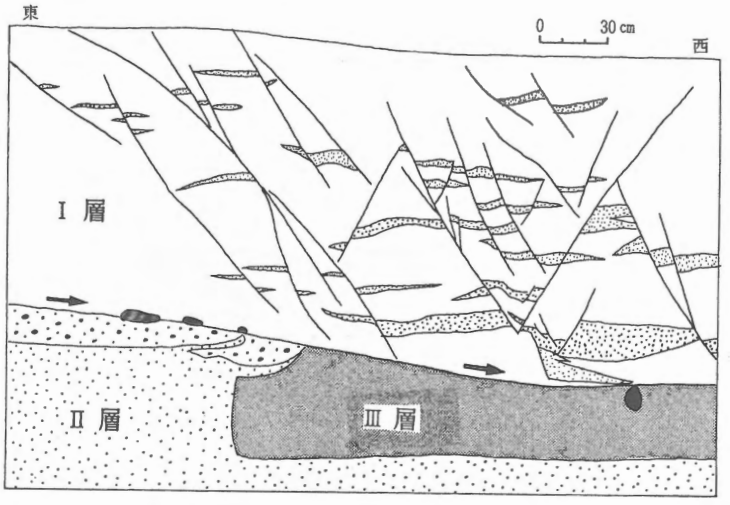
１６世紀末の地震の痕跡

1　志水町遺跡　2　木津川河床遺跡　3　内里八丁遺跡　樟葉野田遺跡　5　鹿谷遺跡　6　栄根遺跡　7　田能高田遺跡
8　西三荘・八雲東遺跡　9　西鴻池遺跡　10　水走遺跡　11　池島福万寺遺跡　12　高屋城跡　13狭山池遺跡
14　坊ケ塚遺跡　15　西求女塚古墳　16　兵庫津遺跡　17　佃遺跡



川西市栄根遺跡の地割れと小断層跡

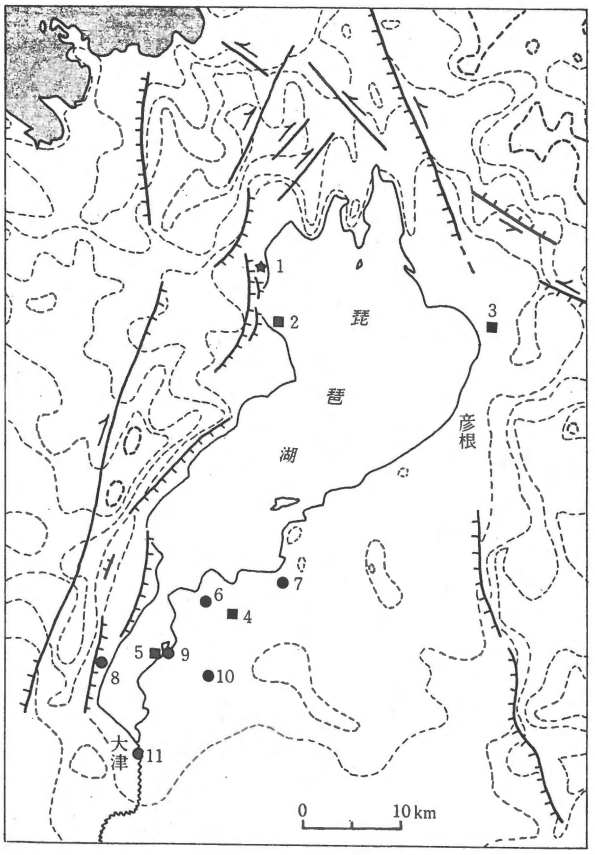
Ⅰ層：現在の耕作土　Ⅱ層：江戸時代の遺物を含む層　Ⅲ層：16世紀後半の遺物を含む層



神戸市西求女塚古墳の地滑り跡

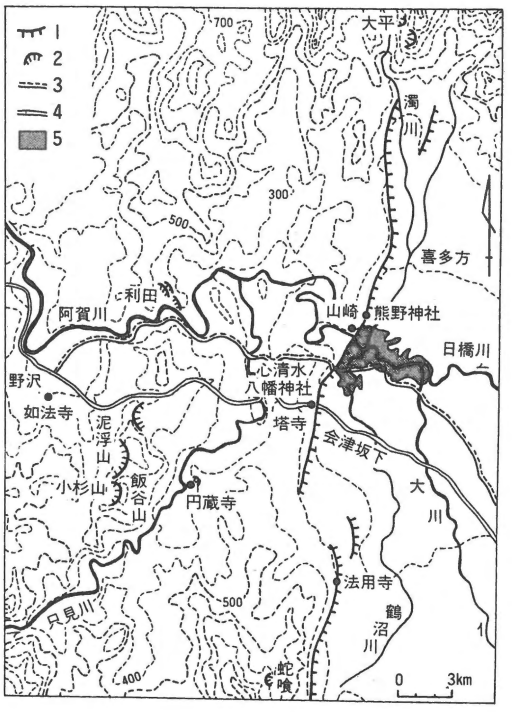
Ⅰ層：滑り動いた墳丘　　Ⅱ層：地山（粗粒砂）

Ⅲ層：16世紀後半の遺物を含む耕作土



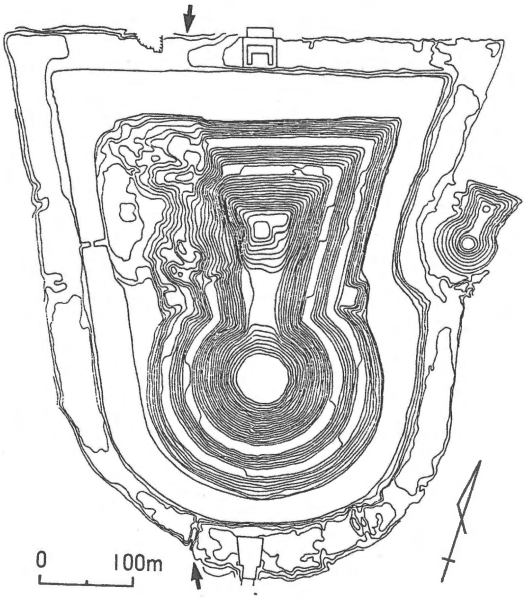
琵琶湖周辺の遺跡で検出された地震跡

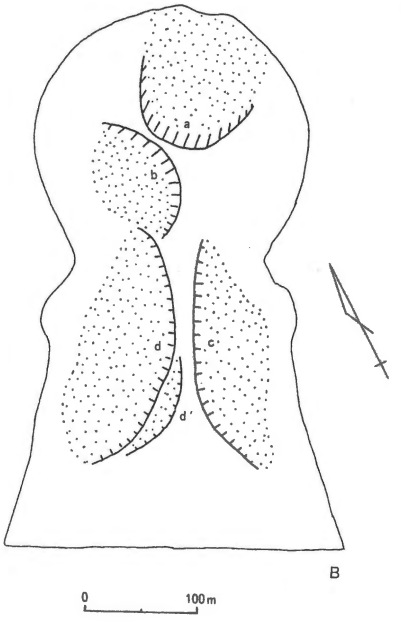
活断層：ケバをつけた側が相対的に下降，矢印は横ずれの方向を示す.） 1 北仰西海道遺跡, 2 針江浜遺跡, 3 正言寺遺跡, 4 湯ノ部遺跡, 5 津田江湖底遺跡, 6 堤遺跡, 7 加茂遺跡, 8 穴太遺跡, 9 烏丸崎遺跡, 10 野尻遺跡, 11 蛍谷遺跡(1：縄文時代晩期前半中ごろ，2〜5：弥生時代中期中ごろ，6〜11：中〜近世).



1611年会津地震で生じた地変

1 会津活断層系 2 崩壊地形 3 旧越後街道
4 新越後街道 5 山崎新湖





前方後円墳の変形跡

IntelligentPadシステムを用いた歴史学研究支援データベースの構築
Implementation of IntelligentPad-based database systems to support historical researches

赤石 美奈, 中谷 広正, 伊東 幸宏, 阿部 圭一, 田村 貞雄
Mina AKAISHI, Hiromasa NAKATANI, Yukihiro ITOH, Keiichi ABE and Sadao TAMURA

静岡大学情報学部
〒432　静岡県浜松市城北3－5－1
Faculty of Information, Shizuoka University
Johoku 3-5-1, Hamamatsu 432, Japan

あらまし: 本論文では、コンピュータ上の多様な情報や、アプリケーションを統一的に扱えるIntelligentPadシステムを基盤とし、歴史学研究で用いられる多種多様な史料を管理・検索するためのデータベースの構築について述べる。IntelligentPadシステムでは、テキスト、図表、画像、映像、音声などからなるマルチメディアドキュメントがパッドと呼ぶメディア・オブジェクトとして実現される。また、計算、作図、編集などの機能を持つ各種のアプリケーションもパッドとして実現される。本研究では、歴史学研究における多種多様な史料をパッドとして表現することにより、史料の種類にかかわらず、統一的に扱うことが可能とする。また、史料の編集・可視化など各種のアプリケーションをもパッドとして実現する。これにより、1867年の伊勢神宮・秋葉三尺坊大権現などのお札降りを発端とした「ええじゃないか」に関する各地の伝承や史料を収集した「ええじゃないかデータベース」を構築する。

Summary: This paper describes construction of a database system to support historical researches. The database system is implemented by the IntelligentPad system, where multimedia documents including texts, figures, movies and sounds are represented by media objects called pads, and applications such as calculation, drawing and editing are also represented by pads. Thus we can deal with multimedia information and applications in the same way. In this study we represent a variety of historical materials by pads, and we also represent such system functions as editing and visualizing historical materials by pads. Then we construct the Eejanaika Database which stores historical stories and materials concerning a historical event "Eejanaika" which was triggered by scattering of protecting charms in 1867.

キーワード: ツールキット・システム, マルチメディア・データベース, フォームベース
Keywords: toolkit system, multimedia database, formbase system.

1 はじめに

地域史研究の発展に伴い、伝承の採録や史料の発掘が盛んになってきた。歴史学研究においては、それらの多様な史料に基づき、独自の仮説・理論を展開する。さらに、各種の史料を基に、定説を再検討し、新たな仮説を生成していく。しかし、これらの仮説を裏付けるためのデータの整理・検索などは手作業に頼っており、多大な労力と時間を費やしている。また、各地で収集された史料や情報の共有が促進されていない。このため、歴史学研究において重要なプロセスである仮説検定や、全体像・具体像の把握が困難であり、コンピュータによる支援が望まれている。

本研究においては、１８６７年（慶応３年）の伊勢神宮・秋葉三尺坊大権現などのお札降りを発端とした「ええじゃないか」 [1, 2, 3] に関する各地の伝承や史料を収集した「ええじゃないかデータベース」を構築し、情報の共有を図るとともに、歴史学のニーズに特化したシステムを構築する。また、多種多様な史料に対して、さまざまな視点からこれらを検証する方法について検討する。

歴史学研究においては、多種多様な史料を扱う。「ええじゃないかデータベース」に格納される史料は、古文書、書き下し文、解釈文、音声、動画（踊り）等が挙げられる。これらの各種メディアを統一的に扱うためには、コンピュータ上に新たなメディアが必要であると考える。これにより、雑多な情報を統一的に扱い、均一なプロトコルを提供することができる。

IntelligentPad システム [4, 5] では、テキスト、図表、画像、映像、音声などからなるマルチメディアドキュメントがパッドとして表現されるのみならず、データベースシステム、メールシステムなどのシステムサービスプログラム、各種アプリケーションプログラムなどもすべてパッドとして表現される。本研究においては、この IntelligentPad システムを基盤システムとして用い、「ええじゃないかデータベース」を構築するとともに、歴史学研究における情報活動を支援するための統合環境を提供することにより、歴史学研究の促進を図ることを目的とする。

本論文は、以下のように構成される。第２章において、歴史学支援のための統合環境を構築する基盤システムである IntelligentPad システムの概要について述べる。第３章においては、メディアデータを管理するためのフォームベースシステムについて述べる。第４章においては、「ええじゃないかデータベース」について述べる。第５章において、まとめを述べる。

2 IntelligentPad システムの概要

IntelligentPad システムでは、コンピュータで扱えるマルチメディアデータ、ユーザの定義したアプリケーション・プログラム、システムにより提供される各種サービス・システムを紙のイメージを持つパッドとして統一的に扱うことが可能である。さらに、パッドの貼り合わせにより、個々のパッドの持つ機能を合成し、新たな機能を定義することができる。本章では、IntelligentPad システムの概要について述べる。

2.1 パッドの内部構造

システムにより提供される基本部品であるプリミティブパッドの内部機構は、Smalltalk-80 [6] により提案されている MVC 構造に基づく。M は、データの保持、管理、加工などを定義した Model オブジェクトを表す。V は、データの表示形態を定義した View オブジェクトを表す。C は、マウスクリックやキーボード入力などのユーザのイベントに対する反応を定義した Controller オブジェクトを表す。パッドは、これらの Model オブジェクト、View オブジェクト、Controller オブジェクトからなるメディアオブジェクトである。

図 1に、パッドの内部構造を示す。コントローラは、ユーザからのイベントをきっかけとして、ビューにメッセージを送る。ビューは、モデルの状態を変更したり、モデルに保持されるデー

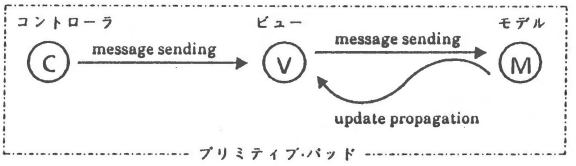


図 1: パッドの内部構造

タを読み出すために、モデルへメッセージを送る。モデルは、保持しているデータの変更などに伴い、更新伝搬メッセージをビューに送る。

2.2　パッドの機能合成

各パッドは、種類に応じて固有の機能を有する。テキスト入力パッド、数値入出力パッド、ボタンパッドなど、システムにより提供されるプリミティブパッドは、単純な機能を有する。これらの複数のプリミティブパッドの合成により、複雑な機能を有する合成パッドを定義することが可能である。パッドの合成は、パッド同士を画面上で重ね合わせ、「貼る」ことにより行われる。貼り合わされたパッド間では、データの授受やコマンドの起動をかけることができる。これにより、個々のパッドが持つ機能を連携し、複雑な機能を持つパッドを構築することができる。パッド間のインタフェースは、統一してあるため、任意のパッドを組み合わせて合成することが可能である。各種のメディアを扱うためのパッドを貼り合わせることにより、複合メディア文書を作成することができる。貼り合わされたパッド間のデータ授受や、コマンドの起動は、各パッドのスロットを通じて行われる。

図 2に、IntelligentPad の画面ハードコピーを示す。図中の矩形領域がすべてパッドであり、それらの合成によりさまざまなアプリケーションが構築されている。

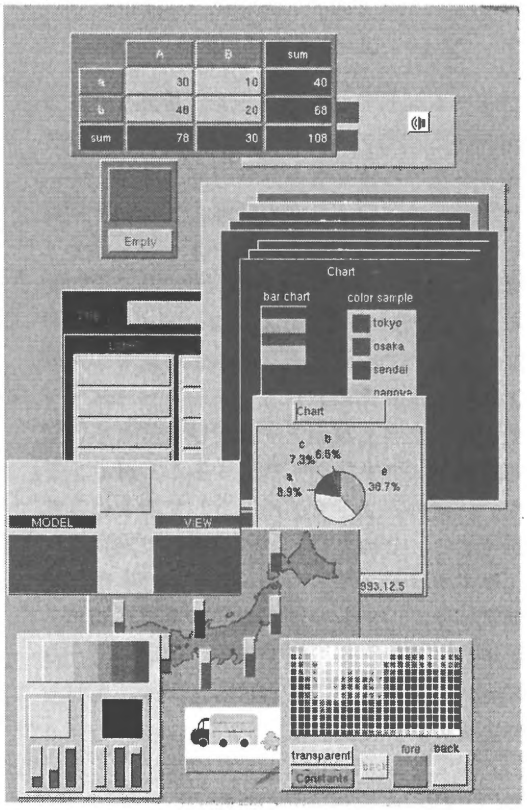


図 2: IntelligentPad の画面ハードコピー

2.3　パッドの複製

パッドは複製機能をもつメディアである。複製により、パッドの保持する情報を再利用して新たなパッドを定義する部品として利用することができる。パッドの複製には、内部状態を共有する共有コピーと、共有しない非共有コピーがある。共有コピーされたパッドは、コピーの後も、同じ情報を参照する。非共有コピーは、コピーされた後は、独立した部品として動作する。図 3に、プリミティブパッドの共有コピーと非共有コピーの構造を示す。図の中央のオリジナルのパッドを共有コピーした場合、オリジナルのモデルを共有した複製の構造が図の右に示されている。また、オリジナルを非共有コピーした場合の複製の構造が図の左に示されている。

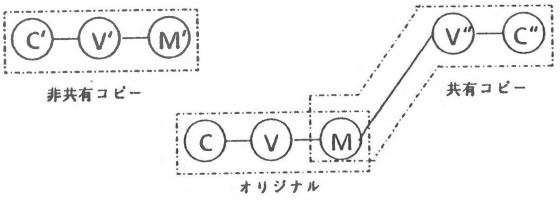


図 3: プリミティブパッドの複製

3 フォームベース・システム

3.1 フォームベースの概要

フォームベースは、フォーム（書類）として定義された定型フォーマットのデータを管理・検索するためのデータベースシステムである。データベースへのデータの登録、削除、検索は、ディスプレイ上に表示されたフォームを通じて直接的に行う。FORMANAGER [7] や FORMAL [8] などの従来のフォームベース・システムは、文字列や数値データのみを対象として開発された。本研究におけるフォームベース・システムは、文字や数値データだけではなく、定型の構造データを管理・検索することができる。また、任意のパッドをデータとして取り扱うことができる。このため、パッドとして表現される、マルチメディア・ドキュメントのみならず、各種アプリケーションプログラムもデータベースに格納することが可能である。本章では、IntelligentPad システムにおける、フォームベースの実現について述べる。

3.2 フォームベースの構成

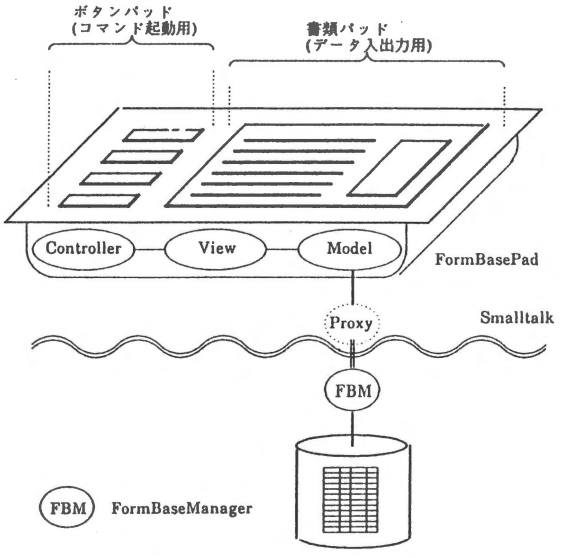


図 4: フォームベースの構成

図 4にフォームベースの構成を示す。フォームベースは、定型フォーマットの定義と、定型データの入出力機能を提供するフォームパッド部分と、フォームの管理・検索機能を持つデータベースにアクセスするためのインタフェースの役割を持つフォームベースパッドと、データベースを起動させるトリガーとなるボタンパッドにより構成される。

フォームは、フォームパッドに任意のパッドを貼ることにより定義される。扱うデータに応じて、テキストパッド、数値パッド、画像パッド、パッドデータ化パッドなどをフォームパッド上に貼る。各項目のレイアウトは、ユーザが自由に決定できる。フォームパッドは、各項目の項目名とデータの対からなる連想リストをスロットに保持する。

フォームベースパッドは、フォームパッドとスロットを通じて入力されるデータを管理するためのデータベースへのインタフェースの役割を果たす。また、ボタンパッドが押された場合には、データベースに対し、データの登録、検索、削除、更新の要求をおくる。任意のフォームパッドを定義し、フォームベース・パッドと合成することにより、さまざまなフォームベースを構築することができる。

4 ええじゃないかデータベースの構築

4.1 ええじゃないかデータベースの概要

歴史学研究においては、膨大な史料を対象とし、これに基づき、仮説の提唱、仮説の検証、仮説の修正というプロセスを繰り返し、独自の理論を展開する。この仮説の裏付けとなる、史料収集、整理には手作業による多大な労力を要する。また、仮説の検証段階では、この史料の中から必要なものを探し、各種の分析を行わなくてはならない。さらに、史料は、各地に分散しており、物理的な制約の下にさらにそれらの共有が不十分である。これをコンピュータにより支援するためには、歴史学研究で扱われる多様な史料データを統一的に扱うためのメディアと、それらを加工・編集する技術、及び統合管理・検索手法の提供を統合環境のもとに構築することが必要である。

１９６７年（慶応３年）の「ええじゃないか」は、お札降りをきっかけに祝祭がおこなわれ、それが過熱化して民衆の狂喜乱舞をまきおこし

図 5: パッド化された史料

た事件である。その際の最初のお札降りがどこでおこったのか、それにたいする対応がどうであり、それがどのようにして「ええじゃないか」に発展していったのかをあきらかにする過程において、具体的にコンピュータがどのように支援をしていくかを検討していく。

本研究においては、「米価データベース」、「政治日程データベース」、「史料データベース」を構築し、これを用いて、政治や経済との関係や、それらが、「ええじゃないか」に与えた影響を検討する過程において必要なツールを開発し、歴史学のニーズに特化したシステムを構築する。

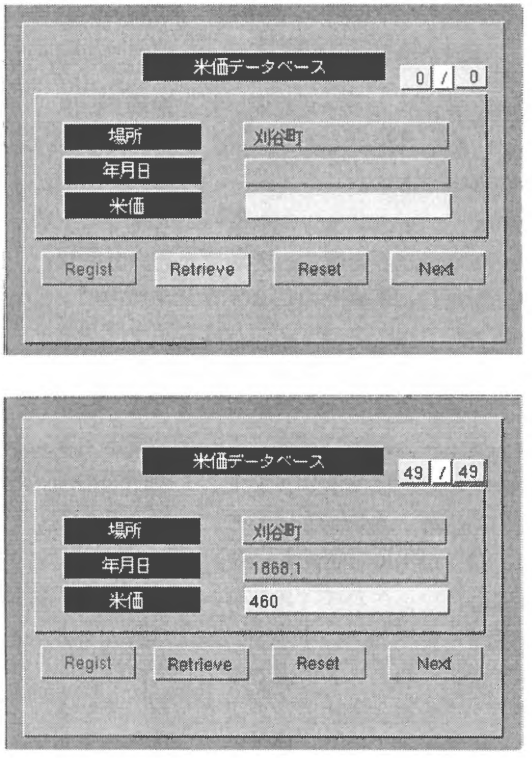


図 6: 米価データベース

4.2　史料のパッド化

図 5に、パッドとして表現された、史料を示す。 それぞれの史料に応じて、各種のプリミティブ・パッドを利用し、それらの合成により任意のデータ構造を持つ合成パッドが定義される。図 5の左には、御札降りについて記述した書物の実物写真のパッドと、書物の内容を活字にしたものを表すパッドを 1 枚の台紙に貼った

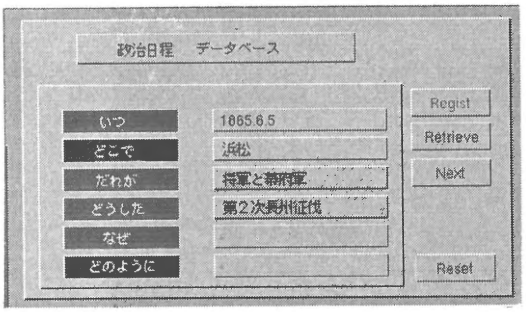


図 7: 政治日程データベース

合成パッドが示されている。また、図 5の右には、御札の写真を示したパッドに、その説明のコメントが貼られたパッドが示されている。また、御札降りの様子が描かれた画像もパッドである。これらの史料は、すべてパッドとして IntelligentPad システムに取り入れられており、図 2に示したパッドとの合成が可能である。また、史料データの種類や構造にかかわらず、パッドデータ化パッドにより、データとしてフォームの一項目に入出力可能である。

4.3 史料の管理・検索

現在、フォームベースを用いて、米価データベース、政治日程データベース、史料データベースを構築中である。

図 6に、フォームベースを用いて試作した「米価データベース」を示す。米価データベースは、場所、年月日、米価を項目とするフォームパッドとフォームベースパッドの合成により実現される。各項目を埋め、登録ボタンが押されると、そのデータがデータベースに登録される。検索条件もフォームの項目を埋めることで指定する。図 6の上には、検索条件として、'場所'の項目を'刈谷町'と指定している。検索された結果は、図 6の下に示されている。

図 7に、フォームベースを用いて試作した「政治日程データベース」を示す。政治日程データベースは、'いつ'、'どこで'、'だれが'、'なぜ'、'どのように'、'どうした'という情報を管理する。図は、第 2次長州征伐に進軍中の将軍と幕府軍の位置を検索したところを示す。

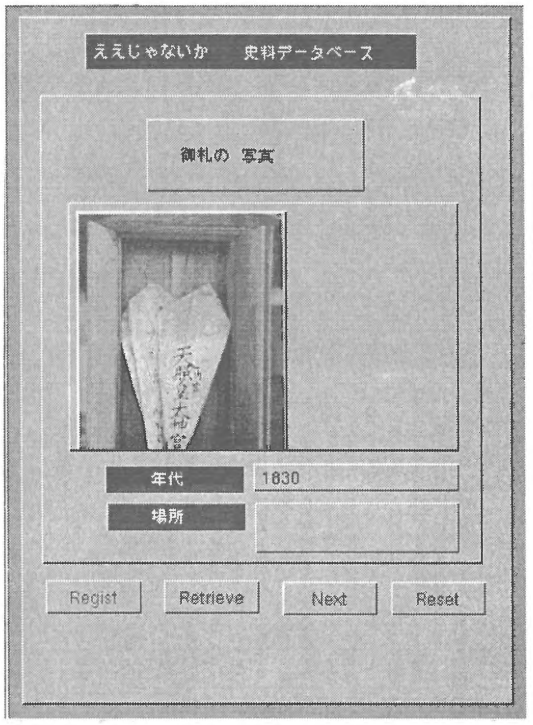


図 8: 史料データベース

図 8に、「史料データベース」の画面ハードコピーを示す。コンピュータ上に取り込んだ史料は、パッドとして統一的に扱うことが可能である。扱う史料は、その種類やデータ構造に応じてさまざまなパッドで表現されている。これらの史料はパッドデータ化パッドにより、フォームの一項目としてデータベースに格納される。フォームには、史料のタイトルと、年代、場所に関する情報を記述する項目がある。また、各史料に固有の情報は、パッドとして直接史料に付加して格納されているものとする。

現在、史料のタイトル、年代、場所の情報から検索を行うことは可能である。これらを検索条件にして、必要な史料パッドをデータベースからとりだすことができる。史料パッドの内容を検索条件として用いる検索機構は、構築中である。

4.4 仮説検証過程の支援

検索された情報を、適切に表現するためのツール群を利用することにより、全体像・具体像の把握・理解を支援する。検索されたデータは、

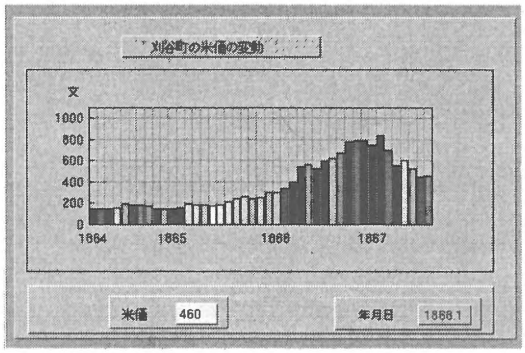


図 9: 米価の変動

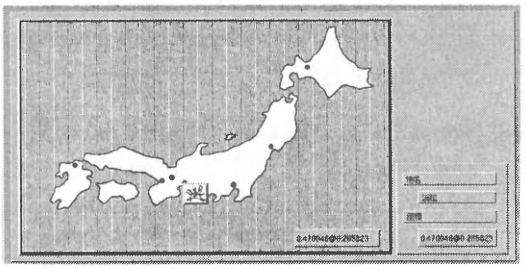


図 10: 幕府軍の位置

IntelligentPad に用意されている既存のツールを用いて、自由に表現することができる。データの変動をグラフで表現したり、各地のできごとを地図上にマッピングすることにより、全体像の把握を容易にする。データの授受は、データを保持するパッドをデータを入力したいパッドに貼るという直接操作により可能となる。

図 9に米価データベースから検索された刈谷町の米価と年代のデータをグラフパッドを用いて、棒グラフにしたものを示す。これにより、１９６７年頃に米価が大きく変動していることが見て取れる。図 6の米価データベースから検索された米価の値をグラフの縦軸に入力し、年月日の値を横軸に入力する。データの転送や授受は、パッドの貼り合わせを通じて行うため、転送したいデータを持つパッドの共有コピーをとり、これをグラフパッド貼り、値を入力している。

図 10は、政治日程データベースから、米価の変動に直接的な契機と考えられる第二次長州征伐の幕府軍の進軍日程を引き出し、それを地図上にマッピングしたものである。幕府軍の位置が地図上で示されている。図中の右下に示されているパッドは、地名を入力すると、それに対応した地図上の位置座標を返す。日本地図の描かれたパッドの上には、与えられた座標に移動する機能を持つパッドが貼られている。これらのパッドの合成により、地名を与えると、その位置にパッドが移動する合成パッドが実現される。

5　おわりに

本論文では、[ええじゃないかデータベース]の構築に関して述べた。本システムを IntelligetnPad システム上に構築することにより、歴史学で扱うさまざまな史料をパッドとして統一的に扱うことが可能であることを示した。また、これにより、パッドの貼り合わせにより、既存のシステムとの合成が容易におこなえる。「ええじゃないか」について論じるためには、ええじゃないかに関する史料のみではなく、政治や経済などの動きなど、さまざまな情報を必要とする。これらに関する情報も、フォームベースを通じてパッド化することにより、統合環境のもとで、多様な視点から史料を検討することが可能となる。また、郷土歴史家により、収集されている既存の史料データベースを、本システムに統合することにより、既存の異種データベースに格納されている史料や、各種データを有効に利用することができると考える。

参考文献

[1] 田村貞雄：ええじゃないか始まる, p.240, 青木書店 (1987)

[2] 田村貞雄：三重県域の「ええじゃないか」, 東海近代史研究会　東海近代史研究, 12, pp.86-105 (1990)

[3] 田村貞雄：岐阜県域の「ええじゃないか」, 東海近代史研究会　東海近代史研究, 13, pp.16-27 (1991)

[4] Tanaka, Y.: A Toolkit System for the Synthesis and the Management of Active Media Objects, Proc. 1st Int. Conf. Deductive and Object-Oriented Databases, pp.269-277 (1989)

[5] Tanaka, Y., Nagasaki, A., Akaishi, M. and Noguchi, T.: A Synthetic Media Architecture for an Object-Oriented Open Platform, Proc. IFIP 12th World Computer Congress, pp.104-110 (1992)

[6] Goldberg, A. and Robson, D.: Smalltalk-80: The Language and its Implementation, Addison Wesley (1983)

[7] Yao., S. B., Hevner, A. R., Shi, Z., and Luo, D. :FORMANAGER: An Office Forms Management System, ACM Trans. of Office Information Systems, Vol.2, No.3, pp.235-262 (1984)

[8] Shu, N. C. :FORMAL: A Forms Oriented Visual Directed Application Development System, IEEE Computer, Vol.18, No.8, pp.38-49 (1985)

4次元歴史空間システムにおける地理情報処理について

Geographical Information Processing of the Four-Dimensional Historical Space System

小林 努、加藤 常員、小沢 一雅
Tsutomu KOBAYASHI, Tsunekazu KATO, Kazumasa OZAWA

大阪電気通信大学、寝屋川市
Osaka Electro-Communication University
Neyagawa-shi, Osaka, 572, Japan

あらまし: 本稿では、4次元歴史空間システムと名づけた考古学研究支援システム上での地理情報処理について述べる。まず、本システムの基本コンセプト、プロトタイプシステムの構成、データベース、操作システムについて概要を紹介する。さらに、本システムにおける地理情報処理機能に関連する時間次元と空間次元の表現、特別な検索機能などについて述べる。

Summary: In this paper, geographical information processing of the four-dimensional historical space system has been presented. First, the basical concept and structure of the system have been outlined. The database structure and database managing system have also been described. Secondly, representation of the time and spatial dimensions have been discussed in relation to the geographical information processing. Finally, some specialized geographical retrieving functions have been introduced and their experimental results have been presented.

キーワード: データベース、考古学、地理情報処理

Keywords: database,archaeology,geographical information processing

1 はじめに

ワークステーションや高性能パソコンの発展により、大量のデータを手軽に扱える環境が整いつつある。また、柔軟なソフトウェアや周辺機器の出現により考古学に限らず人文科学系の研究室でも大規模なデータベースやシステムの構築が容易に行なえるようになった。

考古学においても昨今さまざまなデータベースの構築が試みられている。本稿では、4次元歴史空間システム [1][2] と名付けた考古学研究支援を目的としたシステムの開発について述べる。

本システムは、もともと時間軸を有する考古学データと3次元地形データを結合することによって遺跡・遺物・遺構などのデータ単位を4次元的に把握することをめざしている。

本稿では、このうち、地理情報処理機能について紹介する。

2 4次元歴史空間システム

2.1 システムのコンセプト

4次元歴史空間システムの基本構想は、考古学が対象とする古代世界をコンピュータ内にデータ空間として再現しようとするものである。

4次元歴史空間システムは、3次元の地形データを背景に遺跡データを時代と地理に関して横断的に操作する。

一般に研究支援システムは特定の限られた問題に対して高度な研究支援機能を提供するシステムとして開発され、研究対象のデータベース化と研究目的にあった検索やデータ処理をめざすところにあるといえる。4次元歴史空間システムも基本路線は同じであるが、考古学者の考察空間を広げて、問題の発見や着想のきっかけを与えようとするところに本システムの特徴がある。

2.2 プロトタイプシステムの構成

システムの構成は、データ空間を作り出すデータベースとそのデータベースを効率よく取り扱うための操作システム (図 1) からなる。このデータベースのことを「4次元歴史空間ベース」また、操作システムのことを「4次元歴史空間操作システム」と名付けている。

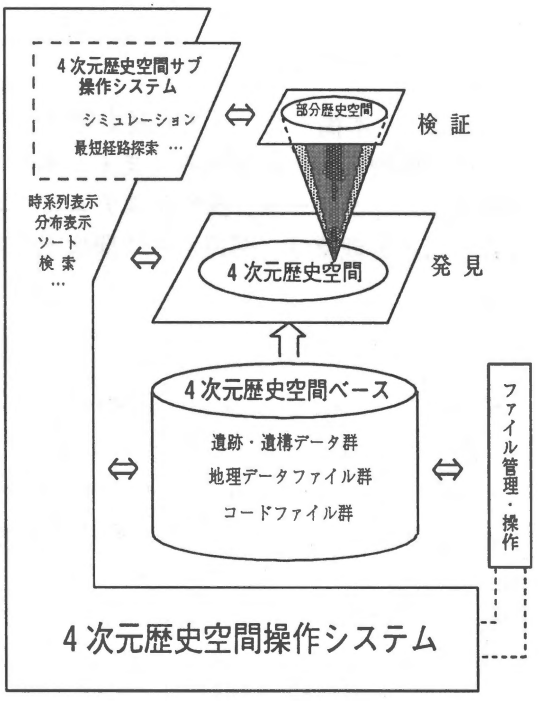


図 1: 4次元歴史空間システムの概念図

2.3　4次元歴史空間ベース

4次元歴史空間ベースは、遺跡データファイル群、日本国土の地理データファイル群、コードのデータファイル群で構成されているデータベースである (図2)。

遺跡データファイルは、従来から考古学分野でデータ化されてきた遺跡名、所在地、年代等のデータベースと基本的に相違がない。しかし、異った種類の遺跡データベースを統一することは難しい。そこで、現在は、すべての遺跡に共通している項目と特定の遺跡の項目に分けて、データベースを構築している。

現在格納している遺跡データファイルは、高地性集落遺跡データ (約５７０件) および、前方後円墳データ (約４３０件) の２種類である。

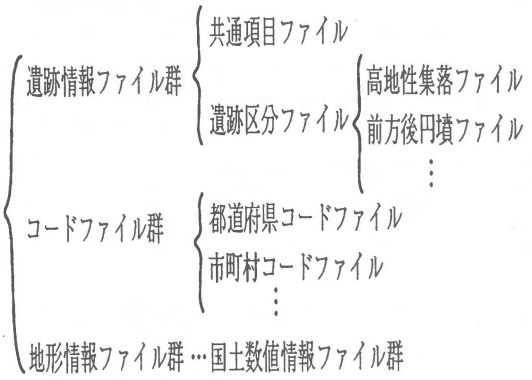


図 2:　4次元歴史空間ベースの基本データのファイル体系

地理データファイル群は、遺跡データをビジュアル的に表すための３次元地理データである。日本全土を標高値及び海岸線をデジタル化した地形データベースである。地理データファイル群には、建設省国土地理院が作成した国土数値情報 [3] の海岸線 (KF-5) および標高 (KS-110-1) データファイルを一部改編し使用している。

コードデータファイル群は、遺跡データファイルに格納される項目値の一部に採用するコード表現の対象データのファイルの集合である。コードファイル群の採用は、システムの柔軟性を保持するため重要である。コードファイルを用いることで、統一的な操作が容易になるとともにファイルのメンテナンスや移植が的確に行え、システムには依存しない形式が確保できる。

2.4　4次元歴史空間操作システム

4次元歴史空間操作システムは、考古学者の思索過程を無理なく支援する機能を具体化することを目標にしている。現在、この操作システムは項目操作群と地理情報処理機能の２種類に分けられる。

図 3:　ＱＢＥ検索画面

地理情報処理機能については、次章で紹介する。項目操作群は、従来から一般に行なわれている個々の項目データに対しての検索、ソート、一覧表示等の操作である。検索は、ＱＢＥ (Query by Example：例示検

索) 方式 (図 3) を採用している。データの画面表示は、1 画面に 1 遺跡の情報がカード形式で表示される方式 (図 4) を採用している。

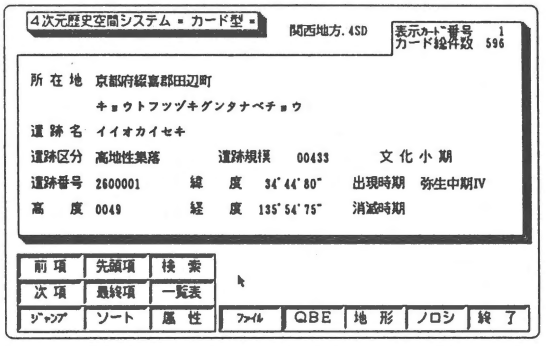


図 4: カード表示画面

3 地理情報処理機能

本システムにおける地理情報処理はすべて、彩色標高地図を表示し、その地形図に遺跡分布を描画することにより行なわれる。地形図は、画面上に経度方向 2 度 (約 1 6 0 km)、緯度方向 4 0 分 (約 8 0 km) の範囲で表示することができ、位置指定または、スクロールにより日本全国の任意の場所に移動することができる。

現在、この地形図上で実現している機能には、遺跡位置表示、円内検索および可視判定検索がある。

3.1 空間と時間

過去の事象は、時間と空間が固定した点的な存在ではなく、時間的にも空間的にも広がりを持った存在である。

歴史家は、時間次元と空間次元からなる 4 次元歴史空間内の事象の因果関係をはじめ、歴史に存在する傾向や法則を見出そうとする。こうした探索的な思考を積極的に支援するためには、時間と空間を一体化した表現ができる処理機能が必要となる。

表 1: 文化小期

文化小期	旧石器
	縄文
	弥生前期 I
	弥生中期 II
	弥生中期 III
	弥生中期 IV
	弥生後期 V
	弥生後期 VI
	古墳前期
	古墳中期
	古墳後期
	古墳終末期
	奈良前期
	奈良後期

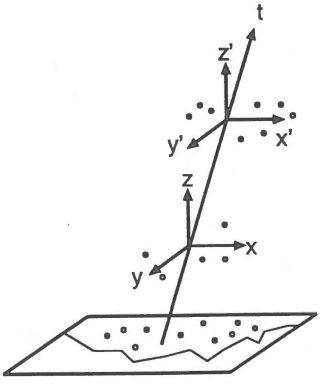


図 5: 時間次元の概念図

本システムの地理情報処理機能において

は、国土数値情報の標高データファイルを用い地形図を描画する。時間次元 (図 5) については、本システムでは文化小期 (表 1) を用いる。高地性集落等の遺跡の出現時期、消滅時期あるいは存在期間などの時間的データは確定的ではあり得ない。さらに、時間次元と絶対年代との正確な対応付けも難しい。そこで、時間的な幅をもつ文化小期によって時間次元を与える。

3.2　遺跡位置表示

遺跡位置表示機能 (図 6) は、地形図に遺跡位置の分布を表示させる。

この機能には、全遺跡を無条件で地形図上に表示させるものと、任意の文化小期を選択することで表示させる遺跡を時間的に制約するができるる。

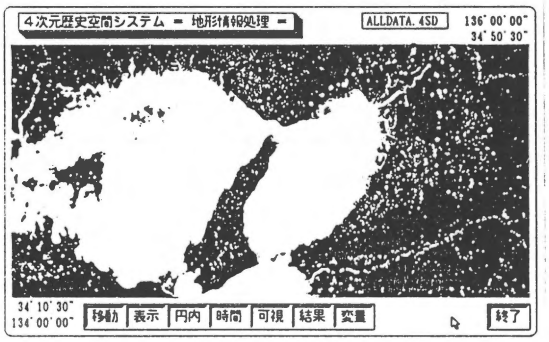


図 6: 遺跡位置表示

3.3　円内検索

円内検索 (図 7) は、ある地点を中心として指定された距離内にある遺跡を検索し、表示する機能である。現在、指定遺跡を中心とした任意距離の円内検索と、任意位置と任意距離での円内検索ができる。また、文化小期を指定することで同時期に存在したと考えられる遺跡のみを表示する。

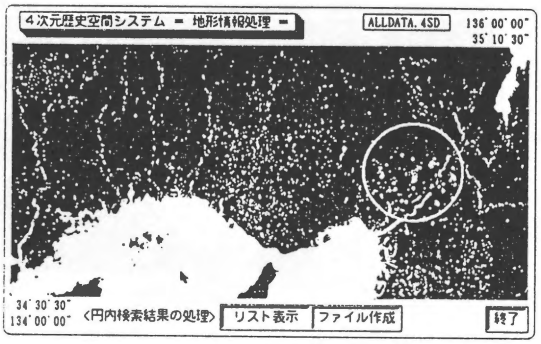


図 7: 円内検索

3.4　可視判定検索

可視判定 (図 8) とは、遺跡の位置とその標高から遺跡相互間の見通し可能性を判定することである。

判定を行なう際のパラメータとして、標高嵩上げ量、可視距離および文化小期を設定する。標高嵩上げ量は、1 つの遺跡内でも標高に差のある場合などの補正のためである。また、肉眼で見通せる距離は、季節や気象条件、さらに生活環境によっても大きく異ると考えられるので遺跡間見通し判定を行なう際、可視距離の制限を設定出来るようにしている。

可視判定検索は、遺跡位置中心および任意位置中心による可視判定と、2 つの任意位置間の可視判定の検索・検証が行なえる。さらに、任意の 1 地点からの可視可能領域を表示させることができる。また、検索を

する際に文化小期による時間制限を与えることにより同時期に存在したと考えられる遺跡間での検証が可能である。

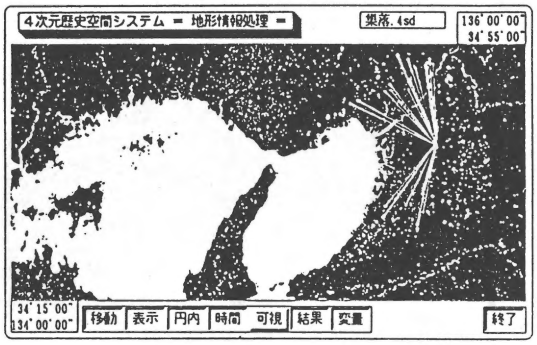


図 8: 可視判定

4 おわりに

地理情報処理機能の強化により、地形図上で即時の検索・検証が可能になった。しかし、本システムは現在のところ十分に目標に達しているとは考えていない。システムをさらに成長させるため、考古学者による本システムの試用を通して出される意見や要求を考古学者とともに検討していくことが必要であろう。

考古学研究支援をめざす意味では、こうした学際的共同研究が不可欠と考えられる。

参考文献

[1] 加藤常員, 小沢一雅, 都出比呂志: 4次元歴史空間システムの構想, 情報処理学会研究報告, CH-7 Vol.92, No.19, pp.1-7(1992).

[2] 加藤常員, 小沢一雅, 都出比呂志: 4次元歴史空間システムの構成, 情報処理学会研究報告, CH-23 Vol.94, No.78, pp.25-32(1994).

[3] 建設省国土地理院: 数値地図ユーザーズガイド, p.494, 日本地図センター, 東京.

視点に依存した属性付け機構を持つ木簡研究支援システム

— 構造進化型データベースの概念 —

Incremental Database System Handling Multiple Viewpoints for Historical Materials

森下 淳也[†], 上島 紳一[††], 大月 一弘[†††]

Jun-ya MORISHITA, Shinichi UESHIMA, Kazuhiro OHTSUKI

† 姫路獨協大学, Himeji Dokkyo University. E-mail:*morisita@himeji-du.ac.jp*.

†† 関西大学, Kansai University. E-mail:*ueshima@res.kutc.kansai-u.ac.jp*.

††† 神戸大学, Kobe University. E-mail:*ohtsuki@cs.cla.kobe-u.ac.jp*.

キーワード: 木簡研究支援システム、科学データベース、半構造化データ、段階的構造化、視点

keywords: wooden slips, scientific database, semi-structured data, incremental data organization, viewpoint

あらまし: 本稿では、研究者の活動を支援する過程をデータベースの構造化の過程とみることにより構造進化型データベースの概念を形成する。そのために必要な拡張概念として、段階的構造化、カテゴリ、視点について議論し、これらを包含するデータモデルを考案する。これらを、我々が構築している木簡研究支援システムにおける木簡データの段階的構造化に沿って述べる。

本システムは、木簡オブジェクトに対して，研究者の任意の視点から属性/属性値付けの可能な構造進化型データベースシステムとして実現されており、未整理な状態にある木簡データを、自由な視点から段階的に分類・集約することができる。本手法は、半構造化状態にあるデータに対してそのまま適用することができ、膨大な量のマルチメディアデータの取り扱いや、未整理なデータの整理、分類作業を格段に効率化することができる。

Summary: In this paper, we build a new concept of incremental database system by regarding researcher's working process as the process of database generation. We discuss the notions of incremental data organization, categories, and viewpoints, and propose a new data model which built-in these notions. Then, we introduce a mechanism for incremental data organization of semi-structured data in handling ancient Chinese wooden slips data. The objective of the mechanism is to support scientists' incremental and hypothetical work processes (object/type identification, classification and verification/ abstraction from users' multiple viewpoints). We have developed a prototype incremental database system based on an Object-Oriented DBMS.

Our system can be used to organize all kinds of semi-structured data incrementally. Users can treat a large volume of multimedia data, and unclassified data efficiently with this system.

1 はじめに

最近、文献情報データ、地球環境データ、ヒトゲノムデータ、静止・ビデオ画像などを対象とした科学データベース（*Scientific database*）が注目されている [1]。科学データベースが問題としている点は、

- テラバイトに届こうとしている膨大な情報を如何にデータベースとして扱うか。
- ネットワーク上に分散したデータをどう扱うか。
- マルチメディアデータなどの複雑なデータ構造をデータベースの中に如何に収容し、管理するか。
- 物理的に格納されたデータ形式の異なるデータをどう扱うか。

などである。これらの問題は、例えば、CADの分野などで活発な議論がなされておりオブジェ

クト指向データベース (OODBMS) などによる解決への糸口を見出している [2]。いずれもデータベースシステム (DBMS) の限界を越える多くの試みがなされている。

我々は科学データベースを、このような議論とは異なる観点で捉える。実際の研究活動の中でのデータベースの果たすべき役割を考えると、研究者の持つデータは、研究者が考える対象を表わすには不完全な状態から始まる。研究者の知的活動の結果、適切な枠組が研究者の目前に開かれ、完全なものになると考えられる。このような活動に対して、従来、DBMS はその活動の外側にあり、新しい概念の導入や発見が起こる毎に、新しいデータベースを再構築するという、手間のかかる道具に過ぎなかった。

DBMS の大きな特徴である、完全性や完備性は、間違いがないという大きな安定を我々にもたらしてくれるが、言い換えれば、これはどのような瞬間にもデータベースが正しくなければならないという自由度のなさを示している。データベースができる事は、常に完全であるとみなされる固定された枠組に対して、情報を流し込むことだけである。

研究者が研究に用いる科学データベースでは、このような堅牢さが重要な分野とは異なり、常時の完全性よりも仮説や思い付きを自由に記述できる柔軟さや新しい枠組を随時試すことのできる軽快さが重要である。このような観点から、我々は科学データベースを

- 研究者の発見による新しい情報を、枠組を越えて同じデータベースの中に、随時、取り込むことができる。
- 研究者の新奇な思い付きをそのままデータベースに反映できる。
- 研究者の知的活動自体が、データベースをより完成度の高い状態へと導く、データベースの形成過程となる。

などが実現された構造進化型 DBMS と考える。

我々は、中国において出土された木簡の研究を行う研究者の協力を得て、従来の DBMS を用いて、木簡研究用データベースの試作を行なった。研究者は、データベースの外側で、自分の目的に応じた意味付け作業を行い、意味付けの結果は、もとのデータベースに反映されるのではなく、別に新たなデータベースを構築してそこに格納する。この経験を通して、データベースの枠組が利用者のデータベースの利用目的や利用方法に大きく依存する、かつ、それをデータベース設計時に確定することが困難であることを確認した。

このような研究者の作業に関して、出発点となるデータベースの状態を**半構造化状態** (*semi-structured*) と呼ぼう。これはデータベースの立場から言えば一つの完成されたものであるが、研究者の利用目的や方法が反映されていない状態のデータベースを意味することになる。

構造進化型 DBMS とは、これらの研究作業を支援するシステム、つまり、半構造化状態のデータベースに対して、利用者が様々な視点から行う分類作業やデータの付与などの意味付けを支援する機能を備え、作業の過程や結果を格納するため、データベースの構造が柔軟に進化する DBMS のことである。

言い換えれば、構造進化型 DBMS で最も特徴的な点は、利用者は、単に問い合わせ操作などを行なうのではなく、問い合わせ操作などに必要なデータの付与、ならびにデータ構造やスキーマ自体の生成を利用者の手で行い、その個々の過程でデータベースの持つ特質を失わないものである。

半構造化状態のデータを扱った研究として、ファイル内のデータに対してデータベースビュー機構を実現する試み [6]、また、ファイル内の文書データからデータベースを自動生成する Rufus システム [7] などがある。また、Zdonik は、ファイルに格納されたマルチメディアデータの部分情報の扱いの重要性を指摘している [8]。但し、本研究は、研究者の作業過程をデータベースの形成過程としてモデル化する点、予めスキーマ構造を定義しないという点でこれらとは異なる。また、木簡研究支援の観点からは、奈良国立文化財研究所、台湾中央研究院のデータベースがよく知られている。

本稿では、2 節で研究支援データベースへの我々のアプローチの仕方を説明する。3 節で、木簡研究の作業過程をモデル化し、データベース

形成との対応を述べる。また、木簡データの段階的構造化についての具体的な例を示す。この作業モデルに基づいて、DBMSのためのデータモデルを4節で述べる。ここで我々のデータモデルの実現のための基盤を、OODBMSにおく。3節でも触れるように、OODBMSでは、オブジェクトがデータベースの単位として扱われ、我々のデータモデルを表現するのに適切な形式を持っている。このOODBMSにおいて、段階的構造化、カテゴリ、視点の拡張概念の表現法を論ずる。

2 研究支援データベース

2.1 枠組の固定されたデータベースの限界

通常、データベースを構築する一般的な方法は、既に確立されたデータの分類体系に基づいて、データベースの枠組を決定する。つまり、データベース設計者が、既存の分類体系に基づく項目（フィールド）を定義し、レコード単位にデータを分類して格納する。代表的なDBMSである関係データベース(RDBMS)では関係テーブルが、OODBMSではクラス及びクラス階層が、枠組を与える。いずれの場合も、格納するデータや利用方法に応じて、データベース設計者が枠組を定義する。

データベースを利用して研究を進める際、格納されている分類形式とは異なる分類法でデータを利用したい場合が起こる。この場合、上記のDBMSでは、利用者が、データベース設計者が定義した枠組とは別に枠組を定義して、データを分類することができない。

また、新たに必要になった項目や新しく発見した項目を残し、そこに利用者がデータを書き込んで、後に再利用したい場合がよく起こる。このような場合、従来のデータベースでは、適当なフィールドを定義しておき、そこに随時、コメントとしてデータを書き込むしか方法がない。書き込むデータを構造化したり、また、データ相互を関係付けて保存しておくことは不可能である。

つまり、従来のDBMSでは、トップダウンに枠組の定義が行われており、利用者にとっては、データベースの枠組は固定されており、変更することはできない。利用者は、固定された分類により格納されたデータを必要に応じて取り出して、データベースの外側で、解析したり、新しい発見を行う。

2.2 半構造化状態のデータ

研究作業では、最初に格納されているデータが、研究の出発点となる基本データである。研究者は、これらの基本データを基に様々な角度から作業を進め、必要なデータをデータベースに付与していく。つまり、最初にデータベースに格納されているデータは、研究の進展に応じて次第に構造化されるデータであり、その意味で半構造化状態にあると言える。

また、半構造化状態のデータを扱う場合、研究者は、様々な利用目的で、着目するデータやその部分データを試行錯誤的に抽出して再利用したり、様々な角度から分類したい場合が多い。部分データを抽出する場合、データは値として存在するため、再利用する為には独立なレコードとして、属性/属性値を付与して再度データベースに格納する必要がある。この場合、属性/属性値は、利用目的や視点に依存して定義できる必要がある。

従来のDBMSでは、枠組が固定されている為、既に格納されたデータの部分を自由に取り出して再利用することはできない。また、抽出した部分やデータを様々な視点から分類する為に、実行時に利用者がデータベースの枠組を生成することも容易でない。

2.3 構造進化型データベース

我々のデータベースシステムは、研究過程に生成される様々なデータの保管庫を持つ研究支援ワークベンチとして位置付けている。即ち、本データベースシステムは、従来のデータベースとは作成目的や利用目的が異なる。

ワークベンチとしては、研究作業を支援する機能を持ち、研究者の作業に応じて作業の経過や結果を格納するため、データベースの枠組自

体が柔軟に進化するデータベースであることが必要である。また、利用者のデータの利用目的は、個々に異なるものであり、また、個々の利用者がデータベースを使用する場合においても様々な視点をもとに、様々な角度から資料に対して意味付けを行う。この利用者の様々な意図（即ち、視点）自体をデータベースに収容し、効果的に利用できる必要がある。

また、このようなDBMSでは、従来のDBMSに比較して、利用者が自由にオブジェクトやスキーマを生成・更新する為、利用者の行う操作の整合性やデータベース内のデータの一貫性を保証する機構が必要である。スキーマは段階的に生成される為、従来のDBMSのようにデータベースの設計時に予めルールを枠組に定義しておくことができない。つまり、利用時にルールを順次、定義する機構が要求される。

以上のような要求を満たし、研究の進行とともにデータベースの枠組自体が動的に進化するデータベースシステムを**構造進化型DBMS**と呼ぶ。つまり、構造進化型DBMSとは、利用者による段階的な部分データの定義機構、データ構造更新機構、スキーマ生成・更新機構、スキーマ更新の監視機能などを備えたDBMSである。

また、利用者がデータベースに対して行うこれらの操作を総称して、半構造化データの**段階的な構造化操作**（*Incremental Data Organization*）と呼ぶ。

3 木簡研究者の作業過程

本節では、研究者の作業過程をモデル化し、木簡データの段階的構造化の流れについて考察する。

3.1 木簡オブジェクト

木簡は、古代に文書として用いられた歴史資料である。通常、木簡は断片化して出土するため、文書の部分であることが多く、各断片をデータ操作の基本単位としてデータベースに格納する。木簡情報データベースとしては，この段階で，複数種類の識別番号，出土地，木簡画像，釈読

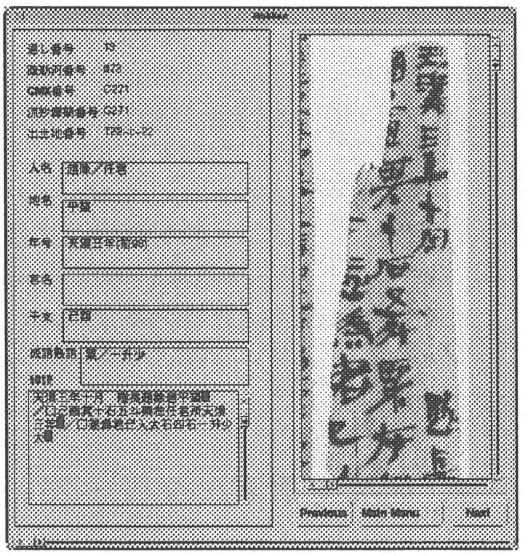


図 1: 木簡画像と属性群

(木簡画像の解釈文)、釈読に含まれる地名、人名などのキーワードなどの属性群が定義されている (図 1)。研究者は、様々な観点からこれらの属性群をもとに作業を進める。作業の基本となるこれらの属性群の付与された木簡を**半構造化状態**にある木簡オブジェクトと呼ぶ。

文書データとしての木簡の持つ大きな特徴は次の２点である。

(1) 通常の文書と異なり、木簡文書の型（タイプ）を予め想定できない．つまり、木簡データ間の上位/下位関係や属性構造などを決定できず、木簡データ群に対するクラス構造を予め定義できない。

(2) 木簡の取り扱いが研究者の視点によって異なる。つまり、木簡に与えるべき属性構造が研究者の視点に大きく依存する。

3.2 研究者の作業過程

図 2の左円に示すように、研究者の作業過程は、同定、分類、検証の３つの過程から構成される。研究者は、これらの過程を試行錯誤的に繰り返し、木簡文書の型を推定する。

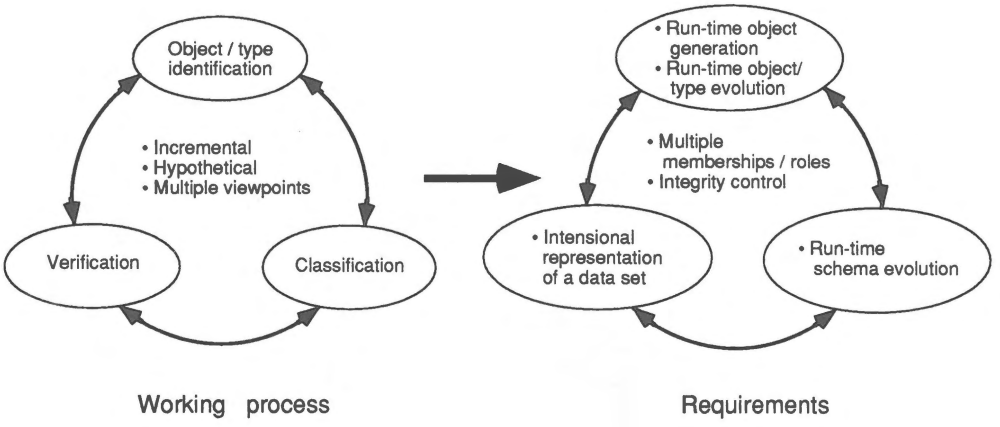


図 2: 研究者の作業過程

3.2.1　オブジェクト / 型の同定

データベースに格納されたデータの意味のある単位や着目する部分列を発見的に認識し、新しく属性/属性値を追加・変更したり、解釈やコメントを書き加えた後、独立なオブジェクトとしてデータベースに格納して再利用する。これをオブジェクトの同定という。

一方、型の同定は、オブジェクトのデータ型を推定する。例えば、ある研究者がある木簡を **"手紙"** の一部と推定し、また別の研究者は **"領収書"** の一部と推定したとする。研究者は、それぞれの推定に基づいて独自の属性/属性値を付与する。(図 3) 研究者は、値 **"XXX"** を **"人名"** と認識し、新しいオブジェクト O_2 を生成して属性/属性値 "人名:XXX" をプロパティとして与える。同時に、もとの木簡オブジェクトの型を **"領収書"** と推定し、木簡オブジェクトに **"種類:領収書"** と付与する。つまり、利用者の視点が異なれば、オブジェクトの属性名が異なり (**属性名の多様化**) , 属性, 属性値も異なる (**属性／属性値の多様化**)。つまり、利用者は多重に定義された視点から属性／属性値を定義する。別の研究者は、異なる視点からこの木簡オブジェクトを扱う為、木簡オブジェクトが異なるビューを持つ必要がある。即ち、異なる属性/属性値構造を持つことが許される機構が必要である。図 4は、同じ木簡オブジェクトに対して、**"手紙"**、**"領収書"** の２つのビューを与えている状況を示す。

3.2.2　分類

研究者は、格納したオブジェクト群や未構造化情報を段階的に関連付け、集約・分類して新たに発見した事実を書き加えながらデータベーススキーマを構成的に生成する。その際、新しく生成したオブジェクトや新しく付与した属性/属性値に対して、仮定や推測に基づいて問い合わせを行いながら分類する。この分類過程においても、分類結果は利用者の視点によって異なる。

オブジェクトを分類するものを**カテゴリ**と呼ぶ。通常、オブジェクトは複数のカテゴリにも分類される為、オブジェクトの多重分類を許す必要がある。カテゴリは更に詳細な分類である**サブカテゴリ**に分類される。こうようにして生成されるカテゴリ階層は、*a-kind-of* 階層を意味し、半順序関係を与える。この階層は、研究者によって試行錯誤的に生成される。

3.2.3　検証

研究過程で、研究者は、直感や推定に基づいて発見的にオブジェクトを分類していく為、科学データベースは、分類結果を解析する機能を備えることが必要である。この機能により、集約・分類したオブジェクト群の正当性を評価し、オブジェクト群の持つ属性の抽象化し、また、そこから新しい概念を獲得することを試みるこ

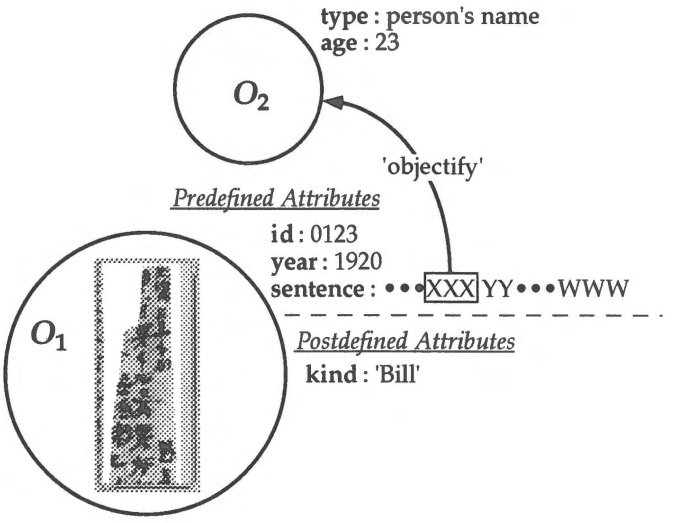


図 3: オブジェクト/型の同定

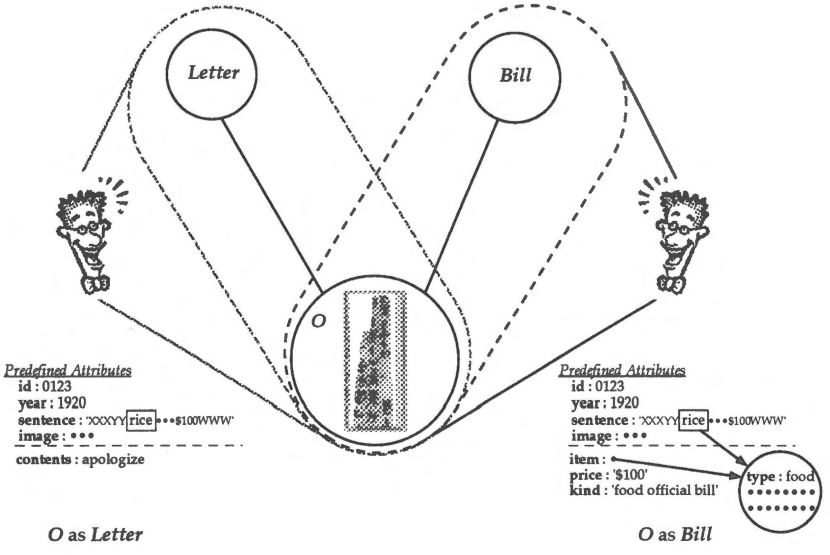


図 4: 木簡オブジェクトの多重ビュー

とができる。

4 データモデル

3.2節で述べた研究者の各作業過程を支援する為、構造進化型データベースが備えるべき機能は、データを格納・検索する機能に加えて、半構造化状態のデータに対して段階的構造化を支援する機能であった。本節では、段階的構造化を行うための基本機能を具体的に述べる。その際、前節でも述べたように、OODBMSを基盤として考える。オブジェクトとしてデータを扱うことで、RDBMSよりも柔軟な対応ができる。また、プロトタイプシステムとして試作した構造進化型DBMSには市販のOODBMS, GemStone[1]に基づく*TextLink/Gem*を用いた。

4.1 段階的構造化

研究者の作業過程（図2の左円）に対応してデータベースシステムが備える機能は図2の右円に示される。各機能は以下の通りである。

- **値の逐次オブジェクト化**
  木簡オブジェクトの持つ属性値の部分を利用者の作業時に（即ち、動的に）独立なオブジェクトとしてデータベースに定義する機能である。

- **オブジェクトの枠組進化**
  利用者の作業時にオブジェクト自身に、自由に属性/属性値を追加・削除したり、オブジェクトを別の分類へ移動する機能である。オブジェクトの集合を表わす外延が互いに異なる属性構造を持つオブジェクトから構成できることも必要である。

- **オブジェクトの重複分類と多目的利用**
  オブジェクトを複数のカテゴリに分類したり、複数の視点からオブジェクトに異なる役割（role）を与えることができる。

- **スキーマ生成／進化機能**
  分類を積み重ねることで実現するスキーマも逐次、段階的に生成され、改良できなければならない。

- **オブジェクト／スキーマ進化の監視機構**
  データベースの本来の機能を失わないため、段階的構造化における利用者の操作を監視し、操作結果の一貫性や無矛盾性を保つ機構。

- **データ集合の内包的表現の生成**
  利用者の推定や検証を支援するため、与えられたデータ集合から、逆に質問文の生成する機能。

以下で、これらの具体的な実現法について述べる。

4.2 基本オブジェクトモデル

段階的構造化を実現するための基本となるのは、カテゴリやオブジェクトの表現方法である。ここでは、その全てがインスタンスオブジェクトであるようなオブジェクトの統一モデルを考える [2, 4, 5]。即ち、レコードをオブジェクトとして扱うのみならず、オブジェクトの集合を持つカテゴリもまた、オブジェクトとして実現する（カテゴリオブジェクト）。

これらのオブジェクト間にスーパー／サブオブジェクトの関係 (*relationship*) を導入し、*a-kind-of*、*an-instance-of*、*set-membership* といった、様々な相互関係を表現できるようにする。相互関係の間に、属性／属性値を付与できるようはからうことで、オブジェクト間の関係を用いた、値による継承を実現する。

4.3 視点の表現

視点はカテゴリオブジェクトにより表現されるものとする。まず、着目するオブジェクトに対してスーパーオブジェクトを生成し、視点とする。次に、視点自身の持つ属性／属性値はスーパーオブジェクトに、また、視点とオブジェクトの関係の持つ属性／属性値はその関係に付与する。これらの属性／属性値をスーパーオブジェクトから着目するオブジェクトに対して動的に継承させることで、その視点から見たオブジェク

[1] GemStone は Servio Corp. の商標である。

トの構造を生成することができる。複数のスーパーオブジェクトを生成し、これを切り替えることによって, オブジェクトに異なる属性構造を持たせることができる。図4では多重の視点を表現している。

4.4 分類機構

我々のシステムでは、木簡オブジェクト群や, 木簡オブジェクトの部分領域オブジェクト群を柔軟に分類するために, **カテゴリ**と呼ばれるオブジェクトを定義し、分類の階層構造を段階的に定義できるようにしているが、ここでは、上記の視点オブジェクトとしてカテゴリオブジェクトを用いている。図5の右下ウインドウにカテゴリオブジェクト間の関係を表わす階層構造を示す。利用者はこれをデータベーススキーマとして動的に生成し、改良していくことができる。

4.5 オブジェクトの同定とアンカーオブジェクト

木簡画像中の任意の部分領域やテキスト情報の部分文字列をオブジェクトとするためには、単に、その情報を引き写す以上の機構が必要である。そのため、**アンカーオブジェクト**を新しく導入する。アンカーオブジェクトは利用者が動的に木簡オブジェクトの属性値を随時、オブジェクト化するための機構であり、これにより段階的・動的なオブジェクトの同定と生成が可能となる。図5の左ウインドウで、木簡画像の部分領域画像に対応するアンカーオブジェクトが、付箋のような形で視覚化されている様子が見られる。また色情報等などの属性構造を右ウインドウに示す (図5参照)。

アンカーオブジェクトとそれが指す部分領域との対応関係の設定を多対多の対応関係とすることができる. 対応できる部分領域は、直接的にも、間接的にも指定することができる。前者は、アンカーが指す領域を利用者が明示的に指定する方法で、後者は、アンカーに書かれる問い合わせ文により実行時に指定する方法である。木簡画像の部分領域に対しては、前者の形でアンカーオブジェクトを生成している。

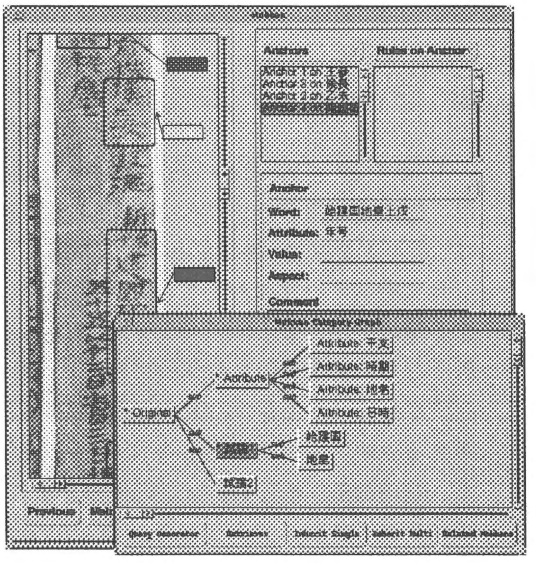


図 5: 木簡研究支援システム

4.6 段階的構造化における ECA 機構

ECA ルールは、イベント E、コンディション C、アクション A の三つ組みからなり、ある出来事が起こったときに、データベース中である状況が整えば行動を起こす機構である [9]。ECA ルールは, 従来アクティブデータベースの分野で用いられてきたが、本システムでは, 段階的構造化における利用者の各種の操作を監視し、システムの一貫性を保つ役割を果たす。ここでは、ECA ルールおよびイベントをオブジェクトとして表現し、ECA ルールを次のようなオブジェクトに対して付加 (束縛) する機能を実現している。

- **木簡オブジェクトへの ECA ルールの束縛**
  この ECA ルールは、
  - 束縛された木簡オブジェクトへのアンカー生成、
  - 一部の領域に対する ECA ルールの束縛、
  - 木簡オブジェクト自身の削除や移動や属性の付与・削除、

  といったイベントを監視し、イベントが検出されると必要なアクションを起動する。

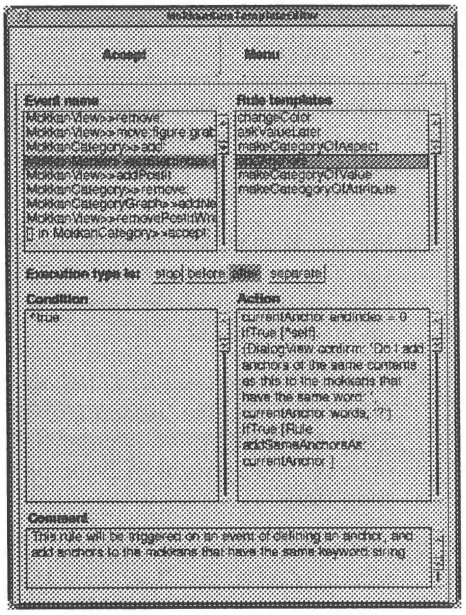


図 6: ECA ルールのアンカーオブジェクトへの束縛

- **木簡画像の部分領域に対応するアンカーへの束縛**
  アンカーに束縛された ECA ルールは、アンカーオブジェクトの持つ属性 (対象木簡、領域、釈読文の対応位置、属性名、色など) に対する操作が起こるとイベントが発生し、必要なアクションが起動される。(図6参照)

- **カテゴリオブジェクトへの束縛**
  カテゴリオブジェクトに対して束縛された ECA ルールは、カテゴリの削除やカテゴリの階層構造の変更操作、カテゴリへの木簡オブジェクトの追加・削除操作などを監視し、必要なアクションを起こす。例えば、あるカテゴリ下に入れることのできる木簡オブジェクトの個数が規定個数以上になった場合、その旨を利用者に通知する ECA ルールが生成され束縛することができる。

イベントオブジェクトの生成は各メソッド中で陽に記述される。また、ECA ルールの評価のタイミングとして、メソッド実行前 (before モード、stop モード)、メソッド実行直後 (after モード) などの区別が可能で、これもメソッド記述中で陽に記述される形で実現している。

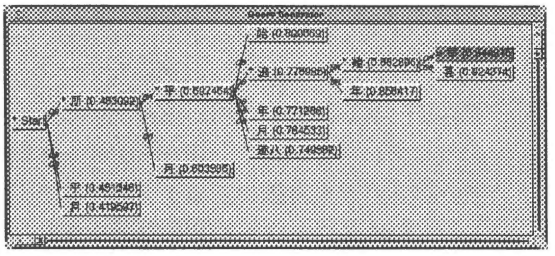


図 7: 代表的な単語によるデータ集合の特徴付け

4.7　データ集合の内包的表現の生成

研究者の検証過程において、データ集合の解析ツールを用いて、カテゴリオブジェクトへの分類の正当性を評価したり、集約したオブジェクトから抽象化された概念を抽出できる機能が有用である。我々のシステムでは、そのような解析ツールとして、データ集合の内包的表現（質問文表現）を得る発見的なアルゴリズムを作成している。このアルゴリズムは、与えられたテキスト集合をそこに含まれる頻出単語の論理和により特徴付ける。特徴付けの評価基準は、情報検索の分野でよく知られる再現率と適合率に重みを与えたものを用いている。

以下、実行例を示す。木簡情報データベースに格納されている約 400 の木簡の釈文やキーワード（単語）のテキスト情報を用いた結果を図 7に示す。適当な木簡集合を入力としてアルゴリズムに与えると、図のような頻出単語の木構造が出力される。与えられた集合は、5 つの単語 { 所, 平, 迹, 始, 幸 } で約 0.944915 の評価で特徴付けていると解釈できる。このアルゴリズムは、実行可能な時間で動作する。

5　おわりに

本稿では, 我々が構築している木簡研究支援システムにおける木簡データの段階的構造化手法について述べた。本システムは、木簡オブジェクトに対して属性／属性値付けの可能な構造進

化型データベースシステムとして実現されており、未整理な状態にある木簡データを、自由な視点から利用者が段階的に分類･集約することができる。本システムで用いられている中心的概念として、オブジェクト統一モデル、木簡データの部分データを抽出するアンカーオブジェクト、木簡オブジェクトやその部分を集約・分類するカテゴリオブジェクト、視点オブジェクト、スキーマ生成・更新の監視機能、データ集合の解析機能などについて述べた。

本手法は、半構造化状態にあるデータに対してそのまま適用することができ、膨大な量のマルチメディアデータの取り扱いや、未整理なデータの整理、分類作業を格段に効率化することができるものと考えられる。

本システムでは、視点オブジェクトはカテゴリオブジェクトにより表現されており、また、視点間には、*a-part-of* の関係のみが導入されている。視点自体や視点集合の表現能力を高める為新しい構造の導入や別の表現方法を検討する必要がある。これらは今後の検討課題としたい。

本研究の一部は、文部省科学研究費補助金による。

参考文献

[1] IEEE Computer Society, *Special Issue on Scientific Databases*, Bulletin of the Technical Committee on Data Engineering, Vol.16, No.1, March 1993.

[2] Tanaka, K., Nishio, S., Yoshikawa, M., Shimojo, S., Morishita, J., and Jozen, T., *Obase Object Database model: Towards a More Flexible Object-Oriented Database System*, Proc. of the International Symposium on Next Generation Database Systems and Their Applications (NDA'93), pp.159–166, September 1993.

[3] J.D.Ullman, *Principles of Database and Knowledge-base Systems*, computer science press, 1988.

[4] Ueshima, S., K. Ohtsuki, J. Morishita, Q. Qian, H. Oiso, and K. Tanaka, *Incremental Data Organization for Ancient Document Databases*, in Proceedings of the DASFAA95, pp.457–466, Singapole, April 1995.

[5] 上島紳一, 大月一弘, 森下淳也, 田中克己, 歴史的資料を対象としたサイエンティフィックデータベースのシステム設計, 電子情報通信学会研究会技術 研究報告 **DE93-47** 9 – 16, 1993.

[6] Abiteboul, S., Cluet, S., and Milo, T., *Querying and Updating the File*, Proc. of the 19th Intl. Conf. on Very Large Data Bases, pp.73–84, August 1993.

[7] Shoens, K. et al.,*The Rufus System: Information Organization for Semi-structured Data*, Proc. of the 19th VLDB Conference, Dublin, pp.97–107, Ireland, 1993.

[8] Zdonik, S., *Incremental Database Systems: Databases from the Ground Up*, Proc. of the 1993 ACM SIGMOD International Conference on Management of Data, pp.408–412, May 1993.

[9] McCarthy, D. R. and Dayal, U., *The Architecture of an Active Data Base Management System*, Proc. of ACM SIGMOD Symposium on the Management of Data, pp.215–224, Oregon, 1989.

古典籍とＪＩＳ漢字
－テキストの本文校訂との関係において－
Classical text and JIS-Kanji

當山　日出夫
Hideo TOUYAMA

花園大学(非常勤講師)
HANAZONO University

631 奈良市二名6-1492 王龍寺
Nara-shi, Nimyou 6-1492 Oryu-ji, 631

キーワード:JIS漢字, JIS X 0208, 本文校訂

Keywords　:JIS-Kanji, JIS X 0208, Text-revision

あらまし:JIS漢字について考える時、次の点を考慮しなければならない。

1．テキストの文字の何%が、JIS漢字で表記可能であるか否か、ということだけに目を奪われてしまってはならない。大切なのは、量ではなく質である。
2．索引作成を目的として電子化テキストを作成する場合、検索方法と漢字の認定は、密接な関係がある。
3．既存のテキストを無批判に信用するのではなく、学問的な批判が必要である。

Summary :We have to think about the JIS-Kanji, keeping the following points in mind.

1. We should not pay attention only about how many percentage of the letters of the text can be represented with JIS-Kanji. The point is the quality, not the quantity.
2. When you make an electrolyzing text to make an index, the authorization of the way of finding and the kanji is related closely.
3. As for the classic, you should not trust existing texts without criticizing at all but criticize them scholarly.

【１】はじめに－JIS漢字の論点－

JIS漢字をめぐっては、これまでに各所で様々に議論されてきている。筆者もまたいくつかの問題提起をおこなってきた。その論点を整理すると次のようになる。(注1)

(1). 字体・字形・書体・フォントなどJIS漢字を論じるための諸概念の整理
(2). 収録字数・字種の範囲の問題
(3). 78年版と83年版以降の改訂をめぐる新旧JIS漢字の問題
(4). そもそもJIS漢字とは何を規定したものなのか
(5). これからのJIS漢字はどうあるべきか
(6). JIS漢字の有効利用と漢字検索の問題

このような、JIS漢字についての議論は、古典テキストをあつかう人文学研究においてコンピュータ利用がはじまった時から既に、最大の問題点であった。研究者が自分の研究目的でコンピュータを使うようになった時、まず直面したのは、「必要とする文字がない」「近い字はあるが字体が不満である」という、JIS漢字の字種・字体についての問題であった。それが、近年になってようやく上記のように整理されてきた段階、と言えるであろう。だが、依然として、JISに無い字をどうするかというのは、未解決のまま残された問題点である。

【２】JISに無い字

JISに無い字への対処策としては、現時点では、次

のようなものがある。

(1). 外字作成。最近では、アウトラインフォントを簡単に自作できるようになってきた。

(2). 代用。アルファベットや仮名文字、あるいは、大漢和番号など、JISコード内の文字で記述可能な方法で代用する。

(3). あきらめる。JISに無い字は、「＝」(活版印刷でいうところのゲタ)とでもしておき、紙に書いた一覧表を附属させる。

(4). 他の文字コード系を使う。現に、中国文学研究者などは、日本のJISコードで書く日本語文と、中国の簡体字を使う中国語文を、一つの文書に同居させることを行っている。(注2)

どの方法がよいのか、それぞれの研究目的・利用目的によって異なる。

だが、ここで改めて考えてみるべきことは、そもそもJISに無い字とは何であるのか、ということではないだろうか。一見、分かり切ったことのように思えるテーマであるが、考えてみると、広範囲にわたる歴史的考証と厳密な論理構成が要求される、きわめてやっかいな問題である。以下、『和漢朗詠集』を材料として、JISに無い字とはいったい何であるのか、あらためて考えてみたい。

【3】『和漢朗詠集漢字索引』の方針

『和漢朗詠集』は、平安時代、藤原公任の撰になる秀句集で、およそ800程の和歌と漢詩句を、テーマごとに分類して編纂したものである。王朝貴族文化の精髄であるとともに、その後の文学作品に多大の影響を与えた作品として、文学史上に名をとどめている。多数の写本・版本のテキストが伝えられているが、最も標準的に用いられるのは、伝藤原行成筆御物本である。現代の校訂注釈本の多くは、このテキストを底本に採用している。

この写本(複製)に基づき、2種類の校注本(岩波日本古典文学大系・新潮日本古典集成)を参看して、本文の表記にもちいられる全ての漢字を検索することを目的として、漢字索引を作成した。

凡例から、字体の認定にかかわる箇所を引用する、

3　本文は、次の方針による。

a　①の影印本文を本行とする。

b　なるべく正字体をもちいるが、原本(影印本)での書写字体にも配慮する。

c　異体字については、索引としての検索のため、原則的に正字体に統一する。

d　②③の校訂本文をともに(　)で示す。ところにより、校注者の判断の違いがあるが、それらを区別することはしない。また、必ずしもすべての異同を示すこともしない。わずかな字体の違いなどで、索引としての検索に支障のないものについては、採用しなかったものもある。

6　字体

1　本索引は、その作成はパーソナルコンピュータによったものである。印刷については、レーザプリンターで印字したものを、そのままの形でオフセット印刷してある。したがって、使用字体および字種は、すべて原則的にJIS規格(JIS C 6226 78)によらざるをえなかった。

2　JIS規格に無い漢字については、編者が外字として作成した。

3　そのため、すべての漢字を正字体に統一することは不可能であり、結果として部分的に新字体・正字体の混用が生じている。しかし、索引として、「和漢朗詠集」の漢詩句の漢字を検索するのに、実用上の不都合は生じていないはずである。

5　編者が外字として作字したのは、「和漢朗詠集」本文の漢字では次の63字である。

(＊本稿末の一覧表を参照。)

最終的に、上記のような方針にもとづく漢字索引を作成したのであるが、古典テキストのコンピュータ処理という観点からは、次の問題点を指摘できる。

(1). 字体の統一処理

多くの場合、古典テキストの校訂にあたっては、字体を正字体に改めて統一的に処理することが多い。古典を、いわゆる「正字」(まさしく「正しい文字」)で表記するのは当然のことのように思われるかもしれない。しかし、実際の写本等では、必ずしも「正字」ですべて書かれているということはない。むしろ、当時の通行の書写字体で書かれるのが普通であり、現在の『康熙字典』を規範とする正字意識を杓子定規に適用すると、かえって、テキストの本来の姿を理解できないことになってしまう。

(2). 索引としての字体処理

古典テキストのコンピュータ処理が、電子化テキストとして、文字・語彙の検索に利用することを目的と

する場合、あるいは、索引の作成を目的とするような場合、検索の便宜のための字体処理が必要になる。具体的には、

A. テキスト全体を統一的に処理する。

B. 検索システムに、異体字シソーラスのようなものを用意する。

この何れかが必要になる。

(3). 本文校訂

凡例３のｄとした、校注本の字まで検索の対象とするか否かである。『和漢朗詠集』を一般の研究者が読む場合、実際には通行の校注本によっている。校注者により、本文校訂の判断が異なる場合があり、現実には、校注本に独自の本文が生じることになる。これらをどこまで採用するかが問題である。また、底本(写本)の誤写とおぼしき箇所であっても、簡単に廃棄してよいかどうかも疑問である。

以上の3点について総合的に考えたうえで、また、索引としての実用的性格にも配慮して、実際の漢字索引作成となる。

【４】『コンピュータであつかえない漢字』

『和漢朗詠集漢字索引』の刊行にさきだって、筆者は、『コンピュータであつかえない漢字－「和漢朗詠集」の場合－』という論文を発表した(注3)。これは、索引作成にあたって、準備的に考えたこと、特にJIS漢字関係の問題点を、まとめたものである。この論文を書いた時、筆者の使ったパソコンに搭載されていたJIS漢字の規格は、「JIS C 6226 78」であった。現在では、「JIS X 0208 90」になっており、この間に、78年版から83年版への変化、いわゆる新旧JIS漢字の改訂がおこなわれたことは周知の事実である。

しかし、古典テキストの本文校訂とJIS漢字についての、もっとも基礎的な問題点については、現在にいたるまで十分に議論がつくされたとは言い難い状況であり、今から、8年前に書いたものではあるが、この論文についてふりかえってみたい。なお、この論文の全文は、今日の視点から自注を加えて新たに出してみたいと思っている。(注4)

結論として、『和漢朗詠集』における非JIS漢字として、64字を認定した。その一覧は本稿の末尾に掲載してある。

一般に、ある文献におけるJIS漢字・非JIS漢字を問題にする時、まず問題になるのは、JIS漢字でどの程度入力が可能か、非JIS漢字はどれぐらい出てくるのか、ということであろう。それについてみれば、『和漢朗詠集』の総字数は約10,000字、異なり字数は約1,800字である。したがって、総字数に対する非JIS漢字の割合は約0,7%であり、異なり字数に対しては約3,5%ということになる。

だが、筆者は、このような数字にそれほど意味があるとは思わない。まったく無意味とするわけではないが、重要なのは、何%かという数字ではなく、その中身である。JIS漢字・非JIS漢字と判定した、具体的な字の種類であり、さらには、その判定基準である。

表には全部で64種類の漢字が並んでいる。おそらく古典テキストの本文校訂の現場に疎い、強いていえば理系のコンピュータ研究者の目からは、どれも同じようにJISに無い字と見えるかもしれない。

たしかに、誰がどう見ても、JISに無い字(非JIS漢字)であると判断する字はある。そして、それが一覧表に示した非JIS漢字のかなりの部分をしめることは事実である。だが、中のいくつかの字については、そう単純に非JIS漢字であると判定してしまうには躊躇される、微妙な学問的判断が要求されている。テキストの本文校訂とかかわる場合である。

【５】所属部首とJIS漢字

まず、筆者の作成した『和漢朗詠集』の電子化テキストは、最終的には部首画数順配列の漢字索引を作成することを目的とした。そのため、異体字等については、次のような原則でのぞむことになった。

(1). なるべく正字体をもちいることにするが、原本(平安時代の写本)で使用される字体にも配慮する。

(2). その字の所属部首が変わってしまうなどの特にいちじるしい支障がないかぎりはJIS漢字をもちいる。

ここで注目してもらいたいのは(2)の方針である。もし、部首画数順ではなく、他の配列方式、例えば、四角号碼配列や音訓配列による索引を作成するのであったならば、また異なった判断を下すことになったからである。具体的には、次の諸例である。

A. 所属部首が同じであるので、統一処理した例。

[1]. 嵐　所属部首は「山」。「山」を上に書く字(JIS)と、左傍に書く字(非JIS)とがある。所属部首が同じであるので、「嵐」に統一的に処理することにした。

[2].楫　所属部首は「木」。原本では、この字の右傍が「戈」になった字体をもちいている。しかし、所属部首は、「木」であるので「楫」に変えた。

[3].苑　所属部首は「艸」。原本では、全8例のうち2例が「苑」。残りの6例は、「艸冠」の下に「ウ」ないし「ワ」を書く字。特に区別して用いた形跡も認められないので、「苑」に統一。なおこの字について、現代の校注本の書き下し本文は、「苑」を使用している。

B．反対に、所属部首が同じであっても、統一せずに別の字とした例がある。

[5].寧　原本では普通には「寧」をもちいる。しかし、1例だけ下部を「用」に作る字が使用されている。それは、固有名詞(人名)で使用されている。こと固有名詞の用字については、単純に改めるわけにはいかず、原本どおりの字体(非JIS)を使用することになる。ただし、この箇所、所属部首という点では、同じ「ウ冠」に属する。

C．似通った字ではあっても、所属部首が異なるために別の字として処理することになる例。

[6].徊　「徘徊」の熟語で使われるが、「徘徊」と書けばJIS漢字で表記可能。この「徘」の「彳」を「人偏」に作る「俳」はあるが、「徊」を「人偏」に作る字はない(非JIS)。一般に、古写本の書写字体において、「彳」と「人偏」とは、さほど厳格に区別されているとは言い難い場合が多い。特に行書や草書で書かれた場合は、同じように書かれてしまう。

しかし、明朝体で印刷する場合、区別される字となるし、また、所属部首が異なってしまう。なお、この箇所、新潮古典集成では「徘徊」としている。

ただし、最終的な漢字索引においては、「徘徊」とした。それは、その当時使用したパソコン(NEC PC9801M2)の能力では、作成できる外字数に厳しい制限があったからである。やむをえず、「徘徊」の例に限って、原則を曲げて妥協することとした。

以上は、所属部首とJIS漢字の判定がからんでいる例である。原則は、原本で使用される異体字(非JIS漢字)が、部首を同じくするならばJIS漢字で処理し、異なるならば外字を作る、ということである。これはあくまでも部首画数順配列の漢字索引作成という目的に則して考えた場合のことである。四角号碼配列索引であるならば、[1]の「嵐」や[2]の「楫」は、別の位置に配列されてしまうことになる。しかし、[6]のような「俳・徘」は同じ位置になる。

【6】本文校訂との関連

本文校訂の点で問題になる例がある。

[7].僻　「僻」の「人偏」を「土偏」に作る字が原本では使われている。これは誤写である。単純な誤写と判定してしまい「僻」に本文を改めるならば、「土偏」の字(非JIS)は不要になる。原本の表記を尊重するが故に、非JIS漢字が必要となる箇所。なお、「土偏」の字は普通の漢字辞典には見いだせない。

[8].護　正しい本文は、偏が「言」ではなく「水(サンズイ)」。上の例とは逆に、原本どおりの表記で済ませるならば、JIS漢字の「護」だけで足りる。

これらの例のうち、[7]の「僻」については、一見すると無意味な処置と思えるかもしれない。単純誤写について、おそらく架空と思われる字を、外字として作る必要があるのかと、疑問に思うのが普通かもしれない。しかし、そうしなかったのは、次の例があるからである。

[9].瑩　現代の校注本（岩波古典大系・新潮古典集成）ともに、この字である。しかし、原本の字は、「玉」ではなく「火」に作る字である。『和漢朗詠集』で「瑩」はこの箇所の他に5例見いだせるが、それらを観察すると、あきらかに「玉」と「火」は区別しなければならない。

ところで、この字を含む漢詩句は、『白氏文集』からの引用である。平安時代撰述の『和漢朗詠集』は『白氏文集』から多くの漢詩句を採用しているが、平安時代に実際に日本で読まれたテキストが、幸いなことに現存している。その古いテキストを、見ると該当箇所の原文は「火」に作っている。これは、現代の校注本が誤ったさかしらな校訂を加えてしまった箇所である。(注5)

つまり、[9]の例について、校注本をそのまま信用して、「火」につくるのは誤写と判定してしまい、「瑩」を本文とするならば、外字作成は不要になる。

だが、本文校訂としては誤った処理になってしまう。一見すると誤写のように見なしがちな箇所であっても、調べてみると、意外とそれが本来の正しい本文を伝えているという例は、古典テキストの研究において稀ではない。したがって、あえて誤写の本文であっても、作字の必要があるのである。

『和漢朗詠集』の漢字索引を作成するといっても、ただ機械的に漢字をデータ入力して、索引作成のプログラムを走らせればよいというものではない。あくまでも、『和漢朗詠集』研究に資するものを、目指さねばならない。そのためには、既存の校注本を盲信するのではなく、学問的批判の姿勢が不可欠である。

【7】字が増えれば解決するか？

以上、指摘してきた『和漢朗詠集』のJIS漢字をめぐる問題は、コンピュータで使える字が増えさえすれば解決する……と見なされがちかもしれない。だが、そう簡単な問題であろうか。

『和漢朗詠集』の非JIS漢字64字のうち、上述の9字を除く55字については、その字がコンピュータで使える字として存在しさえすれば、文句無く解決する。コンピュータで使える字は、多ければ多いほどよい。

しかし、上記に指摘した9字については、単に使える字が増えれば解決するという質の問題ではない。仮に非JIS漢字とした字が使えたとしても、索引作成を目標とした本文校訂という視点からは、問題が無くなるわけではない。

ここにあげた字は、たまたま必要とすべき字が非JIS漢字であったために、出てきてしまったものにすぎない。JIS漢字内部で処理可能であるが故に、JISに無い字として、議論の対象にならないが、しかし、本文校訂の問題としては重要な問題をはらんだ文字が、これ以外にも多数存在するのである。

上記の「嵓」に類似した例として、「峯」と「峰」の場合がある。「峯・峰」は、ともにJIS漢字であるために、非JIS漢字の問題点にはひっかかってこないが、テキストの本文校訂と字体処理については、「嵓」とまったく同じように考えるべきものである。

このような本文校訂にかかわる問題は、どんなに使える字が増えても解決しないし、また、自由な外字作成が可能になったとしても解決しない。

コンピュータで古典テキストをあつかう時、JIS漢字の制限のことは必ず問題になる。それに対して、どれだけの字数があれば十分かという方向での議論、つまり、何字あれば何%をカバーできるかというたぐいの発想では、本文校訂の質を論ずることは出来ない。そして、古典テキスト研究者にもとめられるのは、より良質の本文校訂にもとづく電子化テキストの作成である。この研究の原点を常に確認しておかねばならない。

【8】文字とコンピュータ

最近のコンピュータと文字をめぐる議論は、「ISO10646」およびその日本規格である「JIS X 0221」に話題が集中しているように見受けられる(本稿執筆の時点では、まだ規格が決まった段階で、それを実装したコンピュータは登場していない。)

ISO10646について考える前に、そもそもコンピュータと漢字をめぐる議論では、次の各レベルを区別して論じる必要があることを見ておきたい。

(1). 文字集合

例えば、「JIS漢字・教育漢字・常用漢字」などのように人為的に定めた文字集合(キャラクターセット)である。

ここでは、どのような文字を、どのような字体で、どれほどの数、集めるのが妥当かどうか、が議論される。

異体字については、教育漢字・常用漢字などは、意図的に、その文字集合内部に限定する限り、排除する方針である。JIS漢字の場合は、(その方針が無定見とはいえ)ある程度の異体字を取り込んでいる。

(2). コード系

(1)で定められた文字コード表を、具体的にどのような形でコンピュータで使うかである。今日一般に使われている「シフトJIS」などがこれにあたる。

(3). 文字属性

漢字というものが「形音義」でなりたつものである以上、(1)の設定の段階で、既にある判定が下されている。そうでなければ、文字の選定は出来ない。だが、一旦決まった文字集合に対して、ユーザの側がどのような認識を持って対応するかは、また別の問題である。

JIS漢字についても、すぐに全体を問題にするから議論が混乱するのであって、実際的な利用法

として、中にふくまれる教育漢字だけを使う、常用漢字だけを使う、という選択的利用は可能である。

常用漢字中心の使い方をした場合、例えば「余・餘」は新字体・旧字体の関係として認定されるが、正字体中心の使い方でのぞむならば、「余・餘」は別の字として使い分けることになる(「余」は「自分」の意味、「餘」は「あまり」の意味。)また、「芸」を「藝」の新字体として使うか、「ウン」と読んで植物の名称として使うかは、利用者の解釈に依存する。

あるいは、コード表の制定は、それがISOであれJISであれ、個々の利用者の漢字に対する属性認識まで拘束するものではない、と考えるべきであろう。言い換えれば、自然言語の表記として、人間は自由に文字を使うのである。

(4). 文字検索

文字集合(コード表)にある字であっても、利用者が実際に探せなければ、存在しないに等しい。JISにある漢字であるにもかかわらず、ワープロの仮名漢字変換で出せなかったためか、その字の箇所だけ手書きになってなっている文書を、目にすることは稀ではない。

文字集合が現実に有効に活用されるためには、そこにふくまれるすべての漢字を、もれなく速やかに検索可能なシステムを必須とする。場合によっては、それに無い字については、無いことの確認が求められる。

特に、古典テキスト研究の場合、JISに無い字が頻出するので、JISに無いことを確認する作業が、重要な意味を持ってくる。

(5). フォント

コンピュータのディスプレイ上で、または、プリントアウトで、実際に目にする字のかたちである。

上述の(1)文字集合、(3)文字属性、(4)文字検索、これらは、「字体」レベルで、ある程度抽象的な概念として文字をとらえることになる。だが、現実には、文字は個々の具体的な文字デザインとして存在する。

具体的な文字デザインと、抽象的な字体概念とは、少なくとも概念的には区別して議論しなければならない。

以上、各レベルの諸問題は、コンピュータで人為的な文字集合を使用する場合、必然的に発生することである。これは、JIS漢字だけに固有の問題ではない。将来、ISO10646に的確に対応するためにも、上記の各概念についての認識は不可欠である。

【9】ISO10646(JIS X 0221)

ISO10646の実現によって、古典テキスト研究は、どのように変わるであろうか。はたして多大の恩恵をこうむるであろうか。

確かに、ISO10646が使えるようになれば、使用可能な字は増える。非JIS漢字を外字作成でしのぐような例は、かなり減るに違いない。しかし、だからといって、100%大丈夫かといえばそんなことはないであろう。古典テキストで、これまで「非JIS漢字」が問題になったのと同じように、これからは「非ISO10646漢字」という問題が発生することは必至である。

いや、実は、問題の本質は、字が足りるとか足りないとかではない。いわば自然言語としての文字表記に対して、人為的な固定的文字集合(JISやISO)で対応する場合の、文字論・表記論にかかわる原理的問題が未解決なままなのである。論点を絞っていえば、

(1). 数の不確定。

人為的な固定的文字集合では、絶対に数が不足する。なぜなら、漢字は、部品(部首)の組み合わせによって、新たに作ることが可能な文字である。

(2). 属性のゆれ。

漢字の属性(形音義)は、一義的に決定不可能である。異なる文字集合間で、漢字属性の整合性をどのように考えるか。

このようなことについて、文字論的考察の視点が定められていないのが現状である。(注6)

これは、上記の(1)文字集合および(3)文字属性にかかわる問題であるが、その他にも、(4)文字検索が、ISO10646でどうなるのかも、不安材料である。

JIS漢字でも、実際に研究者が、古典テキストの研究に使うためには、オリジナルのコード表「JIS X 0208」そのものだけでは不十分で、市販のJIS漢字コード辞典の類や、JISコード付の漢字辞典を、座右におかねばならない。

JIS漢字は、一応、日本の漢字を対象としているので、漢字の属性認定については、そう大きな問題は生じない。しかし、ISO10646になれば、中国・台湾・韓国といった、漢字文化圏とはいっても、日本とは異なる言語文化における漢字までをも対象にしなければ

ならなくなる。日本の既存の漢字辞典の中に、どうやって外国の漢字をとりこめばよいであろうか(注7)。

【10】おわりに－ある提言として－

JIS漢字(及びISO10646)については、近年、特に東洋の古典テキストを扱う学問領域においては、重要な議論のテーマとなっている。

ISO10466は、研究者に福音をもたらすであろうか、それとも、より混沌とした混乱状態を招来するに終わるであろうか。期待もある反面、なかなかその実像がつかめないでいる。

今、ここで我々がなすべきことは、過度に期待することでもなければ、問題点の指摘に終始することでもない。人為的文字集合(キャラクターセット)とはいったい何であるのか、異なる文字集合間における漢字属性の整合性とはいったい何であるのか、という課題について、冷静に文字論・表記論的に考察をすすめることである。

人文学の研究者の世界である。研究者の数だけ考え方がある。文字についても、必要とする文字はそれぞれに異なっている。研究者としての文字の共有化は困難かもしれない。しかし、文字についての、基礎的概念の共有化であれば、可能かもしれない。我々に求められているのは、そのための資料収集であり、考えることである。

今まさに、ISO10646(JIS X 0221)が登場し、同時に、JIS漢字(X 0208)の改訂も進行している。今後、コンピュータと文字は、どうなっていくか予断を許さない。だが、どのような状況を迎えようとも……少なくとも「この字が無いのは問題だ」と言うだけにとどまるのは、もう止めにしておこうではないか。

(注1)
以下いずれも當山日出夫。

1. 漢字コードをめぐる諸概念について、情報処理学会／人文科学とコンピュータ研究会、1994年9月16日、於東北工業大学
2. 漢字コードと漢字検索システム、文献情報のデータベースとその利用に関する研究・日本語テキストデータベースの利用法に関する研究、平成6年度合同研究会、1995年3月10日、於統計数理研究所
3. 漢字検索システムの諸問題、情報処理学会／人文科学とコンピュータ研究会、1995年5月25日、於総合大学院大学
4. 漢字の情報管理－JIS漢字と漢字辞典－、情報処理学会／人文科学とコンピュータ研究会、1995年9月15日、於上越教育大学
5. JIS漢字と辞書－漢字検索システムとJIS漢字コード辞典－、(第8回)語彙・辞書研究会、1995年11月25日、於三省堂文化会館

(注2)
多言語の同時使用の場合、現時点では、マッキントッシュの利用が一般的である。

(注3)
當山日出夫、コンピュータであつえない漢字－「和漢朗詠集」の場合－、汲古、第12号、昭和62年12月

(注4)
『人文学と情報処理』第10号に掲載の予定。

(注5)
『白氏文集』の我が国伝来の旧鈔本である「金沢文庫本白氏文集」である。本稿は、古典籍を専門にあつかうことを主眼とした学会のものではないので、詳しい解説は省略することにする。

(注6)
この問題は、旧来の印刷(活版)においても同様に存在する。活版の場合、無い活字を特別につくるぐらいのことは日常的に行われてきた。が、何よりも、活版の段階では、電子化テキストのコンピュータ処理などということは無かった。やはり、コンピュータを契機として発生したあらたな学問的課題というべきであろう。

(注7)
日本も中国も同じ漢字を使っているのだから……というのは安易な認識である。確かに「文字」は共通するものが多いが、「言語」は異なるのである。この問題については、『日本の漢字・中国の漢字』(林四郎・松岡榮志、三省堂、1995年7月刊)に詳しい。

和漢朗詠集　非ＪＩＳ漢字一覧表

(1) 芼 34
(2) 蕙 34・271・619
(3) 莢 34
(4) 遖 42
(5) 塹 69・407
(6) 庾 90・106・374
(7) 繆 90
(8) 媿 97
(9) 擊 107・587
(10) 愁 108・423・489・557
(11) 翎 130
(12) 熒 160
(13) 婕 162
(14) 妤 162
(15) 栟 171
(16) 颺 172
(17) 遷 246・483
(18) 飡 264
(19) 鄰 269
(20) 峭 274
(21) 喓 327
(22) 跎 353
(23) 酷 362
(24) 酳 362
(25) 擾 371
(26) 鼕 376
(27) 侷 376
(28) 滹 389
(29) 聾 391
(30) 漪 412
(31) 扃 417
(32) 藪 437
(33) 濩 438
(34) 鼯 449・493
(35) 忿 449
(36) 壀 458
(37) 緱 462・746
(38) 関 466
(39) 繳 470
(40) 摭 475
(41) 詹 475
(42) 縈 489
(43) 紉 494
(44) 舴 514
(45) 艋 514
(46) 闡 522
(47) 牖 530・561
(48) 瀼 532
(49) 裛 534
(50) 敲 554
(51) 忉 592
(52) 閒 660
(53) 屐 662
(54) 庳 666
(55) 畢 666
(56) 甯 675
(57) 涇 679
(58) 顯 698
(59) 頴 698
(60) 褰 716
(61) 峴 746
(62) 閒 779
(63) 蓻 802
(64) 皤 803

手書き文字時系列筆跡パタンの一解析と今後の計画

An analysis into pen-trace patterns of handwritten characters and a future plan

東山孝生、山中由紀子、澤田伸一、中川正樹

Takao Higashiyama, Yukiko Yamanaka,
Shin-ichi Sawada, Masaki Nakagawa
東京農工大学電子情報工学科
〒184　小金井市中町2－24－16

Dept. of Computer Science, Tokyo Univ. of Agriculture and Technology
2-24-16 Naka-cho, Koganei, Tokyo, 184
phone: 0423-88-7144, fax: 0423-87-4604, e-mail: hig@cc.tuat.ac.jp

あらまし: 我々は文章形式、字体制限なし、などを特徴とするオンライン手書き文字時系列筆跡パタンデータベースを作成し、それを利用した字体変動解析を行っている。オンライン筆跡パタン採集の対象とする文章は新聞から抜き出し、頻出の字種を中心にJIS第一水準1,537字種からなる約1万文字の文章列を作成した。そこに含まれなかったJIS第一水準文字は最後に個々に筆記してもらい、合計3,345字種を収集対象とした。現在までに80人分の収集が終了し、最終的に110人分がデータベース化される予定である。また、データベースの次期バージョンの準備も始めている。本稿は初期に収集した30人分のデータを中心に、手書き文字の筆画数変動、筆順変動の解析について報告する。また、次回の筆跡パタンデータベース作成計画と今後の字体変動解析の方針について述べる。

キーワード：オンライン入力、時系列筆跡パタン、データベース、字体変動、筆画数変動、筆順変動

Summary: A database of on-line handwritten character patterns sampled in a sequence of sentences without any instructions has been made. The sentences for which character patterns are sampled have been picked up from newspapers with the result that they are composed of about 10,000 characters and include 1537 kinds of JIS 1st set character categories. The rest of the JIS 1st set categories are written one by one at the end of the above text. Our laboratory collected 30 people's patterns. We proposed common usage of this database with each offering patterns from 5 people. 15 collaborators have joined this project. Recently, we added 5 sets. Stroke number and order variations have been analyzed from the initially collected 30 people patterns.

Key words: on-line input, pen-trace pattern, database, stroke number and order variation.

1. はじめに

ペン入力の実用化の流れの中で、オンライン文字認識の高度化を望むには、現実的な字体変形が含まれる大量の筆跡パタンのデータベースが必要となる。しかし、これまでそのような形のデータベースは存在しなかった。そのことをふまえ、我々は大量の手書き文字時系列筆跡パタンを収集し、共同利用可能な大規模データベースを作成した[1]。

現在当研究室で収集した 30 人分のパタンに加えて、各機関 5 人分のパタン提供を呼びかけた結果、約 80 人分のパタン収集が終了している。さらに 6 機関の参加希望があり、それを含めると合計 110 人分のパタンが収集される予定である。

本データベースでは字体変形を含む自然な筆跡パタンを収集するという方針から、筆記者には字体に注文を付けず、指定した文章列を筆記してもらった。同時に筆記者の個人情報を記録してある。また、オンライン入力パタンは筆点座標が時系列で採集されるので筆跡過程を追うこともできる。このようなことからさまざまな面からの解析が可能である。

本稿は手書き文字の筆画数変動の解析[2]、筆順変動、字体変動解析についての報告とデータベース収集の今後の計画について述べる。

2. 筆画数変動

標準画数別の実際の筆画数分布を図 1に示した。分布を見ると、どのグラフでも標準画数で書かれたものをピークとし、画数が減少する方へ傾斜した形になっている。また、図 2に標準画数別に文字パタンの最小画数、最大画数を示した。ここで、斜めの直線が交わる点がそれぞれの標準画数である。標準画数が増加するにつれて分布の幅が広がることがわかる。例えば、標準画数 20 画の場合、最大 22 画、最小 3 画で筆記されたパタンが存在する。実際に筆記された画数が標準画数より減少している原因はストローク間の続けが起きていることがあげられる。標準画数が増加するに従いストローク間の続けがより多く起きていることがわかる。しかし、21 画以上の高画数のパタンの分布の幅は狭くなっている。この

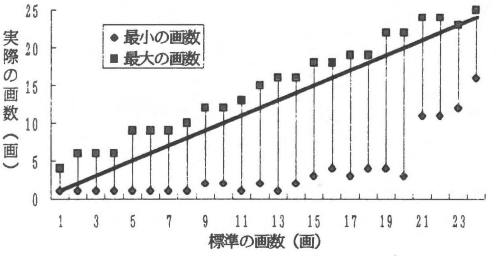


図 2　筆画数分布の範囲

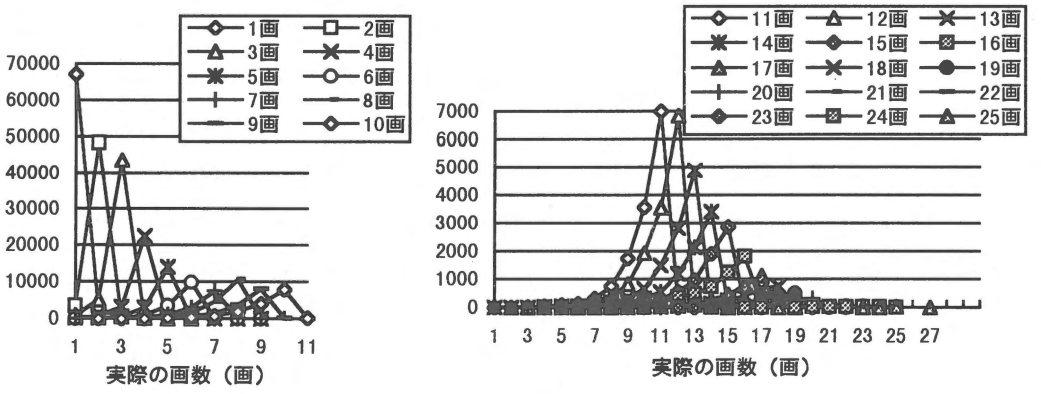


図 1　標準画数別の実際の筆画数分布(全 358,860 パタン)

原因はパタン数の絶対数が少ないことがあげられる（標準画数15〜20画：670×30パタン、21画以上：31×30パタン）。また、標準画数21画以上の文字は筆記しなれていない文字が多いために続けが起こりにくいことが考えられる（表 1）。

表 1　標準画数21画以上の文字

鰯艦顧轟鐸鶴纏灘鰭魔躍鑓露鰻驚轡讃襲鱈聾鰹鑑讐鱒鷲驚鷹鱗麟欝

画数別で全体的に見ると前述のような分布をしているが、各字種別で見ると、表 2に示したように標準画数で筆記されたパタンが必ずしも最も多いわけではない。全3,345字種中731字種では標準画数で筆記されたパタンよりも、標準ではない画数で筆記されたパタンの方が多い。例えば、部首『幺』は標準画数は3画だが続けて筆記され1画または2画で筆記されることが多い。部首『糸』も同様である。表 3に示したように部首『幺』を含む文字『磁』は標準の14画よりも7画で筆記されたパタンのほうが多い。他にも部首『幺』『糸』を含む字種はその影響からか標準画数より少ない画数のパタンが多く見られる(繰磯幾機轡縮など)。部首『口』やしんにょうなどを含む筆跡パタンについても同様なことが起きている。このことから、ストローク間の続けが起きやすい部首が存在していることがわかる。

また、実際に筆記された画数が標準画数よりも多いパタンもわずかながら存在した。この原因としてストロークの切れやゴミなどの影響によるものがあげられる。この他の例として、図 4 に2つのパタン例を示した。『比』は本来4画だが、5画で書かれているものが多く見られた。『比』の左側の部首は本来2画であるが、ほとんどの場合3画で筆記されている。このことは筆記者が普段から筆記していることも考えられるが、表示されているフォントを見て、その形をそのまま筆記していることも考えられる。『篦毘枇庇』や『鼠』などはフォントの影響が大きいと思われる。このような影響を取り除くためにフォントを見せずに音声による指示が考えられる。しかし、実際には見ないとすぐに筆記できない文字が余りにも多い。また、音声からでは仮名で筆記すべきか、漢字で筆記すべきか判断できない。このような点から、フォントを表示する方法を採用している。

表 2　標準画数以外の画数で筆記されたパタンがもっとも多かった字種

731字種（3,345字種中）

画数の差	-7	-5	-4	-3	-2	-1	1
文字種数	1	1	8	19	111	575	16

（ここで画数の差とは、ある字種において標準画数と実際に筆記されたパタンの画数の中で最も多い画数との差である）

表 3　筆跡パタン『磁』の画数

総パタン数 30個　標準画数 14画

実際の画数	7	12	11	14	13	9	8	10	4,5,6
パタン数	6	5	3	3	3	3	2	2	各1

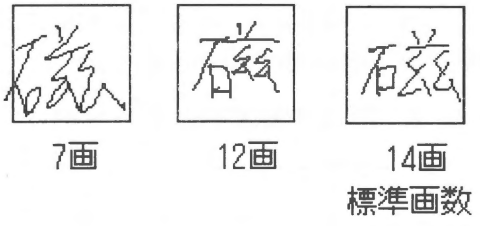


図 3　『磁』の筆跡パタン例

表 4　筆跡パタン『鼠』 の画数

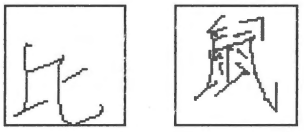
総パタン数 30個　標準画数 13画

実際の画数	14	13	15	12	16
パタン数	16	9	2	2	1

表 5 筆跡パタン『枇』の画数

総パタン数 30 個 標準画数 8 画

実際の画数	8	7
パタン数	16	14



比 鼠

図 4 標準画数より画数が多い例

3. 同一筆記者の筆順変動と字体変動

3. 1. 検査方法

データベースの文章部 10154 文字は 1537 文字の JIS 第一水準文字と 11 文字の JIS 第二水準文字により構成されていることからわかるように、同じ文字が何度も繰り返し出現している。出現数別の文字カテゴリ数を表 6に示す。ここに示されている 2 回以上出現する 857 文字カテゴリを用いて同一筆記者の筆順変動と字体変動を調べる。

表 6 出現数別文字カテゴリ数

出現数	1	～5	～10	～20	～100	101～
カテゴリ数	680	554	181	56	49	17

筆順変動、字体変動を調べるために本研究室で作成されたオンライン手書き文字認識システム[3][4][5]を用いる。この認識システムの特徴としてストロークのつながりに寛容であるために筆画数の変動は認識結果に大きな影響を及ぼさないことがあげられる。

次のような方針で変動を調べる。同一筆記者において、ある文字カテゴリ内で最初の筆跡パタンを辞書パタンとして登録する。登録の終了後、2 番目以降の筆跡パタンに対して前述の認識システムを用いて認識を行う（図 5）。これによりあるカテゴリ内の最初の筆跡パタンとそれ以外の筆跡パタンとの類似度が得られる。この類似度は 0～1000 で表され数字が大きいほどパタン間の差が少なく、類似度が 1000 の場合は全く同じパタンである。よって類似度が大きいパタン間では筆順変動、字体変動がほとんど起きていないと考えられる。類似度の小さいパタン間では筆順、字体の変動が大きく起きていると考えられる。この類似度に一定のしきい値を設定して、しきい値以下のパタンの筆順、字体を実際に目で見ることにより筆順、字体の変動を調べることができる。

また、認識を行うことにより、あらかじめ辞書に登録されているパタンとの類似度を得ることができる。

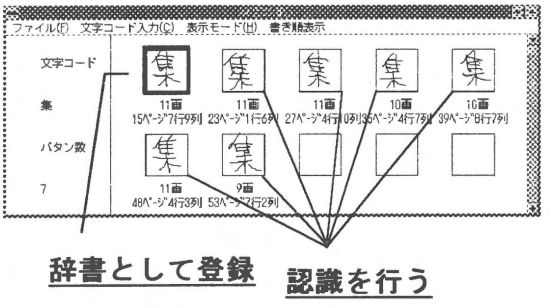


図 5 筆順変動、字体変動の検査

3. 2. 同一筆記者の筆順変動と字体変動の例

筆記者が異なると文字の筆順や字体が異なることは多く見られたが、同一筆記者内ではこれらの変動はほとんど起きていない。

しかし、そうした例がないわけではない。次に同一筆記者内で筆順変動の起きている例をあげる。図 6に示すように『馬』は同一筆記者によって 3 通りの筆順で筆記されていることがわかる。1 つ目の例は正しい筆順の 6 画目を筆記している途中で 5 画目を重ね書きしている。その後に、6 画目に継ぎ足しが行われている様子がわかる。このパタン画数は標準の 10 画から、2 画増えて、12 画になっている。2 つ目の例は正し

い筆順の1画目と2画目の順序を逆に筆記していることがわかる。3つ目の例は正しい筆順で筆記されている。このように同一筆記者においても筆順の変動が起きていることがわかる。

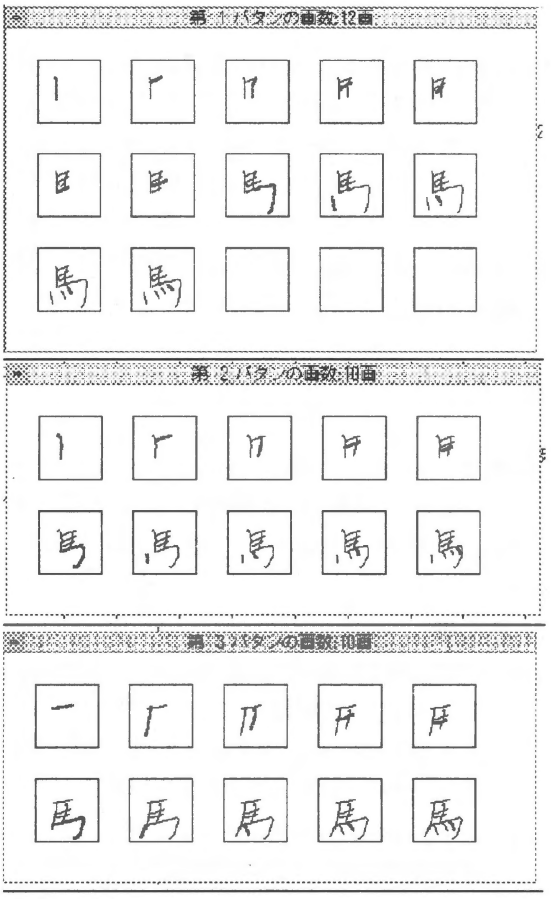


図 6　同一筆記者の筆順変動の例

図 7は同一筆記者における字体変動を示している。『第』を標準の字体とその略字体で筆記している。表示しているフォントは標準の字体の『第』であることから、この筆記者は普段は略字で筆記しているが、例として示されているフォントを参照しながら文字を筆記しているために、標準の字体で筆記したと考えられる。なお、本データベースでは自然な筆跡パタンの収集を目的としているために、略字体の制限も行っていない。

この例のような筆順変動、字体変動の起こる割合などを出すことが今後の課題としてあげられる。

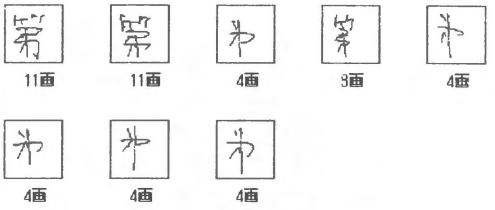


図 7　字体変動の例

4. 今後の計画

今回のデータベースの版でいくつかの問題点があがった。それらを考慮した新しい筆跡パタンデータベース作成の計画について述べる。

4. 1. 例文表示フォント

特にデータベース最後の文字の羅列の部分においては、一般的に目にすることの少ない文字が含まれている。そのために筆記者によっては知らない文字、筆記したことのない文字が含まれている。特に画数の多い文字は表示しているフォントが見難いことが多く、このような場合今回の版ではあらかじめ印刷した例文を見ながら筆記してもらった。しかし、このことは視線の移動や印刷された例文から筆記すべき文字を探すために、大きな筆記の中断が起き、自然な筆跡パタンを収集することの妨げとなる。次の版では表示フォントをペンでタップすることにより拡大されたフォントが表示され（図 8）、筆記の中断を最小限にすることができ、筆記が容易になる。

この他にも上記 2 でも触れたが、表示フォントの形の問題がある。表 7の『備』『心』は明朝体では通常筆記する字体と異なる字体であるが、正楷書体では通常筆記する字体である。ところが、『継』の場合は明朝体が通常筆記する字体で正楷書体が異なる字体である。このような問題が起こるためにすべての文字が通常筆記する字体で正しく表示されるフォントが必要となる。

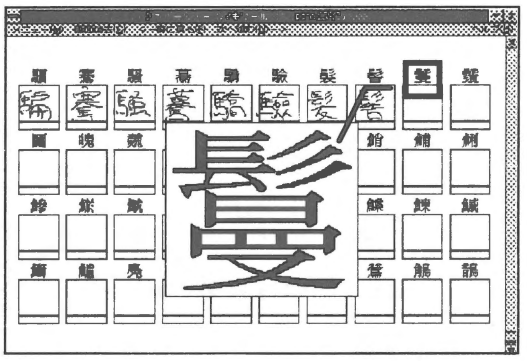


図 8　拡大された例文フォント

表 7　表示フォントの問題

MS 明朝	備・心・継
HG 正解書体	備・心・継

4. 2. 収集パタン座標

今回の版ではマウスの割り込みによりマウス座標で筆点を収集していた。このためにある点でペンを止めたとき、止めている時間によらず、ある点のデータは 1 つしか得ることができなかった。次の版では Windows95 で標準装備されるであろう penwin.dll を利用することによって、一定間隔の割り込みによる筆点の収集が可能となった。よって筆記者がある点でペンを止めたときは、止めている時間だけ、同じ座標を記録することができる。これにより筆記の過程がより正確に追跡できることになる。また、タブレット座標で採種することが可能となるためにさらに細かい分解能で筆跡パタンの収集が可能になる。

4. 3. 第二水準文字の追加

新しいデータベース収集システムの作成に伴い、筆記する例文も新しく作成した。主な特徴として、JIS 第二水準文字を文字の羅列として約 1,500 文字追加した。また、文章部の一文あたりの字数に制限を設けた。現在のペン入力は単語や短いメモ程度の文章の入力に用いられている。このことから現実にあわせて、できるだけ字数の少ない文章により例文を構成することを目的としている。

4. 4. 誤字・脱字検査ツール

収集されたパタンに対して、誤字・脱字の検査を行っている。検査の後により正確なデータベースを作成するために筆記者に訂正を依頼している。今回検査ツールは誤字・脱字の記録は筆記者に訂正を依頼するための簡単なメモ程度の記録しかしていなかった。また、明文化された検査基準がなかったために、検査を行った者の主観に頼る面があった。次の版では今回の誤字・脱字検査の記録を元に可能な限り、検査者の主観の入らない検査基準を設け、細かな記録を残す検査ツールの作成を行う。

謝辞

本研究は試験研究 05558027、および、重点領域研究 07207207 の一部補助による。

参考文献

[1] 中川、東山、山中、澤田、レー、秋山：文章形式字体制限なしオンライン手書き文字パターンの収集と利用、信学技報PRU95-110、43-48(1995.9).

[2] 山中、東山、澤田、中川：オンライン手書き筆跡パタンの収集とその一解析、情処人文科学とコンピュータ研資28-1、1-6(1995.11).

[3] 秋山、中川：ストロークのつながりに寛容なオンライン手書き文字認識、画像の認識・理解シンポジウム（MIRU'94) I、67-74(1994.7).

[4] M.Nakagawa and K.Akiyama:A Linear-time Elastic Matching for Stroke Number Free Recognition of On-line Handwritten Characters, Proc. 4th IWFHR, 48-56(1994.12).

[5] レー、秋山、中川：ストローク数非依存の高速オンライン手書き文字認識手法、情処第50回全大、2-61/62(1995.3).

絵画ＤＢとイメージ検索
―― 浮世絵の線画表現とデータ圧縮効果 ――

Image Searching on Picture DB
―― Line Drawing and Data Compression of Ukiyoe Prints ――

濱　裕光、志賀 直人
Hiromitsu HAMA, Naoto SIGA

大阪市立大学 工学部 情報工学科、大阪市
Information and Communication Engineering,
Faculty of Engineering, Osaka City University, Osaka, 558

キーワード：絵画データベース、浮世絵、線画、Bézier曲線、データ圧縮

Keywords: picture database, ukiyoe print, line drawing, Bézier curve, data compression.

あらまし：マルチメディアＤＢ（データベース）におけるイメージ検索のために、従来の文字キーワードに替わるイメージ・キーワード作成アルゴリズムの開発を行う。イメージには、絵、音、味、暖かさ、痛み、雰囲気、気分など人間の五感に対応して多種類のイメージがあるが、ここでは主に絵画ＤＢを検索するためのイメージを中心に扱う。画像圧縮ではＪＰＥＧなど多くの手法が提案され標準化されてきているが、いづれも原画を忠実に再生することを目的にしており、本研究の用途にはそのままでは適用できない。人間の視覚に直接うったえる絵画ＤＢ検索用キーワードとしての画像イメージの作成方法やイメージ相互を関係づけ、リンクを張る方法の開発を行い、人間にとって優しいしかも示唆に富んだＤＢシステムの構築を目指す。本稿では、題材として浮世絵を用いて非常に少ないデータによるイメージ表現を試みる。また、その加工と部品化の可能性についても検討し、現在までに得られた結果をまとめる。

Summary: Algorithms for image searching on picture DB (data base), using image keywords instead of conventional text keywords, are developed here. There are many kinds of images corresponding to human five senses, for example, picture, sound, voice, taste, warmth, pain, atmosphere, feeling and so on. Many methods for image compression such as JPEG, MPEG and so on, have been developed and standardized, but they are aiming at the goal to reproduce the original faithfully and don't serve our purpose. In this paper, image expressions of ukiyoe prints using a few data are descrived. Furthermore, the good possibility and results obtained by now, for processing such as deformation and for making module using parts, are reported.

1．まえがき

絵画ＤＢ検索では多くの絵を同時に高速に検索できること、いわゆるパラパラめくりができることが必須である。このためには原画を単に縮小して表示するだけでは駄目であり、必要なデータ量と表現できるイメージ（量）

の比を考えると効率が悪く絵の品質もよくない。ここでは単に縮小でなく、原画のイメージを効率よく高速に提示できる手法を開発する。一方、画像圧縮ではＪＰＥＧなど多くの手法が提案されているが、いづれも原画を忠実に再生することを目的にしており、本研究の用途にはそのままでは適用できない。

キーワードとしての役割を果たすには、Ｓ／Ｎの意味では正確でなくても、目的に到達するための羅針盤の様な役割、すなわち「作者：ＸＸ、作品名：○○○○」に導いていくための道案内ができることが重要である。ここで、一つ大切なことは、ユーザが道案内をして貰う途中で目的の絵以外にいろんな他の絵のイメージがあることに気付き、知識を広めながら本当に自分が欲しい情報を獲得できる点である。そこでは、最終的には「自分が最初に考えていた絵とは違う絵が本当の要求に合った」というような発見だって十分にあり得るし、このようなことができることが知的ＤＢ検索システムに必須の機能であるともいえる。このようなことは従来型の文字キーワード検索ではけっして期待し得ないことである。

人間の感性を重視することは最近よく言われる“柔らか頭”のコンピュータを設計することであり、そこではヒューマン・インタフェース、さらには一歩進んで、ヒューマン・コミュニケーションが重要なテーマの一つとなる。人間の感性とＤＢを結びつける研究はまだ緒についたばかりで、現在多方面で模索中である。人間の記憶は完全ではないが、何らかの手がかりを残しており、それをいかに引き出し補完していくかが重要な課題となる。

「あれ！、あの絵！」と、頭の中では分かっているのだが、作者や作品名が浮かばないときがよくある。このようなときに、いくつかの原画を順番にそっくり提示していく方法だと、時間がかかるし、多くのメモリが必要となる。例えば、浮世絵のように線画要素の強い絵に関して言えば、自由曲線を用いたイメージ表現が有効であり、高速化と省メモリに役立つ。しかし、水彩画のように色合いやふわっとしたぼけやにじみの要素が大切な絵には他の感性表現、例えば色、形、テキスチャ、筆使いなどを用いることが必要となる。また、イメージ・キーワードの上位概念である感性キーワードなども検討する必要がある。すでに、色による感性の受取方の違いなど、色彩と形状による感情効果については心理実験によりいくつかの知見を得ている[1]-[4]。また、自由曲線のドット展開法についても高速かつ簡単な方法を提案している[5]。

自由曲線による最良近似、すなわち最適制御点の自動抽出が本年度の研究のメインテーマとなる。処理手順としては、まず原画から雑音除去、エッジ抽出、細線化などの前処理を経て、線画化を行い、そこで得られた線要素を部分曲線に分割し自由曲線による近似を行う。開発されたアルゴリズムは、実際の絵画を用いた計算機実験により検証される[6]。

基本的な考えとして、適当に設定された初期値から始まって、制御点の位置を摂動させながら、２曲線間の近似度評価関数に従って、山登り法により最適解に近づけていく。一般には始点と終点は曲線上にとるのが普通であるが、曲線の接続や表現されるイメージなどを考えた場合、必ずしもそうするのが最適であるとは言えない。ここに、従来の工学的アプローチではあまり考慮されてこなかった人間の感性に対する配慮が必要になる。

自由曲線には多くの表現方法があるがその選択も重要である。結果としては、得られた制御点だけを記憶すればよいことになる。このことは曲線をそのままビットマップ・データとして記憶するのに比べてはるかに少ないデータ量となる。しかも、その利点は、グローバルなアフィン変換（平行移動、拡大縮小、回転など）やローカルな歪など各種変形に耐え得ることであり、すなわち大きな画面でも小さな画面でも、多少変形していてもそれに応じて品質の良い線画イメージを提供（作成）できる点である。単なる縮小では拡大したときに粗さが目立つが、本方式に従えばアウトラインフォントのように融通のきくイメージ・キーワードが作成できる。

最適近似曲線が求まった後では、積極的に制御点の特徴を生かして次のようなことも考

えられる。一つは、絵の共通的な部分の部品化であり、もう一つは、変形によるデフォルメや誇張表現である。また、線画だけでなく、原画上でマッピングにより対応付けを行ない、制御点を動かすことによって生じる自由曲線を用いて、スムーズな原画の変形を行なうことができる。

2．自由曲線とその表現方法

関数の近似をはじめ、補間、データ平滑化、曲線・曲面の設計、その他の多くの分野において広い意味でのスプライン関数の柔軟性と局所性が注目されている。その局所的な性質によって、スプライン関数は、多項式で近似するのが困難であるような複雑な形状を表現できる。そのためＣＧやＣＡＤ、あるいは高品質文字出力のためのアウトラインフォントなど、多方面で使われるようになってきた[7]-[12]。

自由曲線を生成する場合、２つの方法がある。まず、与えられた点列を全て通過する滑らかな曲線を張る方法があり、その代表的なものにスプライン曲線がある[8]。もう一つの方法は、与えられた点列を制御点として用いるだけで、全ての点を必ずしも通らない曲線を構成する方法があり、その代表的なものにＢスプラインやBézier曲線がある[8],[9]。Bézier曲線は制御点を１箇所変更すると、曲線全体に影響が及ぶ。この困難を避けるため、局所的な台を持つ基底を用いてスプライン関数を表現したものがＢスプラインである。そのため、制御点の変更はその点の近くのみの局所的な影響に留まる。

このように自由曲線にはスプライン曲線、Bézier曲線、有理Bézier曲線、Ｂスプライン曲線、円錐曲線など色々な表現方法があり、それぞれに一長一短があるが、いずれも有理Bézier曲線に変換できる。本論文では制御性に優れた有理Bézier曲線を対象にし、輪郭線の近似に用いる。ほとんどの自由曲線を使ったシステムにおいて、その表現式が求まった後に、ドットに展開する必要性がある。一般にドットの数は非常に多く、実用化のためには高速化は不可欠である。文献[5]では、離散的な空間上でBézier曲線を代表とする自由曲線の高速生成法とその応用を提案している。

自由曲線を表現するのに３次あるいは４次曲線が用いられることが多い。それは２次曲線では表現力にやや乏しく、また５次以上の曲線になると、一般に解析解を求めるのが困難になり計算も複雑になるからである。

Ｂｅｚｉｅｒ曲線は、$n+1$個の制御点$\mathbf{P}_0,\mathbf{P}_1,\cdots,\mathbf{P}_n$をもとにしてパラメータ$t$の$n$次多項式によって定まる曲線$\mathbf{P}(t)(0\le t\le 1)$のことをいう。Bézier曲線の表現には、穂坂表現[13]、パラメータ表現、陰関数表現など色々あるが、中でも一番よく使われるのは次式のパラメータ表現である。

$$\mathbf{P}(t)=\sum_{i=0}^{n}{}_nC_i(1-t)^{n-i}t^i\cdot\mathbf{P}_i \qquad (1)$$

式(1)は、有利Bézier曲線の分母が１の場合である。アウトラインフォントの表現などによく使われるのは、通常$n=3$の場合、すなわち４個の制御点$\mathbf{P}_0,\mathbf{P}_1,\mathbf{P}_2,\mathbf{P}_3$を使った３次Bézier曲線である。はじめに平面曲線を考える。曲線の座標値$\mathbf{P}(t)$は、

$$\mathbf{P}(t)=(1-t)^3\cdot\mathbf{P}_0+3t(1-t)^2\cdot\mathbf{P}_1+3t^2(1-t)\cdot\mathbf{P}_2+t^3\cdot\mathbf{P}_3 \qquad (2)$$

で求まる。平面上の曲線だけでなく、同じ式で空間中の曲線も表すことができる。ここでは、Bézier曲線によるイメージ表現を試みるが、主には、輪郭線の近似に用いる。

3．Bézier曲線による輪郭線の近似

図１に、原画の浮世絵から制御点を求め、線画によるイメージ表現を行う過程を示す。まず、前処理として、イメージスキャナーあるいはＣＤ－ＲＯＭから読みとったディジタルデータを、ソーベルのオペレータなどによる微分処理を行った後、あるしきい値で２値化する。次に、摂動させながら、学習により最適自由曲線に近づ

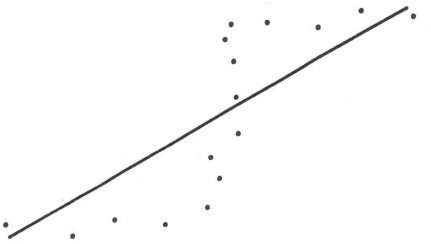
図１　浮世絵原画から制御点の求め方

けていくのであるが、ここでは、輪郭線をBézier曲線で近似するのに２通りの方法を試みる。

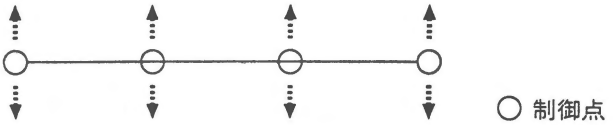
【１】近似曲線（制御点）を高速に得る方法

まず、前処理で得られた点列から、２乗誤差最小の直線を計算する（図２(a)）。次に、両端点を始点$\mathbf{P}_0$、終点$\mathbf{P}_n$とし、その間を$n-1$等分した点を$\mathbf{P}_1,\cdots,\mathbf{P}_{n-1}$とする。このとき、以後の学習を簡単にするため、回転と平行移動により直線とｘ軸を重ねる（図２(b)）。ここで得られた初期値から出発して、摂動による学習を開始する。制御点間の中央のｘ座標により、点列を分割し、各制御点の分担範囲を決める。制御点と分担範囲内の点とのｙ軸方向の距離（平均距離あるいは最大距離を用いる）を求め、その距離が許容範囲内になければ、その制御点を今求めた距離に従って摂動させる（図２(c)）。移動量は、始めのうちは大きく、学習が進むにつれて小さくとると効率がよい。全制御点の摂動が終わったら、曲線全体でもう一度ｙ軸方向の距離を計算し、その値が許容しきい値以下あるは最小になれば学習を終了する（図２(d)）。そうでなければ、学習を続ける。

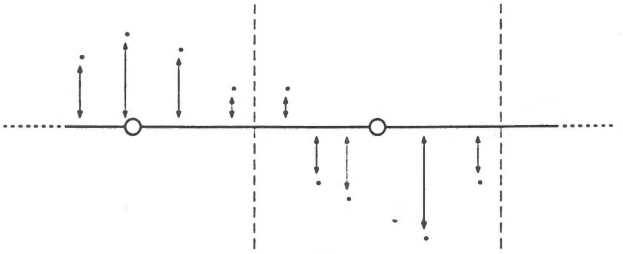
以上の手法の有効性を確かめるため、次に実際の浮世絵を用いた実行例を示す。学習に際して、移動量を次のように決める。ｙ軸方向距離が初期Bézier曲線（直線）の長さの10%以上のときは、その距離だけ移動し、10%以下のときはドット単位で移動する。移動量を変えることにより学習の効率化を図っている（図３）。終了しいき値は、初期Bézier曲線（直線）の長さの１%にしている。経験的にこれくらいの値をとっておくと肉眼で見たときに区別がつかなったので、採用した。原画として写楽の役者絵を用い、以上の手法を適用した結果を図４に示す。図(a)は２値化後の画像であり、図(b)は図(a)の顔の輪郭線の一部に上記の手法を適用して近似を行った結果である。白丸は接続点、黒丸は制御点の位置を示している。図(c)、図(d)は今後の部品化、変形を考える上で、制御点を意識的に変えることによるデフォルメの効果を確かめたものである。鼻のとがった人、顎の出っ張った人の表情が得られているのがわかる。部品化には、局所的な変形と同時に部品全体にわた



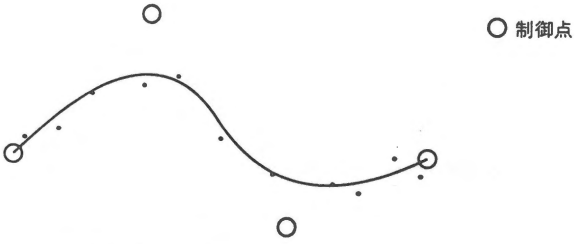
(a)直線近似



(b)制御点の初期値

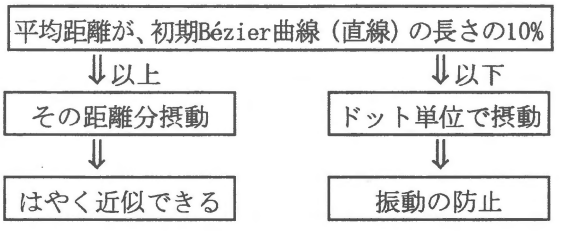


(c)制御点の分担範囲



(d)自由曲線全体にわたる誤差計算

図２　Bézier曲線による近似手法１



・終了しきい値は、初期Bézier曲線（直線）の長さの１%にする

図３　学習時の移動量と終了しきい値

る大局的な変形（アフィン変換などによる）が必要である。ここでは、４次のBézier曲線、すなわち５個の制御点を用いて一つの曲線セグメントを表現している。そのため２０点、接続点の重なりを考慮すると正確には１７点で顔の輪郭が表現できたことになる。しかも、自由に変形可能であり、本手法の有効性が確認できた。しかし、感性のイメージ表現という初期の目的のためにはまだまだ多くの問題が残されている。

【２】接続点における滑らかなつながりを考慮に入れた近似方法

上記の方法は、簡単かつ高速に近似曲線を求める手法であるが、接続点における滑らかなつながりが考慮されていないので、画像によっては不自然なぎざぎざが発生する可能性がある。そこで、Bézier曲線の特性を利用し、曲線セグメント間の滑らかな接続のできる近似手法を考える。数学的には滑らかさとは、高次の微分までを含めて連続であることを意味するが、ここでは、「１次微分（曲線の傾き）が接続点で連続」であるとき、「２つの曲線セグメントが滑らかに接続している」ということにする。

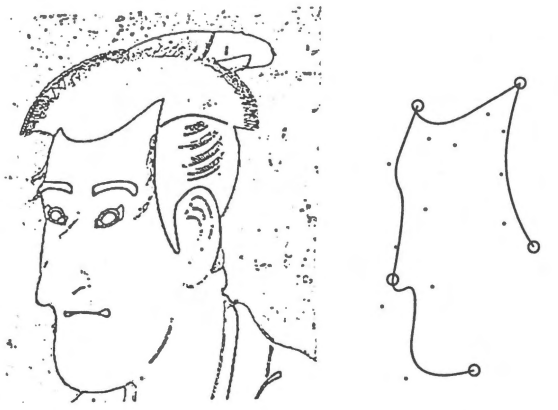
まず、曲線の傾きを計算しながら輪郭線を辿っていき、曲率（２次微分）の変化点を曲線セグメントの区切りとする。得られた曲線セグメントに対して、前述と同様両端点がx軸上にくるように正規化する。その曲線セグメントに含まれる点列の中でy座標の最大をy_Hとする。ここでは、簡単のため３次Bézier曲線を考え。仮に、パラメータ$t=1/2$のときが、Bézier曲線の値が最大値をとると仮定し、

$$y(t=1/2)=y_H \qquad (3)$$

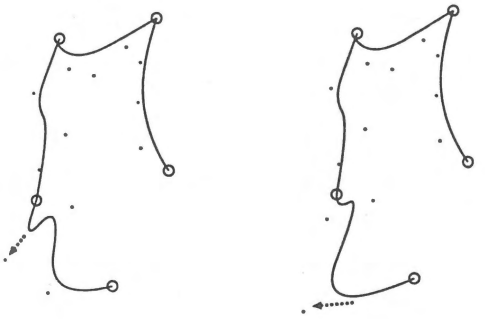
となるように、制御点の初期値を選ぶ。式(2)より、

$$\mathbf{P}(1/2)=1/8(\mathbf{P}_0+3\cdot\mathbf{P}_1+3\cdot\mathbf{P}_2+\mathbf{P}_3) \qquad (4)$$

が得られる。(3)と(4)より、$\mathbf{P}_1$と$\mathbf{P}_2$のy座標は$4/3\cdot y_H$となる。$\mathbf{P}_0$と$\mathbf{P}_3$は曲線セグメントの両端点にとるので、そのy座標は０となる。$\mathbf{P}_0$と$\mathbf{P}_3$はこの時点で確定する。両端点における１次接続を保証するため、制御点$\mathbf{P}_1$と$\mathbf{P}_2$は、両端点における接線上を動くことにする。両端点における接線と$y=y_H$との交点を$\mathbf{P}_1$と$\mathbf{P}_2$の初期値



(a)原画(2値化後)　(b)Bézier曲線による近似



(c)デフォルメ１　(d)デフォルメ２

図４　浮世絵を用いた本手法の適用例
○ はBézier曲線の接続点を表す。

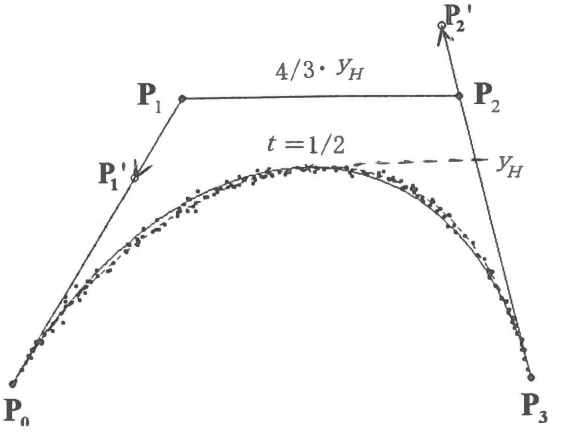


図５　滑らかな接続と制御点の学習

とする。初期値から出発して、$\mathbf{P}_1$と$\mathbf{P}_2$は両端点における接線上を動きながら、点列との距離が最小になるように学習は進んでいく。図５において、実線は制御点の初期値$\mathbf{P}_0,\mathbf{P}_1,\mathbf{P}_2,\mathbf{P}_3$に対するBézier曲線であり、破線は摂動後の制御点$\mathbf{P}_0,\mathbf{P}_1',\mathbf{P}_2',\mathbf{P}_3$に対するBézier曲線である。また、黒点は、２値化画像の模擬点である。この図より、接続点における、滑らかさを保ちながら点列への曲線近似が行われている様子がわかる。

４．まとめ

絵画のイメージ表現に向けて第１歩として、浮世絵を用いた線画イメージの表現方法について述べてきた。自由曲線としてBézier曲線を用いた近似手法を提案し、かなりの自由度を持って浮世絵のイメージ表現ができる可能性を示した。しかし、曲率の大きい場合、各制御点の曲線全体への影響、滑らかな接続、交差・コーナーへの対応、部品化と部品加工の方法など多くの問題が残されており、これらの解決は今後の課題である。手法２では、その一部が解決されているが、実際の応用にはまだまだ不十分である。さらに、具体的なＤＢ作りに向けて検討を続ける必要がある。

発展的形態として、絵画のイメージ表現を線画だけでなく、自由曲線部分、フラクタル部分、テクスチャ部分、プリミティブ部分（基本形状）に大別し、対象となる絵画に適したイメージ表現方法がある。そこでは構成要素抽出が大きな問題となる。例えば、「北斎の神奈川沖浪裏富士」で波頭の部分を自由曲線でいちいち近似していたら非常に効率の悪いものになる。そのためその基本構成要素を見つけてフラクタル表現し、少ないデータ量で波のイメージを表現する事を試みる。また、原画では基調カラーである水色が印象的でありキーカラーとして重要である。ここでもやはり正確さよりもイメージの表現が大切で、多少細かい所の省略や間違いがあってもよい。自由曲線以外にも、テキスチャやプリミティブによる表現を考え、色彩や形状と人の感じ方を関係づけるため心理実験を行い、色や形と「暖かい、きらびやかな、・・・」などの形容詞による修飾表現との関係、言い替えれば、イメージとテキストの関係とその表現法を検討していきたい。

以上述べてきたように、自由曲線を用いた浮世絵イメージの線画表現のための理論的道具作りという意味では初期の目的はある程度達成できたが、まだ実用に向けて多くの問題が残されており今後の検討課題としたい。

〔参考文献〕

[1]佐藤、皆川：“形状と色彩の感情効果に関する研究（第１報）－日本とヨーロッパの統文様を例にして－”、日本色彩学会誌、**18巻**、２号、pp. 137-146、1994

[2]佐藤、皆川：“立体形状物体の表面色への光の演出効果に関する研究（第１報）－ドレープ布への心理効果について－”、繊維製品消費科学会誌、**17巻**、９号、pp. 27-36、1993

[3]佐藤、安宅、皆川：“形状と色彩の感情効果に関する研究（第３報）－単色としての感情効果と幾何学的模様への配色の影響－”、日本色彩学会誌、**17巻**、１号、pp. 37-38、1993

[4]佐藤、皆川：“色彩・形状シミュレーションシステムによる模様と色彩の感情効果に関する研究（第１報）－日本とヨーロッパの伝統文様の２、３を例にして”、日本色彩学会誌、**16巻**、１号、pp. 79-80、1992

[5]濱、奥本：“Bézier曲線の高速生成法とアンチエリアシング”、テレビジョン学会誌、**47巻**、12号、pp. 1629-1636、1993

[6]志賀、柳原、濱：“絵画データベースにおけるキーワード作成のためのBézier曲線による輪郭線の近似”、電気関係学会関西支部連合大会、G12-32、p. G322、1995

[7]大野：“ＤＰＴのためのアウトライン・フォント”、PIXEL、No. 100、pp. 132-135、1991

[8]市田、吉本：“スプライン関数とその応用”、教育出版、1986

[9]長島：“ＣＧのための図学(1)～(10)”、PIXEL、No. 67～No. 78、1988-1989

[10]斉藤、穂坂：“拡張２次有理Bézier曲線による高品位文字フォントの生成とその特徴”、情報処理学会論文誌、**Vol.31**、No. 4、pp. 562-570、1990

[11]寅市、関田、森：“高品位文字フォントの自動圧縮”、信学論、**Vol.J70-D**、No. 6、pp. 1164-1172、1987

[12]西田、中前：“Bézier曲線で囲まれた領域の走査変換法－アウトラインフォントへの応用－”、情報処理学会グラフィックスとＣＡＤ研究会、**45**、1990

[13]穂坂、木村：“３次元自由形状設計制御理論とその手法”、情報処理、**Vol.21**、No. 5、pp. 481-492、1980

[14]岸田、岑、濱：”Basic Studies on Visual Information Processing Modelization by Binocular Stereopsis Vision”, 3D Image Conference '94, pp. 99-104, 1994

[15]江、濱：“A quick method for extracting corners”、電気関係学会関西支部連合大会、G12-31、p. G321、1995

画像データベースの自然言語インタフェースについて

On Natural Language Interface for Image Database

伊東幸宏，中谷広正
Yukihiro ITOH, Hiromasa NAKATANI
静岡大学情報学部，浜松市
Faculty of Information, Shizuoka University,
Hamamatsu, Shizuoka, 432

あらまし:　本稿では，画像データベースの自然言語インタフェースの構築について述べ，その構築経験から人文科学分野の研究・思考を支援するためのユーザインタフェースの枠組について考察する．画像データベースの自然言語インタフェースの構築では，自然言語で表現される直観的印象概念を解釈して画像特徴に関する制約を生成するための特徴空間の構築が必要となる．この特徴空間構築プロセスを形状特徴空間の構築を例に説明し，このプロセスが一つの思考プロセスのモデルとなりうることを述べ，人文科学研究支援システムの一つの可能な形態について述べる．

Summary: In this paper, we describe a method to construct a natural language interface for an image data retrieval system, and discuss a framework of an interface for a computer aided research system. In order to construct a natural language interface for an image retrieval system, we should construct a feature space on which subjective words representing user's intuitive impressions are arranged, and through which the system transforms user's inquiry sentences into representations of restriction on image features. We illustrate the process to construct a shape feature space to retrieve images of chairs. Then we discuss similarity between the process and thinking process of researchers of humanities, and show one possible framework of interface for a computer aided research system.

キーワード: 画像データベース，自然言語インタフェース，思考支援
Keywords: Image database, Natural language interface, Computer aided research

1　はじめに

多くの人文科学領域で画像データを含むマルチメディアデータベースが構築されつつある．しかし，それを使いこなす技術，すなわち，広い意味でのユーザインタフェース技術は，十分に検討されているとは言えない．

人文科学研究に積極的にデータベースシステムが活用されるようになるためには，まず，データベースの非専門家にとって扱いやすいインタフェースが提供されることが必要である．しかし，単に表面的な取り扱いにくさを解消するだけでは十分ではない．データベースシステムを積極的に利用してゆくことによって，人文科学の研究者の思索が触発されるような環境，言い換えれば，ある種の思考支援機能を含んだ環境が提供されることが望ましい．このうち，使いやすいインタフェース構築の問題に関しては，ＧＵＩの研究を含め多くの研究がなされてきている．しかし，思考支援機能を含んだインタフェースについては，まだ研究が緒についたところであるというのが現状である．

我々は，これまでに，データベースの非専門家であるユーザにとって使いやすいインタフェースとして，自然言語インタフェースシステムを検討してきた．このシステムは，椅子の電子化カタログから，色や形の情報を用いて椅子画像を検索するという例題を想定して試作されている．このシステムは，「柔らかい色」「陽気な形」といったような，ユーザの主観的印象を表す語彙を用いることができる，「もっと淡い色」「もっとシャープな形」等，比較表現を用いて対話的に検索を進められる，という特徴をもつ．

このシステムの試作の際に，我々は椅子の形状特徴についての言語表現を解釈して，椅子画像から画像解析によって抽出しうる特徴量に関する制約条件を生成するための形状特徴空間の構築を行ってきた．この特徴空間構築プロセスは，データベース中の画像データを直感的な印象概念に基づいてクラスタリングしうる空間と，その空間上での座標値と画像特徴の間の関係を定義する作業である．我々は，このプロセスを基にして画像データベースを用いた研究・思考プロセスの一つのモデルを想定できるのではないかと考えている．

そこで本稿では，このインタフェースシステムの概要と特徴空間構築プロセスについて述べ，その枠組みに基づいて，思考支援機能を含めた

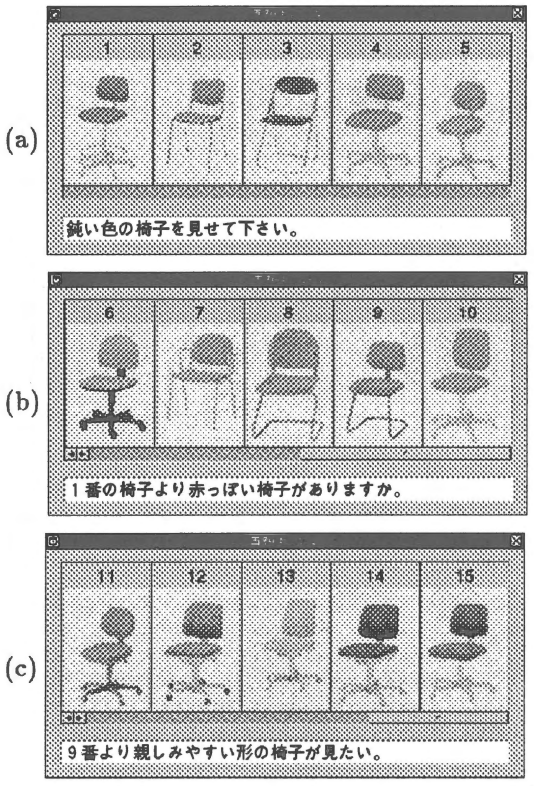


図 1: 検索例

人文科学研究にとって有用なインタフェースシステムへの拡張の方向性について述べる．

2 システムの概要

2.1 動作例

システムの検索例を図 1に示す．まずユーザには日本語で欲しい椅子を入力してもらう．例えば，「鈍い色の椅子を見せて下さい。」と入力したとする．すると，システムは 5 枚の候補画像を提示する (図 1(a))．ユーザは続けて，「1番の椅子より赤っぽい椅子がありますか。」と提示された画像との比較表現を用いて検索が行うことができる (図 1(b))．提示された画像に満足がいかない場合はさらに続けて，「9番より親しみやすい形の椅子が見たい。」のように，対話的に要求を入力することができる (図 1(c))．このように，ユーザは対話的に検索を行うことにより，欲しい画像の検索が行える．

2.2 処理の流れ

次に，処理の流れを簡単に示す．なお，ここでは議論を簡単にするため，色特徴を言及する表現を例にとって説明する．

ユーザから入力された自然言語文は形態素解析，構文解析が施され，意味表現 [1] が生成される．この意味表現において我々は，「暖かい」，「かわいい」といった感性語句をそれぞれ，「暖かさ」，「かわいさ」といった感性的属性概念を用いて定義した．

次にこれらの感性的属性概念は色特徴空間上の検索領域に変換される．我々はこの過程を意味表現の解釈と呼ぶ．この解釈は予め感性的属性概念ごとに用意された領域，ならびに，ピーク点に基づいて行われる．この領域とは多くの人がその色に対して感性的属性概念で表されるような印象を持つ色特徴空間上での色の範囲である．また，ピーク点とは最もその印象を表すと思われる色特徴空間上での点である．

最後にシステムは指定された検索範囲内の色をもった椅子画像を探索し，候補画像として提示する．なお，画像中の椅子の色の値は，画像解析を行って自動的に抽出される．

以上で説明した「意味表現の解釈」処理は，ユーザの直感的印象を表現した概念表現を，以下のような特徴空間上での検索条件表現へと変換するプロセス，と捉えることができる．

(1) 人間の印象を反映する．すなわち，同様な印象をもつ概念同士は空間上においても近い領域を占める．

(2) 画像解析により特徴空間上の座標が求められる．

このような条件を満たす特徴空間として，色特徴に対しては色相，明度，彩度軸で構成するＨＳＶ空間がよく知られている．しかし，形状特徴に対しては，これまでのところそのような空間は提案されていない．そこで，我々は，以下のような方法を用いて形状特徴空間の構築を行った．詳細は第 3 章で述べるが，我々は，この特徴空間構築のプロセスが，データベースシステムによる人文科学の思考支援の一形態を示唆していると考えている．そこで，次章では，その形状特徴空間の構築プロセスについて述べる．

3 属性空間の構築 - 形状空間の場合 -

我々は，以下のようにして形状特徴空間を構築した．まず，イメージ空間の測定には SD 法 [2] を用いることにした．SD 法とは，心理実験を基に因子分析をするといった手法であり，抽象的概念から具体的な物品まで，さまざまな刺激が及ぼすイメージ・印象・雰囲気・感情といった側面での，心理効果を測定する方法として知られている．ここで簡単な流れを記しておく．

図 2: ３３種類の椅子

Chair Number 1

直線的な	—:-o-:—:—:—	曲線的な
落ち着いた	—:—:—:—:-o-	落ち着かない
陰気な	—:—:-o-:—:—	陽気な
	:	
まとまりのある	—:—:—:-o-:—	まとまりのない

図 3: アンケート用紙 解答例

Step1 なるべく形の異なる対象（椅子）の画像を収集する．我々は実験では３３種類の椅子の２値画像を用いた (図 2).

Step2 ＜陰気な－陽気な＞，＜暖かい－涼しい＞といった形容詞対からなるアンケート用紙を作成する (図 3). 我々は２０種類の形容詞対を用いた．また，それぞれの対は５段階で評価してもらうことにした．

Step3 被験者にアンケート用紙を用いてＯＨＰにより提示された画像の形についてその印象を評価してもらう．我々は２６名の被験者（男子学生）に対して実験を行った．

Step4 アンケートの結果から因子分析を行い，イメージ空間の軸を決定する．

表 1は因子分析の結果である．この結果から椅子の形の心理的意味は，互いに独立した３つの因子に依存しているといえる．また，この因子負荷行列を用いて因子スコアを推定し，それをプロットしたものを図 4に示す．

図 4は，横軸にそって第１因子を，縦軸にそって第２因子を表している．この図では左にいくほど「まとまりのある」，「親しみやすい」，「使いやすい」，「落ち着いた」，「単純な」，「上品な」，「地味な」，「安定した」，といった第１因子に含まれる印象が強くなることを示している．逆に右方向にいくほど反対の印象が強くなることを示している．また同様に，上方向にいくほど第２因子に含まれる「軽やかな」，「安っぽい」，「冷ややかな」，「硬そうな」，「すっきりした」，「直線的な」，といった印象が強くなっており，下方向にいくほどその反対の印象が強くなっていくことを示している．

SDスケール	因子負荷量			共通性
	第１因子	第２因子	第３因子	
まとまりのある-まとまりのない	0.93	-0.04	0.14	0.89
親しみやすい - 親しみにくい	0.92	0.10	-0.08	0.86
使い易い - 使い難い	0.76	0.13	0.41	0.76
落ち着いた - 落ち着かない	0.76	0.41	0.39	0.90
単純な - 複雑な	0.72	-0.60	-0.00	0.88
上品な - 下品な	0.68	0.39	-0.18	0.65
派手な - 地味な	-0.66	0.19	-0.66	0.91
安定した - 不安定な	0.63	0.50	0.47	0.87
軽やかな - 重厚な	0.02	-0.92	-0.36	0.98
安っぽい - 豪華な	0.23	-0.91	0.01	0.87
暖かみのある - 冷やかな	0.33	0.87	-0.24	0.93
硬そうな - 柔らかそうな	-0.23	-0.85	0.06	0.78
すっきりした - ごてごてした	0.60	-0.73	-0.24	0.94
直線的な - 曲線的な	-0.15	-0.51	0.27	0.35
かわいくない - かわいい	-0.16	0.09	0.93	0.90
陰気な - 陽気な	0.20	0.05	0.89	0.83
かっこいい - かっこ悪い	0.09	-0.04	-0.84	0.71
古臭い - 新しい	0.54	-0.21	0.77	0.93
趣味的な - 実用的な	-0.59	0.19	-0.72	0.90
一般的な - 個性的な	0.66	-0.22	0.68	0.95
Commonality	8.29	5.22	3.73	17.2
共通性 (%)	41.4	26.1	18.7	86.1

表 1: 因子分析結果

このように，因子スコアは因子にどの程度依存しているかを表しているもので，例えばこの図では２７番の椅子は第２因子のプラス方向に対する依存が高い．すなわち，２７番の椅子は「まとまりのある」，「落ち着いた」形をしているという印象を与えると考えられる．逆に，８番の椅子は反対方向に依存が高いため，「かわいくない」，「陰気な」形の椅子といった印象を与えると考えられる．

次に，我々は構成したイメージ空間と相関の高い物理的特徴空間を構成することにした．まず，我々は意味空間の因子と対応すると思われる物理的特徴概念を抽出することから始めた．

構成したイメージ空間の軸上に椅子を並べ，観察を行い，そこからその軸に影響すると思われる物理特徴を調査した．その結果，我々は椅子の形の印象は主にその椅子の脚の形に依存していると推定した．そこで，椅子の形に基づき４つのタイプに分類した (図 5). タイプ１は脚が一本脚の回転椅子，タイプ２は２本脚の椅子，タイプ３は折りたたみ椅子，タイプ４は４本脚

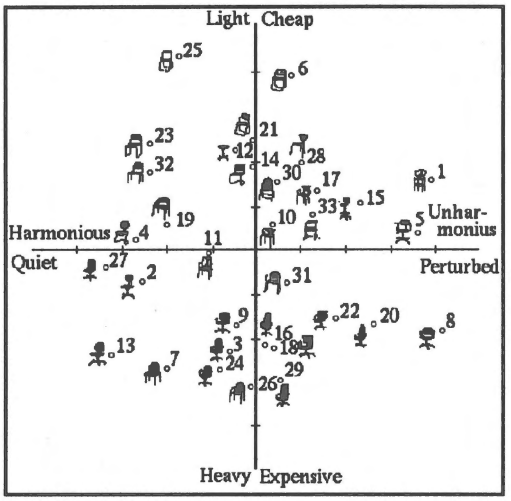


図 4: 構成したイメージ空間（断面図）
(横軸: 第 1 因子, 縦軸: 第 2 因子)

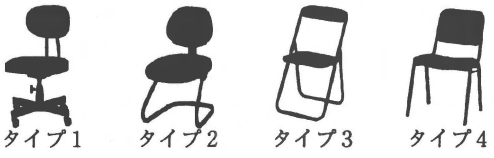


図 5: 椅子のタイプ分け

の椅子，といった 4 つのタイプに大別した．

しかし，脚のタイプだけではイメージ空間に対応付けることは難しい．そのため，さらに分類した椅子のタイプごとに調査を行った．例えば，図 6 は，タイプ 4 の椅子を第 2 因子の軸上に並べたものである．この図から左側の椅子ほど黒い部分の面積が大きいことが分かる．そこで，我々はタイプ 4 の椅子が第 2 因子に影響を与える特徴として式 1で表される特徴概念を定義することにした．ここで，「椅子の面積」は図 7のように画像を 2 値化し，その画像から抽出される黒い部分の面積を指し，「椅子の凸包の面積」とは図 7のように 2 値化した画像の凸包の面積を指す．

$$-\frac{\text{椅子の面積}}{\text{椅子の凸包の面積}}, \qquad (1)$$

他のタイプもそれぞれ同様にして定義した式を表 2，に示す．我々はこれらの式の値をそれぞれ変形させ，イメージ空間軸上の値と対応させるようにした．

図 8，9にイメージ空間と形状特徴空間を示す．図 8に我々の心理実験により構成されたイメージ空間の 3 つの軸を示し，図 9に我々が定義した式を用いて算出した形状特徴空間の 3 つの軸を示す．心理的イメージ空間と形状特徴空間のそれぞれ軸の相関係数は，第 1 因子が 0.99，第 2 因子が 0.96，第 3 因子が 0.95 と高い相関が得られた．

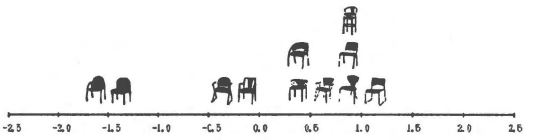


図 6: 第 2 因子軸上のタイプ 4 の椅子

図 7: 椅子の面積とその凸包の面積

4 思考支援に向けて

研究者が画像データベースを用いて行う思考には様々なタイプが考えられるが，そのうちの一つのタイプとして，次のようなものを考えることができよう．それは，研究者は雑然と集められたデータ画像を俯瞰すべき着目点を定め，その視点に基づいてデータを整理して，ある種の一般性，普遍性を見いだす，というプロセスを踏む思考である．このタイプの思考を行う場合，データを常にデータベース作成者が定めた座標軸に拘束されて取り扱うのでは柔軟な発想が妨げられる恐れがある．

データベースを用いた研究者の思索を支援するためには，データを整理すべき独自の座標軸の設定を支援すること，その座標軸を用いて直接データを取り扱えることが必要であろう．

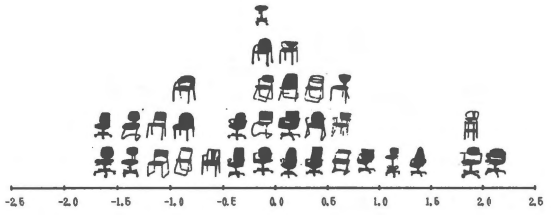
すなわち，画像データベースを利用した人文科学研究のプロセスモデルを以下のように考える．

（1） 大量のデータを俯瞰すべき視点についての着想を得る．

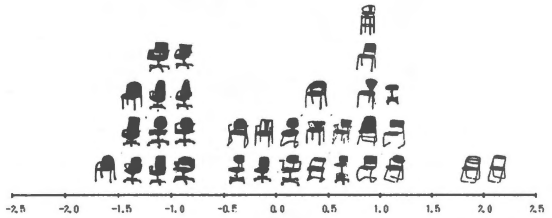
（2） その着想に基づいて，緒データを整理すべき空間を想定する．その空間上では，緒データが，上述の視点から見て特徴的な様々な概念に対応したクラスターを構成し，それらの概念間の関係が明瞭に判断できなければならない．

（3） そのような空間上に整理されたデータから，一般的，普遍的な概念間の関係を導き出す．

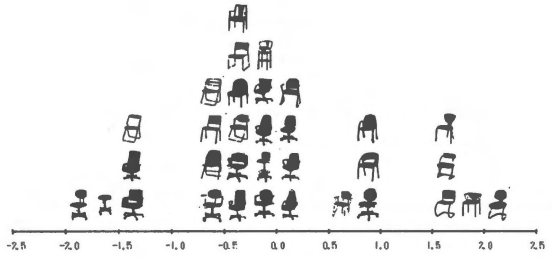
このうち，（2）の空間の想定プロセスは，前述の特徴空間の構築プロセスとほぼ同一のア



第 1 因子

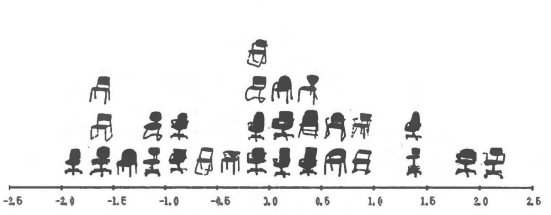


第 2 因子

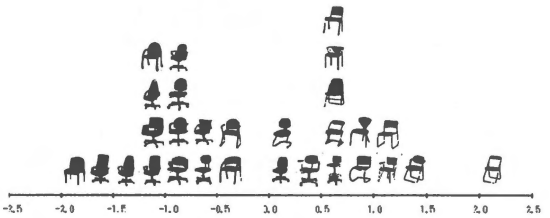


第 3 因子

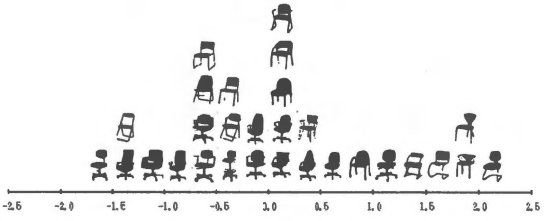
図 8: 心理実験で得られたイメージ空間



第 1 因子



第 2 因子



第 3 因子

図 9: 画像特徴より算出した特徴空間

プローチで考えることができると思われる．すなわち，研究者は，まず，緒データを俯瞰すべき着目点を想定し，その着目点から見て特徴的な概念を拾い出す．データベース中のテストデータに対してそれらの概念を基準とした評価を行い緒データを整理すべき座標空間に関する仮説を生成する．次いで，その空間と画像データがもつ緒特徴量とを関係づけた上で，仮説を検証するために，より広い範囲のデータに対し概念記述を用いた検索を試行して評価する．その結果に応じて仮説の修正を繰り返すことによって，（3）のステップに耐えうる特徴空間を想定できるのではないだろうか．

このように考えると，前章で述べたような特徴空間の構築支援機能を持ち，概念表現を直接入力できるインタフェースシステムが，人文科学支援の一つの形態をなすと考えることができるのではないだろうか．

我々は，ユーザの直感的印象概念を整理するための座標空間の構築にＳＤ法を用いた．この方法は，複数の被験者による直感的評価に基づいている．場合によっては，このような方法は適用できないことも考えられるので，この他の方法論を用いてこのフェーズを支援するためのツールも必要となろう．また，特徴空間と画像特徴との関係を定義する際には，座標軸に沿って実画像を配置してそれを人間が観察して抽出すべき特徴量と，座標値と画像特徴量との間の関係式を求めるという方法をとった．このフェーズは，このままでは画像処理や数式処理に不慣れなユーザでは実行できない．従って，このフェーズを支援するためのツールが必要である．そのためには，例えば，ニューラルネットや機械学習の理論を援用することなどが考えられる．

5　まとめ

画像データベースの自然言語インタフェース構築について述べ，そこで考えた特徴空間構築プロセスが，画像データベースを利用した人文科学研究支援の一つのプロセスモデルと考えられることを示した．

今後，まず，この考え方を，実際の人文科学研究に具体的に適用することが可能か否かを検証することが必要である．さらに，課題として，前章の最後でふれた２つのフェーズの支援ツールの設計・開発を進めてゆくことが挙げられる．

タイプ分け		第1因子の関数	第2因子の関数	第3因子の関数
タイプ1	肘なし	$\frac{\lvert Lu-Ll\rvert}{Lu+Ll} - \frac{Se}{Sf}$	$-\frac{\lvert Lr-Lb\rvert}{Lr}$	(b)
	肘あり	(a)	$-\frac{Sa}{Sh}$	
タイプ2		$\frac{2(W-H)}{MAX(W,H)} - \frac{4\pi Sb}{Lb^2} - \delta$	$\frac{\lvert Lb-Le\rvert}{Le} - \frac{Sa}{Sh}$	δ
タイプ3		$\frac{H-W}{MAX(W,H)} - \frac{Sb}{Sa}$	$-\frac{Sa}{Sh}$	$\frac{4\pi Sb}{Lb^2} + \delta$
タイプ4		(c)	$-\frac{Sa}{Sh}$	$\frac{Su-Sl}{4Sb} + \frac{\lvert Lu-Ll\rvert}{4Lb}$ $-\frac{\lvert Le-Lb\rvert}{Le} + \frac{\lvert Lr-Lb\rvert}{Lr}$

(a)

タイプ分け		関数
$\frac{W-H}{MAX(W,H)}$ (=A)> 0.3		$\frac{\lvert Lu-Ll\rvert}{Lu+Ll} - \frac{Se}{Sf}$
A<= 0.3	$\frac{Se}{Sb} < \frac{Sr-Sb}{Sb}$	$\frac{Se}{Sb}$
	$\frac{Se}{Sb} > \frac{Sr-Sb}{Sb}$	$\frac{Sr-Sb}{Sb}$

(b)

タイプ分け	関数
$MIN\left(\frac{\lvert Le-Lb\rvert}{Le}, \frac{\lvert Lr-Lb\rvert}{Lr}, \frac{\lvert Lt-Lb\rvert}{Lt}\right)$(=B)< 0.03	$\frac{W-H}{100MAX(W,H)}$ +B
B >= 0.03	$\frac{W-H}{(100MAX(W,H)}$ −B

(c)

タイプ分け	関数
$\frac{\lvert Lb-Le\rvert}{Le} < MIN(\frac{\lvert Lb-Lr\rvert}{Lr}, \frac{\lvert Lb-Lt\rvert}{2Lt})$	$\frac{Sb}{Sa} + \frac{10Sd}{ScN} - \frac{\lvert Lb-Le\rvert}{Le}$
$\frac{\lvert Lb-Le\rvert}{Le} > MIN(\frac{\lvert Lb-Lr\rvert}{Lr}, \frac{\lvert Lb-Lt\rvert}{2Lr})$	$\frac{Sb}{2Sa} + \frac{5Sd}{ScN} - MIN(\frac{\lvert Lb-Lr\rvert}{Lr}, \frac{\lvert Lb-Lt\rvert}{2Lt})$

表 2: 定義した関数一覧

参考文献

[1] 高木, 伊東: “自然言語の処理,” 丸善, 1987.

[2] 岩下豊彦: “SD法によるイメージの測定”, 川島書店, 1983.

[3] 原田, 杉浦, 大庭, 中谷, 伊東: “自然言語による画像データベースの検索”, 人工知能学会全国大会 (第7回) 論文集,pp593-596, 1993.

[4] 杉浦, 原田, 中谷, 伊東: “自然言語による画像データベース検索”, 電子情報通信学会技術研究報告 Vol94 No.51,pp55-62, 1994.

[5] H.Nakatani and Y.Itoh: “An Image retrieval system that accepts natural language,” AAAI-94 Workshop Notes on Integration of Natural Language and Vision Processing, pp7-13, 1994.

[6] S.Harada, H.Sugiura, H.Nakatani and Y.Itoh: “On constructing pictorial feature space for image retrieval,” IJCAI-95 Workshop Notes on Representation and Processing of Spatial Expressions, pp103-117, 1995.

[7] 原田, 田中, 伊東, 中谷: “画像検索のための形状特徴空間の構築”, 電子情報通信学会技術研究報告 Vol95 No.320,pp7-12, 1995.

多視点距離データを用いた 3 次元形状モデリング
3D Object Modeling Using Multiple-View Range Data

横矢 直和 †　増田 健 ‡
Naokazu YOKOYA† and Takeshi MASUDA‡

† 奈良先端科学技術大学院大学 情報科学研究科、生駒市
† Graduate School of Information Science, Nara Institute of Science and Technology
Ikoma-City, Nara 630-01

‡ 電子技術総合研究所 知能情報部、つくば市
‡ Machine Understanding Division, Electrotechnical Laboratory
Tsukuba-City, Ibaraki 305

あらまし: 遺物、美術品、工芸品等の文化財データベースから成る仮想博物館の構築においては、通常の文書情報や画像情報に加えて、仏像等の彫刻に代表される複雑な 3 次元物体の形状を如何に計測・記録・提示するかが重要な課題となる。すなわち、複雑な自由形状を有する文化財の表面の形状を如何に計測するか、計測されたデータを如何に表現するか、形状データを目に見える物あるいは触れる物として如何に提示するかという問題である。本稿では、最初の問題である 3 次元形状データベース構築のための実物体の計測に基づく形状モデリングにおける課題と問題点を整理し、具体的な研究事例を紹介する。まず最初に、(1) 多視点 3 次元計測、(2) 多視点データの位置合わせ、(3) データ統合による表面形状記述の必要性について述べ、次に、(1)、(2) の課題に対する具体的なアプローチとして、筆者らが研究を進めている光投影型レーザレンジファインダによる距離画像取得とランダムサンプリングと ICP アルゴリズムを用いた多視点距離画像の位置合わせを実験結果を交えて紹介する。

Summary: In constructing a virtual museum composed of databases of cultural properties such as historic remains and artistic handicrafts, important problems are measurement, recording and display of their three-dimensional (3D) shapes. In other words, the problems are measurement, representation, visualization and realistic presentation of complex free-formed 3D object shapes. This paper addresses the problem of obtaining 3D shape descriptions of real objects by measurements. First we divide the problem into three subproblems: (1) multiple-view measurement of a 3D object, (2) multiple-view data registration, and (3) data integration for obtaining an entire surface representation of the object. We then present an approach for the stages (1) and (2) which consists of multiple-view 3D surface measurement by using active range sensors such as laser rangefinders and automatic range image registration based on a technique of random sampling and the ICP (Iterative Closest Point) algorithm.

キーワード: 3 次元形状モデリング、多視点計測、距離画像、距離データの位置合わせ、データ統合
Keywords: 3D object modeling, multiple-view measurement, range image, range data registration, data integration.

1 はじめに

従来、遺物、美術品、工芸品等の記録は実測図や写真などの2次元的な記録が中心であるが、コンピュータによる文化財管理、計算考古学、美術研究、仮想博物館構築等においては、文化財の3次元情報の記録・保存が重要であると考えられる。特に、仏像等の彫刻に代表される複雑な3次元物体については、通常の文書情報や画像情報に加えて、実物体の計測に基づく客観的かつ高精度な3次元表面形状情報が有力な手がかりを与えることが多い。3次元形状の取得とそのデータベース化によって、例えば、

- 定量的な形状解析、形状比較
- 臨場感のある立体的な画像提示
- 光造形やNC加工によるラピッドプロトタイピング

等が可能となり、学術研究や文化財鑑賞に新たな手段を提供することができる。

このような目的での3次元形状の計測とデータベース化においては、3次元画像計測 [1] と計測データに基づく形状モデリング (幾何モデル生成) が有効であるが、通常の3次元計測装置では $2\frac{1}{2}$ 次元情報 (ある特定の方向から見える物体表面の基準面からの距離情報) しか観測できないため、物体の全表面の形状データが得られないという問題がある。したがって、表面形状記述を取得するためには、

1. 全ての物体表面を観測するための多視点3次元計測、
2. 多視点距離データの (半) 自動位置合わせ、
3. 多視点距離データの統合による表面形状記述の生成

というアプローチが必要になる [2, 3, 4, 5]。

本稿では、実物体の計測に基づく3次元形状モデリングのための上記のアプローチにおける課題と問題点について考察するとともに、1.、2. の課題に対する具体的な研究事例として、筆者らが研究を進めている光投影型レーザレンジファインダによる距離画像取得とランダムサンプリングとICPアルゴリズムを用いた多視点距離画像の自動位置合わせ法 [6] を実験結果を交えて紹介する。また最後に、残されている今後の課題について簡単に述べる。

2 3次元実物体のモデリングにおける課題と問題点

2.1 表面距離データの取得

非接触でシーンの3次元情報を得る方法は受動的手法と能動的手法に大別できる [7] (図1参照)。

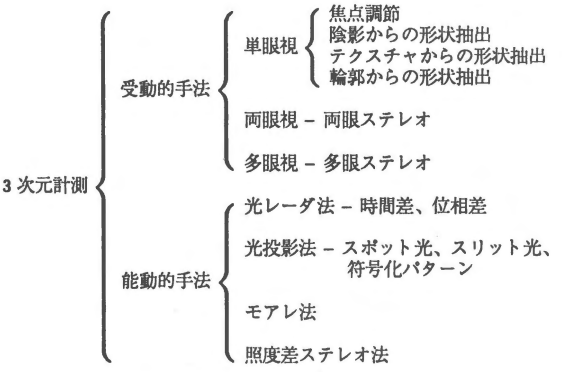


図 1: 非接触3次元計測手法

この中で、レーザ光を用いた光投影法に基づく能動的距離センサは、シーンの密な距離データ (距離画像) の取得が比較的容易であるのに加えて、

1. 計測の高精度・高分解能化
2. 高速化の追求による実時間計測
3. 距離データと表面カラーデータの同時取得

の各側面からの研究が進み、実用的な商用機器も登場しているため、現状では、3次元形状モデリングのための計測装置として最も有力である。しかしながら、現状の装置では以下のような問題が残っている。

(1) 鏡面反射物体や黒色物体については計測できない (物体の表面材質)。

(2) レーザ照射方向と面の法線ベクトルがずれるにしたがって計測の信頼度が落ちる。

(3) 複雑な形状については死角が生じ、計測できない部分が生じる (物体の形状)。

(4) 1つのセンサを用いた1回の計測で物体の全ての表面を観測するのは不可能である。

最初の2つの問題点は計測方式に由来するもので、この計測方式を採用する限り、解決は難しい。後の2点はある意味では装置構成に原因があるとも考えられ、工夫の余地がある。以下では (3) と (4) の問題について考察する。

近年、比較的広範囲の表面を計測できる装置として、(i) 対象物体を可動ターンテーブルに載せて回転させる方式や (ii) センサ部が物体の回りを回転する方式によって物体側面の全周を計測できる装置がいくつか登場している。いずれも、回転軸方向と回転角度をパラメータとする円筒座標系の距離データが得られる (図 2 参照)。例えば、図 3 は (ii) の方式のレーザレンジファインダで木彫モアイ像を計測した例である (表面カラーデータも同時に取得)。しかし、このような方式でも、上部と下部の計測精度が低いことに加えて上面と下面は全く計測できないことが多い (図 3 でも頭頂部と底面は計測できていない)。

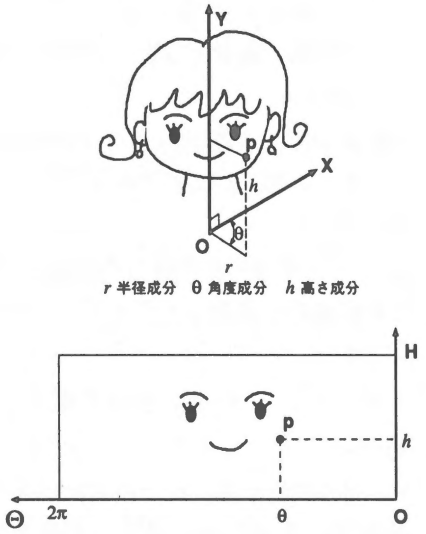


図 2: 全周計測と円筒座標系画像 [8]

このため、物体の置き方を変えて、異なる視点から複数回計測せざるを得ない。このような多視点計測においては、物体の回りに複数の距離センサを配置する方式も考えられる。いずれの場合も、計測データはセンサ中心座標系 (あるいは視点中心座標系) で表現されており、複数の距離データから 1 つの表面記述を得るためには、データ間の座標変換が必要になる。複数センサ間や複数視点間の相対位置を予め正確に設定できるとは限らないので、

- 計測データに基づく複数距離データの位置合わせ

が必要になり、位置合わせされたデータの統合 (貼り合わせ) によって最終的に物体の表面記述を生成することになる [2, 4, 5]。

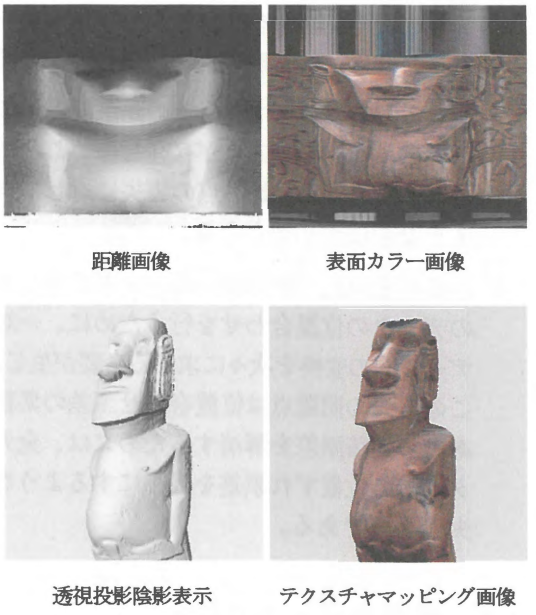


図 3: 全周計測型レンジファインダによる木彫モアイ像の距離データとカラーデータの同時取得

2.2　多視点距離データの位置合わせ

ここでは、剛体物体を複数の異なる視点から観測した距離データの自動位置合わせについて考える。剛体を表す距離データ間の位置合わせ問題はデータ間の剛体運動パラメータを求める問題であり、アプローチは次の 2 つ手法に大別できる [9]。

1. データ間の点対応に基づく方法
   データ間での点対応が与えられれば、全対応点間の距離の自乗誤差を最小にするような変換パラメータを求めることができる。このための手法には、回転行列を単位 4 元数で表現して問題を線形化する方法 [10] 等がある。実際には、対応点探索が問題となる。観測方向に不変な局所特徴である曲率等を手がかりに対応点を探索する手法が提案されているが、対応点を安定に求めることは容易ではない。
2. 反復最近点選択アルゴリズム
   予め点対応を確定するのではなく、アルゴリズム内で最近点を仮の対応点とし、仮の対応に基づく剛体変換プロセスを繰り返す方法である [11](詳細は後述)。仮の対応から変換パラメータを求めるところでは上述の 1. の手法を用いることができる。この ICP(Iterative Closest Point) アルゴリズムは初期変換が適切でないと局所解に陥るという問題がある。

複雑な3次元実物体を計測した複数距離データの位置合わせにおける問題点を以下に整理する。

- 視点の移動に伴って奥行きの隠蔽関係が変化し、データ間で対応点が存在しないことが起こる。このため、対応点の喪失に対してロバストな手法が必要となる。
- ある特定の視点から撮ったデータに対して他のデータの位置合わせを行うために、一対のデータ間の変換を次々に求める必要が生じる。この場合の問題点は位置合わせ誤差の累積である。累積誤差を解消するためには、全データ間での位置ずれ誤差を最小にするような工夫が必要である。

2.3 多視点データの統合

位置合わせされた距離データを統合して表面形状の記述を得る方法には次の2つがある。

1. 複数視点から観測した全ての距離データ (3次元座標データ) の集合から三角形メッシュ表現等の形状記述を生成する。
2. 予め三角形メッシュ表現等の形状記述が得られている複数視点データを接合し表面全体の記述を生成する。

データ統合における課題は、

- 異なるデータ間の境界を如何に滑らかに接合するか、
- 計測方式やデータ表現形式が異なる複数のレンジファインダで取得したデータの整合性を如何にとるか、

である。また、複雑な形状に対しては接触式センサを併用せざるを得ない場合もある。

3 多視点距離画像の位置合わせアルゴリズム

距離画像には雑音や隠蔽などの現象による外れ値が含まれているので、それらの影響を受けにくいようにアルゴリズムを構成しなければならない。ここでは、予め対応を与えることを必要とせずに、2枚の剛体の距離画像から3次元剛体変換パラメータを求める頑強なアルゴリズムを提案する [6]。

3.1 アルゴリズムの概要

単位4元数を用いた3次元剛体運動推定を利用したICPアルゴリズムを、ランダムサンプリングおよびLMS(Least Median of Squares) 推定と組み合わせることによって、頑強なパラメータ推定を行う。以下では、異なる視点から観測した2枚の距離画像を R^I、R^{II} で表し、R^I から R^{II} への3次元座標 $\mathbf{x}$ の剛体運動 $\mathbf{T}$ を

$$\mathbf{T}(\mathbf{x}) = \mathbf{R}\mathbf{x} + \mathbf{t} \qquad (1)$$

と表す。ただし、$\mathbf{R}$ は 3×3 の3次元回転行列、$\mathbf{t}$ は平行移動ベクトルである。

提案アルゴリズムの流れを以下に示す。

1. R^I から $N_S(\geq 3)$ 個の点をランダムに選択し、この点の集合を P_i^I と表す (i は繰り返しの回数)。
2. 点集合 P_i^I と距離画像 R^{II} の間の剛体変換パラメータ $\mathbf{T}_i$ をICPアルゴリズムによって求める。
3. $\mathbf{T}_i$ によって R^I を変換した画像と R^{II} との自乗誤差の中央値 (メディアン) $\mathrm{MS}(\mathbf{T}_i)$ を計算する。
4. ステップ1〜3を N_T 回繰り返す ($i = 1 \sim N_T$)。
5. N_T 回の試行から、最小の変換誤差を与える変換 $\mathbf{T}_{i^*}$ を R^I から R^{II} への最適な変換と決定する ($\mathrm{MS}(\mathbf{T}_{i^*}) = \min_{1 \leq i \leq N_T} \mathrm{MS}(\mathbf{T}_i)$)。

3.2 ランダムサンプリングの効果

本方法では、ロバスト統計 [12] における外れ値検出の考え方を導入することにより、確率的な手法によって、前述の3次元的な隠蔽による対応点の喪失と距離画像において避けることのできない雑音に対して頑強な位置合わせを可能にしている。

距離画像中で雑音や隠蔽によって対応点が正しく求まらない点 (外れ値) の割合を ϵ とすると、1つの点を無作為に選んだとき、その点が外れ値でない確率は $1-\epsilon$ である。N_S 個の点をランダムサンプリングしたとき、選ばれた全ての点が外れ値でない確率は $(1-\epsilon)^{N_S}$ である。したがって、N_T 回の試行において、外れ値を全く含まないサンプルが少なくても1回は選ばれる確率 p は

$$p(\epsilon, N_S, N_T) = 1 - \{1 - (1-\epsilon)^{N_S}\}^{N_T} \qquad (2)$$

となる。これが正しく位置合わせが行われる確率に相当すると考えることができる。外れ値の存在確率が $0.1(\epsilon = 0.1)$ の場合に確率 $p(\epsilon, N_S, N_T)$ が描く曲面を図 4 に示す。ただし、横軸が N_S、縦軸が N_T、高さ方向が p を表す。この確率は、ϵ が 0 に近い場合、N_S が小さい場合、あるいは N_T が大きい場合に 1 に近づく。しかし現実には、N_S が小さいと ICP アルゴリズムによる推定の信頼度が落ちる、N_T が大きいと計算時間がかかるという性質があり、対象とする形状と独立に最適な N_S と N_T を決めるのは難しい。

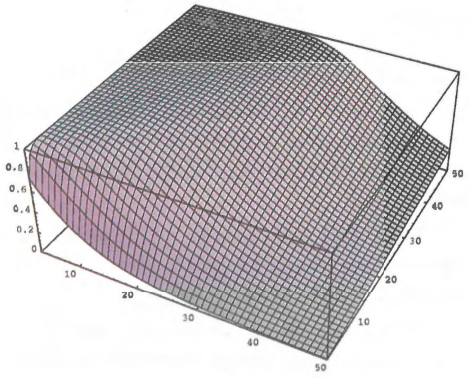


図 4: 確率 $p(0.1, N_S, N_T)$ が描く曲面

3.3 ICP アルゴリズム

ここでは、上記のアルゴリズムのステップ 2 における ICP アルゴリズムについて述べる。ICP アルゴリズム [11] は 2 つの 3 次元形状間の変換 (データ形状からモデル形状への変換) パラメータを推定するアルゴリズムであり、次の 2 つの手続きから構成される。

手続き (a): 推定されたパラメータで変換された各データ点を最も近いモデル点へ対応づけることによって、仮の対応づけを行う。

手続き (b): 仮の対応づけに基づいて、データ形状からモデル形状への剛体変換パラメータを推定する。

前述のアルゴリズムのステップ 2 においては、P_i^I をデータ形状、R^{II} をモデル形状として扱う。手続き (a) で求まる P_i^I に対する R^{II} 内での仮の対応点を P_i^{II} とすると、手続き (b) では

$$\sum \|P_i^{II} - \mathbf{T}(P_i^I)\|^2 \qquad (3)$$

を最小にする変換 $\mathbf{T}$ を推定する。ICP アルゴリズムでは、手続き (a) と (b) を変換が収束するまで繰り返す。変換に伴う対応点間の距離の自乗誤差

$$e = \sum \|P_i^I - P_i^{II}\|^2 \qquad (4)$$

は反復に関して単調減少することが証明されており、誤差 e の変化があるしきい値以下になった時点で反復を終了する。

ここでは、手続き (b) での変換パラメータの計算に単位 4 元数を用いた方法を採用する。単位 4 元数は、4 次のベクトル $\mathbf{q} = (q_0, q_1, q_2, q_3)$ で、$q_0 \geq 0$ かつ $q_0^2 + q_1^2 + q_2^2 + q_3^2 = 1$ を満たすものである。これを用いて式 (1) の回転行列 $\mathbf{R}$ は次のように表される。

$$\mathbf{R}(\mathbf{q}) = \begin{bmatrix} {q_0}^2 + {q_1}^2 - {q_2}^2 - {q_3}^2 & 2(q_1q_2 - q_0q_3) & 2(q_1q_3 + q_0q_2) \\ 2(q_1q_2 + q_0q_3) & {q_0}^2 + {q_2}^2 - {q_1}^2 - {q_3}^2 & 2(q_2q_3 - q_0q_1) \\ 2(q_1q_3 - q_0q_2) & 2(q_2q_3 + q_0q_1) & {q_0}^2 + {q_3}^2 - {q_1}^2 - {q_2}^2 \end{bmatrix} \qquad (5)$$

単位 4 元数表現を用いて、パラメータ推定問題は

$$\sum \|P_i^{II} - \mathbf{R}(\mathbf{q})P_i^I - \mathbf{t}\|^2 \qquad (6)$$

の最小化問題になる (具体的な解法については文献 [6] 参照)。

3.4 実験と考察

一方向計測型レーザレンジファインダで観測した複数視点距離画像の位置合わせ実験の結果を紹介する。図 5 はビーナス石膏像を 6 つの異なる視点から観測した距離画像 (陰影表示) である。本実験では、連続する 2 視点画像に対して提案手法を適用し、2 画像間の剛体運動パラメータを計算した。本実験において、位置合わせ後の連続 2 画像間の残差はレンジファインダの計測精度の範囲内に収まっており、アルゴリズムの有効性が確認された。図 6 は、図 5 の 6 枚の距離画像について求めた 5 組の剛体変換パラメータによって距離画像を合成したものである。

実験を通して、提案手法では、初期変換として恒等変換を採用しているため、2 つの視点間のずれが大きい場合 (特に、回転角度が大きい場合) には、ICP アルゴリズムによる式 (6) の最小化が局

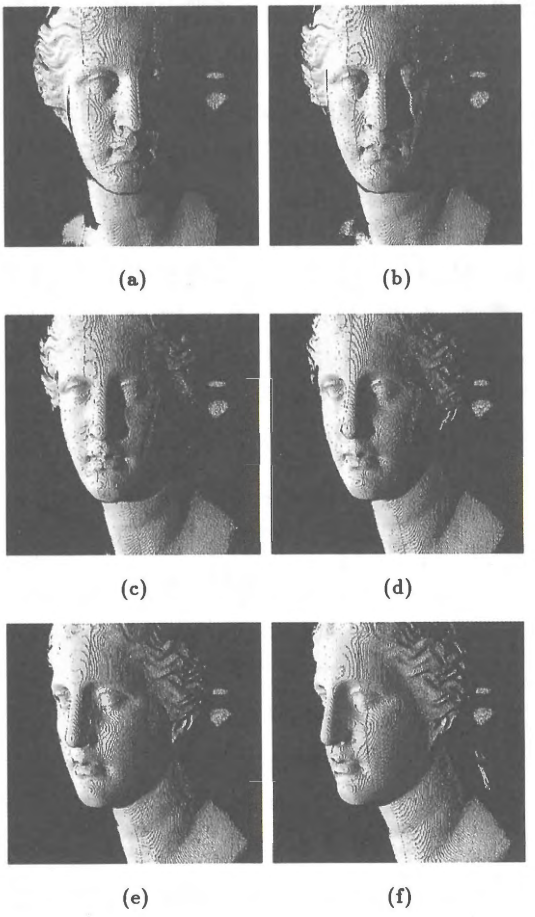


図 5: 6 つの視点から観測した距離画像 (陰影表示)

所解に陥り、正しい変換パラメータが得られないことが明らかになった。しかし、形状モデリングのための 3 次元計測では、視点間のお大雑把な位置関係が分かっていることが多く、また人手による指示も可能であるので、特に問題にはならないと思われる。

4 おわりに

本稿では、実物体の 3 次元計測に基づく形状モデリングにおける課題と問題点について考察するとともに、その中での重要な課題の 1 つである多視点距離画像の自動位置合わせ法について述べた。今後は、位置合わせされたデータを統合し三角形メッシュ表現による全表面の記述を生成する。また、全周計測型センサと一方向計測型センサの併用についても検討する。

謝辞 本研究の一部は文部省科研費補助金 (No.07207214) による。

図 6: 6 視点距離画像の合成結果 (観察方法は図 5(c) と同じ)

参考文献

[1] 井口, 佐藤: 三次元画像計測, 昭晃堂, 1990.

[2] Y. Chen and G. Medioni: "Object modeling by registration of multiple range images", *Image and Vision Computing*, Vol.10, No.3, pp.145-155, 1992.

[3] G. Turk and M. Levoy: "Zippered polygon meshes from range images", *Proc. SIGGRAPH '94*, pp.311-318, 1994.

[4] R. Bergevin, D. Laurendeau and D. Poussart: "Registering range views of multipart objects", *Computer Vision and Image Understanding*, Vol.61, No.1, pp.1-16, 1995.

[5] 藤木, 山本, 田村: "幾何形状モデル生成のための異種距離画像データの接合", 信学技報, PRU95-163, Nov. 1995.

[6] T. Masuda and N. Yokoya: "A robust method for registration and segmentation of multiple range images", *Computer Vision and Image Understanding*, Vol.61, No.3, pp.295-307, May 1995.

[7] 佐藤, 横矢: "測定手法の種類と基本原理 — 能動的手法を中心として —", 計測と制御, Vol.34, No.6, pp.435-439, Jun. 1995.

[8] 辰野由美子: "頭部全周計測距離データを用いた表情解析とその顔表情アニメーションへの応用", 奈良先端科学技術大学院大学修士論文, NAIST-IS-MT351056, Mar. 1995.

[9] 増田, P. ボランジャ: "多視点距離画像の空間的統合による全周計測", 計測と制御, Vol.34, No.6, pp.449-452, Jun. 1995.

[10] B.K.P. Horn: "Closed-form solution of absolute orientation using unit quaternions", *J. Opt. Soc. Am. A*, Vol.4, No.4, pp.629-642, Apr. 1987.

[11] P.J. Besl and N.D. Mckay: "A method for registration of 3-D shapes", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol.14, No.2, pp.239-256, 1992.

[12] P.J. Rousseeuw and A.M. Leroy: *Robust Regression and Outlier Detection*, Wiley, New York, 1987.

ハイパーメディア・コーパスの構築と言語教育への応用について

Hypermedia Corpus and Its Application to Language Education

上村隆一
Ryuichi UEMURA

福岡工業大学工学部（福岡市東区和白東3-30-1）

Fukuoka Institute of Technology
3-30-1, Wajirohigashi, Higashi-ku, Fukuoka-shi 811-02
E-mail: uemura@ipc.fit.ac.jp

あらまし：近年、言語研究の分野においても、統計的な分析や資料整理の道具としてのコンピュータ利用が定着してきた。欧米では、大規模な言語データベース（以下コーパス）を用いて、文脈を伴った任意の語彙、語法の出現頻度を調査したり、文学作品、新聞・雑誌記事など特定の分野の言語資料を独自に電子化テキストとして作成し、言語学的な分析結果とともに公開する例が増加している。さらに、最近数年間にインターネットが急成長し、わが国でもその具体的な利用方法が注目されるようになるにつれて、日本から世界へ向けての情報発信の必要性が急速に高まってきた。特に、日本語・日本文化等に関するデータベースを作成し、インターネットを通じて情報を提供することは、わが国に対する国際社会の理解を助け、同時にわれわれ自身が自国の言語・文化を理解し、評価する際の基礎資料としても重要な意味をもつと思われる。

本稿では、著者が研究代表者として開発を進めている日本語会話コーパスの構築プロジェクト（平成7年度文部省科学研究費補助重点領域研究「人文科学とコンピュータ」公募研究（課題番号07207124））について、とくに外国人向け日本語教育への応用の観点から研究成果を中間報告する。

Summary: The object of our joint project started in 1991 is to create an original hyper-media corpus of spoken Japanese. We have collected 'live' spoken data from actual conversation between experts of teaching Japanese as a foreign/second language and native / non-native speakers of Japanese, based upon a testing format known as OPI (Oral Proficiency Interview). The whole contents of experimentation recorded on high precision video tapes and magneto-optical disks were converted to digital video and/or sound data files. The sample version of our corpus is now being transferred to WWW server on our campus (URL: http://corpus. fit.ac.jp) with HTML-tagged texts and is expected to be available over the Internet. This project is authorized and sponsored by Japanese Ministry of Education as one of the 1995 research programs on priority areas (Project No. 07207124). The author is in charge of the research organization (consisting of 5 members including an advisory professor affiliated with U.S. institution) as a whole. It is expected to be completed in 1998.

キーワード：ハイパーメディア、コーパス、日本語、会話分析、インターネット

Keywords: *hypermedia, corpus, Japanese, conversational analysis, Internet*

1．研究経過

本研究は、日本語母国語話者(以下NS)と非母国語話者(以下NNS)の現実発話に含まれる言い誤りの類型を比較分析することを目的として、1991年度より開始した試験研究の延長線上にある。研究当初から、分析対象となる一次言語資料の絶対量不足を痛感したため、われわれはまず、インタビュー実験形式によ

る会話データの収集と、それに基づくコーパスの構築作業から開始することにした。

平成７年度は主としてNNSのデータ収集を行うことになり、７月に東京都内の民間日本語学校と国際基督教大学、10月に米国のプリンストン大学においてそれぞれインタビュー実験を実施し、約70名分の会話データを得た。（被験者の内訳は**表１**のとおり。）

表１　インタビュー実験被験者の内訳

国籍	米国 *26*　韓国 *25*　中国 *5*　日本 *3* ロシア・オーストラリア 各 *2* ドイツ・オーストリア・タイ 各 *1*
性別	男 *29*　女 *37*
年齢層	20代 *58*　30代 *7*　40代以上 *1*

日本語学校の被験者は１名を除いて全員が韓国籍で、年令は全員20歳代であった。大半が就学生で、日本語学習歴、日本滞在期間、生活環境も似通っている極めて均質のグループであった。次に、国際基督教大学では夏期日本語講座の受講生を被験者としたため、被験者の国籍は多岐にわたり、母語や言語背景も多様であった。この傾向はプリンストン大学についても同様で、欧米系とアジア系がほぼ同数であった。年齢層、男女比なども適当に分散していたと思われる。

会話データの収録に際しては、事前にこの研究の趣旨とリスクに関して説明を行い、第１回目は口頭での承諾、２回目以降は同意書への署名いう方法をとった。確認事項は次の通りである。

1. この会話データ収録は15～30分の日本語によるインタビューである。
2. このインタビューは録音、録画される。
3. このデータは研究目的以外には使用されない。
4. このデータはインターネット上で公開される。
5. 4に関しては現時点で予測不可能な犯罪行為を含むリスクについては、研究者および実験者は一切責任を負わない。

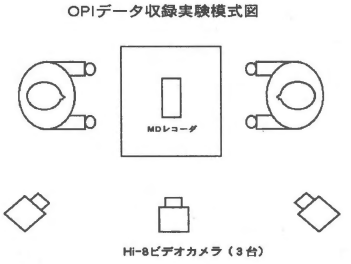
個人データのインターネット上での公開に伴うリスクについては、現時点では不透明な部分が多く、法律的な安全策も未だできていない。このような状況の下では、今後も事前に被験者に十分な説明を行うことは研究者のすべき最低限の配慮であろう。その上で被験者の了解を明確な形で得ておくことが研究計画の成功には不可欠であろうと考える。(注1)

２．実験方法とデータ処理

会話データの収録形式としては、NNSの会話能力判定方法として知られるOPI(Oral Proficiency Interview)を採用した。これは会話モードとロールプレイモードの二つの要素によって構成されるインタビュー形式のテストであって、本来は外国語学習者の口頭表現能力を総合的に評価することを目的としたものである。従って、他の多くの会話体データの収集方法と異なり、実験者はできるだけ被験者に自発的に多く話をさせるように配慮した。

データ記録媒体については、画像データを高画質8ミリビデオテープに、音声データを光磁気ディスク（MD,ディジタル録音専用メディア）にそれぞれ収録した。収録時間は被験者１人につき20-30分程度である。（**図１**参照）

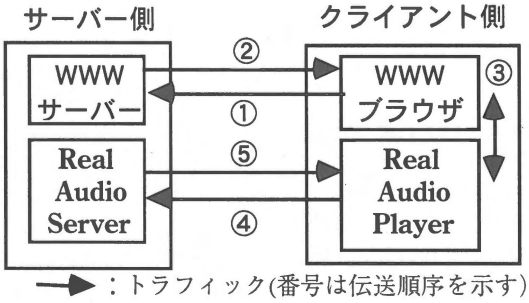
図１　実験機器類セッティング



現在までに、音声データからテキストデータへの書き起こし作業の一部（約10名分）が完成し、圧縮した音声データとともにインターネットのWWWサーバー(http://corpus.fit.ac.jp)およびFTPサーバー上で順次公開している。

特に、音声データの公開方法に関しては、最近インターネット上でのオン・デマンド型音声サーバー技術として注目されているReal Audio Serverをいち早く導入し、本年９月から実験的にインタビューの内容の一部（フリートーキングとロールプレイについて各々約20人分）を転写テキストと一体化した形で提供している。（**図２**参照）

図2　オン・デマンド音声サーバ概念図



ディジタル変換された画像データ（動画）は被験者1人当り200～300MB（圧縮ファイル）に達するので、現時点ではインターネット上では公開せず、追記型光ディスク（CD-R）に保存している。また、このコーパス作成作業と同時に、繋ぎ語や代名詞類の言語学的分析を同時に進めており、それらの研究成果の一部は、情報処理関係の学会で発表し、さらに言語学関連の研究論文集等で公刊している。

3．本研究の特色

本研究プロジェクトにおいて実現をめざす日本語会話コーパスは、従来のテキスト・データベースと異なり、文字・画像（動画）・音声データを統一された使用環境(GUI)で同時に利用可能にする、いわゆるマルチメディア型データベースである。われわれのコーパスの主な特徴は、

①分析対象が話し言葉である場合、文字化しにくい音調、強勢、ポーズなどの諸特徴をそのまま音声情報の形で提供できる。その結果、テキストデータに特殊な音調記号等を付与する必要がなくなる。

②画像（動画）データを提供することにより、非言語情報（身ぶり、手ぶり、顔の表情など）をテキストと同時に利用できるようになり、会話の状況、話者の特徴、周囲の雰囲気などを分析の手がかりにすることができる。

③画像・音声データ自身をディジタル化することにより、ランダム・アクセスが可能になるので、テキスト検索に加えて、画像・音声データの検索等を容易に実行できる、などである。

上記のコーパスの特性を十分に活用することにより、これまで研究対象から事実上除外されてきた非言語情報を含む会話分析が可能になる。また、従来のコーパスのように、転写記号等に関する専門知識を必要としないので、研究用資料としてだけでなく、外国人向けの日本語教育の資料・教材としても十分に活用できるであろうと思われる。

4．日本語教育からみた本研究の意義

本研究プロジェクトが日本語教育の現場に与えるインパクトは限りない。その一つを挙げれば、本年度実施した2回の会話データをみるだけでも、第二言語の習得に関わる諸要因についての多くの示唆を読み取ることができるだろう。（転写テキストの具体例は本稿末尾の**サンプル・データ**を参照。）さらに数多くのデータが得られれば、個々の要因についての研究も可能になることが期待される。さらに会話データの中のロールプレイモードのヴァリエーションは、社会的文化的に適切な言語使用を教師がどのように指導すべきなのかという問題に対する解決の糸口を与えれくれる。次年度以降に予定している、NSの会話データ収集が進めば、現在多くの日本語教師が持っている社会的・文化的に適切な言語使用のイメージの見直しを迫られることになるかもしれない。(注2)

5．今後の研究計画

1996年度は本年度に引き続いて、会話データの収集、テキスト転写作業を行う。また、情報処理、日本語教育関連の学会において、本研究の第2次中間報告を発表する。実験データの収録作業については、同年度内に次の2点を実行する。

1）米国在住の日本人研究協力者（牧野成一プリンストン大学教授）の指導の下に、国内と同一の実験条件で現地在住のNS,NNSそれぞれの会話データを収録。

2）研究分担者と国内2大学の協力の下に、主としてNSの会話データを収録。

音声データについては、オーディオテープ(DATを含む）やMDに録音した場合、データベース作成時に膨大な量の変換作業を必要とするため、別途ノートブック型パソコンと16ビットサンプリング可能な音声入力・編集ソフトを用いて直接ディジタル録音を試みる。このことにより、書き起こし（転写）作業から音声データベース部分の構築に至る作業工程を大幅に短縮することができる。

実験で収録したデータは、日本語教育関係の研究分担者が大学院生等に委託した形で転写作業を行った後、責任を持って校閲する。転写作業にあたっては、16ビットサンプリングによりパソコン上で直接ディジタル録音したファイルから、ディジタル録音・編集機を用いて適宜分割・再編集した音声データを用いる。

1997年度は研究班をコーパス作成班とコーパス分析班に分けて、適宜協力しながら研究計画を遂行する。まず、「作成班」は前年度内に収集した音声データと転写テキストデータについて、任意の文字列から当該個所の音声データを検索するプログラムの開発を試みる。特に、時系列データを扱う検索用言語としては、HyTime処理系等を参考にしながら、SGML／DTDの拡張を考える。同時に、ディジタル動画データ(MPEG形式）の検索方法も検討する。なお、コーパス本体は上記のインターネット上で提供するネットワーク版と、CD-ROMで提供するスタンドアロン版の2種類を作成する予定である。また、CD-ROM版

の検索ソフトウェアの更新及び追加情報の提供等はすべてインターネットを利用し、国内外の言語研究者および日本語教育関係者に公開する。　次に、「分析班」は言語学と日本語教育の立場から、NSとNNSの発話内容を比較検討し、段落形成、連接名詞、代名詞類、繋ぎ語等の各トピックについて会話分析を試みる。データの蓄積が一定レベルに達した後、当初の研究目的であった「言い誤り」の類型に関する分析を開始する。年度内に本研究プロジェクトの最終報告を行うが、論文とデータは通常の印刷物に加えて、電子化テキストの形式で作成し、インターネット上のFTPサーバーからも利用可能にする。

6．おわりに

コーパスを用いた言語分析は、欧米ではすでに確立した研究手法であり、分析対象も文語体のテキストにとどまらず、会話内容を転写したデータから成る口語体テキストにまで及んでいる。日本でも、最近コーパスの重要性が認識され、欧米のLOB,London-Lund, Brown, BNC等の大規模コーパスを利用した英語の文法・語法などの研究例が増加しているが、日本語コーパスについては、いまだ欧米ほど確立した大規模なものがなく、まとまった研究成果も報告されていない。従って、われわれの研究は、日米間にまたがる大規模な日本語会話コーパスの構築プロジェクトとしては前例のないものであり、インターネット環境およびマルチメディア型データベースを使用した言語研究としても、先駆的な役割を果たすもの、といえる。

*注）1,2ともに共同研究者の村野良子氏（国際基督教大学・日本語教育）の指摘による。

参考文献

Armstrong, S.(Ed.) : *Using Large Corpora.* MIT Press. (1994).

Buck, K. (Ed.) *The ACTFL Oral Proficiency Interview Tester Training Manual.* The American Council on the Teaching of Foreign Languages. (1989).

Uemura, R. : "Hypermedia Corpus Project of Japanese Conversation - Interim Report," *Language and Information Processing* **6**, pp. 1-5. Fukuoka Institute of Technology. (1995).

上村隆一「コーパスによる日本語会話分析ー指示詞の使用について」小泉保教授古稀記念論文集『言語学の展望』 pp.93-105. 大学書林. (印刷中).

（参考）サンプル・データ

（1：は実験者、2：は被験者を示す。カッコ内は相づち、／は長いポーズを表す。）

A. 会話モード（フリー・トーキング）

1： ああーそうですか。（2：ええ）最近の日本のもん、ま、色々問題を抱えていると（2：そうですね）思いますけれどもね、そのー、どうですか。まあ、まよく聞かれることだと思うんですけど、オウム真理教、あたりがね、そのどうして日本の社会に、ああいう今までにないようなね、（2：ええ）規模の、おー、ほんとに、なんて言うか、極悪、（2：うん）非道なね、（2：うん）ああいう、一つの、そのグループが出来たっていうのは、どういうことでしょうね。（2：すー）なんか歴史的な観察、ありますか。
2： ええ。まあ歴史的な観察は／まあ歴史にも、そんなグループが、ないんですけど、でも似てるグループがあると思いますよ。（1：うん）日本の社会は、アメリカ人の考えは、日本の社会は平和で、えーといつも、目上の人に従う社会ですね。でも、ま、歴史の立場からみるとー、ま、色々な、えーとほんとに、強く、反対した人が、いましたねえ。
1： しかしだけど、オウム真理教とは（2：オウム真理教と）ちょっと、比較するとちょっとまずいんじゃ（2：ちょっと違いますねえ）ないんですか。

B．ロールプレイ・モード

2： あのー／。あのー／。うん。
1： うん、だあれー、お姉ちゃん？
2： あ、あのー、／うん、／あのー、あなたの名前てー、／山田ー／ていうのかなあ？
1： うん山田ーだけど。ぼく山田一郎。
2： あ、そう。（1：うん）じゃあ、あの、お父さん、／ちょっといらっしゃるー？
1： うん、お父ちゃんいないよ。
2： ああ、あ、そう。じゃあ、あのー、うんー、どうしようかなあ。わたしー、お父さんにちょっとー、あのー、お父さんいらしたら、ちょっとー、うん、お会いしたいと思うけどー、お父さん何時ごろ、お帰りになるの？
1： お父ちゃんねー、なんか、明日帰ってくるっていってたよ。
2： あ、そっかー。じゃあー、お母さんは？いらっしゃー、らないの？
1： 今ねえ、お母ちゃんねえ、今、買い物行ったみたい。
2： そっか。じゃあ、あのー、ここでちょっと、待っててもいいかなあ？
1： うーん、たぶん、うん待っていていよ。うん、なんか、一緒に遊んでくれる？
2： あ、いいよ。
1： あの今パソコンやってんだけど。
2： あ、じゃあ、あの、パソコンやりながら、ちょっと、お母さん、待つわ。

Hypermedia Corpus of Spoken Japanese

PREVIOUS

MENU

 T.K.(Japanese)  H.L.(American)

 K.S.(Korean)

 C.Y.(Korean)  B.P.(American)

 M.S.(Austrian)

 D.N.(German)  M.I.(Russian)

Click on any picture icon to access each data.

Please feel free to send your comments to:
uemura@ipc.fit.ac.jp

CORPUS SAMPLE (Interview Setting)

MENU

Click on any picture icon to hear each sound data.

Free Talking

Role Play

Free Talking

1：はい、こんにちは。
2：こんにちは。
1：あの、私の名前は牧野と申しますが。
2：牧野先生ですか。
1：はい、牧野です。そちらは、お名前は。
2：ブライアン・プラットです。
1：あ、じゃ、ま、ブライアンさんて呼んでいいですか。
2：う、ブライアンでいいです。
1：ああ、そうですか。
2：はい。
1：ブライアンさんは、アメリカ人ですか。
2：ええ、そうです。
1：ああ、どこ、アメリカのどこからいらっしゃったんですか。
2：えと、出身はウエスト・バージニア州ですが、（1：あ）あの、今イリノイ大学の大学院生です。
1：あ、そうですか。ああ、私はイリノイ大学で前教えていたんですけれども。
2：ああ、わかりますよ。
1：あ、知ってますか？ああ、そうですか。ああ、そうですか。ウエスト・バージニアっていうのは小さな州（2：そうですね）ですよね。ええ。私はあすこに一度行ったことがあるんですけれども。
2：ああ、そうですか。ありますか。
1：うーん、まあ、行ったことがない人にどういうふうな州だって説明できますかね。
2：そうですね。
1：ちょっと説明してみてください。
2：まあ、アメリカの文化からちょっと離れていると思いますね。（1：うん）アメリ、ウエスト・バージニア州で。（1：うん）ええと、まあ、丘がたくさんあるし、（1：うん）ええと、離れている町が多いんですね。（1：うん）だから、その、離れている町に住んでいる人は、（1：うん）まあ、世界のことか、アメリカのこと、ま、よく知らない人が多いんですね。だから、えーと、ま、外国人に会ったことがない人もいるし、（1：うーん）えーと、まあ、世界のことを全然構わない人もいるし、ちょっと、も、面白いところだと思いますね。

Role Play

「歌物語」語彙の数量的分析と研究

Quantitative Research and Analysis on the lexicon of "Uta Monogatari"

西端 幸雄
Yukio NISHIHATA

大阪樟蔭女子大学学芸学部国文学科
577 東大阪市菱屋西4-2-26
4-2-26, Hishiyanishi,
Higashiosaka-City, Osaka, 577

キーワード：歌物語，コード，数量的分析

Keywords　: Uta Monogatari, Code, Quantitative Analysis

あらまし：「歌物語」とは、平安時代初期に成立した『伊勢物語』『平中物語』『大和物語』の３作品の総称である。これらの作品は、物語の展開に関係ある歌を適宜配することにより、る和歌とが融和した形で物語全体の展開を円滑にしているという点で共通している。また、作品全体の構成は、男女の恋を中心テーマとした短編の集まりとして成り立っている点で共通し、また、各作品相互に類似した内容を持つ短編物語が少なからずあるという点でも共通している。こうした共通点を持つという点からも、「歌物語」と総称される由縁があろう。

ところで、これらの作品についての研究は、各作品個別には、国文学・国語学それぞれの分野において、長年行われてきた。ただ、「歌物語」と総称されるにしては、これら３作品を総合的に捉えようとした研究は、国文学・国語学の両分野ともに、少ないのが現状である。しかし、本研究の準備段階における試験的な研究においては、「歌物語」３作品の使用語彙を相互に比較してみると、『伊勢物語』『大和物語』の２作品は、極めて類似した語彙を使用している傾向が強いのに対して、『平中物語』だけが、使用語彙としては、かなり異質な側面を示している点が明らかになった。

そこで、本研究では、これら３作品の使用語彙に対して相互に数量的・統計的処理を施すことにより、それぞれの作品の使用語彙の特色を明らかにすることを目的とする。

その目的のうち、次の４項目について、重点的に研究を行う。

ア　各作品間の語彙的な面での類似点と相違点を明らかにする

イ　それぞれの作品間に有る類似した短編物語の中の使用語彙を数量的・統計的に比較し、各短編物語の語彙的な面での類似点と相違点を明らかにする。

ウ　それぞれの作品内の散文箇所での使用語彙と和歌での使用語彙の特色を明らかにする。

エ　平安時代に成立した他の仮名文学作品（『源氏物語』『枕草子』等）と比較することにより、「歌物語」３作品の使用語彙の特色をより一層明らかにする。

Summary:Uta Monogatari is the collective title of 3 works, "Ise Monogatari", "Heichu Monogatari" and "Yamato Monogatari", written at the beginning of the Heian period. By properly arranging the poems which relate to the development of the story, They are similar in that they integrate story portions written in prose with waka poetry verse po

rtions to smoothly develop the entire story. Other commonalities include the fact that the entire works are structured as collections of short pieces on the theme of male-female love, and that there are at least a few similar themes which appear in all of the short tales. It was on the basis of these commonalities that these works came to be known collectively as Uta Monogatari.

Although research on these individual works has been conducted in the fields of Japanese language and literature for many years, at present there are few comprehensive studies in these fields of these three works as a whole. But in the experimental research conducted in preparation for this project, an overall comparison of the lexical items employed in the works revealed a strong resemblance between the items used in "Ise Monogatari" and "Yamato Monogatari", but differences between these two and "Heichu Monogatari".

The purpose of this research is to clarify the characteristics of the lexical items used in each work through quantitative and statistical analyses. The research is based on the four points outlined below.

(a) to determine the similarities and differences among the lexical items used in the three works

(b) to quantitatively and statistically compare the lexical items used in those short tales which are common in all three works in an effort to determine the lexical similarities and differences among the different tales

(c) to compare the prose and verse portions in each work to determine the characteristics of the lexical items in the two

(d) to determine the lexical characteristics of the three works in Uta Monogatari through a comparison with other works (Genji Monogatari, Makura Soˆshi) in the kana literature which developed during the Heian era.

1 「歌物語」について

「歌物語」とは、平安時代初期に成立した『伊勢物語』『平中物語』『大和物語』の３作品の総称である。

これらの作品は、物語の展開に関係ある和歌を適宜配することにより、散文である物語部分と韻文である和歌とが融和した形で物語全体の展開を円滑にしているという点で共通している。また、作品全体の構成は、男女の恋を中心テーマとした短編の集まりとして成り立っている点で共通し、また、各作品相互に類似した内容を持つ短編物語が少なからずあるという点でも共通している。こうした共通点を持つという点からも、「歌物語」と総称される由縁があろう。

ところで、これらの作品についての研究は、各作品個別には、国文学・国語学それぞれの分野において、長年行われてきた。ただ、「歌物語」と総称されるにしては、これら３作品を総合的に捉えようとした研究は、国文学・国語学の両分野ともに、少ないのが現状である。しかし、本研究の準備段階における試験的な研究においては、「歌物語」３作品の使用語彙を相互に比較してみると、『伊勢物語』『大和物語』の２作品は、極めて類似した語彙を使用している傾向が強いのに対して、『平中物語』だけが、使用語彙としては、かなり異質な側面を示している点が明らかになった。

そこで、これら「歌物語」３作品の使用語彙のうちの自立語を数量的に比較することによって、それぞれの作品の語彙の特徴を明らかにすると共に、これら３作品を「歌物語」と称して、あたかも等質な作品のように取り扱ってきた、これまでの国文学・国語学の分野における姿勢に対しての疑義も提示してみたい。

なお、本稿で用いる「歌物語」３作品の語彙データについては、平成６年度科学研究費補助金「研究成果公開促進費」の交付により刊行した『歌物語総合語彙索引』（西端幸雄・木村雅則共編　勉誠社刊）に付載したフロッピー内の語彙データベース（約48000語）を使用することとする。

II　語のコード化

ところで、作品に使用されている語彙を比較する場合、その語彙そのままで比較を行ったとしても、語形の違いや使用されている語種の違い程度が明らかになるだけで、厳密な比較は行えない。そのため、品詞や意味といった語の性格を表す情報を語に対して付加しなければならない。

ということで、筆者は、これまで、古典文学作品に使用されている語に対して、以下に詳述するようなコードを付けることにより、それぞれの語の性格を明示できるようにしてきた。現在、平安時代を中心にした散文・和歌作品１９作品での使用語彙に対して、コード化を完了している。

その語のコード化の拠り所とした資料は、『分類語彙表』（国語研究所編・秀英出版刊）である。この『分類語彙表』は、現代語を文法的・意味的性格に分類しているため、古典語を適用する場合、様々な問題点はあったが、『分類語彙表』の分類方法に照らし合わせて、古典語の分類についても、できる限り厳密な作業を行なった。

いま、そのコード体系の一端を摘記すると、下記の通りである。ただし、この基準のうち、現代語と古語との性格の違いによって、『分類語彙表』では取り扱われていない「固有名詞（人名）」「枕詞」の分類が問題となったが、本稿においては、この「固有名詞（人名）」を大分類の＜１＞類、小分類の＜-2＞に、「枕詞」を、便宜的に大分類の＜４＞類に入れることにした。

大分類（１０００の位・文法的性格）
- 1　体の類　名詞
- 2　用の類　動詞
- 3　相の類　形容詞・形容動詞・連体詞・一部の副詞
- 4　その他　接続詞・感動詞・一部の副詞・枕詞

小分類（１００の位・意味的性格）
- -1　抽象的関係（人間や自然のあり方のわく組み）
- -2　人間活動の主体
- -3　人間活動－－精神および行為
- -4　人間活動の生産物－－結果および用具
- -5　自然－－自然物および自然現象

なお、『分類語彙表』では、各語句を、さらに細かく意味別に分類し、全体として、３～４桁のコード番号で、それぞれの語句の文法的・意味的性格を表しているが、本稿においては、小分類以下の位の分類（１０の位以下）は、特に断わらない限り、取り扱わないこととする。また、大分類の＜４＞類については、『分類語彙表』では、＜４１＞＜４３＞といった小分類を施しているが、本稿においては、その小分類は施さず、大分類のみにとどめた。

さて、上記の基準によって、それぞれの語に対して、文法的・意味的性格付けを行なったのであるが、それぞれの分類に該当する具体例を示すと、以下のようになる。

・文法的・意味的性格付けの具体例
- 11　あかつき（暁）、あき（秋）
- 12　あま（海人）、あかし（明石）
- 13　あきのこころ（秋心）、あきのわかれ（秋別）
- 14　あかぢのにしき（赤地錦）、あきた（秋田）
- 15　あかつきつゆ（暁露）、あきやま（秋山）
- 21　あかす（明）、あぐ（上）
- 23　あかしくらす（明暮）、あく（飽）
- 25　ある（荒）、いく（生）
- 31　あさし（浅）、あし（悪）
- 33　あさまし（浅）、あぢきなし（味気無）
- 35　あかし（明・赤）、きよし（清）
- 4　あかねさす（茜）、あな（感動）

以上のような語のコード化を通して、これまで以下のような点を考察してきたのであるが、その中で、このコード化したデータをもとに当

該作品の使用語彙を見てみると、その性格をかなり鮮明に捉える得るということも実証できた。

- 「八代集和歌語彙の性格―その意味的性格と語彙史的位置づけを探る」（樟蔭国文学29・1992年）
- 「和歌と語彙―語彙の変遷と八代集和歌の変遷」（日本語研究センター報告１・1992年）
- 「和歌と散文の使用語彙の比較」（日本語研究センター報告２・1993年）
- 「語彙史の立場から見た『拾遺和歌集』――使用語句の性格を統計的に見る」（国語語彙史の研究14・1994年）
- 「古典文学作品の使用語彙の性格－『古典対照語い表』データのコード化を通して」（樟蔭国文学31・1994年）

III 「歌物語」３作品間の比較

では、実際に「歌物語」３作品の使用語彙に対してコードを与えたデータをもとに３作品の使用語彙の性格を比較してみることとする。

ここで取り扱う語は、先にも断ったが、付属語を除いた自立語だけとする。この自立語だけに限定する意図は、作者の物の見方・考え方が直接に語として表現されているのが自立語であるということからである。

後掲の表１【作品別使用語彙の性格＜異語数＞】には、「歌物語」３作品の全自立語をコード別に整理し、さらに、それを散文と和歌とに分類したものを掲げた。

まず、この表１の中の《全体》項目の＜体の類の合計＞＜用の類の合計＞＜相の類の合計＞の欄に注目すると、『伊勢物語』『大和物語』に占める＜体の類＞の割合と比較して、『平中物語』に占める＜体の類＞の割合がきわめて低く、逆に両者の＜用の類＞の割合を見ると、『平中物語』の割合が高いのに対して、『伊勢物語』『大和物語』の割合が低いということが分かる。また、両者の＜相の類＞の割合を見ると、『伊勢物語』『平中物語』『大和物語』それぞれに占める割合が、わずかずつ異なっているということも分かる。

この一点だけからしても、これら「歌物語」３作品を等質な作品として、同列に取り扱うことを再考する必要があるということが分かる。

さらに、「歌物語」３作品の使用語彙の性格の違いを明らかにするため、《全体》項目の＜体の類＞＜用の類＞＜相の類＞の内部の細かな点について比較を行ってみると、まず、次のような相違点のあることが分かる。

① ＜１１＞『伊勢物語』に占める割合が高く、『平中物語』『大和物語』は、ほぼ同じ割合である。
② ＜１２＞『大和物語』に占める割合がきわめて高いのに対して、『平中物語』に占める割合がきわめて低い。
③ ＜１３＞３作品に占める割合は、ほぼ同じである。
④ ＜１４＞『伊勢物語』『大和物語』に占める割合が、ほぼ同じであるのに対して、『平中物語』に占める割合が低い。
⑤ ＜１５＞３作品に占める割合は、ほぼ同じである。
⑥ ＜２１＞『伊勢物語』『大和物語』に占める割合が、ほぼ同じであるのに対して、『平中物語』に占める割合が高い。
⑦ ＜２３＞『伊勢物語』『大和物語』に占める割合が、ほぼ同じであるのに対して、『平中物語』に占める割合が高い。
⑧ ＜２５＞３作品に占める割合は、ほぼ同じである。
⑨ ＜３１＞『伊勢物語』『平中物語』に占める割合が、ほぼ同じであるのに対して、『大和物語』に占める割合が低い。
⑩ ＜３３＞３作品に占める割合は、ほぼ同じである。
⑪ ＜３５＞３作品に占める割合は、ほぼ同じである。

以上に掲げた点から、「歌物語」３作品の使用語彙の性格を、総括的に見れば、『伊勢物語』『大和物語』の使用語彙の性格に比べ、『平中物語』の使用語彙は、特に、＜１２：体の類・人間活動の主体＞＜１４：体の類・人間活動の生産物――結果および用具＞＜２１：用の類・抽象的関係（人間や自然のあり方のわく組み）

＞＜２３：用の類・人間活動－－精神および行為＞において、その違いが顕著に現れており、異質な面を持っていると言えよう。

また、このような「歌物語」３作品の使用語彙の性格が、散文における使用語彙が関わって生じているのか、和歌における使用語彙の性格が関わって生じているのかを検討してみると、後掲の表１の中の《散文》項目と《和歌》項目の＜体の類の合計＞＜用の類の合計＞＜相の類の合計＞の欄に注目すると、総括的に見た場合、『平中物語』において、＜体の類＞の占める割合が低く、＜用の類＞の占める割合が低いという傾向は、和歌にも『伊勢物語』『大和物語』とは異質な面が見られるが、それ以上に散文の方に比較的顕著に現れていることが分かる。つまり、『平中物語』の使用語彙は、同じ「歌物語」と称される『伊勢物語』『大和物語』とは異なり、これら３作品を「歌物語」として同列に扱うことは、少なくとも、国語学的な観点から見た場合、危険であると言えよう。

Ⅳ　仮名散文作品との比較

前項において、「歌物語」３作品のうち、『伊勢物語』『大和物語』に比べ、『平中物語』の使用語彙の性格が異質である点を指摘したが、では、これら「歌物語」３作品、特に、使用語彙が特異だとした『平中物語』は、平安時代、およびそれに近接した時代に成立した他の仮名散文作品と、使用語彙の性格の面でどのような関係にあるのかを検討してみることとする。

ここで取り扱う作品は、後掲の表２に掲げたように、「歌物語」と同時代の平安時代に成立した作品として、『竹取物語』から『大鏡』、その後に続く時代の作品として、『方丈記』『徒然草』、また、参考として、奈良時代成立の『万葉集』を取り上げた。（作品名の項目に『伊勢』『天伊勢』とあるのは、『校本伊勢物語』と『天福本伊勢物語』を表す。なお、本稿で、取り扱っている『伊勢物語』は、『天福本伊勢物語』である。）

表２に掲げた作品のうち、平安時代に成立した主だった作品についての数量的分析については、すでに、「古典文学作品の使用語彙の性格－『古典対照語い表』データのコード化を通して」（樟蔭国文学31･1994年）において行ったが、その中で、物語性の強い『源氏物語』については、次のようにまとめた。

『源氏物語』は、『竹取物語』が示す割合（体の類　４３．６％、用の類　４０．３％、相の類　１４．１％）と関連づけて考えると、純粋な物語作品の特色を如実に表しているものと言えよう。これらの作品に『夜半の寝覚』『浜松中納言物語』を加えて検討すると、それぞれの作品に占める名詞と動詞の割合は、以下の通りである。

	体の類	用の類
竹取物語	４３．６	４０．３
源氏物語	４２．４	４４．７
夜半の寝覚	３５．５	４２．５
浜松中納言物語	３７．０	４２．１

上に掲げた割合が、『源氏物語』と『夜半の寝覚』『浜松中納言物語』との間で隔たりを見せているのは、単位の取り方の違いに起因しているものとも考えられる。（中略）名詞の占める割合より動詞の占める割合の方が高いという点では、上記の三作品における使用語彙の性格は、共通していると言えよう。ということは、本稿において扱っている『源氏物語』や『竹取物語』が示す体の類・用の類の割合は、物語作品共通の性格を表しているものと考えて間違いないだろう。

さらに、本稿で問題にしている「歌物語」３作品のうち、『伊勢物語』については、次のようにまとめた。

これらの作品中で唯一歌物語のジャンルに属する『伊勢物語』に目を移すと、以下に示すような割合を示しているが、

	体の類	用の類	相の類
伊勢物語	54.4	31.9	12.4
古今集	54.8	32.7	11.1
後撰集	52.9	34.7	10.9

土左日記　54.6　30.3　13.6

この『伊勢物語』の示す割合は、上に掲げたように、和歌集の『古今集』『後撰集』と近似しており、さらに、ジャンルは異なるものの、所収和歌の比較的多い『土左日記』とも近似している。（中略）『伊勢物語』の体の類・用の類・相の類の割合が、『古今集』や『後撰集』と近似し、中でも、『古今集』とは、ほとんど同じ割合を示していることは、興味深い点である。この『伊勢物語』と『古今集』の関係については、『伊勢物語』の成立を、『古今集』所収の業平歌との関連でとらえようとする見方があり、室伏信助氏の「勅撰たる古今がなければ生まれ得ない物語が伊勢物語ではなかったか」といった意見が示唆に富むものである。

以上の点から、まず、『平中物語』に注目すると、この『平中物語』に占める＜体の類＞の割合43.8%、＜用の類＞の割合40.8%というのは、表２の中の『竹取物語』におけるそれぞれの割合＜体の類＞の割合43.6%、＜用の類＞の割合40.3%と近似していることが分かる。そして、前記引用中で述べたように、この『竹取物語』は、『源氏物語』と近い、極めて物語性の強い作品である。ということからすると、『平中物語』は、物語内部の形式としては、物語と和歌とを適宜配して、物語の展開をはかっているという点では、「歌物語」と称されようが、その作品の使用語彙の性格からすれば、『竹取物語』や『源氏物語』といった純粋な物語作品に極めて近いものであると言えよう。

一方、『伊勢物語』『大和物語』の方は、全体的に見た場合、＜体の類＞の割合55.0%と55.9%、＜用の類＞の割合30.7%と30.8%と、両作品の使用語彙の性格が類似している点、また、『伊勢物語』の使用語彙の性格は、上記引用中にも述べたように、『古今集』『後撰集』といった和歌集や和歌を多く配して日記を進行させている『土左日記』と近いという点で、和歌に比重を重く置いた、文字通り「歌物語」と呼ばれるにふさわしい作品であると言えよう。

Ⅴ　最後に

以上、「歌物語」３作品の使用語彙の性格から、それぞれの作品の特色を明らかにし、また、特に、『平中物語』については、「歌物語」というジャンルに置いて、『伊勢物語』『大和物語』と同列に取り扱ってよいのかという疑義を提示した。

ただ、本稿は、まだ試論段階で、それぞれの作品の使用語彙に対して付けたコードの上２桁の範囲でしか、問題の解明を図らなかった。今後は、それを、コード全体で検討することによって、「歌物語」３作品の使用語彙の特色をよりいっそう明らかにしてみたいと思う。

表1　作品別使用語彙の性格＜異語数＞

上段＝語数　下段＝割合（%）

	伊勢			平中			大和		
	全体	散文	和歌	全体	散文	和歌	全体	散文	和歌
11	240	162	78	169	130	39	266	175	91
	14.6	14.8	14.2	12.7	14.8	8.7	12.4	11.9	13.6
12	253	205	48	140	103	37	436	367	69
	15.4	18.7	8.8	10.5	11.7	8.2	20.4	24.9	10.3
13	113	85	28	82	57	25	133	89	44
	6.9	7.8	5.1	6.2	6.5	5.6	6.2	6.0	6.6
14	122	79	43	55	30	25	144	110	34
	7.4	7.2	7.8	4.1	3.4	5.6	6.7	7.5	5.1
15	176	72	104	129	37	92	218	82	136
	10.7	6.6	19.0	9.7	4.2	20.4	10.2	5.6	20.3
体の類の合計	904	603	301	575	357	218	1197	823	374
	55.0	55.1	54.9	43.3	40.6	48.4	55.9	55.9	55.9
21	242	168	74	260	160	100	313	222	91
	14.7	15.3	13.5	19.6	18.2	22.2	14.6	15.1	13.6
23	223	151	72	245	193	52	303	226	77
	13.6	13.8	13.1	18.4	22.0	11.6	14.2	15.4	11.5
25	40	18	22	37	13	24	43	18	25
	2.4	1.6	4.0	2.8	1.5	5.3	2.0	1.2	3.7
用の類の合計	505	337	168	542	366	176	659	466	193
	30.7	30.8	30.7	40.8	41.6	39.1	30.8	31.7	28.8
31	132	86	46	116	82	34	144	89	55
	8.0	7.9	8.4	8.7	9.3	7.6	6.7	6.0	8.2
33	62	47	15	49	43	6	78	56	22
	3.8	4.3	2.7	3.7	4.9	1.3	3.6	3.8	3.3
35	16	13	3	18	13	5	26	19	7
	1.0	1.2	0.5	1.4	1.5	1.1	1.2	1.3	1.0
相の類の合計	210	146	64	183	138	45	248	164	84
	12.8	13.3	11.7	13.8	15.7	10.0	11.6	11.1	12.6
4	24	9	15	29	18	11	37	19	18
	1.5	0.8	2.7	2.2	2.0	2.4	1.7	1.3	2.7
総数	1643	1095	548	1329	879	450	2141	1472	669

表2　仮名散文作品の使用語彙の性格＜異語数＞

上段＝語数　下段＝割合（%）　＜『万葉集』は参考に掲載＞

	竹取	伊勢	天伊勢	平中	大和	土左	蜻蛉	枕	源氏	紫日	更級	大鏡	方丈	徒然	万葉
11	179	251	240	169	266	171	502	605	1196	369	303	845	213	545	629
	13.7	14.8	14.6	12.7	12.4	17.4	14.0	11.5	10.5	15.0	15.5	17.5	18.6	12.9	9.7
12	124	247	253	140	436	134	295	718	1118	300	236	1096	120	670	1161
	9.5	14.6	15.4	10.5	20.4	13.6	8.2	13.7	9.8	12.2	12.1	22.7	10.5	15.8	17.8
13	88	111	113	82	133	64	291	443	1097	228	111	498	85	604	302
	6.7	6.6	6.9	6.2	6.2	6.5	8.1	8.4	9.6	9.2	5.7	10.3	7.4	14.2	4.6
14	76	130	122	55	144	45	261	507	637	184	117	343	83	314	615
	5.8	7.7	7.4	4.1	6.7	4.6	7.3	9.7	5.6	7.5	6.0	7.1	7.2	7.4	9.5
15	104	182	176	129	218	123	341	518	796	160	188	282	154	359	1162
	7.9	10.8	10.7	9.7	10.2	12.5	9.5	9.9	7.0	6.5	9.6	5.9	13.4	8.5	17.9
体の類の合計	571	921	904	575	1197	537	1690	2791	4844	1241	955	3064	655	2492	3869
	43.6	54.4	55.0	43.3	55.9	54.6	47.0	53.2	42.4	50.3	49.0	63.6	57.1	58.8	59.5
21	259	264	242	260	313	143	724	920	2439	415	329	615	162	588	1143
	19.8	15.6	14.7	19.6	14.6	14.5	20.1	17.5	21.4	16.8	16.9	12.8	14.1	13.9	17.6
23	233	234	223	245	303	137	545	756	2297	364	288	537	144	574	700
	17.8	13.8	13.6	18.4	14.2	13.9	15.1	14.4	20.1	14.7	14.8	11.1	12.5	13.5	10.8
25	36	41	40	37	43	18	101	141	364	57	75	93	28	84	204
	2.7	2.4	2.4	2.8	2.0	1.8	2.8	2.7	3.2	2.3	3.8	1.9	2.4	2.0	3.1
用の類の合計	528	539	505	542	659	298	1370	1817	5100	836	692	1245	334	1246	2047
	40.3	31.9	30.7	40.8	30.8	30.3	38.1	34.6	44.7	33.9	35.5	25.8	29.1	29.4	31.5
31	106	128	132	116	144	90	257	299	760	188	149	248	94	252	277
	8.1	7.6	8.0	8.7	6.7	9.1	7.1	5.7	6.7	7.6	7.6	5.1	8.2	5.9	4.3
33	60	63	62	49	78	32	175	203	496	131	86	160	31	165	129
	4.6	3.7	3.8	3.7	3.6	3.3	4.9	3.9	4.3	5.3	4.4	3.3	2.7	3.9	2.0
35	19	18	16	18	26	12	62	87	148	47	49	58	15	48	55
	1.4	1.1	1.0	1.4	1.2	1.2	1.7	1.7	1.3	1.9	2.5	1.2	1.3	1.1	0.8
相の類の合計	185	209	210	183	248	134	494	589	1404	366	284	466	140	465	461
	14.1	12.4	12.8	13.8	11.6	13.6	13.7	11.2	12.3	14.8	14.6	9.7	12.2	11.0	7.1
4	21	22	24	29	37	15	43	34	67	16	19	43	19	35	106
	1.6	1.3	1.5	2.2	1.7	1.5	1.2	0.6	0.6	0.6	1.0	0.9	1.7	0.8	1.6
総数	1311	1692	1643	1329	2141	984	3598	5246	11421	2468	1950	4819	1148	4240	6505

高次辞書データベースのための語彙知識自動獲得システム

Automatic Word Knowledge Acquisition System for Advanced Dictionary Database

亀田弘之[1]・藤崎博也[2]
Hiroyuki KAMEDA[1] ・ Hiroya FUJISAKI[2]

1. 〒192 東京都八王子市片倉町1404-1 東京工科大学工学部
2. 〒278 千葉県野田市山崎2641　　東京理科大学基礎工学部

1. Faculty of Engineering, Tokyo Engineering University,
   Katakura 1404-1, Hachioji-City, Tokyo 192, JAPAN
2. Faculty of Industrial Science and Technology,
   Science University of Tokyo,
   Yamazaki 2641, Noda-City, Chiba 278, JAPAN

キーワード：知識獲得，未知語，辞書データベース，自然言語処理
Keywords： knowledge acquisition, unknown word, dictionary database, natural language processing

あらまし： そもそも自然言語の語彙は、人間が世界を表現・記述するために必要に応じて創造するものであり、いわば“開いた集合”である。従って、新しい単語が創造される度毎に、辞書に登録する必要があるが、これを機械により支援するシステムはまだない。本稿では、真に実用的な自然言語処理技術の実現に寄与するとともに、言語学・辞書学等の人文科学の研究にも役立つことを意図して構築中の、単語辞書と統語規則とを用いてべた書き日本語文から未知語を抽出するとともに、新たに得られた単語を新語として単語辞書に自動的に登録することのできる未知語獲得システムについて述べる。また、本システム構築のために補助的に作成したユーティリティの概略についても述べる。

Summary: Vocabulary of a language is in general an "open set," for human creatively to express and describe the world. This fact forces natural language systems to have an ability to provide new words (henceforth: unknown words) with the system dictionary when they are found. But no natural language systems can still support automatic word registration to the system dictionary. In this paper, automatic word knowledge acquisition sysytem, which both aims at realizing apractically useful natural language processing system, and also supports wide areas of studies on the Humanities, e.g. linguistics and lexicology, is presented as well as auxiliary utilities for implementing the system.

1. はじめに

自然言語は、人間相互の意思疎通のための手段であるとともに、人間が自己の内外に生起する様々な事象を認識・記述し、さらにそれらを素材として知識を形成・蓄積し、かつ、思考を巡らせるための媒体である[1]。これゆえに自然言語の語彙は特に、社会や時代の進展・変化にともない、必要に応じて創造され、いわば“開いた集合”となっている。

一方、自然言語を機械により処理する研究は、計算機の黎明期から積極的に手がけられており、現在では機械翻訳システムやワードプロセッサ等が商品化されるに至っているものの、既存の自然言語処理システムでは、辞書・文法はいずれも予め限定されており、新しい表現、すなわち機械にとっての未知の表現に対する処理能力を欠いている[2]。

未知語の取扱いに関する研究は従来から部分的になされており[3-11]、著者らもこの問題を解決し、真に実用的な自然言語処理技術を完成させることを主目的として、未知語処理の研究に早くから従事している[1,3]。本稿ではその内、本格的な自然言語処理用の辞書データベースのための語彙知識自動獲得システム、すなわち、予めシステムに準備された辞書データをもとに、機械が未知語を検出し、その品詞等の語彙知識を推定・獲得することのできる未知語獲得システムについて、その概要と動作例について述べる。

2. 未知語の定義・分類と未知語処理の定義

2-1. 未知語の定義[3]

人間は、言語を媒体として相互の意思疎通を行う場合、状況や背景的知識等を適宜利用して、見かけ上同一の表現であっても多様な意味を表現・伝達することができるとともに、さらには必要に応じて新たな単語や表現を創造する。これに対して、その受信者は、多くの場合何の支障もなく、それらの単語や表現に担われた発話者の意図する意味を円滑に推察し理解することができる。人間がこのようなことをなし得るのは、人間が多義的な表現に対して発信者の意図した意味を推察することができるとともに、新たに創造された初見の表現に対しても、ほとんどの場合その意味を推察することのできる能力を持っているからである。つまり初見（未知）の単語であっても、文字・形態素・造語法・統語情報・談話情報等の言語的知識や、文脈・背景的知識等の言語外知識を用いて、初見であることすら気付くことなく迅速かつ適切にその意味を理解することができる。人間はこのような優れた能力を持ち合わせているために、“人間にとっての未知語”という概念は、必ずしも明確には定義することはできない。

一方、現在の機械は、上述したような高度な処理能力を持っておらず、一般的には、単語辞書と、単語の配列を規定する文法規則（統語規則）とを主たる知識として言語処理を行っているため、多義的な表現や新しい創造的な表現は、十分に処理することができない。特に、システムの単語辞書に予め載っていない単語は、システムにとっては未知となる。本稿ではこのような観点から、「機械にとっての未知語」とは未登録単語のことであるとし、以下の議論ではこの定義によるものとする。

2-2. 未知語の種別

未知語には、大きく分けると３つの種別があり、本研究ではそれらを第一種の未知語・第二種の未知語・第三種の未知語と呼ぶこととする。以下にそれらの定義と実例とを示す。なお、この定義に関する詳しい説明は、参考文献[3,12]を参照されたい。また、以下の未知語の実例（下線の付されたもの）はすべて、広辞苑（第４版・岩波書店）の主見出しとして記載されていないものである。

2-2-1. 第一種の未知語

【第一種の未知語の定義】 単語自体は辞書に登録されているにもかかわらず、表記が辞書のものと異なるために、辞書検索に失敗する単語（異表記同義語）のこと。

この種の未知語は、日本語における表記の多様性によるものであり、例えば、単語“慶ぶ”は、広辞苑（第４版）の主見出しに記載されている表記“喜ぶ”あるいは“悦ぶ”とは一致しないために、この辞書をもとにした処理では、未知語（未登録語）扱いとなる。

第一種の未知語はさらに細かく分類することができる。以下(1)～(5)にその例を示す。

(1) 漢字異表記： 異なる漢字で表記されているために生じる未知語。
例：「喜ぶ」と「慶ぶ」
(2) 送りがな異表記： 送りがなの付け方の違いにより生じる未知語。
例：「行う」と「行なう」
(3) 混ぜ書き異表記： 漢字と平仮名等の混ぜ書きにより生じる未知語。
例：「飛び込む」と「飛びこむ」
(4) 片仮名異表記： 片仮名表記の異いにより生じる未知語。この種類の未知語はさらに３つに分類できる。
・大小文字異表記： 片仮名大小文字の異いにより生じる未知語。
例：「ソフトウェア」と「ソフトウエア」
・長音記号異表記： 長音記号により生じる未知語。

例：「コンピューター」と「コンピュータ」
・外来語異表記：外来語表現の異いにより生じる未知語。
例：「バイオリン」と「ヴァイオリン」
(5) 記号の異表記：　数字表記や単位表記の異いにより生じる未知語。
例：「百」と「１００」、「１個」と「１コ」

第一種の未知語はこのように、表記におけるゆれ・慣用的用法・学術（専門）用語の表記規約等に起因するものの他に、「みんなでガンバロー！」のように特定の単語・表現を強調する等の特殊な用法に起因するものもある[13]。

2-2-2. 第二種の未知語

【第二種の未知語の定義】　単語の各構成要素は辞書に登録されているが、その単語自体は辞書に登録されていない単語（既知語を用いて造語された複合語）のこと。

第二種の未知語の例としては、以下のようなものがある。

例：「情報学」　（情報 ＋ 学）
「数学辞典」（数学 ＋ 辞典）
「再試験」　（再 ＋ 試験）

日本語においては、必要に応じてさまざまな複合単語が日常的に造語され利用されるので、この種の未知語の処理は重要である[14]。また、第二種の未知語を機械処理するためにはさらに、いくつかの分類を行う必要があると考えられる。このような観点から、例えば以下のような下位分類が得られる。

(1) 複合語の単語構成要素中に付属的な形態素があるか否かに着目する分類方法。付属的な形態素とは、その単語構成要素自体は単語としての機能を持たず、他の単語（名詞等）に付属することにより、正否、肯否定、是非、程度、性質、状態、範囲、時間、省略等の意味を表す単語構成要素のことをいう。また、接頭語・接尾語もこれに含まれるものとする。以下に例を示す。

例：（付属的な形態素を含む複合語）
「不採用」、「正社員」、「非常識」、
「弱酸性」、「電子化」、「確実性」、
「例外的」、「政府内」、「各国」、
「処理時」、「～等」
（注：強調文字部分が付属的な形態素）
例：（独立した単語構成要素のみから成り、付属的形態素を持たない複合単語）
「証券取引」、「主旨説明」、「最低気温」

(2) 複合語中の各形態素の意味を合成することにより全体の意味が得られるか否かに着目する分類方法。

例：（意味合成可能単語、すなわち、それぞれの意味を合成することで全体の意味が得られる単語）
「人権侵害」、「転送時間」、「最終決定」
例：（意味合成不可能単語、すなわち、それぞれの意味を合成しても全体の意味が得られない単語）
「湾岸支援」、「企業行動憲章」、
「関税貿易一般協定」

2-2-3. 第三種の未知語

【第三種の未知語の定義】　単語の構成要素として、単語辞書に登録されていないものが含まれるもの。

第三種の未知語も以下のように下位分類することができる。未知語の例とともに示す。

(1) 辞書にない単語構成要素（強調文字で表記）を部分的に含む単語
例：「ＩＰアドレス」
「トラブル・シューティング」、
(2) 単語構成要素すべてが辞書にない単語
例：「ボリス・パンキン」
「インターネット」
(3) その他
例：（省略により第三種の未知語となるもの）
「フ諸島」（フォークランド諸島のこと）
「阪大」（大阪大学のこと）

2-3. 未知語処理の定義

“自然言語処理”という用語は、自然言語の理解と生成との両面を指すのに用いられるのと同様に、“未知語処理”も、未知語の理解と生成とを一般には指すが、本稿では、理解の側面のみに着目し、未知語処理を、未知語の検出、内部構造推定、意味推定の３段階に大別する。

3. 各種未知語の処理方法

機械による自然言語処理の場合には、上述した種々の未知語に対して、一般には様々な処理方法が有り得る[3]。以下では、本研究で試作したシステムにおいて、動作を確認した部分に関連するものに重点を置いて述べる。

3-1. 未知語の検出（各種の未知語に共通）

本研究で作成したシステムでの未知語の検出は、原則的には、入力文の統語解析処理の枠組みにおける下記の処理によって行われる。

①入力文から文字列を切出し単語候補とする。
②単語候補に関して辞書検索する。
③辞書に載っていれば既知語であり、未知語の検出は不成功裏に終了する。

④辞書に載っていなければ未知語候補とみなす。
⑤未知語候補の内部構造を調べ品詞を推定する。
⑥品詞推定に失敗する場合には、処理①へ飛ぶ。
⑦推定した品詞が統語解析に矛盾を生じさせなければ、統語解析は成功裏に終了し、その結果、推定が妥当であるとともに、当該の未知語候補は真の未知語であると確認され、未知語検出は終了する。
⑧推定した品詞が統語解析に矛盾を生じさせるならば、統語規則に基づき、他の品詞の可能性も調べ、その可能性があれば処理⑤へ、なければ入力文から新たな文字列を切出して処理②へ飛ぶ。

なお、上述の内部構造推定処理は、各種の未知語あるいは品詞毎に異なる。

3-2. 第一種の未知語の処理

第一種の未知語は、すべての可能な表記を予め辞書に登録することにより処理するか、未知語が出現する度毎に辞書項目に記載されている形(表記)を推定し処理する方法とがある。これらは、出現頻度や未知語処理の処理量との兼ね合いを考慮して最終的に決定することが必要であるが[15]、本研究では、すべての異表記を予め網羅的に知ることが一般には不可能であること、また、単語辞書が不用意に増大しないようにとの配慮から、辞書項目の表記を推定し、異表記同士の文字列照合を行う方式を採用した。

筆者は既に、図1に示す多重照合方式を提案し、C言語によりそのシステムを作成し、基本的有効性を確認しているが[16]、本研究のシステムには、現在、片仮名照合部分のみインプリメントされている。具体的には、片仮名文字列における部分文字列の書換え機能が実装されており、例えば、「イタリヤ」を「イタリア」と書換え辞書検索することができる。

以下、図1の用語について簡単に説明をするが、詳しくは、参考文献[16]を参照されたい。

- **表記照合**：単語の辞書表記をキーとする検索方法。
- **読み照合**：単語の読み（実際のシステムでは、平仮名表記）をキーとする辞書検索方法。
- **普通照合**：表記照合と読み照合の総称名。
- **大小文字照合**：片仮名表記における片仮名小文字と大文字とを同一視して辞書検索する方法。
- **長音記号照合**：片仮名表記における長音記号の有無を無視して辞書検索する方法。
- **外来語照合**：外来語の表記の場合に生じるゆれとして、上記の大文字小文字と長音記号以外にも、「ヴァ」と「バ」、「ヴィ」と「ビ」等がある。これらの異表記を同一視して辞書検索する方法。
- **片仮名照合**：上記、大小文字照合、長音記号照合、外来語照合の総称名。
- **外国語文字照合**：現代日本語の場合、文章中に英語等の外国文字単語が現れることがある。このような異表記に対応するために、外国文字単語（外国語単語）を、対訳辞書を用いて、日本語の単語に変換し辞書検索を行う方法。
- **送り仮名照合**：送り仮名のゆれによる異表記を同一視して辞書検索する方法。
- **混ぜ書き照合**：漢字と平仮名の混ぜ書きによる異表記を同一視して辞書検索する方法。
- **多重照合**：上記すべての照合方法の総称名。

3-3. 第二種の未知語の処理

第二種の未知語の処理では検出の他に、内部構造の推定と意味推定とを行う。以下では、まず、これらの処理に必要な知識を概観した後に、本研究で採用した第二種の未知語の処理方法とそのインプリメント方法について述べる。

3-3-1. 未知語の意味推定に関連する知識[17]

未知語の意味を推定するためには、種々の知識を必要とする。以下に、そのための知識を列挙する。まず知識は、「言語に関する知識」・「発話行為に関する知識」・「発話内容に関する知識」・「発話者に関する知識」の4つに大きく分類される。

「言語に関する知識」とは、媒体としての言語自信に関する知識のことである。また、言語は、語彙や文法規則が相互に密接に関係し合って構築されているという体系(langue)としての側面を有するとともに、個々の具体的発話として現れた諸例(parole)の側面をも持つ。具体的には、体系としての知識と

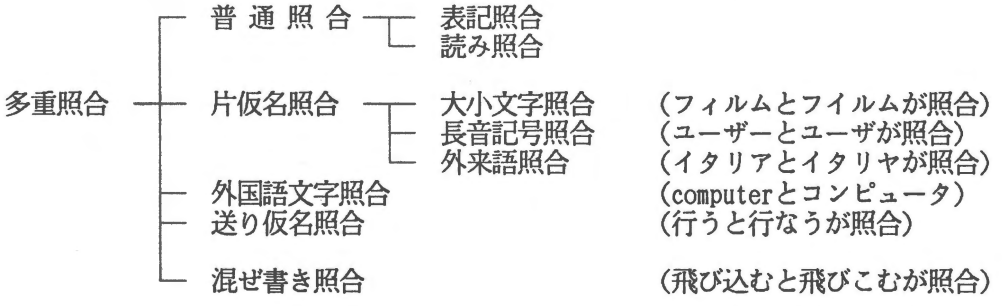


図1. 多重照合の種類

しては、文字セット・文法（統語規則）・各単語の用法等があり、一方、具体的諸例に関する知識としては、用語や文等に関する用例がある。

「発話行為に関する知識」とは、発話の目的・意図に応じて、その目的・意図を達成するための発話計画・方略に関する知識のことである。これはいわゆるコミュニケーション（意思疎通）のノウハウに関するものであり、発話の相手（大人か子供か、専門知識を持っている人か、聴く意志のある人か等）・発話の形態（直接対面、電話、手紙等）などに応じて、伝達したい内容をどの様に整理し、どの様な順番で、どの様な用語・表現を用いれば、より効率的・正確な意思疎通を行ない得るのか、に関する知識のことである。

「発話内容に関する知識」とは、発話により述べられる事実に関する知識のことであり、専門的知識（百科事典的知識）・背景的知識（個別的出来事に関する知識）・常識等がある。

「発話者に関する知識」は、具体的な発話状況において、発話の際の主体者（発話者と聴き手）自体に関する知識のことであり、上記の発話に関する知識を実際に運用する際に利用されるべきものである。

上述の内容をまとめたものを下図に示す。

①言語に関する知識：
- 体系としての言語に関する知識
  （文字・文法・用法等）
- 事例（用例）としての言語の知識
  （用語・例文等）

②発話行為に関する知識：
- コミュニケーション行為のノウハウに関する知識

③発話内容に関する知識：
- 情報伝達媒体としての言語により伝達される情報・知識自体の内容に関する知識
  （対象に関する知識・専門的知識・世界に関する知識・背景的知識等）

④発話状況に関する知識：
- 具体的発話者の特性 等

図2. 未知語の意味推定に関連する知識

未知語の処理を行うためには、このように様々な知識を必要とするが、筆者らは、現在のところ、まず上記①の言語に関する知識のみに着目する処理方法に重点をおき、規則に基づく処理方法と、用例からの類推に基づく処理方法とを取り上げて研究を行っている[18-25]。本研究で作成したシステムでは、そのうち規則に基づく処理が現在組込まれている。

3-3-2. 規則に基づく処理方法[18-20, 23-25]

第二種の未知語の多くはいわゆる複合語であること、また、筆者の行った語彙調査では[3]、未知語の多くは単語構成要素が2個の第二種未知語であることから、本研究では、まず単語構成要素が2個の未知複合語に処理対象を限定してシステムを作成した。このシステムでは、単語構成要素間の関係を規則化することにより、未知語の内部構造と意味とを推定する。

(1) 語内文法

従来の文法（文文法）を参考にして、単語構成要素とそれらの相互関係を分析し、表層レベルと深層レベルの2層の形式で整理した。

①表層レベル：　このレベルでの単語カテゴリとして、名詞的要素、動詞的要素、形容詞的要素、副詞的要素の4つを設定し、このうち、名詞的要素と動詞的要素を中心的に分析、整理した。最終的にインプリメントした規則は以下の通りであり、プログラミング言語prologにより記述されている。

```
表層構造関係(名詞的要素,　動詞的要素).
表層構造関係(動詞的要素,　動詞的要素).
表層構造関係(形容詞的要素, 名詞的要素).
表層構造関係(副詞的要素,　動詞的要素).
表層構造関係(名詞的要素,　名詞的要素).
```

図3. 表層構造に関する規則の例

②深層レベル：　動詞的要素とその他の要素との意味的関係に着目し、まず後者に対して、深層格と意味カテゴリとを設定した。また、動詞的要素に対しては、意味カテゴリと意味パタンとを設定した。なお、意味カテゴリとは、単語構成要素の担う意味の分類カテゴリのことであり、意味パタンとは、動詞的要素に関わる（格文法の意味での）意味構造のことである。以下に、深層格と動詞的要素の意味カテゴリを列挙する。

- 深層格：動作格、対象格、源泉格、目的格、経験者格、受け手格、関係格、受益格、道具格、方法格、役割格、比較格、程度格、場所格、時間格、期間格、原因格、結果格、手段格、目的格、条件格、内容格、範囲格（23個）
- 動詞的要素の意味カテゴリ：存在、属性、占有、関係、知覚状態、感情状態、自然現象、物理的遷移、占有遷移、属性移動、身体動作、生産、社会動作、精神遷移、知覚動作、感情動作、思考動作(17個)

これらの整理された知識をもとに、次頁の図4と図5とに示すような規則をインプリメントした。

```
深層構造関係(表記(Element1), 品詞(Cat1),
      意味(Meaning1), 表記(Element2),
      品詞(Cat2), 意味(Meaning2),
      深層構造(Deep_str), 総合的意味(M))
 :-  Cat1 = 名詞的要素, Cat2 = 動詞的要素,
      動詞パタン(動詞的概念(Element2),
        動作主格(Element1)),
      Deep_str = 動詞パタン(動詞的概念
        (Element2), 動作主格(Element1)),
      M = [Meaning1, が, Meaning2, する].
```

図4. 深層構造に関する規則の例

```
動詞パタン(動詞的概念(旅行), 動作主格(Agnt))
        :- 人間(Agnt).
動詞パタン(動詞的概念(旅行), 場所格(Plc))
        :- 場所(Plc);建造物(Plc).
動詞パタン(動詞的概念(旅行), 時間格(Time))
        :- 時間(Time).
動詞パタン(動詞的概念(旅行),道具格(Instr))
        :- 交通手段(Instr).
```

図5. 動詞的要素の意味パタンの例

(2) 単語構成要素辞書

単語構成要素は多くの場合、単独で単語となり得るので、単語辞書で兼用することも可能であるが、単語と単語構成要素とを明確に区別して取り扱うために、便宜上、単語構成要素は単語構成要素辞書に分離記述した。なお、単語構成要素は現在のところ、６９個登録されている。

```
登録要素(表記(烏), 品詞(名詞的要素),
        意味(カラス)).
登録要素(表記(烏), 品詞(形容詞的要素),
        意味(カラスのような)).
登録要素(表記(見学), 品詞(動詞的要素),
        意味(実地に見て知識をひろくする
        (見学する) こと)).
登録要素(表記(見学), 品詞(名詞的要素),
        意味(見学)).
登録要素(表記(方式), 品詞(名詞的要素),
        意味(やり方)).
登録要素(表記(３０日), 品詞(名詞的要素),
        意味(３０日)).
登録要素(表記(小), 品詞(形容詞的要素),
        意味(小さな)).
```

図6. 登録した単語構成要素の例

3-4. 第三種の未知語

入力文から切り出される文字列は、先にも述べたように、まず、単語辞書に登録されているか調べられ、登録されていれば既知語として処理され、もし、登録されていなければ、未知語候補として処理される。この際、未知語候補は、まず、第二種のものと仮定されて処理が続行され、処理に矛盾が生じた場合には、第一種の未知語として処理される。さらに矛盾が生じた場合には、第三種として処理する。なお、この仮定にも矛盾が生じた場合には、切り出された文字列は単語ではないと判断する。

4. 未知語獲得システム

4-1. システムの概要

上述した未知語処理方法を統合し、未知語を獲得するシステムを、DEC製 ノート型パーソナルコンピュータDigital HiNote CT475（主記憶20MB、ハードディスク350MB）上に、Arity/Prolog（Version 5.1, ライフボート社製）を用いてインプリメントし、以下の処理モジュールと知識ベースからなる。

A．処理モジュール
- 主処理モジュール部
- テキスト入力モジュール部
- 統語解析モジュール部
- 第一種未知語処理モジュール部
- 第二種未知語処理モジュール部
- 第三種未知語処理モジュール部
- 単語獲得モジュール部
- 補助関数モジュール部

（これら全体で約２，４００行）

B．知識ベース
- 単語構成要素データベース
- 単語辞書データベース
- 造語規則データベース
- 統語規則データベース

4-2. 処理の流れの概要

本システムの処理の流れの概要を次頁の図７に示す。処理制御の流れは、prologシステムに依存しており、トップダウン的に処理する。以下に図７に準じてアルゴリズムの概略を述べる。

①テキスト読込み： 漢字仮名交じりべた書きの日本語文を、文字列としてキーボードから読込む。

②文リスト構築： 読込まれた文字列を文字毎に分解しリスト構造の形式に変換する。変換結果を以下、文リストと呼ぶ。

③文法による先頭単語品詞の推定： 文法（統語規則）を参照しながら、文リストの先頭に位置する単語の品詞を、トップダウンに推定する。

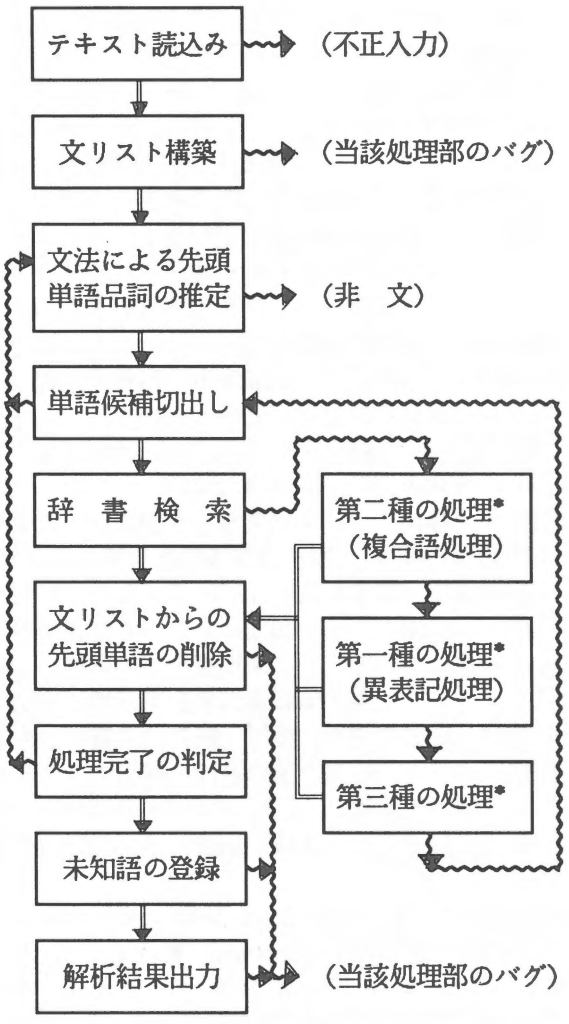


図7. 未知語獲得システムの処理の流れの概要

④**単語候補切出し：**　文リストの先頭側から単語候補として、部分リストを切出す。

⑤**辞書検索：**　単語候補としての部分リストが、辞書に登録されているか検索する。検索に成功すれば、これは未知語ではなく、処理は⑨に移る。検索に失敗する場合は、これを未知語候補とみなし、処理は次の⑥へ移る。

⑥**第二種の処理：**　未知語候補が、第二種の未知語かどうか調べる。第二種の未知語とみなし得る場合には、処理は⑨へ、そうでなければ⑦へ移る。

⑦**第一種の処理：**　未知語候補が、第一種の未知語かどうか調べる。第一種の未知語とみなし得る場合には、処理は⑨へ、そうでなければ⑧へ移る。

⑧**第三種の処理：**　未知語候補が、第三種の未知語かどうか調べる。第三種の未知語とみなし得る場合には、処理は⑨へ、そうでなければ④へ移る。なお、上記⑥～⑧が未知語処理の中核部分である。

⑨**文リストからの先頭単語の削除：**　文リストの先頭に位置する単語を削除し、残りのリストを新たな文リストとする。

⑩**処理完了の判定：**　文リストが空リストか調べる。空リストならば、統語解析の処理は完了しているので⑪へ、そうでなければ③へ移る。

⑪**未知語の登録：**　統語解析の際に、未知語が検出されていれば、推定品詞等の情報も統合して、新たな辞書項目として辞書に登録する。

⑫**解析結果出力：**　統語解析結果をディスプレイ上に表示する。

4-3. 動作例

例えば、“イタリヤ”（第一種の未知語）、“見学旅行”（第二種の未知語）、“シェーンな”（第三種の未知語）を含む文として、「シェーンなイタリヤの見学旅行に行った」を入力すると出力として、

『文(主部(名詞句(名詞句no(連体詞(プチな，第三種未知語)，名詞句(名詞(第一種未知語(名詞(イタリア)))、助詞(の)))，名詞句(未知複合語(見学旅行)、助詞(に)))，述部(動詞句(動詞(行く)))))』

が表示される。この例では、統語規則の知識とともに、“シェーンな”が連体詞の語尾「な」を持っていること等から連体詞と推定され、“イタリヤ”が“イタリア”の異表記単語として照合され、さらに、“見学旅行”は“見学”と“旅行”が複合語の構成要素になり得るとの知識および造語規則から、第二種の未知語と推定されている。

5. その他のユーティリティ

上述した未知語獲得システムを実際に辞書データベースのために稼働させるためには、単語辞書と統語規則とを充実させる必要がある。本研究では、そのために、以下のようなユーティリティを作成した。

5-1. 単語辞書作成のためのユーティリティ

CD-ROMに格納された広辞苑（第4版）のデータを素材として、prologプログラム形式の単語辞書を作成するユーティリティを作成した。記述言語は、文字列操作言語jgawkである。なお、処理の一部は、テキストエディタVZ(Version 1.6)の文字列検索・置換機能を用いており、現在単語辞書には約6万単語が登録されている。

5-2. 統語規則作成のためのユーティリティ

光学的文字読取り装置(OCR, Optical Character Reader)により電子化した単文(「スペイン語基本文2000」, 大学書林)を、テキストエディタVZを利用して、手作業により単位切りおよび品詞情報のタグ付けを行い、そのタグ付きテキストから品詞列情報およびprolog形式の統語規則を自動抽出することのできるユーティリティを作成した。記述に使用した言語は、jgawkである。現在、これらを利用して統語規則作成のための基礎的資料を蓄積・分析中である。

6. おわりに

本稿では、辞書データベースのための未知語獲得システムの概要とその試作結果および、単語辞書・統語規則作成用ユーティリティについて述べた。

なお、本研究の一部は、文部省科学研究費補助金試験研究(B)(1)(課題番号:07558274, 研究代表者藤崎博也)により行われた。

<<参考文献>>

[1] 藤崎:“言語的思考過程の定式化”,林大(編), 講座 現代の言語2「言語と思考の発達」第2章, 三省堂(1984).

[2] 藤崎・亀田:“知識獲得研究の展望”, 電子情報通信学会第二種研究会「言語獲得と概念形成」, LA90-1, pp.1-8(1990).

[3] 藤崎・亀田 他:“人間の言語処理過程のモデルに基づく自然言語理解システムの構築”, 昭和63年度科研費特定研究「言語情報処理の高度化のための基礎的研究」第6班研究発表資料集(1989).

[4] 吉村・武内・津田・首藤:“未登録語を含む日本語文の形態素解析”, 情報処理学会論文誌, Vol.30, No.3(1989).

[5] 永瀬:“形態素タイプを用いた日本語構文解析前処理”, 情報処理学会第41回全国大会(1990).

[6] 大沢・藤崎:“未知語を含む文の形態素解析システム”, 情報処理学会第42回全国大会(1991).

[7] 植田・小松・横尾・宮崎:“部分複合語による複合名詞構造解析”, 情報処理学会第43回全国大会(1991).

[8] 石川・伊藤・牧野:“文節オートマトンを用いた未知語処理法”,電子情報通信学会第二種研究会「言語獲得と概念形成」,LA92-17,pp.1-8(1993).

[9] 山田・山村・佐川・大西・杉江:“英文における未登録語の意味推定の検討”,情報処理学会研究報告, Vol.93,No.1, 93-NL-93, pp.63-70(1993).

[10]神岡・安西:“仮説生成機構を用いた未知語を含む文の解析”, 人工知能学会誌, Vol.3, pp.627-638(1988).

[11]P. Norvig:“Paradigms of Artificial Intelligence Programming”,Morgan Kaufmann(1992).

[12]亀田・藤崎・森田・倉島:“未知語の分類とその処理に関する考察”, 情報処理学会第36回全国大会講演論文集, 5T-5, pp.1195-1196(1988).

[13]富田隆行・眞田和子:“表記”, 教師用日本語教育ハンドブック2 改定版,国際交流基金(1988).

[14]荻野綱男:“名詞辞書に含まれるべき見出しの範囲 ── 特に複合名詞の扱いをめぐって ──”, ソフトウェア文書のための日本語処理の研究-8 ── IPAL補完文法 ──, 情報処理振興事業協会, 61技-072, pp.207-221(1987).

[15]亀田:“未知語処理機能をもつ自然言語処理システムにおける語処理の効率について”, 電子情報通信学会第二種研究会「言語・知識の運用と獲得」研究発表資料, LK92-4(1992).

[16]亀田・藤崎:“日本語文章の形態素解析における未知語の獲得”,電子情報通信学会第二種研究会「言語獲得・概念形成」,LA90-15,pp.1-8(1991).

[17]亀田:“用例からの類推にもとづく知識の獲得と一般化について ── 未知複合語の獲得を中心にして ──”, 電子情報通信学会第二種研究会「言語・知識の運用と獲得」研究発表資料(1993).

[18]亀田:“言語知識の獲得過程を解明するための心理実験と未知複合語解析システムの試作”, 平成3年度科研費重点領域「知識科学」成果報告書(1992).

[19]KAMEDA:“A Processing Method of Class-2 Unknown Japanese Compound Words of Two Components with Use of In-Word Grammar and Its Prototype System Implementation”, IEEE, TENCON'92, pp.710-714(1992).

[20]亀田:“未知語の意味推定過程解明のための実験と未知複合語の意味推定システム基本処理部の試作”, 平成4年度科研費重点研究「知識科学」成果報告書(1993).

[21]波多野・小嶋:“未知語の意味の推定・獲得の過程”, 平成4年度科研費重点領域「知識科学」シンポジウム論文集, pp.45(1993).

[22]亀田・波多野・小嶋:“用例からの類推に基づく未知語意味推定システム”, 人工知能学会第8回全国大 会23-7,pp.653-656(1994).

[23]亀田・桜井:“語内文法に基づく未知複合語意味推定システムの作成と評価”,人工知能学会第8回全国大会, 23-8, pp.657-660(1994).

[24]亀田・桜井:“べた書き日本語文からの未知語獲得システムの作成”, 電子情報通信学会「思考と言語」技報TL94-11, pp.17-24(1994).

[25]亀田・桜井:“統語解析処理にもとづく未知語獲得システムの試作”, 電子情報通信学会総合大会講演論文集「基礎・境界」, pp.474-478(1995).

社会調査結果の視覚化データベース

Visualized Database of Social Survey Results

吉田　光雄

Mitsuo YOSHIDA

大阪大学　人間科学部

Human Sciences, Osaka University

キーワード: *Mathematica*, 探索的データ解析, 情報処理教育

Keywords: *Mathematica*, EDA(Exploratory Data Analysis), Education for Information Processing

あらまし: 社会調査結果を計算機内に保存し、必要に応じてデータを取り出し、*Mathematica* を用いて、基本統計量を計算すると同時に、ヒストグラム、散布図等のグラフを描き、データの様相を視覚的に検討することのできるデータベースをワークステーション上に構築した。すなわち、保存されたデータをテキストとして取り出せると同時に、それらを視覚的に提示し、探索的にデータの内部構造を探り、データの持つ情報を十分に引き出して分析を深化させるために用いることができる。

本システムは小規模の試作用データベースであるが、メモリー・ハードディスクの容量や高速演算などの十分な計算機資源が得られれば、さらに大規模のデータにも適用することができる。快適に動かして大量のデータを処理するためには、かなりな量のメモリを必要とするが、統計処理のメニュの提示や選択、さらにはより美しいグラフィックスの描画や3次元表示法に対する工夫など、ユーザー・インターフェイスの改良は今後の課題である。

Summary: Social survey results should be preserved as database, so that one can consult it afterwards and extend it to another analysis. If it can be browsed in a visual format, such a database would be an efficient and convenient tool for many kinds of social surveys.

This report summarized how to save data in *Mathematica* as a database, and how to make statistical treatments using graphic techniques, such as histogram, bar-chart, pie-chart, scatter diagram, three dimensional graphic and others in it.

Mathematica seems to be suitable for this purpose, because of its wide coverage of mathematical and statistical usages.

This database seems to be useful, not only to preserve all of the data, but as a visual database that contributes as an EDA(Exploratory Data Analysis) and VDA(Visual Data Analysis) revealing new and hidden structures of the data.

An example of social survey results concerning computer education at Osaka University was demonstrated.

1. はじめに

Mathematica ^{TM}[3] は1986年、Wolfram Research 社から発売され、現在世界中で広く使用されている数式処理ソフトのひとつである。操作の容易さ、機種互換、グラフィックス化等に特徴があり、例えば、式の展開、因数分解、方程式、数値計算、行列演算、微分・積分、微分方程式などの数式処理の他、関数のグラフ表現、統計計算など、数学的処理を容易に行うことができる。強力なパッケージ[2]を多数含み、"A System for Doing Mathematics by Computer" として評判の高いものである。

統計学に関するパッケージも多く含まれているので、さまざまな統計処理にも活用することができる。しかも、グラフィックス機能が豊富であるので、目で見る統計学として有効である。ワークステーション (NeXT) 版およびマッキントシュ版はユーザー・インターフェイスのよい Notebook を介して使用可能であり、コマンド入力、プログラミング、出力、グラフィックスがすべて同一ウインドウ上で実行され、しかも両機種間で高い互換性を有している。

統計パッケージ SAS を用いた社会調査結果の視覚化データベース、ならびにS言語を用いたものも可能であるが[5]、本稿では *Mathematica* を用いて行う方法について報告する[1]。また、まだモジュール化されていない、いくつかの統計的手法については、内蔵の言語を用いてプログラミングを行ったので、それらについても報告する。追加された各種の統計演算はライブラリとして保存されている。

Turkey,J.W.[4]により探索的データ解析 (EDA, Exploratory Data Analysis) の方法が言われて以来、統計データを視覚的に分析する方法 (VDA,

[1]本稿は、文部省科学研究費補助金・重点領域研究、「情報化社会と人間」研究成果報告書、および日本計算機統計学会第8会大会にて報告したものの一部である。

Visual Data Analysis) が注目され、いろいろな方法が提案されている。本データベースはあくまでもデータの統一的保存が目的であるが、保存されたデータをテキストとして取り出せると同時に視覚的に提示し、全体像の把握を容易にすることをも目指す。そして、更に必要な場合には再度統計処理も行うことが可能なものとする。統計処理はすでになされているが、様々な角度からより探索的にデータの内部構造を探り、データの持つ情報を十分に引き出し、分析を深化させるために用いることができる。

大阪大学では、情報処理教育センターを中心にNeXT を端末とする分散処理システムのネットワークが構築され、学生の情報処理教育に活用されている(1)。教材の提供・課題の提示・レポートの収集などが容易に行える教育支援のためのアプリケーションが開発されており、それを活用することにより、登録学生は容易に本データベースにアクセスすることができる。

2. データの作成と *Mathematica* の基本

2.1 保存および参照

Mathematica を起動する前に、カレント・ディレクトリにデータをテキスト (Ascii) 形式で保存しておく。データの形式は、各行にサンプル (ケースまたはオブザーベーション)、各列に変数 (調査項目) をあて、行列データの入力形式に従って、各行ごとに { } を使用する。

データは連続量でもよいし、数字化された属性データでもよい。ただし、処理や結果の解釈に際して、データの属性に注意する必要があり、現在のところデータベースのシステム内に、可能な統計処理をチェックする機能は組み込まれていない。

調査票の項目一覧も必要であり、*Mathematica* 内の変数番号と、調査票の項目番号の対照も必要である。本計算機システムがNeXT ワークステーションであるため、多重ウインドウを開くことが可能であり、項目対照を明示するウインドウを常時開いておくと、操作が容易となる。項目の選択は変数名やラベルを使用するのではなく、通し番号で行う。

調査票は多くの場合、回答者のバイアスを避けるためランダムに並べられているが、計算機に保存する際には、予め質問項目番号、回答カテゴリの方向性などを整理し、整合性を保って保存しておいた方が便利であろう。調査項目一覧もデータに合わせて並べ換えておく。

メモリなど、計算機資源が豊富な場合には、全データを一括して読み込ませておくと迅速な処理が可能であるが、十分でない時は、必要な項目を処理の都度読み込めばよい。

一旦保存されたデータは修正ができないよう、ライトプロテクトをかけておく。データを加工しても、保存はせず、必要な修正はその都度行うようにしておかないと、データベースの機能を果たさなくなる。従って、以後の修正が無用となるよう、最初に計画的に保存しておく必要があろう。

現在のところ NeXT 版 *Mathematica* (Ver. 2.1) は日本語対応とはなっておらず、出力に日本語が使用できないのは難点である。ただし、処理結果をワープロや作図ソフトに取り出して編集すれば、日本語は使用可能である。

2.2 *Mathematica* の基本

Mathematica を起動し、ハードディスク上のカレント・ディレクトリ (~) に保存されているデータ・ファイルを *Mathematica* に読み込む。必要に応じて、特定の項目に関する全サンプルの回答を抽出し、編集しなければならないが、そうした作業を便利に行うことのできるコマンドも多く用意されている (表1)。また、統計処理を行うためには、統計パッケージを予め読み込んでおかねばならない(2)が、これらの一連の初期作業をルーチン化してプログラミングしておいてもよいし、*Mathematica* との対話の内容は *Notebook* に保存されているので、それらを利用して copy & paste しつつ、再度実行することも可能である。

3. 統計処理

3.1 基本統計量の算出

統計パッケージ (記述統計学) を用いることにより、平均・メディアン・モード・分位数などの代表値、分散・不偏分散・標準偏差・範囲・四分偏差などの散布度のほか、歪度・尖度なども直ちに計算することができる (表3)。

3.2 区間推定・検定

パッケージ (信頼区間、検定) を用いて、母分散や母平均に関する区間推定・検定を行うことができる (表4)。データは項目番号を指定して粗データをベクトルとして取り出す。その際、オプションを用いて、有意水準、両側・片側などを設定する。

表 1: データ読み込み・編集のためのコマンド

パッケージ　*Statistics`Master`* *Statistics`DataManipulation`*	
`ReadList["~/file1.dat",Number]`	
`ReadList["~/file1.dat",Number, RecordList->True]`	
`TableForm[data]`	`Column[data, n]`
`Column[d,{n1,···}]`	`ColumnTake[d,n]`
`ColumnDrop[d,{n1,···}]`	`ColumnJoin[d1,d2,···]`
`Row[data, n]`	`RowTake[d,n]`
`Row[d,{n1,···}]`	`RowDrop[d,{n1,···}]`
`RowJoin[d1,d2,···]`	
`BooleanSelect[d,sel]`	`TakeWhile[d,pred]`
`LengthWhile[d,pred]`	

表 2: データ集計のためのコマンド

パッケージ　*Statistics`DataManipulation`*	
`Frequencies[data]`	`QuantileForm[d]`
`CumulativeSums[d]`	`BinCount[d,{min,max,dx}]`
`RangeCount[d,{c1,c2,dx}]`	
`CategoryCounts[d,{cat1,cat2,···}]`	
`BinLists[d,{min,max,dx}]`	
`RangeLists[d,{c1,c2,dx}]`	
`CategoryLists[d,{cat1,cat2,···}]`	

表 3: 基礎統計量算出のためのコマンド

パッケージ　*Statistics`DescriptiveStatistics`*	
`Mean[data]`	`Median[d]`
`Mode[d]`	`GeometricMean[d]`
`HarmonicMean[d]`	`Quantile[d,q]`
`Quantiles[d]`	`LocationReport[d]`
`Variance[d]`	`VarianceMLE[d]`
`StandardDeviation[d]`	`MeanDeviation[d]`
`StandardDeviationMLE[d]`	`SampleRange[d]`
`QuartileDeviation[d]`	`Skewness[d]`
`DispersionReport[d]`	`Kurtosis[d]`
`CentralMoment[d]`	`ShapeReport[d]`
`covariance[`$x1, x2$`]`	`correlation[`$x1, x2$`]`
`multipleCorr[`x, y`]`	`partialCorr[`x, i, j`]`
`canonicalCorr[`$x1, x2$`]`	`contTable[`$x1, x2$`]`
`association[`$x1, x2$`]`	

表 4: 推定・検定のためのコマンド

パッケージ　*Statistics`Confidenceintervals`*
`MeanCI[data]`
`MeanCI[d, KnownVariance->v]`
`MeanDifferenceCI[d1,d2]`
`MeanDifferenceCI[d1,d2, KnownVariance -> {v1,v2}]`
`VarianceCI[d]`
`VarianceRatioCI[d1,d2]`
パッケージ　*Statistics`HypothesisTests`*
`MeanTest[d,mu]`
`MeanTest[d,mu, KnownVariance->var]`
`MeanDifferenceTest[d1,d2,diff]`
`VarianceTest[d1,var]`
`VarianceRatioTest[d1,d2]`

パッケージ (連続型分布) には、予め正規分布、t−分布、F−分布、χ^2−分布 などの標本分布が組み込まれているので、上側・下側・両側確率、上側・下側・両側パーセント点を求めることができる (表 5)。従って、任意の値の自由度についての統計数値が計算可能であるし、プログラミングすることにより、粗データからではなく、標本平均・標本分散などから、直接区間推定や検定を行うこともできる。

4. データの視覚化

グラフ用パッケージを読み込み、グラフを描くことができる (表 6)。BarChart で描くのは棒グラフであるが、ヒストグラムの代用とすることもできる。

標本統計量から分布のパラメータを求め、確率分布をあてはめて、密度関数・分布関数などのグラフを描くこともできるし、データの分布と重ね

表 5: 標本分布

パッケージ　*Statistics`ContinuousDistributions`*
`NormalDistribution[0, 1]`
`ChiSquareDistribution[df]`
`StudentTDistribution[df]`
`FRatioDistribution[df1,df2]`
NonCentralChiSquareDistribution[df,lambda]
NonCentralStudentTDistribution[df,lambda]
NonCentralFRatioDistribution[df1,df2,lambda]
`CDF[dist,x]`
`Quantile[dist,q]`

表 6: グラッフィクス

パッケージ *Graphics`Graphics`*	
`BarChart[data]`	StackedBarChart[d1,d2,···]
`PieChart[d]`	`ListPlot[d]`
`TextListPlot[d]`	`LabeledListPlot[d]`
`DisplayTogether[plot1,plot2,···]`	
`DisplayTogetherArray[{{p1,p2,···},` {···}, ···, {···}}]	
パッケージ *Graphics`Graphics3D`*	
`BarChart3D[data]`	`ScatterPlot3D[d]`
パッケージ *Graphics`MultipleListPlot`*	
`MultipleListPlot[d1,d2,···]`	
MultipleListPlot[d1,d2,···,PlotJoined->True]	

表 7: ユーザー・グラフィックス関数

1 変数データの統計図	
`11.msd[`*dname, item*`]`	平均・標準偏差
`12.mvec[`*vec*`]`	平均値ベクトル
`13.psing[`*vec*`]`	比率
`14.distsing[`*vec*`]`	度数分布
2 変数データの統計図	
`21scatter[`*vec*`]`	散布図
`22.ctab[`*vec*`]`	相関表
`23.cortab[`*vec*`]`	相関表 (多数)
多数データの統計図比較	
`31.hist6[`*vec*`]`	ヒストグラム
`32.mvprof6[`*vec*`]`	平均値プロフィール
`33.pcomp6[`*vec*`]`	比率の比較
`34.distcomp6[`*vec*`]`	度数分布の比較
`35.mpc6[`*vec*`]`	平均 (比率) の比較

て描き (表 6)、確率分布のあてはめの程度を検討することもできる。

調査項目 (変数) をいろいろ抜き出し、グラフを描くことによって、データ全体を見渡すことが出来るし、かつ outlier などの発見も容易となる。

グラフィックス・コマンドはオプションが豊富に用意されており、凡例・カラー・線の太さと種類・ラベル・テキスト挿入などを自由に設定することにより、様々な美しいグラフを作成し、プレゼンテーションの効果を上げると同時に、EDA としても活用することができる。データから探索的にさらに多くの情報を引き出すことが、本データベースの目的でもある。

5. ユーザー・グラフィックス関数

データの種類、例えば、連続量データかカテゴリデータかにより、描くグラフの種類も異なってくる。どのような種類のデータに対して、どのようなグラフが描けるかを、先の標準パッケージをもとに、描画し易いようにユーザー関数としてプログラミングを行った。

社会調査の場合、属性データの分類に際しては、0,1,2,··· の離散型データが用いられ、名義尺度として数量化されるケースが多い。さらに順序データ、連続量データであっても離散型で処理されるケースも多く、属性データの処理は基本であろう。

連続量データの場合でも度数分布表に集計して、順序尺度として使用されるケースが多い。従って、多用されるのも、カテゴリカルデータのグラフ化である。

統計処理としては、度数の集計、度数分布、分割 (関連) 表などであり、そららの図示として、円グラフ、帯グラフ、分割表の 3D 表示などが用いられる。

連続量データの全体像を見るためには、粗データをそのまま用いて、散布図を描くこともできるが、プレゼンテーションを工夫して、3 次元 (立体) 散布図を描くことも可能である。

さまざまなグラフ化に対応するため、標準グラフを活用したユーザ関数を作成した。パッケージを使用しているので、それらの読み込みが必要である。

データセットから項目 (変数) を指定して、平均・標準偏差を算出する関数が msd である。データ名を *dname* で与え、項目を *item* で選出する。*item* が複数に及ぶときは、mvec を使用する。このとき関数の引数は *vec* のみであるので、関数を呼ぶ前にデータセット dname[1]=*datasetname* と項目番号 *vec=itemnumbers* を予め設定しておかねばならない。*vec* はリストである。mvec はグラフを描く。

データが (0,1) の場合は、回答 (yes=1) の比率を求め、ヒストグラム (棒グラフ) を描く。

下位反応カテゴリの度数を集計し、比率を求め、分布状況についてのヒストグラムを描くのが psing である。これらは複数項目の処理が可能であり、事前にデータセット名と処理項目番号を選定しておく。

2 変数の連続量データについて、平均・標準偏差・共分散・相関係数を算出し、散布図を描くのが scatter であり、項目を *vec* で選択する。選択

された vec の項目から2項目づつを取りだして処理する。

離散量データの場合、セルの度数のカウントとなり、散布図では度数が1点としてしか図示されないので不合理である。図示方法を変更し、3次元立体図と3次元表面図で図示するのが ctab である。1項目の単純集計 (marginal distribution)、2変数のクロス集計 (joint distribution) の度数も出力する。単純集計のグラフでは標準関数の制約から、2項目のみの描画ができず、従って1項目を反復して、並べて描くこととする。

ctab は2項目間の処理のみ、cortab は多数項目を vec で与え、その中から順次2項目の組み合わせを取り出して処理するものである。

以上は単独のデータセットについての処理であるが、それが複数となったとき、データ間の比較を行うためのユーザー関数が、番号31以下である。

まず、hist6 は度数分布を集計しヒストグラムを描く。mvprof6 はで指示する項目の平均値を求め、それのプロフィールを描く。pcomp6 は比率の比較、distcomp6 は度数分布の比較である。mpc6 は平均値 (または比率) の比較を行う。

予めデータセット名 (dname[i]=$filename_i$) とその数 ($ng \leqq 6$) を与え、処理する項目を ($vec=i_1, i_2, \cdots$) で与える。データセットの数の上限は6である。あまり多くのグラフを重ねると相互の比較が困難となるが、上限の6をプログラム内で修正することは容易である。

グラフはいずれも折れ線グラフを描き、相互比較の便を期す。

6. 多変量解析

現在のところ、多くの多変量解析の手法は *Mathematica* に組み込まれておらず、他の統計ソフトを使用しないのであれば、自らプログラミングせざるをえない。行列やベクトルの積、逆行列などの行列演算や、固有値・固有ベクトルの計算などは *Mathematica* 内に定義されているので、それらを用いれば多変量解析のプログラムを書くことは、さほど困難な作業ではない。むしろ、*Mathematica* に内蔵のプログラム言語では、ベクトル演算がサポートされており、一般の言語より短く、簡単に書くことができる。

これらを発展させ、多変量解析の例として、重回帰分析、判別分析、主成分分析のプログラミングを行った。必要に応じてデータベースから簡単にアクセスすることができる。さらに、FORTRAN, C, Pascal などの一般の言語で書かれた実行形式のプログラムをリンクさせて、用いることもできる。

7. 具体例

社会調査の一例として、大阪大学生に対して実施された情報処理教育に関する調査結果[1]を使用する。調査項目は以下の通りである。

- 学部・専攻などのフェースシート項目
- 大学入学以前のコンピュータ経験
- 入学後の計算機使用の実態
- 計算機に対する意識
- 情報処理教育に対する意見

全学部1年生のランダム・サンプルとして999名 (在籍学生のほぼ35%(文科系 489 名、理科系 510 名)) のデータを入手し、計算機内に保存し、データベースとした。

調査項目は、ID番号をも含めて No.1 ～ No.106 である。5段階表示の順序データ、(0-1) 表示のカテゴリデータなどが混在している。

サンプルは、例えば (1) 学部別、(2) 文科系・理科系別、(3) 全調査対象を合併したもの、などの構成とすることが可能であり、3群毎にファイルを作成しておけば、必要に応じて目的のデータを読み込むことができる。

基礎統計量の計算、単純・クロス集計、グラフ化が可能であるが、一例として、図1 ～図6 に各種のグラフを示す。

文　献

[1] 松浦敏雄他 (1993) 教育用計算機システムの運用と学生の意識, 行動計量学, **40**, 17-31.

[2] Phillip Boyland (1992) Technical Report, Guide to Standard Mathematica Packages, Wolfram Research.

[3] Stephan Wolfram (1988) Mathematica, Addison-Wesley, 訳書 (1992) アジソン・ウエスレイ.

[4] Tukey, J.W. (1977) Exploratory Data Analysis, Addison-Wesley.

[5] 吉田光雄 (1993,95) 映像情報データベースの開発, 文部省科学研究費補助金・重点領域研究「情報化社会と人間」, 研究成果報告書.

565 吹田市山田丘 1-2, 大阪大学人間科学部
Tel: 06-879-8051, Fax: 06-879-8054
E-mail: yoshida@hus.osaka-u.ac.jp

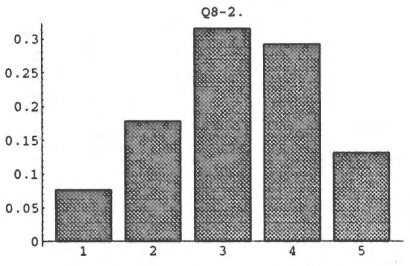


図 1: Q8-2. コンピュータが好きだ (文科系、賛成の程度)

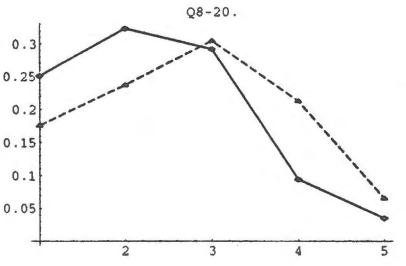


図 2: Q8-20. 卒業後はコンピュータを使用する仕事をしたい (文 (実線)・理 (点線) の比較)

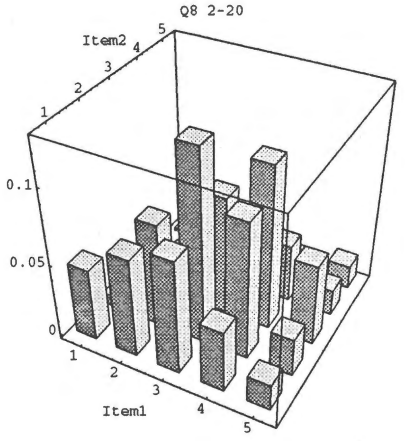


図 3: Q8 2(Item1)-20(Item2) のクロス集計 (文科系)

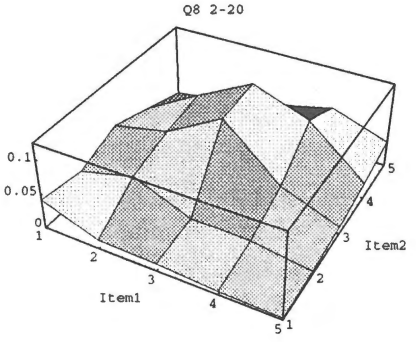


図 4: Q8 2-20 のクロス集計 (理科系)

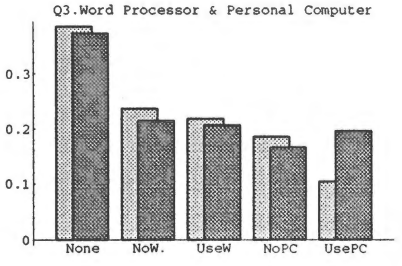


図 5: Q3. 自宅におけるワープロ (W)・パソコン (PC) の有無と使用 (文 (淡)・理 (濃))

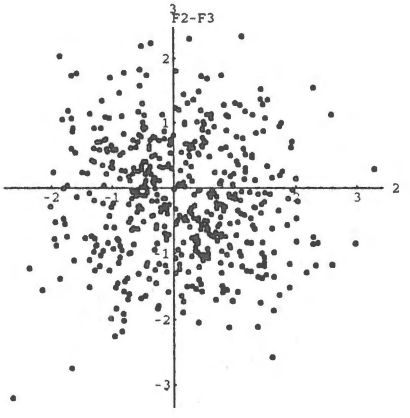


図 6: 主成分得点散布図 (文科系・F2-F3)

「間」に関するデータベースの構築
Construction of a database for '*Ma*'

中　村　敏　枝
Toshie NAKAMURA

大阪大学人間科学部，豊中市待兼山町１－１６
Faculty of Human Sciences, Osaka University
1-16 Machikaneyama, Toyonaka, Osaka 560

キーワード：時間的な「間」，音楽の「間」，スピーチの「間」，「間」と呼吸，「間」の感性

Keywords: Temporal '*Ma*', Musical '*Ma*', Vocal '*Ma*', '*Ma*' and respiration, Sense of '*Ma*'

あらまし：　日本文化の全てにおいて，「間」は非常に重視されてきた．したがって，芸談，文芸論，芸術論，武芸談として，「間」を扱った記述は多数存在する．しかし，断片的なものが多く，「間」について系統的な情報を与えてくれる文献は非常に少ない．また，従来の記述では建築や絵画などにおける空間的な「間」は文字と静止画で言及することが出来た．しかし，会話や音楽における聴覚的な「間」，ならびに，スポーツや舞踊などにおける運動的な「間」を具体的に示すことはできなかった．コンピュータを用いて，文字情報のみならず，映像･音響情報を取り込んだマルチメディアデータベースの作成が望まれる所以である．

既存の「間」のデータを収集し，文字情報，映像・音響情報による「間」の資料を提供して研究の便宜をはかるとともに，それらの資料に基づいて分析・測定と心理実験を行い，そこで得た物理的ならびに感性的な数値データをも盛り込んで，人間の生活場面における快適な「間」の提供に貢献できるようなデータベースを構築することが本研究の目的である．

ここでは，「間」の概念を整理し，時間的な「間」の実例と数量的データをマルチメディアデータベースにまとめる構想について述べる．

Summary: In all the cultures of Japan '*Ma*' has been of great importance. Consequently, as an actor's talk on his art, an essay on literary arts, etc., there are a large number of descriptions that deal with '*Ma*' qualitatively. They don't offer systematic information of '*Ma*', but fragmentary ones. They deal with '*Ma*' in constructions and art objects in letters and still pictures. However, auditory '*Ma*' in conversation and music as well as kinetic '*Ma*' in sports and dancing, have not been expressed in the actual state. These statements mentioned above are reasons for the desire of the construction of a multimedia database that adopts not only letter information, but also picture and acoustic information using a computer.

The purpose of this study is to collect multimedia information on '*Ma*', founded on existing data on '*Ma*', and then add the physical and psychological data obtained by experiment and measurement. This is used to prepare a database which can contribute to the design of a comfortable '*Ma*' in the daily life.

The arranged concept of '*Ma*', actual examples of temporal '*Ma*' and the idea of multimedia database construction are reported here.

1. はじめに

日本の文化はとりわけ「間」を重視してきた．白石（1979）は「間」を含む成句・複合語を150以上挙げているが，「間を合わせる」，「間がいい」，「間が抜ける」，「間伸びした」，「間が持たない」，「絶妙の間をとる」など，「間」に関する多くの言葉がある．その事実が示すように，対人関係，礼儀作法，スポーツ，文学，美術，演劇など我々の生活全般において「間」は重視されてきた．また，「間」の感性は日本人に特有のものというわけではない．国内外のあらゆる文化において「間」の重要性が認められるが，「間」を系統的に論じた文献は極めて少ない．種々の場において多様な形で「間」についての記述が断片的に現われるのである．

また，文字と静止画で表現できる，建築や絵画などの空間的な「間」と異なり，会話や音楽における聴覚的な「間」，ならびに，スポーツや舞踊などにおける運動的な「間」は文献の中で具体的に示すことが出来なかった．コンピュータを用いて，文字情報のみならず，映像・音響情報を取り込んだマルチメディアデータベースの作成が切望される所以である．

また，コンピュータ技術を駆使した現在の人工環境の中に，人間の感性に合わない不快な「間」が蔓延している（例えば，人工アナウンス）ことを鑑み，「間」の数量的なデータを提供することの必要性を痛感している．筆者は以前より「間」の数量的研究に取り組み，「間」の物理量と心理量の間の数量的な法則性を見い出してきた（中村，1987-1995）．そのような，「間」の物理量，感性量をデータベースに加えることによって，人文科学的な文化の研究に貢献するだけでなく，「間」が関わるあらゆる分野で応用的に役立つ資料を提供することが可能になると思われる．

国内外の文化における「間」のデータを集め，文字情報，映像・音響情報による「間」の資料を提供して研究の便宜をはかること，それらの資料に基づいて物理測定と心理実験を行い，そこで得た物理的ならびに感性的な数値データをも盛り込んで，人間の生活場面における快適な「間」の提供に貢献することをめざして，データベース作成の構想をたてた．

現在，下記項目のデータ収集に努めている段階である．

(1) 「間」に関する文献の収集

絵画，建築，音楽，演劇，話術，スポーツなどの分野において，芸談，文芸論，芸術論，武芸談として論じられている「間」の文献を集め，整理する．

(2) 「間」に関する映像データの収集

a. テレビ，映画，ビデオテープ，レーザーディスクなどのメディアによって提供されている映像の中から「間」に関係するものを探し，収集する．

b. 演劇，建造物，絵画などを撮影し，「間」の資料を集める．

(3) 「間」に関する音響データの収集

a. コンパクトディスク，ミニディスク，磁気テープなどのメディアによって提供されている音響の中から「間」に関係するものを探し，収集する．

b. 音楽演奏，演説，環境音などを録音し，「間」の資料を集める．

(4) 「間」に関する数値データの作成

上記 (2) ～ (3) によって収集した素材を用いて，実験ならびに測定を行い，「間」に関する数値データを作成する．

a. 物理データ　画像解析，音響分析により，「間」の物理量を測定する．

b. 心理データ　心理学実験によって「間」の感性を数量化する．

c. 物理データと心理データの間の法則性を追及し，感性情報入力によるデータ検索を可能にするための基礎データを作成する．

以上のうち，主として時間的な「間」について，マルチメディアデータベース作成のためのデータの用意が若干できたので，その一部を報告する．

2.「間」に関するテキストデータベース

2.1. 「間」の定義について

<u>データ例</u>

「--- これらの時空の間に，流動連続するダイナミズムの美意識ではなく，独特なる断絶によって創出される美意識を発見したか

らである．間はこの断絶によっているところが最も重要な条件だと考えられる．つまり間は時間的・空間的に切断された距離感が，独特の断絶によって創出される美意識だといえよう．」
（西山松之助，1983，p.118）

「間」の美意識というものは，何時，いかにして創出されたものなのであろうか．

データ例
「間という言葉，つまり芸の上での重要な条件として意識された用語として，それが文献に見えるのは，今のところ『教訓抄』である．この書は天福元年（1233）に南都楽人狛近真が書いた雅楽の本である．---その後，17世紀前半期の寛永・正保頃，しかもそれが柳生但馬守宗矩と宮本武蔵の兵法書に見えるのである．---以上のように，文献に見える間は，鎌倉中期にチラッと見え初め，以後400年の空白の後，江戸中期の寛永から明暦頃，つまり，17世紀の前半期から中頃にかけて，ほぼ日本の独自な美意識として成立したということが知れるのである．」
（西山松之助，1983，p.119-123）

「間」が日本文化において特別重要となった背景として，日本語のリズムが挙げられる．

データ例
「古典語にせよ近代ヨーロッパ語にせよ，長短とか強弱とか音節そのものに変化があるから，それを素材としてリズムを作ることができるが，日本語には長短も強弱もない．そのために間を置くことが必要になってくる．」
（別宮貞徳，983，p.82）

日本人の「間」に対する感受性は日本文化の根源としての言葉に由来するというのであるから，文字通りの空間的隔たりから時間的な意味，さらには形而上学的な形にまで「間」の概念は拡大し，生活，スポーツ，芸術，意識形態にまでわたる日本文化すべてに「間」が関与していて，定義が難しいこともうなづける．

2.2. 生活における「間」

データ例
「日本人の日常用語としての間は，本来空間的な距りを示すことばであり，住生活では居間，茶の間，応接間など，スペースと機能の両方を表現しているし，間取りは住居のスペースをどう使うかという配慮をあらわしている．床の間の間は，日本人の住生活で独特なスペースであり，そこは足で踏み入ることのできない，心理的にも隔離された空間としての間を感じさせる場でもある．しかし---それが単に住居空間だけではなく，そこに象徴的な意味が含まれていることに気がつく．」
（南　博，1983，p.9-10）

2.3. スピーチにおける「間」

末 利光はアナウンサーとしての30年の経験に基づいて，話ことばにおいて「間」の問題ほど大切なものはない と断言している．

データ例
「単語と単語のあいだの間，助詞の後の間，文節の間，節と節の間，挿入部の間，会話の前後の間など，えにいわれぬ間が文章の全てを支配しているのです．ですから，その間のとり方によってその人がどれだけ深く文章の意味を理解しているか．言わせていただけるならば，感覚のよし悪しから性格までが伝わってくるのです．間とはそういうものなのです．」
（末 利光，1991，p.40-41）

2.4. 音楽における「間」

音楽における「間」は，日本の伝統音楽について論ずる時には，最も重要な要素の一つとして必ず挙げられる．

データ例
「日本音楽では音のないところが問題で，それは休止符ではないわけですね．音のないところにちゃんと意味があるというか，生命があるというようなことですね．---音の長短というよりも“間”の長短ということを真剣に問題にする．その真剣さが日本ほどの国は，東洋にもないのではないかと思いますね．」

(吉川英史・神　正・河竹登志夫，1985，p. 22)

データ例

"間"の問題というのは，非常に大きな日本音楽の特性だと思います．例えばヨーロッパのオーケストラは油絵と同じで全部を塗りつぶして30分なり1時間の1つの世界を作るわけですね．それに対して日本音楽の場合は，特に能楽や尺八音楽などは水墨画や俳句などもそうだと思いますが，ぎりぎりの必要な本質，エッセンスだけを表現し，充実した空白を残しますね．」
(吉川英史・神　正・河竹登志夫，1985，p. 23)

「間」の表現は日本音楽においてのみ現われるわけではない．

データ例

「演奏者は"間"として表現すべき箇所は楽譜上の指示に拘束されることなく，聴取上適切な"間"と感じられる長さで演奏し，聴取者もその意図を理解する．リズム，拍子，テンポといった音楽の時間表現では説明できないもう一つの時間表現"間"が西洋音楽においても存在することを示すものであろう．」
(中村敏枝，1994a，p.155)

2.5. スポーツにおける「間」

例えば，相撲が現代の日本人に特別に愛好されるのは，相撲に「間」があるからではないかという．

データ例

「土俵入り，呼出しの声に応じて土俵にのぼり四股をふみ，塩をまき，蹲踞の姿勢で正対し，仕切をする．その儀礼的，美学的な形式性の中に，特に何回かの仕切直しのうちに，次第に戦意がみなぎり，緊張がたかまってくる．仕切りの阿吽の呼吸．そこのところが日本人にはたまらない魅力なのではないか．テレビの大相撲ダイジェストは便利だが，余分と見える仕切直しの部分が省かれているので，興味も興奮も半減する．一瞬の力動を前にする静の時間，その"間"のとり方，おたがいの心理のかけひき，まさに日本人の"間"好みに適っている．そして突張りの距離的な"間"，前まわし，差し手あらそいの"間"，攻めると引くときの"間"，投げをうつ瞬間の"間"のとり方，相撲ほど時間，空間，人間の"間"を表現してくれる競技はない．仕切の呼吸のとり方の微妙な"間"など，ほかのどんなスポーツでも，芸能でも味わうことができない．」
(奥野健男，1983，p.211-215)

2.6. 舞踊における「間」

舞踊家の「間」に関する芸談は多い．よく引用される例であるが，尾上菊五郎は彼の著書『芸』の中で九代目団十郎の言を伝えて「間」の重要性を述べている．

データ例

「踊の間と云ふものに二種ある．教へられる間と教へられない間だ．取分け大切なのは教へられない方の間だけれど，これは天性持って生れて来るものだ．教へて出来る間は間（あいだ）と云ふ字を書く．教へても出来ない間は魔の字を書く．私は教へて出来る方の間を教へるから，それから先きの教へやうのない魔の方は，自分の力で索り当てる事が肝腎だ」
(尾上菊五郎，1947，p.127)

2.7. 歌舞伎における「間」

狂言作者の竹柴蟹助は柝を入れる「間」の難しさを話した後，歌舞伎を「間の芸術」であると述べている．

データ例

「弁慶が花道にかかって『陸奥の国へぞ下りける』というと，段切れになります．そうしてトントンと足拍子を踏む，そして金剛杖をつくんです．この金剛杖と一緒にチョンとやらなくちゃいけない．これは見ていたんではできない．間なんです．長唄の段切れの間と役者の間と合せてチョンと柝をいれるんです．このひとつの柝で幕がすっとしまってしまうから，しくじったらたいへんです．」
(河竹登志夫，1985，p. 23)

2.8. 落語における「間」

寄席の芸における「間」の重要性については永井啓夫（1983）が詳しく論じている．落語は「間」の取り方ひとつで客の笑いを取れるか否かが決まる．

データ例

「　まがとれぬ----
落語家が自殺（昭和56年7月10日「東京新聞」朝刊）
芸に行きづまった春風亭一柳は，アパート10階の非常階段から飛び降りて，ついに自らの生命を断ってしまった．この引用は事件を報じる記事の見出しで，本文中には『2ヵ月ほど前から噺（はなし）の間の取り方に悩み，眠れないほどノイローゼにかかっていた』という遺族の談話がある．」
（蒲生郷昭，1983，p.132-133）．

桂ざこばが「桂朝丸」と名乗っていた頃の芸の特徴は早口ということであった．

データ例

「うまくお客の波長と合った時には，笑いの爆発が起こるのだが，ひとたび外れると，客とは平行線を描いたまま接近することなくサゲまで行ってしまうことがあった．このころの演芸評論家は，この点を問題にして『朝丸の落語には余裕がない．間がとれていない』と評した．---しかし，---朝丸がざこばになる1年ぐらい前から，確実にテンポの緩急を身につけ，独特の間で笑いをとり出した---」
（小佐田定雄，1991，p.104- 105）．

2.9. 「間」のゆらぎ

芸術においては「間」の合うことが重要であるとともに，基準になる「間」からのゆらぎが必要である．

データ例

「『絶妙の“間”』と呼ばれる芸術的・名人芸的な“間”は基準値からの微妙なずれによって作り出されるのである．必ずしも最適の“間”をとらず，聴取者の予期する“間”から微妙にずれることによって生ずる緊張，弛緩，意外性の上手な組合わせの効果，それが芸術的・名人芸的な“間”ではないだろうか．」
（中村敏枝，1994a，p.168）

データ例

「間には，通常，三味線音楽のリズムに乗って行く間，というのがある．これがなければ舞踊芸術は成り立たないわけで，これは，いわば，基準になる間である．これに全く合わない間は，度外れな間で，芸にも何にもならない．しかし，この基準の間に，あまりぴったり合いすぎると，味もしゃしゃりも無くなってしまう．つまり，こういうのを芸にならないというのであって，舞踊は，間が基準の芸術であることは言うまでもないが，ただ間に合っているだけでもいけないので，そういう間を常間（じょうま，定間とも書く）と言って斥ける．間が基準の芸術でありながら，常間に踊ってはいけないというところが，日本的というか，日本舞踊のむつかしい，奥深いところなのだ．」
（川口秀子，1983，p.169）

2.10. 「間」の科学

以上述べてきたように，「間」に関する定性的な論述は限りなくある．しかし，定量的には全く論じられていない．

データ例

「伝統芸道における「間」とは，---機械的に測れるものでも，数量に換算されるものでもなく---」
（生田久美子，1987，p.62-63）

データ例

「それは休止，沈黙，間隙のような物理的，生理的な過程でなく，形やことばでは表現できない体験であり---」
（南　博，1983，p.19）

データ例

「理屈にあわない，理屈で割りきれないからこそ，間なのだ．」
（川口秀子，1983，p.168）

以上のように考えられてきたのである．たしかに，「間」は量的に掴み所がないように思え

る．しかし，「間」を定量的に扱うことが全く不可能というわけではない．筆者は種々の音声・音楽を用いて心理実験を行った結果，「間」の定量的研究が可能であることを確認した．

データ例

「--- 刺激条件が一定であれば，丁度よいと感じられる“間”の長さは常に安定した値を得られること，刺激条件の変化に伴って「間」の長さも系統的に変化し，“間”の一般的な法則性を追求しうることを本実験結果は示した．」

（中村敏枝，1983，p.168）

2.11. 「間」と呼吸

「間」について語られる時の呼吸・いきという言葉は比喩的に使われる場合も多いのであるが，文字通り生理現象としての呼吸が合う様を，武智鉄二の体験談の中に見ることができる．

データ例

「僕は古靭が（義太夫を）語っているのを聞いているうちに，呼吸をつめることを自然に覚えてしまったんです．---それを覚えてしまうと，呼吸をつめることが習慣になっちゃうのね．で，こっちの呼吸が向こうの呼吸に合わないとか，あっちの呼吸が抜けてくると，こっちは気分が悪くなるんです．生理的に呼吸をつめてるわけですから，それがさっと呼吸をすかされると，こっちのつめた呼吸のいき場がなくなって，それで気分が悪くなるんですね．」

（坂東三津五郎・武智鉄二，1972，p.49）

この話を受けて坂東三津五郎も，次のように語っている．

データ例

「それは義太夫に限らず踊りでも芝居でも，ほんとうに呼吸をつめてやってられると，こちらも呼吸がつまってくるし―．」

（坂東三津五郎・武智鉄二，1972，p.49）

筆者は音楽聴取時の呼吸，演奏時の呼吸を実際に測定する実験を行い，「間」との関係を調べた．

データ例

「その結果，“間”の伝達の背後に呼吸が存在することが確認された．“間”を意図した長い音符や休符の箇所で聴取者の呼吸が同期する傾向のあることを見出した．演奏者は演奏に合わせて呼吸をし，聴取者は呼吸を演奏音に合わせる．呼吸を合わせることによって，演奏者と聴取者の間に意図の伝達が成立する．そのような，呼吸を媒介とした演奏者―聴取者間の関係を本研究結果は示唆した．」

（中村敏枝，1992，p.177）

3. 「間」のマルチメディアデータベース

3.1. 例1 フランツ・シューベルト作曲「Du bist die Ruh」（君こそわが憩い）における「間」

文献データ

「作曲家自身，どんな音をどう持ってきても表せないような間に，譜面上＜休符＞として空白にしている場合がありますが，これなどは，“間”の最たるものだと思います．その代表的な例としてフランツ・シューベルト（Franz Schubert 1797～1828）の「君こそわが憩い」"Du bist Die Ruh"があげられます．---歌曲の王といわれる天才シューベルト26歳の作品です．シューベルトが1小節全休符を置いて，全く音を入れることなく最高潮に達した感情を「間」にゆだねた部分は，---急速に高揚していく感情を間に託しています．そこにはシューベルトがこれ以上感情の高揚を音にする必要がないと判断し，音を超えた音が，感動となって空間に広がって行くのです．そこは歌い手と伴奏者と聴衆が一つになって感動にひたる空間の世界なのです．---これほど率直に，なりふり構わず間を自分のものに取り入れてしまった西洋音楽を，私はほかに知りません．」

（末 利光，1991，p. 192-195）

文献データ

「この曲をフィッシャー・ディスカウがどのように歌っているかを調べるために演奏音

を測定した．Fig. 1 に第 50 小節から第 77 小節までの音の波形を示す．問題の休符部分（矢印）は 2.6sec（第 61 小節）と 3.0sec（第 75 小節）であった．ほぼ一呼吸の“間”を取っていると思われる．ここに示した部分は曲中のクライマックスで，最も高らかに歌い上げている部分である．クレッシェンドに次ぐクレッシェンドで高揚の極みに至った感情は次の“間”によって，深く心に浸透する．音によって創り出された緊張と感動の境地が，音の無い“間”によって更に高められ，心に深く定着するのである．ディスカウは繰り返しの二つの“間”の表現を巧みに変化させている．一回目は第 60 小節の *ff* が創る緊張と感動をさらに高め定着させる“間”，二回目は第 73 小節の *ff* から第 74 小節の *pp* への急激な変化ならびに限界にまで抑えた音量が生み出す緊張と感動を定着させる“間”である．最大の音量と最小の音量，音の創り出すどちらの緊張も，続く無音の存在があってこそ無限の感動の境地に聴衆を導くことができるのである．それは正に日本人が“間”と呼んでいるものにほかならない．」
（中村敏枝，1992，p. 173）

音響データ

下記 CD による．

シューベルト／“美しき水車屋の娘”
東芝EMI（株）
DIETRICH FISCHER-DIESKAU

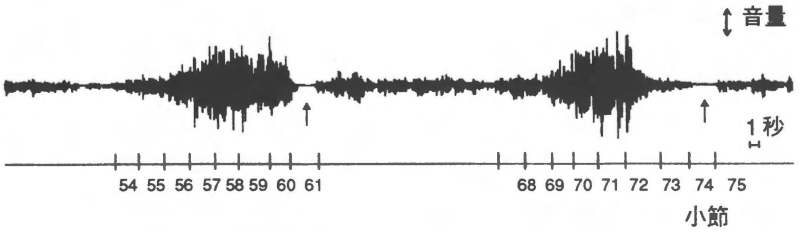


Fig. 1 シューベルト作「君こそわが憩い」（フィッシャー・ディスカウ）における「間」（矢印部分）

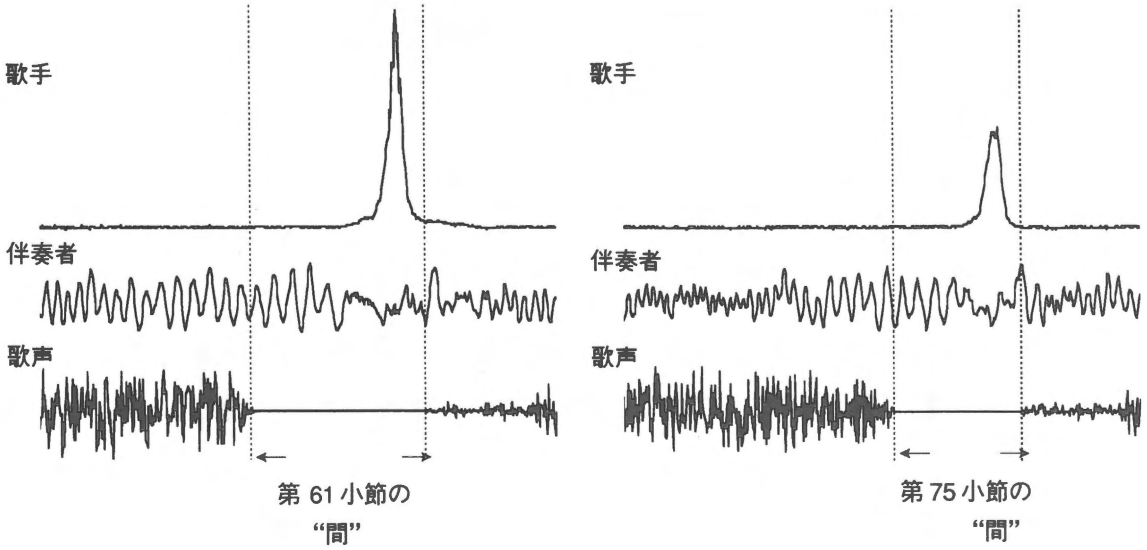


Fig. 2 「君こそわが憩い」を歌唱中の歌手の呼吸ならびに伴奏者の呼吸

シューベルト歌曲集 II
(株) 徳間ジャパンコミュニケーションズ
PETER SCHREIER
シューベルト歌曲集 東芝EMI (株)
ELISABETH SCHUMANN

静止図データ

a. 「Du bist die Ruh」楽譜.
b. 物理データ：上記音響データの物理的分析結果. Fig. 1 に一例を示す（中村敏枝, 1992, p. 175）.
c. 心理データ：例えば, “間” の効果に関する聴取実験データ（中村敏枝, 1994d, p. 114）.
d. 呼吸測定データ：例えば, 演奏者と伴奏者の呼吸の同期に関する実験結果. Fig. 2 に一例を示す（中村敏枝, 1994c, p. 5）.

数値データ

上記実験・測定結果の数値.

3.2. 例2 能「石橋」における「間」

文献データ

「例外的と思われるほどに長い “間” だけれども, しかしそこに日本人の “間” に対する考え方があるように端的に表われているのが, 能「石橋」の後シテの獅子が出る前のお囃しの “間”, かなり激しい小鼓や大鼓や笛や太鼓がぱっと止まった瞬間から長い音の空間がある. そのときの緊迫感というものは音がないけれども, その効果からいえば 100 人のオーケストラがトウッティ (総奏) で演奏した音以上に緊張感を与えるものだと思うんです. その長い空間のなかでポンと出るあの音は, 何か深山幽谷で露のしたたる音からヒントを得たといいますけど, ああいうものは恐らく世界中にないと思いますね.」
(吉川英史・神 正・河竹登志夫, 1985, p. 22)

文献データ

「音の出る “間” を分析すると, 興味深い結果がえられた. Fig. 3 に示すように, ここでとられている “間” は, 予想通りほぼ 3 sec の倍数 (約数) のみである. 一番長い約 12 sec の “間” も一呼吸の 4 倍と考えることができる.」
(中村敏枝, 1994a, p. 164)

映像データ

NHK ビデオ「石橋」 古式 金春流
シテ 白獅子 金春信高
ツレ 赤獅子 金春安明
ワキ 寂昭法師 岡 次郎右衛門
笛 一噌仙高
小鼓 北村 治
大鼓 安福建雄
太鼓 小寺俊三
ほか

音響データ

上記に同じ

静止図データ

上記音響データの物理的分析結果 (Fig. 3).

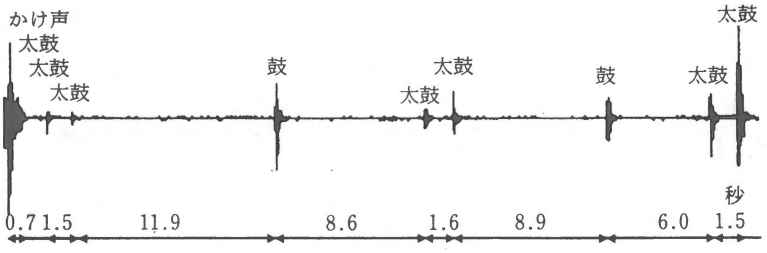


Fig. 3 能「石橋」における「間」

3.3. 例3　徳川夢声の朗読における「間」

文献データ

「徳川夢声老の話術は“マの芸術”だといわれている．週間朝日の編集長をしていたころ，よくお供して『問答有用愛読者大会』というのに出かけたが，見ていて，なるほど，これが“マ”というものかとしみじみ感じた．」

（扇谷正造，1977，p. 29）

文献データ

「そのむかし徳川夢声さんという話の大家がいましたが，その方は常に『ことばは間です．間がいちばんです．いちばんいいのは何もしゃべらないことです』といって笑わせましたが，確かにその通りだと思います．」

（末　利光，1991，p. 18-19）

音響データ

カセットテープ「宮本武蔵」原作・吉川英治，朗読・徳川夢声
企画・販売／日本通信教育連盟
製作・製造／ポリグラム株式会社

静止図データ

a. 物理データ：上記音響データの物理的分析結果．
b. 心理データ：例えば，「間」の効果に関する聴取実験データ（中村敏枝，1995，未発表）．

4. 「間」の文献リストデータベース

データ例は後掲文献欄参照．

5. おわりに

「間」に関するデータベースを構築するために，これまでに収集してきた種々の形式のデータの一部を紹介した．現在，「間」のデータベースはマルチメディアの技術を駆使する必要があることと，その具体的方法の可能性を確認した段階である．今後，デザインを検討し，具体例を作成して，データベース作成の準備態勢を整えたい．一般公開できるようなデータベースを提供するまでには，構成データの著作権の問題や如何にして多数のデータを収集するかなど解決しなければならない問題が残っている．

6. 文献

坂東三津五郎・武智鉄二　1972　芸十夜　駸々堂出版．

別宮貞徳　1983　日本語のリズムと間　間の研究（南　博編），講談社，75-94.

福田　精　1983　運動姿勢と日本人の間　間の研究（南　博編），講談社，39-58.

蒲生郷昭　1983　日本音楽の間　間の研究（南　博編），講談社，131-152.

郡司正勝　1979　おどりの美学―間・移り　いき・いなせ・間（南　博編），現代のエスプリ，141，181-184.

袴田大蔵・中野八十二・志沢邦夫　1976　剣道の間合に関する一考察　武道学研究，9，63-65.

浜口雅行・林　邦夫・白藤一郎・堀山健治　1978　手の内と呼吸相から見た応じ技の分析　武道学研究，11，9-10.

池田守利　1977　空手道の突き動作と呼吸調整との関連について　武道学研究，10，25-29.

生田久美子　1987　「わざ」から知る　東京大学出版会．

石黒節子　1979　おどりの間と呼吸　いき・いなせ・間（南　博編），現代のエスプリ，141，188-195.

金木　悟・松原　章・古川　陵　1978　上段からの正面打撃時における呼吸作用の研究　武道学研究，11，1-2.

川口秀子　1983　日本舞踊の間　間の研究（南　博編），講談社，167-182.

河竹登志夫　1984　竹柴蟹助氏にきく　日本の美学，1 (2)，ぺりかん社，17-24.

吉川英史・神　正・河竹登志夫　1985　日本音楽をめぐって　日本の美学，1 (4)，ぺりかん社，10-26.

小山　哲・林　邦夫・堀山健治・浜口雅行・白藤一郎　1978　呼吸相から見た剣道形の分析　武道学研究，11，15-17.

前林清和　1987　近世武芸における「呼吸」と「気」の諸問題　武道学研究，20，51-61.

南　博　1983　序説―間とは何か　間の研究（南　博編），講談社，7-20.

永井啓夫 1983 寄席の芸（間へのアプローチ） 間の研究（南 博編），講談社，205-220.

中村敏枝 1987 演奏時間における楽譜からの逸脱ー“間”の知覚との関係ー 日本音響学会聴覚研究会資料，H-87-32，1-6.

中村敏枝 1989 音楽、音声における“間（ま）”について 日本音響学会音楽音響研究会資料，MA88-22，24-29.

中村敏枝 1992 コミュニケーションにおける“間”の科学と情報処理ー“間”のもたらす意味とコミュニケーションー IMAGE2-'92ーイメージスクェア'92ー，77-86.

中村敏枝 1992 作曲者・演奏者・聴取者の間の意図の伝達ー演奏音の物理測定と聴取の心理測定に基づいてー 博士論文（大阪大学）.

中村敏枝 1993 “間（ま）”の感性に関する心理学的研究 電子情報通信学会技術研究報告，92，37-42.

中村敏枝 1994a “間”の感性 感性情報処理，オーム社，151-169.

中村敏枝 1994b 「間」の心理学 ゆらぎの人間科学，平成6年度大阪大学放送講座，77-92.

中村敏枝 1994c 音楽における「間」と呼吸について 日本音響学会音楽音響研究会資料，MA94-16，1-8.

中村敏枝 1994d メディアにおける“間”の心理学的研究 感性情報の情報学・心理学的研究文部省科学研究費 重点領域研究 平成5年度成果報告書，111-114.

中村敏枝 1995 「間」における演奏者と伴奏者の呼吸の同期 日本心理学会第59回大会，104.

西山松之助 1983 間の美学成立史 間の研究（南博編），講談社，115-130.

小倉 朗 1977 日本の耳 岩波新書.

奥野健男 1983 “間”の構造 文学における関係素 集英社.

尾上菊五郎 1947 藝 改造社.

大保木輝雄 1984 武芸における気論に関する諸問題（その2）—間との関連から— 武道学研究，16，64-65.

扇谷正造 1977 “マ”ということ 季刊「歌舞伎」，9 (4)，29-30.

小佐田定雄 1991 上方落語・米朝一門・おさだまり噺 弘文出版.

白石大二 1979 間を含む成語の辞典 いき・いなせ・間（南 博編），現代のエスプリ，141，196-207.

末 利光 1991 間の美学—日本的表現 三省堂選書.

田島東海男・坪井三郎 1978 呼吸パターンによる正面打撃動作の相違—竹刀速度・動作・筋作用の相違— 武道学研究，11，3-4.

田中 守 1982 剣道における「間」について—近世武芸伝書を中心に— 武道学研究，15，31-32.

徳丸吉彦 1983 間は拍子かリズムか 間の研究（南 博編），講談社，95-114.

利倉幸一 1977 七世三津五郎 舞踊芸話 演劇出版社.

坪井三郎 1973 剣道における動作と呼吸の研究—正面打撃・各稽古中の呼吸波形— 体育学研究，18，23-29.

方言音声データベースの作成と利用に関する研究

A Study on Making and Using Speech Corpora of Dialects

田原広史、江川清、杉藤美代子、板橋秀一
Hirosi TAHARA, Kiyoshi EGAWA, Miyoko SUGITO, Syuichi ITAHASHI

ＪＣＭＤ作成委員会、大阪樟蔭女子大学日本語研究センター内
Committee of Making Japanese Speech Corpora of Major City Dialects,
in Japanese Language Research Center of Osaka Shoin Women's College,
4-2-26 Hishiyanishi, Higashi-Osaka-City 577 JAPAN

キーワード：韻律的特徴、主要都市方言
検索、言語学

Keywords: prosodic features, major-city-dialects, searching, Linguistics

あらまし：この研究は、全国１３主要都市約２５０名の方言音声データをデータベース化するための作業および流通化の方策についての研究である。研究の要点としては、1)「方言音声データベース」を作成すること、2)検索、分析のためのツールを開拓あるいは開発すること、3)当該分野における利用者を開拓し、利用のためのルール作りをおこない、そのルールに基づいて流通化を促進していくこと、以上の３点が柱となっている。

Summary: This Research is based on one of the results of "Integrated Studies on Prosodic Features of Current Japanese Language with Application to Spoken Language Education", funded by "Grant-in-Aid for Scientific Research on Priority Areas by Ministry of Education, Science and Culture", 1989-1992. The results are thousands of recorded DATs(Digital Audio Tapes), which contain vocal sounds of approximately 500 items, such as words, sentences, short-story-readings, a set of Japanese phoneme, numbers, etc. The speakers are selected from 13 Japanese major cities, about 70-100 people per city, 5 old-males, 5 old-females, 5 middle-males, 5 middle-females, 5 young-males, 5 young-females, 10 junior-high-males 10 junior-high-females, 10 elementary-males 10 elementary-females. We are now making speech corpora from this data from 1993, named "Japanese Speech Corpora of Major City Dialects" funded by "Grant-in-aid for Database-making by MESC". This project continues to 1997, and we will make two types of Speech corpora, one is reading of Weather Forecast Report(about 1 minute per speaker) made of Compact Disk, the other is word-reading, made of CD-Rom. Under these conditions, we take the next three aims in our study. 1) making this corpora more complete, 2) developing software to search and analyze the data, 3) increasing the number of the users by advertising our study, and making rules of utilization.

1．研究の背景

・研究の前身

この研究は、重点領域研究「日本語音声における韻律的特徴の実態とその教育に関する総合的研究」（平成元年～４年度、代表者杉藤美代子、以下「日本語音声」）の中で収集された全国各地の方言音声資料を整備（データベース化）し、より効率的な利用、流通を目指すものである。

「日本語音声」期間中に収集された音声資料のうち大きなものとしては、「全国共通項目調査」と「主要都市調査」と呼ばれる二つがある。「全国共通項目調査」では、単語、文、文章、五十音、数字など約1000項目に及ぶ項目を、全国100地点の高年齢層話者各１名についてディジタル録音したものである。「日本語音声」期間中に１９枚のＣＤと３枚のＣＤ-ＲＯＭとして刊行された。「主要都市調査」では、１３主要都市（札幌、弘前、仙台、新潟、名古屋、東京、富山、大阪、高知、広島、福岡、鹿児島、那覇）において、一都市につき、５世代男女計７０名、約500項目についてディジタル録音された。

この資料に関しては、「日本語音声」終了後、平成５年度より新たに成果公開促進費（データベース科研）を受け、現在「日本主要都市方言音声データベース」（同作成委員会）として５年計画でＣＤ、ＣＤ-ＲＯＭ化のための編集作業をおこなっている。平成５、６年度でＣＤ各２枚、計４枚を刊行し、７年度はＣＤ１枚およびＣＤ－ＲＯＭ１枚を刊行の予定である。

・音声データベースをとりまく環境

音声データの編集については、上記「データベース科研」において鋭意作業中であるが、音声そのものを収録するＣＤと異なり、ＣＤ－ＲＯＭ化にあたっては二つの問題が生じた。それは、音声ファイル形式の問題と、検索システムの開発の問題である。ＣＤはＣＤプレーヤがかなり普及しており、一般研究者でも使える状況にあるが、ＣＤ－ＲＯＭはパーソナルコンピュータがなければ分析はおろか聞くことさえできない。「日本語音声」期間中はＣＤ－ＲＯＭの作製はおこなったが実際に音声を聞いた人はほんの一握りに過ぎず、このような研究を進める環境になかった。

ところが、それから４～５年の間に飛躍的にパソコンの普及が進み、ＣＤ－ＲＯＭドライブを標準搭載したパソコンも出回ってきている。研究のための環境は整ってきたといえる。そのような状況の中で、平成７年度重点領域研究「人文科学とコンピュータ」の一公募班として本研究はスタートした。現在、「方言音声データベース」のよりいっそうの整備、より汎用性のある音声データ形式の模索とデータ変換、７年度予算で購入したマッキントッシュによる検索ツールの開発などに取り組んでいる。

・流通に関する試み

また、本作成委員会が所在する大阪樟蔭女子大学日本語研究センターが中心となり「西日本国語国文学データベース研究会」（ＤＢ-Ｗｅｓｔ）を開催している（年２回、平成７年１２月で７回目を迎えた）。この研究会は国語国文学分野におけるデータベースに関連するノウハウの啓蒙、研究、発表をおこなっているのみならず、データベースに関する情報交換の拠点となっており、作成中のデータベースに関しても流通化のためのルール作り、モニター利用の試み等をおこなっている。

2．研究の目的

このような研究背景をふまえ、本研究では研究の目的として次の三つを設定している。

1)「方言音声データベース」そのものを
　より整備されたものにすること。
2)検索、分析のためのツールを開発すること。
3)当該分野（言語学、音声学、国語学、日本

語教育学等）における利用者を開拓し、利用のためのルール作りをおこない、流通化をよりいっそう促進していくこと。

この３点について研究を進めている。進め方は1)2)3)の順にステップアップしていくのではなく、1)2)3)同時進行で進めることが必要である。その理由は、それぞれの段階が密接に関連しており、フィードバックをおこなうことによって、データベースそのものもよくなるし、使用環境も整備されていくと考えるからである。1)に関しては上に述べたとおり、別途データベース科研を受け、編集作業をおこなっているが、製品化（主にＣＤ-ＲＯＭ化）するに当たって、試作品の作成、手直し等の研究を本研究においておこなっている。

・音声データベースの現状

この分野における「音声データベース」は、上記「日本語音声」において作成されたものが始めてであり、きわめて立ち遅れた状況にある。日本語の音声研究・音声教育では、抽象化した音声的特徴を実際の発音と結びつけ、かなりの音声情報を捨て去った形で研究が進められてきた。

現在では、より生の音声に近いものを対象とした、実験音声学、音響音声学の分野が見直されつつあるが、これには、従来の研究に飽きたらず、意欲的に新しい分野に踏み出して行った研究者、教育者たちの努力によるところが大きいことに加え、ハードウェアの面で音声技術、情報工学の飛躍的な発展があったことも忘れることができない。

近年のディジタルオーディオ技術の進歩によって、高品質の録音資料の収集が手軽にできることになったことは言うにおよばず、ＤＡＴ、ＣＤ、ＣＤ-ＲＯＭのような媒体の登場によって、音声が半永久的に劣化しない形で保存でき、さらに進んでパーソナルコンピュータの普及によって検索等も飛躍的に簡単におこなえるようになった。

本研究では、このような時代の流れの中で、高品質の音声データベースを、全国の研究者が容易な形で利用できるようＣＤ、ＣＤ-ＲＯＭの形に整備し、保存、管理、流通の方法を含め、当該分野における音声データベースというものを総合的に研究している。この研究により、今後、当該分野の音声データベースに関して、作成、利用、流通などを含め、水路づけがなされることになると考えている。

近年、マルチメディアを合い言葉に世界のコンピュータ事情は一変つつあるが、日本でも音声自動認識の性能比較を重要な目的として、音声データベースの検討が続けられ、単語音声についてはＪＥＩＤＡ日本語共通音声データやＡＴＲ音声データベースが公開されている。

ただし、これらはいずれも音声情報処理の分野での利用を前提としたものであり、共通語を対象とした「正しい日本語」の「単語読み」である。したがって、音声の韻律的特徴の実態の把握や、教育への応用については、まったく考慮されていない。この研究で扱っている「方言音声データベース」は、日本語方言の韻律的研究を前提に収録されたものであり、その点で一線を画している。

３．研究の現状

この研究で扱う分野は、大きく次の４つに分けられる。

1)検索用テキストデータの入力、整理、データベース化
2)音声信号データの編集、評価、製品化
3)検索用ツールの開拓、開発
4)流通化に関する調査研究

1)検索用テキストデータの入力、整理、データベース化

検索用テキストデータには、「発声内容に関するデータ」（読み、表記、アクセント型など）と「発声者に関するデータ」（話者の年齢、性別、出身地など）の二つがあり、平成７年度までに「発音内容に関するデータ」

１３地点分、総計８４２８項目、「発話者に関するデータ」約１２００人分について、すべての情報の電子化を終えた。この作業には、市販の日本語データベースシステム『桐』（管理工学研究所）を利用している。

音声データと連結して検索作業をおこなうためには、これらの入力されたデータの整備、改良、実際の検索作業に向けての試行錯誤が必要であり、現在はこの作業を中心におこなっている。

2)音声データの編集、評価、製品化

・文章項目のＣＤ化

現在、データベース科研により編集中である。方法は次のとおり。

1)ＤＡＴに録音された音声資料から目的の部分を別のＤＡＴにダビング編集し、すべての収録者（１地点70人から100人）の音声について検聴をおこなう。

2)機械雑音、環境雑音、読み間違えの回数、声質、方言の程度などについて評価をおこなった結果から、最終的に20人について、ＣＤ化する音声をピックアップする。

3)元テープに帰って採用された音声を再度ダビング編集し、ＣＤ作製業者に送る。その際、ＣＤのレーベル、リーフレット、トレイカードのデザインもおこなう。

4)業者はこの音声を一度アナログに変換した上、左右チャンネルのバランス、全体の音量を調整、マスターテープ、原盤を作製し、ＣＤにプレスする。

成果の一部としてこれまでに、天気予報の朗読文章を以下の４枚のＣＤ（音楽用ＣＤ）として発表しており、今年度も引き続き出していく予定である。

『天気予報 Vol.1 富山市・大阪市』
『天気予報 Vol.2 高知市・福岡市』
『天気予報 Vol.3 名古屋市・仙台市』
『天気予報 Vol.4 札幌市・弘前市』

・短文、単語項目のＣＤ－ＲＯＭ化

次の段階として、文章以外の項目、短文、単語などについて、ＣＤ－ＲＯＭとして実用化する計画を立てている。このための作業は平成５年度から進めているが、具体的な手順を以下に示す。

1)上記ＣＤに採用された人についてのみ編集をおこなう。ＤＡＴからディジタル信号のまま、テープの最初からパソコンに取り込みファイル化する（１ファイル３ＭＢ程度）。

2)１人分終わったら、３０程度になったファイルを再度一つづつ読み込み、短文あるいは単語単位に編集し、それぞれファイルとして書き出す（１人分１日６時間で２～３日かかる）。

3)１人分が終了したら光ディスクに書き込む。

この作業の繰り返しである。現在、編集作業はＮＥＣ製のパーソナルコンピュータで、編集用ソフトウェアは、『音声工房』（ＮＴＴアドバンステクノロジ社）を使い、16ビット、16KHｚで編集している。

ＣＤ－ＲＯＭ化にあたっては、このようにして編集してきた数多くの項目（１人の話者につき500発話以上）から、どの項目をデータベース化するかについての検討、個別の音声の評価、ＣＤ-ＲＯＭ内におけるファイル構造の検討などをおこなう必要がある。その成果を、今年度中に、一枚の試作品ＣＤ－ＲＯＭとして発表する予定である。

3)検索用ツールの開拓、開発

検索のためのツール開発に関しては、「日本語音声」期間中に作成されたＣＤ-ＲＯＭ用に開発した検索プログラムがあるが、汎用性がまったくないものである。そこで、このプログラムの設計思想はこのまま生かし、汎用性のあるものに全面的に作りなおす計画を持っている。ただし、もちろんこの目的に合った検索ツールがすでに市販されていれば、それに越したことはないので、ソフトウェアの開拓をおこなっていく必要もある。

パソコンの機種戦争はまだまだ続くものと思われ、単一機種用でしか使えないようなシステムは望ましくない。少し前までは、音声

ファイルと検索用データベースを連動させた形で利用するという目的には、マルチメディア性が高く、インターフェイス、ソフトウェアが充実しているアップル社のＭａｃｉｎｔｏｓｈが有利であった。ところが、ＤＯＳ／Ｖマシン、ＮＥＣといったＷｉｎｄｏｗｓ側も９５の発売と共にいっそうマルチメディア性が高まっていくと考えられ、今後色々なソフトウェアが発売されていく可能性もある。

このような現状を考え、音声データ形式については、Ｍａｃ－ＯＳ、Ｗｉｎｄｏｗｓがにサポートしている形、具体的にはＷＩＮＤＯＷＳ(.ＷＡＶ)形式を採用することにした。この形式であれば、世界中のほとんどすべてのパーソナルコンピュータで再生可能であろう。

今年度の作業は、これまで３年間にわたって編集してきた大量の音声データファイル（約100人分、ファイル数約50000個、容量約３ＧＢにものぼる）をＷＩＮＤＯＷＳ形式に変換する作業を中心におこない、検索ツールの開発については十分な時間がさけなかった。今後は変換されたファイルを用いて具体的な検索ツール作成、実用化していく予定である。

4)流通化に関する調査研究

データベース科研により作成したＣＤをモニター（データベースを使用、評価してくれる人）に配布し、利用方法、利用状況など流通化に関する調査研究をおこなっている。現段階でのモニターの数は１３０名程度であり、平成７年度は使用目的、音質、話者の妥当性などについてのアンケートを実施した。

現段階におけるこのデータベースに関する規程は、以下のような簡単なものである。

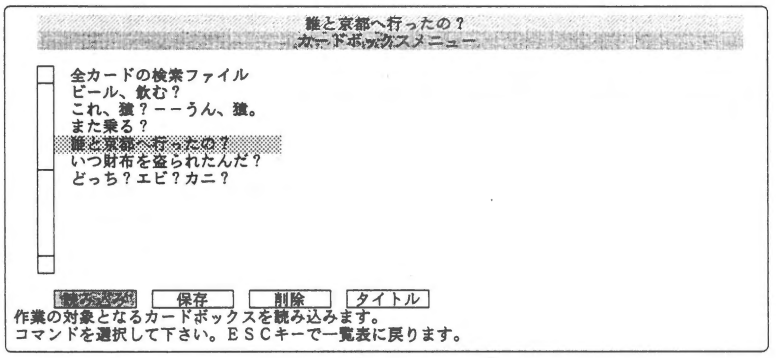
・個人的な使用の範囲を越えてダビングをおこなわないこと。
・研究等に利用した場合は、論文中にその旨を明記すること。
・研究、教育以外の利用はしないこと。
・これらの取り決めを伝えた上で、別の人が利用することは構わないが、モニターアンケート時に報告すること。

現段階における配布媒体はＣＤのみであるが、今後はＣＤ-ＲＯＭに関しても同様の形で調査研究を進めていく予定である。

４．今後の研究

上に述べてきたことを、引き続き進めていく予定である。現状では、まだ研究ではなく作業が中心と言えるかも知れない。研究の目的、計画はほぼ固まっているので、今後はそれぞれの分野を進めながら、適宜フィードバックをおこない、研究を発展させていきたいと思う。

検索結果

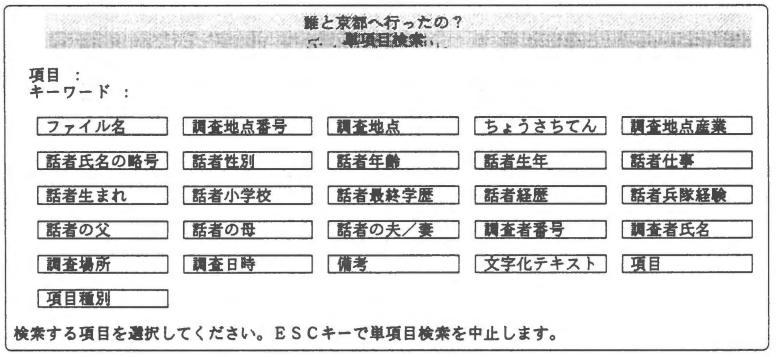


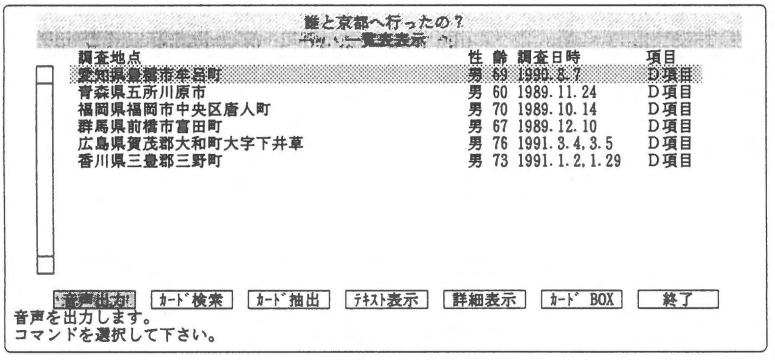
← 基本メニュー

種々の検索項目を登録し、メニュー形式で検索できます。
ここでは「誰と京都へ行ったの？」という単文音声を選択しました。

単項目検索メニュー →

「誰と京都へ行ったの？」という項目からさらに右の話者情報によって絞り込みができます。



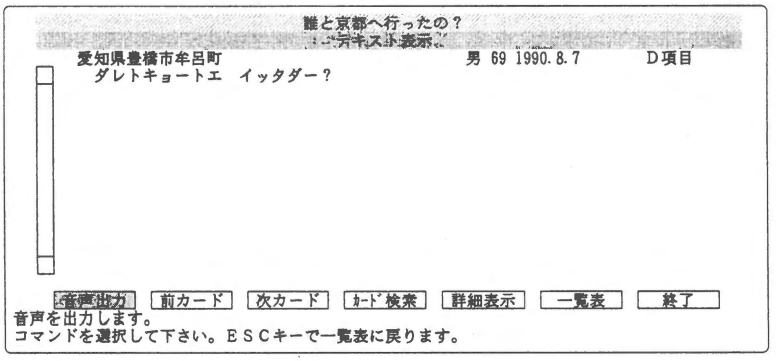


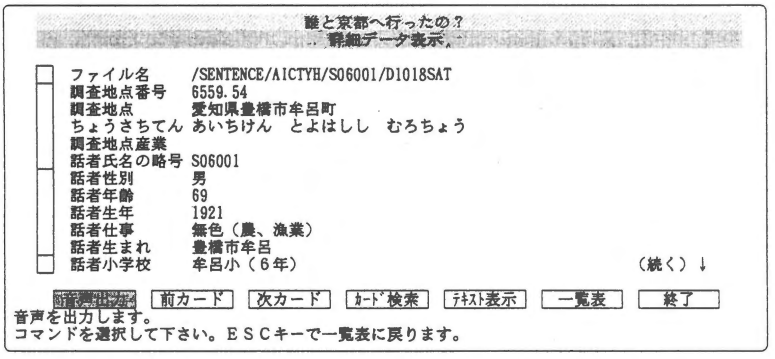
← 選択した結果一覧

ここでは６、７０歳代の男性を選んでみました。
この中からさらに選択し、音声を出力します。
ここでは愛知県の男性を選びます。

テキスト表示画面 →

発話したテキストを表示します。
長い文章だと、このボックスいっぱいに表示されます。
テキストを見ながら試聴することもできます。





← 詳細データ表示画面

話者のさらに詳しい情報、その町の産業、録音した時の状況などを見ることができます。

公開シンポジウム「人文科学とデータベース」プログラム

日時:平成７年１２月２５日(月)、１２月２６日(火)
場所:大阪電気通信大学・寝屋川キャンパス Ｂ３１０大教室

２５日午前　特別講演
「古地震データと活断層」　寒川　旭(地質調査所大阪地域地質センター)

２５日午後　一般講演
○「IntelligentPadシステムを用いた歴史学研究支援データベースの構築」
　赤石美奈，中谷広正，伊東幸宏，阿部圭一，田村貞雄(静岡大学)

○「４次元歴史空間システムにおける地理情報処理について」
　小林努，加藤常員，小沢一雅(大阪電気通信大学)

○「視点に依存する属性付け機構を持つ木簡研究支援システム
　ー構造進化型データベースの概念ー」
　森下淳也(姫路獨協大学)，上島紳一(関西大学)，大月一弘(神戸大学)

○「古典籍とＪＩＳ漢字」當山日出夫(花園大学)

○「手書き文字時系列筆跡パタンの一解析と今後の計画」
　東山孝生，山中由紀子，澤田伸一，中川正樹(東京農工大学)

○「絵画ＤＢとイメージ検索 ー浮世絵の線画表現とデータ圧縮効果ー」
　濱裕光、志賀直人(大阪市立大学)

○「画像データベースの自然言語インタフェースについて」
　伊東幸宏，中谷広正(静岡大学)

○「多視点距離データを用いた３次元形状モデリング」
　横矢直和，増田健(奈良先端科学技術大学院大学)

２６日午前　一般講演

○「ハイパーメディア・コーパスの構築と言語教育への応用について」
　上村隆一(福岡工業大学)

○「『歌物語』語彙の統計的研究」西端幸雄(大阪樟蔭女子大学)

○「高次辞書データベースのための語彙知識自動獲得システム」
　亀田弘之，藤崎博也(東京工科大学)

２６日午後　一般講演

○「社会調査結果の視覚化データベース」吉田光雄(大阪大学)

○「『間』に関するデータベースの構築」中村敏枝(大阪大学)

○「方言音声データベースの作成と利用に関する研究」
　田原広史，江川清，杉藤美代子，板橋秀一(大阪樟蔭女子大学)

文部省科学研究費・重点領域研究「人文科学とコンピュータ」
データベース計画研究班

代表者 小沢　一雅（大阪電気通信大学、情報工学部）

分担者 梅田 三千雄（大阪電気通信大学、情報工学部）

加藤　常員（大阪電気通信大学、情報工学部）

江澤　義典（関西大学　総合情報学部）

吉岡　亮衛（国立教育研究所 教育情報資料センター）

公開シンポジウム「人文科学とデータベース」 １９９５

発行日　１９９５年１２月２５日

発行所　公開シンポジウム「人文科学とデータベース」実行委員会
〒５７２　大阪府寝屋川市初町１８－８
大阪電気通信大学　情報工学部　小沢研究室内
電話　：０７２０（２４）１１３１
ＦＡＸ：０７２０（２４）００１４

印刷・
製本　穂高産業株式会社
京都市右京区西院西高田町１７－１７
電話　：０７５（３１４）７０５１