

第15回公開シンポジウム
人文科学とデータベース
発表論文集

Proceedings
of the 15th Symposium
on Humanities and Database

2009年11月28日(土)
神戸大学国際文化学部 F301 教室

【主催】 第15回公開シンポジウム「人文科学とデータベース」実行委員会

【共催】 文部科学省大学院教育改革支援プログラム
「文化情報リテラシーを駆使する専門家の養成 (A026)」
(神戸大学大学院国際文化学研究科)

神戸大学大学院国際文化学研究科メディア文化研究センター

【後援】 人文系データベース協議会

目次

【自由論題報告】

「継承語教育文献データベースの構築」

田中 順子, 森下 淳也 (神戸大学), 中島 和子 (トロント大学) 1

「日本語の起源, 解明方法と数値評価」

永田 良茂 7

「表現豊かな自然発話コーパスのアクセスについて」

モクタリ 明子 (神戸大学), ニック・キャンベル (ダブリン大学),
田畑 安希子 (神戸大学) 15

【招待講演】

「国立民族学博物館におけるデジタルアーカイブズ」

久保 正敏 (国立民族学博物館) 21

【自由論題報告】

「和歌解析用 MeCab 辞書の開発

—八代集解析済みコーパスによる学習—

山元 啓史 (東京工業大学) 31

「作文指導に向けた表現データベースの構築と表現評価指標の検討」

安藤 秀明, 秋吉 信吾, 竹内 和広 (大阪電気通信大学) 37

「『古事記』学術支援データベースの構築

—系図表示システムにおけるデータ構造の検討—

生田 敦司 (大谷大学), 杉山 正治 (立命館大学), 柴田 みゆき (大谷大学),
宮下 晴輝 (大谷大学) 43

「古事記崩年干支についての疑念」

小沢 一雅 (大阪電気通信大学) 49

第 15 回公開シンポジウム「人文科学とデータベース」プログラム 57

継承語教育文献データベースの構築

田中 順子 森下 淳也 中島 和子
神戸大学 国際文化学研究科 神戸大学 国際文化学研究科 トロント大学 名誉教授

グローバル化に伴って国を越えて移動する年少者や移動先の国で生まれ異言語環境で学齢期を過ごす二世児、三世児が増える一方である。このような子どもの言語形成、人間形成にとって、親の母語・母文化継承が教育上重要な意味を持つ。国内の「外国人児童生徒」にも「継承語教育」が必要であり、長い歴史を持つ「継承日本語教育」の知見がいかされるときが来ている。本研究は、未整理のまま放置されている「継承日本語教育」の文献を収集、整理、統括してデータベース化する試みであり、現在 65 名（24 ヶ国、11 言語）の国内外の母語・継承語・バイリンガル教育研究会会員が協力している。整理した文献をデータベース化した後にはウェブ上で公開する予定である。本稿ではこのデータベース化の背景と目的、および進捗状況について報告する。

キーワード：継承語、継承日本語教育、マイノリティー言語、外国語児童生徒教育、母語保持、マルチリンガルリズム

Making a database of information on Japanese heritage language education in the world

Junko Tanaka Jun'ya Morisita Kazuko Nakajima
School of Intercultural Studies School of Intercultural Studies Professor Emerita
Kobe University Kobe University University of Toronto

Due to globalization, children who transit across borders with their parents, and those who are born in a foreign country and spend their school years there (the second or third generation immigrant children) have been increasing. For the children, inheriting their parents' culture and language is important in terms of their language- and human-development. Japan has many such children who are in need of "heritage language education." The knowledge accumulated in the long history of Japanese heritage language education outside of Japan should be utilized for such education of immigrant children in Japan in order to preserve their mother tongues. The present project aims to make a "Japanese heritage language education database" of resources that have not been synthesized, and to make such resources readily accessible to educators. Currently 65 researchers from 24 countries are collecting, evaluating, and synthesizing information on Japanese heritage language education in 11 languages. In this paper, we will present the background and the aim of making such database, and report the status of the project.

Keywords: heritage language, teaching Japanese as a heritage language, minority language, educating foreign children, mother tongue maintenance, multilingualism

1. 背景

日本から海外へ移住した人達が、自分たちの母語や母文化（日本語や日本文化）を次世代に継承する「継承語としての日本語教育」は 100 年以上の歴史を持っている。しかし、世界各地に散在する文献や情報が入手しにくいことや、教育者不在の零細な教育的営みであり、それを研究の対象とする専門家も育っていないため、「継承語としての日本語教育」は日本語教育の中でもっとも未整理な領域の 1 つである[1] [2]。昨今、日本国内のブラジル系、中国系、フィリピン系、ベトナム系のマイノリティー言語話者の児童生徒に対して、これら様々な言語での母

語・継承語・継承文化教育の重要性が日本国内で認識されつつある。今日、移住にともなって日本人が海外で歩んできた継承日本語教育の軌跡と知見が国内でもいかされるべき時代となった。高年齢化が進み、労働人口減少にともない、外国人受け入れは拡大する一方であり、それに伴って外国人児童生徒の言語・文化継承の課題が浮上しているのである。

日本の公立小中学校で日本語を学習言語として使用するブラジル系、中国系、その他の児童生徒は、圧倒的に優勢な日本語のプレッシャーのもと、家庭で親が使用する母語・継承語は順調に伸びず、親が母語で話しかけても子どもは日本語で答えを返すというアンバランスな二言

語使用になるため、親子のコミュニケーションに支障を来している。一見すると、これは移民児童が母語ではなく日本語を選択して学習しており、日本社会には望ましい状況のように見えるかもしれない。しかし、このような状況は、二つの文化にまたがった健全なアイデンティティを育てる上でも、また日本語を学習言語・学校言語とする公立小中学校で高度の日本語を習得して、年齢相応の学力を獲得する上でも大きな支障となるものである。[3][4]

継承語教育のニーズは子どもが移動する年齢と世代によって異なる。言語形成期後半（9/10歳以降）に移動する「1世児」は、母語がすでに確立されているため学習言語が変わっても母語喪失の危険性は少ない。親の母語が1世児自身の母語でもあり、母語による読み書きの補強をすれば母語と現地語のバイリンガルとして育つ。一方言語形成期前半（9/10歳以前）の「1世児」や現地生まれの「2世児」は、母語が十分育たず現地語が生活言語であると同時に学習言語になるため、放置するとつぎのいずれかになる危険性がある。

(1)「会話型バイリンガル」（読み書きは現地語のみ、会話は両言語）

(2)「聴解型バイリンガル」（読み書きは現地語のみ、聞いて理解するが話すのは現地語のみ）

(3)「現地語モノリンガル」（現地語しか使えない）

(4)ダブル・リミテッド（どちらの言語も年齢相応のレベルまで育たず学習困難に陥る）。

(2)、(3)が陥る親子のコミュニケーションの断絶による情緒不安定、帰属意識のゆれ、また教育上明らかにマイナスである(4)を回避するためにも、継承語教育によって「読み書きも両言語でできる高度バイリンガル」を育て、ホスト国、ホーム国の両方で活躍できる有用な人材を育てたいものである。

継承語教育の文献の所在とその質は国によって異なる。継承日本語教育の長い歴史を持つのは、中南米と北米の一部（例：ハワイ州、カルフォルニア州、ビリティシュコロンビア州）であるが、子どもが文盲にならないようにという配慮から村の長者が子どもに読み書きを教えるという実践に端を発している。その後時代とともに地域の「日本語学校」として組織化されていくが、いずれも零細な寺子屋式の塾的なものであり、研究の対象になることはほとんどなかった。

1960年代になると、日本企業の海外進出に伴って海外子女対象の「全日制日本人学校」と「補習校」が世界各地に設立された。「全日制日本人学校」は母語・継承語保護学校であり、「補習校」は週末継承語プログラムである。両者とも日本への帰国を前提にしたものであるが、

現在は永住予定の子どもが混在するようになり、文科省制定のカリキュラムをこなすことが困難を極めている。従来「海外帰国子女教育」という観点からの研究が多かったが、最近特に北米やEU諸国で継承語教育の観点からの補習校教育研究が増えている。

継承語教育一般に関する文献が豊富なのはカナダと米国である。カナダでは、継承語教育の教育的意義付けが1970年代に始まり、多様文化主義法のもと教育委員会支援の週末継承語プログラムが今でも開かれている。[5] 一方、米国では、国防政策の一貫として外国語に堪能な人材が必要とされるようになり、その関係で2000年ごろから即戦力を持つ継承語学習者が急に注目を浴びるようになった。現在、継承語教育に対する研究と実践がもっとも盛んである。[6][7]

国内においては、母語保持・伸長の必要性は認識されつつあるが、その制度上の位置づけはこれからである。現時点では、専門的に追求する研究者も研究機関も不在である。この空白を埋める役割を担ってきたのが「母語・継承語・バイリンガル教育（MHB）研究会」である。2003年に設立以来、MHBは、年次大会、研究集会、紀要発刊、オンラインディスカッションを通して、補習校児童生徒、日系人児童生徒教育、国内のオールドカマーの民族語・民族文化教育、ニューカマーの母語・継承語教育、ろう児の母語（手話）教育の課題と取り組んできた。本プロジェクト「継承語教育文献データベースの構築」は、MHBが形成してきた国内および世界各地の会員（約450名）をフルに動員して行うものである。

本稿では、(1)プロジェクトの目的と意義を明らかにするとともに、(2)プロジェクト内でどのようなデータベースを作成しようとしているのか、(3)現在の進捗状況と最終形態について報告する。

2. 目的

当プロジェクト[8]の目的は、(1)継承語教育の先進国であるカナダ、2000年以降目覚ましい隆盛を見せている米国、日本人の集団移住先である南米諸国、国内のオールドカマー、ニューカマーを中心に重要文献を収集、整理、統括してデータベース化し、母語・継承語・バイリンガル教育研究者・教育関係者に学会ホームページ上で公開提供すること、(2)上記の作業を通して、継承語教育の専門家や、質の高い継承語教師の育成に寄与すること、の2点である。文献収集に当たっては、国内外の母語・継承語・バイリンガル教育研究会会員から協力者を募り、現在調査協力者として、世界6地域（アジア、中南米、オーストラリア、カナダ、米国、

EU)、24ヶ国、11言語の65名が現在これに当たっている。加えて MHB の理事9人が調査責任者として調査協力者の統括に当たっている。

文献の収集は、現在次の5つの課題を掲げて行っている。

1. ホスト国の言語政策

日本語以外の言語も含んだ継承語全般に対してホスト国がどのような言語政策・施策を持っているか。

2. 歴史的経緯

歴史的にどのような経緯で現在に至っているか。

3. 現在の継承語教育

現時点で特定の地域(国・地域・対象グループ)の継承語教育は、どのような状況にあるか。

4. 継承語教育の成果

継承語教育の成果がどのくらい上がっているか。

5. 将来の課題

継承語教育の将来の課題は何か。

3. データベース作成手順

継承語教育データベースは次のような11段階の手順で製作する。

(1) 仮テンプレートの作成

まず、データベースに含める項目を決定する為に、エクセルで仮テンプレートを作成した。データベース作成作業開始時には、データベースに含める項目の全体像が不明であったため、想定できる資料の種類を以下の5種類に大別し、それぞれについて一つのテンプレートを作成した。このテンプレート作成作業を通じて入力項目を仮決定した。

Template 1. 著書・論文、

Template 2. 口頭・ポスター発表、

Template 3. 教科書・教材・ソフト・作品等、

Template 4. 雑誌・新聞記事、

Template 5. 政府刊行物・法律・条例・(公的)ガイドライン・スタンダード・カリキュラム・学習指導要綱)

(2) 仮テンプレートへの入力

エクセルで作成した仮テンプレートに65名のMHB研究会会員が入力を行う。

(3) 仮テンプレートへの入力項目のチェック

入力した結果を9名の統括者がチェックをして、ルールから外れた入力が行われていないかを検討する。

(4) 仮データベースの稼働

入力チェックを経たテンプレートをデータベースシステムに流し込み、情報がテンプレートからデータベースに移設できるかどうかを確認する。

(5) 作業用テンプレートの完成

現行の入力ルールに従って入力されたデータがデータベースに移設ができることを確認できれば、仮テンプレートを作業用テンプレートとして完成したものとみなす。完成したテンプレートは、(7)以降の段階で入力作業を行うMHB作業部会員限定で公開される。

(6) 入力マニュアルの作成

入力作業を行うMHB会員向けに入力マニュアルを完成させる。illegalな入力が行われないようにするためには入力マニュアルが適切であることが重要である。そのため入力マニュアルの作成作業には細心の注意を払う。

(7) 一部入力済みのテンプレートを skydrive に収納

仮テンプレートを経て完成版となったテンプレートをウェブ上の文書共有システム(skydrive, google document 等)上にアップロードする。

(8) MHB 作業部会員によるデータ入力

文書共有システム上に置かれた完成版テンプレートに、会員が世界中からアクセスをして入力し、データのエントリー数を増やして行く。

(9) 神戸大学継承語データベースに流し込む

次にテンプレート上の情報をデータベースに移設する。入力が終わったテンプレートをデータベースに流し込み、データベースを作成して稼働実験を行う。データエントリー数がどの程度になればデータベースへの移設を行うかはまだ未定である。

(10) データベースの稼働

10段階の最後はデータベースの稼働である。この段階で不具合が出た場合は、再び(3)から(9)の段階をたどる。

(11) ウェブインターフェースを持たせる

データベースの入力をウェブ経由で行えるようにする。

4. テンプレートに含む入力項目

作成を計画しているデータベースの中の情報は3つあり、文献情報、書き下ろしの解説からなる概要データベース、および全文である。少なくとも文献情報と概要データベースを含んだ

データベースの作成を目標としている。著作権を侵害しない範囲において、アブストラクト、フルテキストを掲載する。

著作権の問題がクリア出来れば全文を所収したデジタルアーカイブにまで発展させたい。全文情報の所収が出来ない場合は、全文にたどり着く為のリンク情報を含める予定である。

5. 継承語教育文献データベースシステム

継承語教育文献データベースシステムは、著書・論文、口頭・ポスター発表、教科書・教材・作品の3つの対象を格納するデータベースシステムとして構成されている。このシステムの特徴は、協力者によってデータベースが構築できるようにデータ収集の基本を Web アプリケーションとしている事である。従来なら、既に設計された ID のテーブルや各項目の定義域などが用意されており、どのような分類がなされるかも既知であるとされるが、このような特殊な文献データベースの場合、現時点ではデータの広がりや予測する事ができないため、幾つかの部分で、動的に対処する必要がある。図 1 に Web による著書・論文編集システムの入力画面(部分)を示す。

(1) 入力項目の動的管理

図 2 に対象言語の入力部分を示す。対象言語として「言語の特定なし」も含めて8つの言語の選択肢を用意してあるが、これ以外のもので登録された場合、その頻度に応じて、次の入力表示には項目として出現するように配置する。例えば現時点では「ドイツ語」、「フランス語」などである。これは、データベースに新しい項目が入力された場合、それを対象言語のインデックステーブルに登録して、次の入力システムが呼び出される際には、動的にその項目が選択肢に加わるようにシステムを構築する。

(2) 既存のデータによる入力支援と入力補完

重要な項目については、データ作成時に動的にデータベースを呼び出し、その入力する値の候補となるべきものを提示する事で、入力を簡便なものにする。例えば、図 3 は著者名を協力者が入力する際に、前方一致のインクレメンタルサーチが表示されているイラストである。このように、入力する候補をシステムから提示することで、入力を助けるのみならず、多くの協力者によるために発生する、タイプミスや入力文字のぶれ(旧字体を使用する、空白の挿入)をなくし、統一する事ができる。

実際、著者名は、新しい著者が入力された場合、インデックスに登録され、ID が割り当てられ、論文データとの関係はこの ID を外部キー

として参照する事で、正確さを高める事ができる。

(3) データの完成度の導入

このようなデータベースの場合、どの程度、各データ項目が確立したものであるかというのは、実際に資料として役立てる際には、重要なものとなる。そのため、図 2 の下にあるように備考欄を用いて、入力者、チェック者によって各データの完成度を記述するようにする。

6. 進捗状況

データの収集に関しては、これまで、調査協力者会議第1回(6月4日)、第2回(8月9日)を開催、第3回は、来年3月25日、27日の予定である。調査協力者を参加者としてヤフーグループを作成して連絡網としている。互いに情報交換を行って収集した文献は、Windows Live SkyDrive に貯蔵している。作業の進捗状況に合わせて、定期的にテンプレートのアップグレードをしていく予定である。現在は、テンプレート Version 1 の段階であり、Version 2 へのアップグレードは、12月初めに予定している。2009年11月現在、このテンプレートのアップグレード作業と平行して、エクセルファイルのデータを格納するデータベースを作っている段階である。

7. まとめと今後の展望

継承語教育文献データベースは文献データベースであるが、ネットワークの利点を活かして、関連 URL による原著に対するリンクや著作権の問題が発生しない原著のデジタル化も視野に入れて準備されている。今後はデータベースに、既存のデータによる入力支援やプルダウンメニューを付加していく計画である。

参考文献

- [1] 佐々木倫子(2003)「3代で消えない JHL とは?—日系移民の日本語継承」『母語・継承語・バイリンガル教育研究会』プレ創刊号 12-21
- [2] 中島和子(2003)「JHL の枠組みと課題—JSL/JFL とどう違うか」『母語・継承語・バイリンガル教育研究会』プレ創刊号 1-11.
- [3] 湯川笑子(2006)「年少者教育における母語保持・伸長を考える」『日本語教育』128. 13-23.
- [4] 中島和子(2007)「外国語習得と母語との関係—セミリンガル現象の要因と教育的処置

- に関する基礎的研究」科学研究費補助金基礎研究（B）最終報告
- [5] カミンズ・ダネシ（中島和子・高垣俊之訳）（2005）『カナダの継承語教育- 多文化・多言語主義をめざして』明石書店
- [6] Brecht, R.D., & Ingold, C.W. (1998). *Tapping a National Resource Heritage languages in the United States*. ERIC Digest, EDO-FL-98-12.
- [7] Brinton, M.D., Kagan, O., & Bauckus, S. (2008). *Heritage Language Education: A New Field Emerging*. Manwah, NJ : Lawrence Erlbaum Associates.
- [8] 平成 21 年度科学研究費補助金（基盤研究（B））課題番号：（21320096）「継承日本語教育に関する文献のデータベース化と専門家養成」（研究代表者 中島和子）。

付録

森下 淳也 > 継承語教育文献データベース > 著書・論文編集システム > データ追加 検索に戻る

著書・論文編集システム・データ入力

項 目	入 力 領 域	説 明
タイトル（原文）		タイトルを原文で入力して下さい。タイトル中の年号その他数字、英単語は半角でお願いします。
タイトル（英文又は和文）		日本語・英語の場合、英文タイトルがあれば記入、日本語、英語以外の言語のタイトルの場合は英文または和文タイトルを入れて下さい。数字は半角で願います。
著者・編者名		複数の場合、著者と著者の間は","または"/"で区切ってください。姓と名の間は空白を入れてください。日本語の場合は姓が先、英語名の場合は名が先です。
発行年		数字は半角でお願いします。
出版社・発行元		
出版地		
掲載書・掲載誌		
巻・号		数字は半角でお願いします。
ページ数		論文のみ。数字は半角でお願いします。
版数	☐ 初版 ☐ 2 版 ☐ 3 版 ☐ 不明 ☐ その他 →	書籍のみ。入力選択肢から1つを選択して下さい。その他の場合、記述願います。

図 1. 著書論文システムの入力画面設計（部分）

対象言語	☐ 言語の特定なし ☐ ポルトガル語 ☐ スペイン語 ☐ 英語 ☐ 中国語 ☐ ベトナム語 ☐ 日本語 ☐ 韓国語 ☐ その他 → その他の言語:	対象となる言語を選択して下さい。複数回答できます。その他の場合にはその言語を記述願います。
備考	☐ 全欄完成 ☐ 概要とキーワード以外ほぼ完成 ☐ 一部調査中 	このデータの完成されている度合いを、全欄完成/概要とキーワード以外ほぼ完成/一部調査中の3つの中から選んで下さい。気づいた事柄を自由に書き入れてください。

追加せずに検索に戻る

送信

編集システムに戻る

森下 淳也 > 継承語教育文献データベース > 著書・論文編集システム > データ追加
検索に戻る

図2. 対象言語と備考の入力欄（図1の設計画面よりの抜粋）

著者・編者名	中 中井 美穂 中島 和子 中嶋 知子 中森 明菜 中山 律子
発行年	

図3. 前方一致のインクレメンタル・サーチによる著者名入力の支援

日本語の起源、解明方法と数値評価

The Origin of Japanese Language ; Solving with the Approach Method and Numeric Estimation

永田 良茂 (地名/語源研究家)

Yoshishige Nagata (Place Name & Ethnology Researcher)

神戸市北区泉台 2-9-9

2-2-9, Izumidai, Kita-ku, Kobe-City, Japan

あらまし： 日本語の系統問題は長い間系統不明の孤立語とされて、混合語説が有力視されるも起源問題は未解決である。「言葉とは何？」の本質の追究からこの問題を決着する方法を考え、比較言語学の立場から基礎語彙による親縁関係を見てみると、日本語もアイヌ語も8割近くが南島語族に親縁関係を認めることが出来た。特にアイヌ語は縄文語を引き継ぎ、南から北の北海道に閉じこめられた結果成立した言語、文化であることを示している。近隣諸言語の語彙の親縁関係比較判定には音韻類似度を定義して数値評価を加えた。

Summary : The Japanese Language has been called an isolated Language because it belongs to the unknown linguistic group and recently considered as a blend between Altaic and Austronesian language families , however its origin has not been determined conclusively.

This paper , using the approach method and the question “what are words ?” as a base , indicates that from the view point of comparative linguistics there is an 80% between Ainu and Japanese in the Austronesian language family . The Ainu language had specifically been resume to have reached Kokkaidou by way of the Japanese Islands from the south and had been isolated similarities of the vocabularies .

For the Comparative Judgement of the similarity of the vocabularies of the neighboring Languages, we made use of numeric estimation by the defined rate of the phonological similarity between the two vocabularies of the different languages,

キーワード： 比較言語学、 言語年代学、 縄文語、 日本語、 アイヌ語

Keywords: Comparative Linguistics, Glottochronology, JOMON Languages,
Japanese Languages, AINU Languages

1. はじめに

第12回公開シンポジウムにおいて、「言語年代学における日本語の系統、縄文語を推定する」 [1] という表題の報告をさせていただいたが、今回の発表はそれを発展させたものである。

「言葉とは何？」の言語本質の追究から言語起源の追求方法をまとめ、その方法に基づいて広く、中国語等を除いた南島語族やアルタイ語派の日本周辺の語彙調査を行い、調査結果は『日本語の起源とアイヌ語』 [2] として出版した。

その中から解明方法といくつかの具体的な語彙の

親縁関係調査結果を例示して、調査結果としては基礎語彙における日本語とアイヌ語との親縁関係を数値評価として出すことが出来たので報告する。

M,スワデッシュの言語年代学のうち、彼の試行錯誤した基礎語彙100語を、時代を経ても変わりにくい語彙と考え、その語彙を構成する音韻と意味との範囲を広げて親縁関係を調べたものである。

特にアイヌ語に対しては音韻記号による音韻類似度を定義して、困難が予想される親縁関係判定を評価してみた。

2. 日本語研究の流れ

我が国の言語系統論争において、約百年前に藤岡勝二がウラル・アルタイ語における14ヶ条の特長を示し、そのうち13ヶ条は日本語にも当てはまると主張し、その後橋本進吉などの上代仮名遣い（8母音説）や母音調和の発見などが相次ぎ、14ヶ条が成り立っていたことがわかり、日本語系統のウラル・アルタイ語説が多くの学者に支持されるに至った。

また、明治末年の日韓併合と重なり、金沢庄三郎は『日韓両国語同系論』を発表して、支持する人も多かった。

戦後になると、戦前にアルタイ語を学ぶためにベルリン大留学を行なった村山七郎は、日本語は南島基層語にアルタイ語が重層した、混合語説を主張し今日の主流になりつつある。

近年も（1986年）国立民族学博物館の共同研究テーマとして「日本語系統論の方法に冠する基礎的研究」と題して、言語学、民族学を含めた多くの学者を動員しての検討がなされ、『日本語の形成』[3]として成果報告がなされているが、「日本語の源流」を目指した報告にはほど遠く、各学者の従来の主張を列記したものであり、「サワラヌモノニ、タタリナシ」的な従来の「系統不明な孤立語」としての立場を強調するような結果にしかっていない。

その中で従来の研究の方法を整理し図解した下図はうなづけるものである。語彙論を中心とした研究では南島語族との類似がクローズアップされ、形態論・統語論を取り上げるとアルタイ語派との類似が鮮明となることは明らかである。

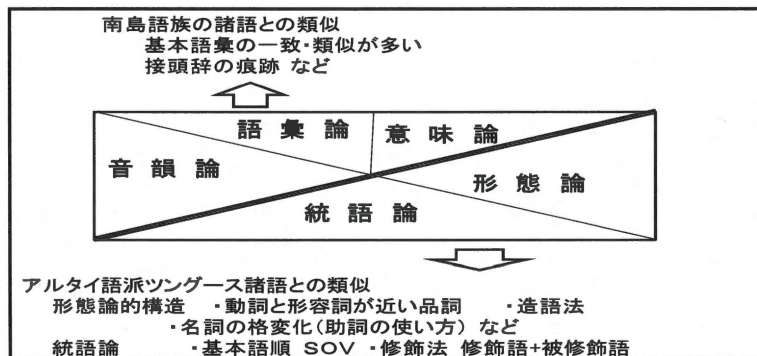


図 1. 日本語系統追求の主な方法論

注記：

『日本語の形成』[3]の中の「類型から見たギリヤーク語ー日本語との関係において」から中の図を借用し、周囲の補足説明を筆者が行なった。

3. 日本人論

国立科学博物館のホームページ「日本人はるかな旅展」において、世界中の現代人は、ホモ・サピエンス（新人）という一つの種に分類され、アフリカに10万年前に誕生し、下図のように広がった説が有力であることを伝えている。

図 2. 新人（ホモサピエンス）の拡散



日本語の起源を取り扱う場合にこれらのどの渡来

が有力な「日本語の源流」となったか、であろう。

4. 日本語起源調査に必要な考え方と方法

4. 1 基本的な考え方と方法

村山七郎によると、ヨーロッパでは19世紀前半に言語の親縁関係という考え方が確立し、仏人メイエの考え方として、「言語記号の任意性」がその親縁関係の土台をなすと、紹介している。[4]

「言葉とは何？」に対する回答が「言語記号の任意性」であり、言葉とは「音」と「意味」の組み合わせからなり、文字記号で表したものが「言語記号」に相当し、「音」と「意味」の組み合わせ関係には必然性がない、無関係であることを「任意性」と言っているが、その組み合わせ関係は、ただ文化遺産としての伝承に基づくところに重要な意味合いを持ち、異なる言語間で同じ言語記号が見つかることは共通祖語から受継がれた共同遺産として保存されてきた可能性が大きいことを示すのだそうである。

この考え方を基に日本語およびアイヌ語の起源を調査する方法・考え方を次のようにまとめた。

4. 2 日本語・アイヌ語の特長など

村山七郎の著作などから、日本語およびアイヌ語の特長がなぜ生じたのか、本源のものかどうか、南島諸語との比較に必要なことなどを若干まとめた。([1]の内容にほぼ同じ。)

(1) 日本語とポリネシア語の類似性から

ハワイ語などのポリネシア語など、日本語と同じ5母音であり、重母音・重子音を許容せず、日本語古語と同じく濁音のない子音のみの簡易な音韻であり、母音終わり（開音節）の単語でありカタカナで書いても通じる言葉である。（大野晋のタミル語起源説のタミル語の特徴とも一致）

しかし、これらの特徴は本源のものではなく、南島諸語の中の中心、インドネシア語派が拡散してメラネシア語派、ポリネシア語派と二次、三次的に拡散した結果であり、日本語も二次、三次的に拡散した結果であることを示しているようである。これらの言語の共通の特徴は民族移動とともに言語接触を繰り返し、より簡易な音韻に落ち着いた結果に思える。

図 3.日本語系統親縁調査に必要な方法の考え方

1.言葉（語彙）

「言語記号の任意性」 → 語彙比較の重要性

2.言葉の組織的な調査方法

言葉は時代と共に変わるが、変わりにくい語として、基礎語彙がある。

M.スワデッシュの基礎語彙100語を採用。

3.伝搬ルートを想定した広域の語彙調査が必要

上記の図 2.の到来ルートが参考になる。

スワデッシュの基礎語彙100語における言語年代学については、第12回公開シンポジウムの「言語年代学における日本語の系統、縄文語を推定する」[1]で取り扱ったが、今回は時代を経ても変わりにくい語彙に重点を置き、語彙の「意味」と「音」との類縁関係を広く調査した。

ポリネシア語の開音節語に関して、単語には露出形（独立名詞）と被覆形（従属名詞）とがあり、露出形は開音節であり、被覆形は閉音節が保存されており、元々は閉音節があったことを示しているそうである。（日本語の被覆形の例；白（シロ）に対して、白（シラ）魚、白（シラ）サギなど。インドネシア語派の場合は露出形の場合にも閉音節語が多い）

日本語の場合も同様に語末尾子音はとれるが被覆形においては保持されているものがあるようである。（動詞語幹など、「開ク」の例；ak-、「聞ク」の例；kik-）

アイヌ語には閉音節の単語が多く、古い形を残している。（台湾諸語も同じ）日本語の場合には被覆形でも開音節であるが、原始日本語では閉音節語（VC, CVCなど）があっただろうと言われている。

また、日本語の露出形、被覆形の単語の古形は双方の音韻を含む場合が見られる。

（例：「ヒ（火）:pii」に対して「ホ・ノ・ホ（炎）:po」から祖語*(a)poi が復元される）

(2) 二音節語と単音節語

日本語は2音節の語（VCV, CVCVなど）を基本としており、ポリネシア諸語などと共通な特徴であると見られており、日本語の単音節語の多くが古語では重母音語であったと言われている。一方、アイヌ語は単音節語を基本としていると考えられている。

しかしこのことも本源的ではなく、南島諸語の場合は接辞と語幹の各1音節の組み合わせの場合が多く、もともとは一音節で、アイヌ語の場合には長い年月の間に2音節語が1音節語に圧縮された「幹収縮」の形を取ったものが多く見られる。その結果、この概念化された語幹語により多くの合成語を形成できるようになったものではないだろうか。「アイヌ語はもとをただせば数百の単語から成り立ったもの」などの田村すす子の言葉が思い出される。

また「アラタ（新）シ」が「アタラシ」になったように、アイヌ語には南島諸語から子音間、母音間、子音・母音間の前後の入れ替わった「音位転換（メタテーゼ）」の例も多く見られる。

(3) その他の南島語的な特徴

アイヌ語は接辞（接頭、接尾辞）を駆使した言葉であり、南島諸語的であり、抱合語に分類されている。日本語においても「マッシロ（白）」、「マッサオ」などの接尾辞“ma-”はインドネシア系言語に見られるものだそうである。また、インドネシア系言語には前鼻音化形の接辞（“[me]N-”など）があり、その痕跡が日本語にも見られるそうである。

また、日本語には少なくアイヌ語には多い語頭の“r”音であるが、南島祖語の“*d”音に対応するもので、日本語ではさらに“t”音などに変わったものであると言われている。

一方、アルタイ語派は接頭辞がなく、接尾辞を駆使し、特に動詞の語尾変化による動詞活用などはアルタイ語的特長である。また、助詞、助動詞の駆使による多様な文の構成などもアルタイ語的で、膠着語と呼ばれる所以である。

5. 基礎語彙調査

5.1 基礎語彙について

M.スワデッシュは考古学における炭素14（¹⁴C）による遺物年代測定にヒントを得て、2言語間の分離年代を測定する言語年代学を1950年代に打ち立てた。当初は13言語の215語の語彙を使用した。最終的に基礎語彙100語を試行錯誤の上、使用した。

我が国で日本語と18ヶ言語間でこれを試みたのは服部四郎で、『日本語の系統』にその結果が報告されている。このような調査方法では日本語系統は決められないと調査の失敗を認めている。

村山七郎は『日本語の研究方法』の中で、服部の日本語と朝鮮語との類似語19語に対して、個々にコメントし4語しか類似が認められないと厳しく批判し、基礎語彙に相当する各語の語彙選択の困難さ、言語年代学の妥当性を疑問視している。

そこで年代の算出をあきらめて、時代を経ても変わりにくい語彙としての基礎語彙の価値に注目し、関連語彙の「意味」と「音」の範囲を広げて関連語彙も考慮して、公正な親縁関係の判定に努めた。

5.2 調査の具体化

スワデッシュの基礎語彙100語の各語彙に対して、それに相当する南島祖語を上部に、日本語、アイヌ語および沖縄本島語・八重山諸語を左欄に、インドネシア・フィリッピン諸語、台湾諸語、メラネシア語派およびポリネシア語派の語彙を右欄に配置し、南島祖語との語彙類似と左右欄の語彙類似の認められるものを線で結び、語彙調査表として図表に示した。（例として 表 1. 参照）

またその語彙の集約結果として、日本語またはアイヌ語語彙群とインドネシア・フィリッピン諸語、台湾諸語、沖縄・八重山諸語、アルタイ諸語およびアイヌ語または日本語との語彙類縁関係判定を○△×（－）の3段階で判断し、表にまとめた。

（例として表 2/4.および表 3/5. 参照）

また判定・評価の根拠を（……>……）などの変移推定で示した。*印は復元南島祖語を示す。音声記号はここでは印字フォントとの整合性を考えて次のように置き換えた。y→R, η→N, ?→'（声門閉鎖音（グロータルストップ））

5. 3 語彙調査の具体例

(1) 基礎語彙No.1 私: I から

表 1. 基礎語彙調査 私：I の例

No.1 私.1 語彙比較

AN祖語

DMP	ZOR
aku	aku' ku

日本語古語

ア
ワ
ミ
ナ

アイヌ語

ku
a
an-/~an
chi
echi
si (自ら、自ずと)
yay (自分)

沖縄諸語

沖縄本島

wan
uga
duu

八重山

anuu
banu
baN

インドネシア・フィリピン

aku	イMLY
kula	イJAV
au	イBAT/ROT
si-aku	比ILC
ako'	比TAG
a	比KAG/MNB

台湾諸語

ku?	台SAT
saku?	
si?	台CAT
kuZing	
ya'ku?	台SED
a'o	台TSO
ku	台KAN
ako	台SAR
iLako	
kunako?	台RUK
ya'ku?	台THA
yaku?	台PAZ
yako'	台SAI
ya'o	
a'qen	台PAI
?ingku?	台PUY
ka'ko?	台AMI
ja'ku?	台BUN
?aiku?	台KUV
yao	台SIR
yake'n	台YAM

ku の音韻類似度

*ku	1.000
*aku	0.800
*aku'	0.667
kula	0.667
au	0.500
siaku	0.571
ako'	0.600
ku?	0.800
saku?	0.571
kuzing	0.500

a の音韻類似度

*aku	0.500
*aku'	0.400
kula	0.400
au	0.667
siaku	0.333
ako'	0.400
a	1.000
saku?	0.333
a'o	0.500
yako'	0.333

Hi の音韻類似度

siaku	0.667
saku'	0.640
si?	0.667

注記:
U ← a
H ← ch
*: 祖語を示す。

この表から考えられる語彙関係を補足する。

・ここで示した大部分の語彙が南島祖語との類似が認められ、その派生と考えられ、祖語からの変化を感じ取ることが出来て、日本語、沖縄語およびアイヌ語もその中にあると思える。

・日本語およびアイヌ語の“a”はフィリピンTAG（タガログ語）またはMNB（マノボ語）と同じであり、独人O.テンブウォルフの南島祖語“aku”またはR.D.ソルクのaku’からの語頭の幹収縮（Stem Shortning）形になっている。

・アイヌ語 ku はゾルクの南島祖語 *ku をそのまま継ぎ、台湾諸語KAN（カナカナブ語）などと同じである。

・八重山諸語の中で anuu, banu, baN は南島祖語 *aku* の音韻変化から引き継がれたものと考えら

れる。濁音化・鼻音化・清音化の中で子音 k は

$$k \rightarrow N (= \eta_s / \eta_g / g) \rightarrow n$$

の変化は一般的であろう。(

また anuu などから幹収縮形としてアイヌ語 an-/-an がある。なおこの an-/-an の用法に関して、村山七郎は所有接尾辞 -an の使い方は単なる所有ではなく、血縁関係や身体部分とかを表わす名詞に限り使用され、必然的な切り離し得ない所属性が表現されるとしメラネシア諸語の所有接尾辞の用法特長と同じであることを指摘されている。

・ 沖縄語 uga は一見、南島祖語 aku と無関係に見えるが、南島語に多く見られる aku のメタテ-ゼ形 uka の濁音化 uga であり、古語の「ワ」、「ワガ」への転訛が十分に考えられる。

*aku → (uka) → uga → ua(=wa)

古語 ワ → waga ワガ

ここの anuu の uu は長母音を示す。(重母音ではない。) また、沖縄・八重山諸語の

wan/banu/baN → wa 古語 ワ

への転訛・幹収縮化も十分に考えられる。

- ・アイヌ語 chi, echi, si- は台湾諸語CAT (チウリ
- ・アタル語) si? などと同じである。
- ・日本語古語 「ミ」、「ナ」はアルタイ系であろうとされている。

上記のような言語間の実証が言語系統問題の解決に不可欠であることを予想できる。

特に、「我が(君)」の古語の変化、「ア・(ギミ)」、「アガ・」、「ワ・」、「ワガ・」などが沖縄諸語の中に変化の痕跡を残しているようである。

これらの語彙類似関係をアイヌ語および日本語(古語)にまとめ次表のような類縁判定にまとめた。

なお、アイヌ語に対しては次式の音韻類似度を定義して語彙の類似度評価を試みてみた。

音韻類似度 =

$$\frac{2X \text{ (対象語と被対象語の音韻記号の一致数)}}{\text{対象語の音韻記号数} + \text{被対象語の音韻記号数}}$$

表 2. アイヌ語の 私：I 類縁判定結果

表 6.1.2 アイヌ語の親縁関係の判断結果

	南島祖語	イ・比諸語	台湾諸語	日本語	沖縄諸語	アルタイ語派
アイヌ語 a /an ku chi/echi si	○ a<*aku ku=*ku	○ a<aku/ako/saku /... .. a=a(比KAG/MNB) si/chi <si-aku(比ILC)	○ a<a'o/ako/... .. ku=ku(KAN) /ku?(SAT) si/chi=si?(CAT)	○ a=ア	○ an-/an < anuu/banu/baN	- bi :MON O.J. ミ<*bi/*man(TU) O.J. ナ<na(アkor)

表 3. 日本語の 私：I 類縁判定結果

表 6.1.3 日本語の親縁関係の判断結果

	南島祖語	イ・比諸語	台湾諸語	アイヌ語	沖縄諸語	アルタイ語派
日本語 ア ワ ミ ナ	○ a<*aku	○ a<aku/ako/saku au のメタテーゼ形	○ a<a'o/ako/... ..	○ a=ア	○ an-/an < anuu/banu/baN wan	○ bi :MON O.J. ミ<*bi/*man(TU) O.J. ナ<na(アkor)

(2) 基礎語彙No.82 火：fire から

村山七郎によると日本語「火」、アイヌ語 ape は下記のような変化の結果であると言われている。

ヒ < pñi < poi < apoi < *'apuj

ape < apoi < *'apui

アイヌ語、日本語はタガログ語レベルまでは同じで、日本語はそれ以降3回以上の変化を経たであろうのに対して、アイヌ語は1回の重母音から短母音への変化のみであり、古い形を保持してきたことが見て取れる。ところが日本語にもアイヌ語にも祖語の形を受け継いでいる言葉もある。祖語の apu- の形は日本語「アブ(灸)・ル」は

abu- < apu-

であり、鹿児島弁では「アブイ」、沖縄語では abujuN である。

アイヌ語 uhuy:燃える は、

uhu- < apu-

である。いずれにしても *'apuj のもとの形からの派生と考えられる。

また、ポリネシア諸語であるハワイやマオリ語では、afi, api と語頭の a- は維持しているが -fi, -hi は日本語が「ヒ」へたどり着いたと同じような簡易な音韻への変化の結果であろう。ただし、こちらから kapura のように -apu- の古い形も残している。

これらのアイヌ語に対する判定結果は下記の通り。

表 4. アイヌ語の 火: fire 類縁判定結果

表6.82.2 アイヌ語の親縁関係の判断結果

	南島祖語	イ・比諸語	台湾諸語	日本古語	沖縄諸語	アルタイ語派
アイヌ語	○	○	○	○	○	—
ape	*apuj	apny など	apoy/hapoy など	ヒ と同源	pii などと同源	tuwa /tuwe :MAN bri /pul :KOR gal :MON
uhuy	*apuj	apu/apoy	hapui など			

表 5. 日本語の 火: fire 類縁判定結果

表6.82.3 日本語の親縁関係の判断結果

	南島祖語	イ・比諸語	台湾諸語	アイヌ語	沖縄諸語	アルタイ語派
日本語	○	○	○	○	○	—
ヒ	*apuj	api/apo'1 など	apui/?a'pui など	ape	hwii/pii	tuwa /tuwe :MAN bri /pul :KOR gal :MON
アブル	*apuj	apu など	apui/?a'pui など	uhuy		

(3) 基礎語集No.72 日: sun から

日本語の「ヒ(日)」に関して、南島祖語の *t'iNaR の t'i- からのインドネシア、フィリピン、台湾、ポリネシアなどへの派生音として si-, di-, hi-, ci- などとある。

八重山の cii, pii 沖縄本島の hwii、日本語「ヒ(日)」へと音韻変化を示しているようである。結局南島祖語の *t'iNaR の間接的な派生語と言えるのであろうが直接的な派生語としてはここでは採用しなかった。

枕詞の「シナテル」の「シナ」、「シノノメ(東雲)」の「シノ」など *t'iNaR の反映とされている。日本語「シロ(白)」もこの祖語が語源である。

日本語「ハリ(晴)」「ハレ」は意味が少しずれるが南島祖語 *vaRi の派生語であろう。マライ語の mata-hari:太陽 の後半音にほぼ同じである。

アイヌ語 nusat:明け方 は意味は少しずれるが *t'iNaR:光 から派生のマライ語 sinar の n 音と s 音が入れ変わったメタテーゼ形である。(最後の r/t の相互入れ替えは起こりやすい。)

表では *t'iNaR の直接的な派生語として線引きしなかったが

POC(原始オセアニア語) *siNaR: 光からの直接的な派生語として評価した。

また「何日」などの「日」を表わすアイヌ語 to は南島祖語 *a(n)dav:日/太陽 から

to < daw < *adav

と変化したと推定できる。

アイヌ語 chup の意味は 日/月/腹 であり、chupu:つぼめる とあり「丸い物」からの意味の派生で、ここでは類縁関係は認められない。

沖縄県竹富島(石垣島の南西)に伝わる恋歌、安里

屋ユンタの歌詞に出てくる

「サー 君は野中の 茨の花か サーユイユイ

暮れて帰れば やれほに 引き止める

マタハリヌ チンダラ カヌシャマヨー」

の最後のハヤシの部分である。「カヌシャマヨー」とは、「神様よ」と分かる。「マタハリヌ チンダラ」部分は下記のように南島語のオンパレードである。古い言葉であろう、正確に訳せている例が無いようであるが、

マタ < *mata: 目

ハリ < *vaRi: 太陽

マライ語では mata-hari: 太陽

又 助詞「の」

チンダ < ts'inda < *t'iNaR: 光

ラ 感動詞

結局、意味としては「(自然の) 目である太陽の光はね、神様であることよ」等となるのではないだろうか。

5. 4 基礎語彙調査結果

基礎語彙100語のアイヌ語および日本語（古語）

の親縁関係を集計した結果は次の通り。（[2]から）

表 6. アイヌ語および日本語の親縁関係集計値（基礎語彙100語中の親縁関係認定数）

アイヌ語	南島語系統（○判定分）					アルタイ語派 （○+△判定分）
	南島祖語	イ・比諸語	台湾諸語	日本語古語	沖縄諸語	
	78/100	71/100	70/100	56/100	39/100	20/100

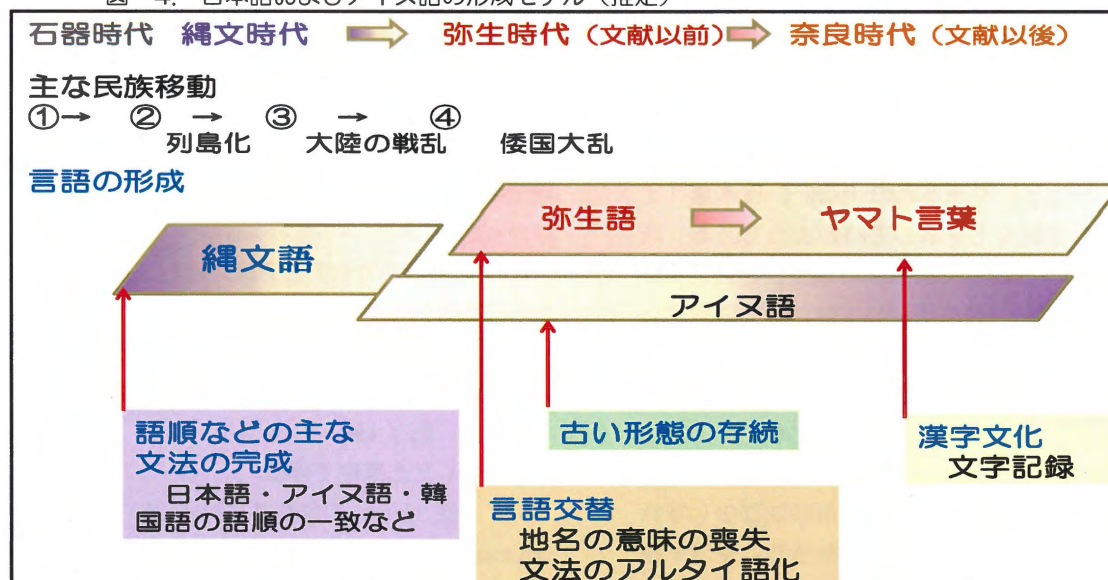
日本語古語	南島語系統（○判定分）					アルタイ語派 （○+△判定分）
	南島祖語	イ・比諸語	台湾諸語	アイヌ語語	沖縄諸語	
	79/100	72/100	63/100	56/100	94/100	40/100

この結果は日本語もアイヌ語も同程度に南島語を起源とすることを意味する。特にアイヌ語は古く南方からの人々と共に入ってきて、縄文時代に北に達して最終的に北海道、南樺太および千島に残った、

縄文語を維持・保存した言語であろう。

図 2.に示した日本列島への民族移動の経緯と合わせて、日本語の成立モデルを下図のように推定できる。

図 4. 日本語およびアイヌ語の形成モデル（推定）



6. まとめ

本報告は拙書『日本語の起源とアイヌ語』[2]の中から紙面の都合上、考え方と方法、具体例および結果のみ示した。日本語の起源を考える上でアイヌ語が鍵を握り、図2.および図4.の日本民族の主な流れのうち②および③の流れが縄文時代の主流をなし、アイヌ語に残され、その後の④の流れが弥生文化をもたらし、日本語がアルタイ語化していったものであろう。

この結果は鳥居龍蔵の言葉「日本の先住民族はアイヌ」[5]の言葉通りであり、従来の歴史学、民族学、言語学などの見直しが必要なることを示すものであろう。

発表の機会を与我えていただいた実行委員会、協議会の先生方に深く感謝致します。

参考文献：

- [1] 永田良茂「言語年代学における日本語系統、縄文語を推定する」
人文科学とデータベース第12回公開シンポ '06/12
- [2] 永田良茂「日本語の起源とアイヌ語」友月書房 '09/03
- [3] 崎山 理編「日本語の形成」三省堂 '90/03
- [4] 村山七郎・大林太郎「日本語起源」弘文堂 '95/05
- [5] 鳥居龍蔵「有史以前乃日本」磯部甲陽堂 '18/07

その他の文献は、[1]の文献に同じためここでは省略。

表現豊かな自然発話コーパスのアクセスについて

モクタリ 明子† 田畑 安希子† ニック・キャンベル‡
†神戸大学大学院総合人間科学研究科 ‡ダブリン大学トリニティカレッジ

本稿では、2000年から2005年までの5年間に渡って収録した大規模自然発話音声コーパス（全収録時間1,500時間以上）の一部を紹介する。コーパスの話者は学生・教員・主婦・子供など様々であり、全ての発話が課題なしの自然発話である。現在このコーパスを公開するための作業を進めているWEBページでは、書き起こしテキストを目で追いながら該当する発話音声聞くことや、特定の語彙を含む発話を検索することが可能である。また、同コーパスのうちすでに研究利用されている他のデータについても、その内容および研究成果を紹介する。

キーワード：自発的発話、インターネットアクセス、発話様式、外国語教育

Access to an Expressive Speech Corpus

Akiko Mokhtari† Akiko Tabata† Nick Campbell‡
† Graduate School of Intercultural Studies ‡ School of Linguistic, Speech and
Kobe University Communication Sciences Trinity College Dublin

This paper introduces a part of a large natural spontaneous conversation corpus (total recording time is more than 1,500 hours) which was collected over the period of five years from 2000 to 2005. Speakers are ordinary people including students, professors, housewives and children, and no task was given to the speakers. Part of the data is available on a website which is periodically updated, where you can listen to utterances with an aligned display of the speech annotation, and where you can also search a particular expression from the data base. This paper also describes other data of the corpus and some of the research results obtained from these data.

Keywords: Spontaneous speech, internet access, speech style, foreign language education

1. はじめに

言語研究や音声認識技術の向上など様々な目的のため幅広い研究分野において音声コーパスの需要は年々高まっている。いずれの場合においてもコーパスをデザインする際に、いかに本来我々が自然な状態で発している発話音声に近いものを、音質を落とすことなく収集できるかということは大きな課題であろう。例えばこれまで感情や態度に伴う音声の変動について、主に音声工学や実験音声学において多くの研究がなされてきた。これらの研究の中には、悲しい物語や楽しい物語などを話者に読ませることによって、できる限り自然な感情・態度が込められた音声を引き出そうとしているものや (Iida et al.: 2003)、飛行機事故の実況中継を伝えるラジオのアナウンサーから得られた恐怖などの感情が込められている音声を分析対象とした研究もなされている (Stevens & Williams: 1972)。また Erickson et al. (2004) は実験室でモノログ的に収録されたものではあるものの、スクリプトなどは用意せず、自発的発話を通して偶発的に収録することができた悲しみの発話音声を刺激音として用いている。しかし依然としてアナウンサーによる朗読音声を基本とする「実験室的環

境」で収録されたデータが大部分の研究において用いられているのが現状である。

実験の主旨や研究の背景となる考え方によって、適切なデータの収録方法は異なってくるため、一概にどの方法が最も適切であるかを指摘することはできない。しかし現実の日常コミュニケーションと実験室的環境で収録された音声データとの間に大きな隔たりがあることが多くの研究者によって意識されているのも事実である (e.g. Cowie: 2000, Maekawa et al.: 2000)。

我々は日々の生活の中で、発話の背後にある意図を表したり、テキストだけでは伝えられない情報を付け加えたりするために音色や発話様式を変化させている。このような表現豊かな発話音声は、コントロールされた実験室的環境で収録されたデータでは到底カバーしきれないものである。

そこで日常の発話音声をもつ表現豊かな音色を反映する対話音声を収集するために自然環境で1,000時間の音声データを収録することを目指したのがESPプロジェクトである。ESPプロジェクトは、ニック・キャンベルをプロジェクトリーダーとし、2000年から2005年までの5年間に渡り、科学技術振興機構(JST)の支援を受け、大規模自然発話コーパス(以下、ESP

コーパス) を構築した。このプロジェクトは、コントロールされたデータの収集や信号処理を行うのではなく、データ数を増やすことで日常会話の中に現れる豊かな音色をカバーすることができるという考え方のもとに進められた。収録に参加した話者は全て一般の人であった。収録された音声データは2人以上の話者によるインタラクティブな対話であり、収録場所はデータによって異なるものの話者の自宅や大学の研究室など話者がリラックスして話せる場所が多い。後述するように、より高音質のデータを収録するために録音ブースで行われた対話も一部含まれているが、いずれの場合も話者に課題は一切与えられず全てのデータが完全に自発的な発話である。

ESP プロジェクトの研究実施体制は、研究項目の異なる7つのサブグループから構成された。データ収集は主に神戸大学国際文化学部の教員・同大学院総合人間科学研究科の院生から成る意味構造グループと、研究代表者のニック・キャンベルおよび国際電気通信基礎技術研究所(ATR)の研究者から成るシステム応用グループによって行われた。本稿では、録音総時間数1,500時間におよぶ全コーパスのうち、意味構造グループによって収集されたデータ(以下、神戸データ)について、その内容および研究利用のためのアクセス方法を中心に説明する。また第3節では、すでに研究利用されているシステム応用グループによって収集されたデータの概要およびその研究成果についても紹介する。

2. 神戸データ

2. 1. 収録

音声データの収録は、意味構造グループのメンバーによって行われた。話者・収録環境・収録機材・課題などについて、詳細を以下に述べる。

<話者>

意味構造グループのメンバーおよびその家族・学内外の友人など、アナウンサーや俳優ではない一般の人たち。



図1: 収録風景

<収録環境>

ほとんどの対話は話者がリラックスして話することができるよう自宅や大学研究室内で収録された(図1参照)。一部の対話は対面式録音ブース(YAMAHA, ANF35S11LL)で収録された。録音ブースでの収録は、2人の話者が別々のブースに入り、ガラス越しに対話をする形で行われた。録音ブースで収録されたデータは各々の声を相手の声と混ざることなく取り出すことができる。

<収録機器>

DAT TCD-T10 (SONY) および DAT WALKMAN TCD-D100 (SONY) を用いた。

<課題>

話者に課題は一切与えられなかった。

<その他>

データを収録するにあたり、話者には法律書式での事前承諾を得た。収録後、個人情報の有無など、法律的に問題が生じる可能性のある発話部分には、ブザー音(ビーブ音)を施した。メンバー以外の話者には原則として謝金を支払ったが、種々の事情で支払われなかった場合もあった。

2. 2. 書き起こし

書き起こしには、言語情報の他に「笑い」「咳」「言い淀み」などの非言語情報を含む31種類のタグが付与されている。表1にタグの全種類を提示する。

表1: コーパスタグ付け記号

番号	タグの内容	英語名	タグ記号
1	笑い	laughing	@W
2	咳	coughing	<CO>
3	叫び	crying	<CR>
4	ささやき/不明瞭音	murmur/uncertain	<MU>
5	舌鼓	smack	<SM>
6	咳払い	hawk	<HA>
7	吸気音	ingressive	@S
8	呼吸音	breathing	<BG>
9	力み声	creaky	<CK>
10	息混じりの声	breathy	
11	震える声	trill	<TR>
12	音変異	variable	<VA>
13	鼻音化	nasalization	<NA>
14	母音の伸長	lengthening	<LE>
15	焦点	forcus	<FO>
16	流暢でない・言い淀み	disfluency	<DF>
17	ポーズ	pause	#
18	沈黙	silence	#
20	上昇	rise	<RI>
21	下降	fall	<FA>
23	始める	start	<-S>
24	終わる	end	<-E>
26	ノイズ	noise	<NO>
27	フィラー	filler	<FI>
28	オノマトペ	onomatopoeia	<ON>
29	繰り返す	repeat	<RE>
30	ビーブ音処理		{固有名詞}
31	書き起こし不可		[?]

2. 3. 公開

ここまで紹介してきた音声データは、現在インターネットで公開する作業を進めている。データの公開を可能にするために、文字書き起こしをチェックし、個人情報および問題発言の有無を再確認した。そして該当箇所があった場合には、ブザー音処理を施し、完全に聞こえないようにした。話者に連絡をとり、双方から承諾を得られた対話だけをネット上に公開することにした。

これらの作業は進行中だが、すでに個人情報等のチェックが済んだデータは <http://www.speech-data.jp/tabu/kobedata/> にアクセスすると、実際に聞くことができる。

<対話音声の再生>

この WEB ページでは、対話音声とその書き起こし、また対話を視覚的に示したバーチャートを同時に表示・再生できるフラッシュを使用している。このフラッシュには 1) 音声再生表示、2) リスト表示、3) 全チャート表示の 3 種類の表示機能があり、切り替えて表示することができる。

まず聞いてみたい音声ファイルをクリックすると、『音声再生表示』の状態では 60 秒分の書き起こしがバーチャートで表示される（図 2 参照）。画面右上の「LIST」と書かれたボタンをクリックすると、『リスト表示』に切り替わる。画面右上の「All View」というボタンをクリックすると、『全チャート表示』に切り替わる。

1) 音声再生表示

『音声再生表示』では、対話音声を聞きながらバーチャートと書き起こしを同時に見ることができる（図 2 参照）。

音声を停止している状態でバーチャート枠内でマウスを動かすとチャート内のマウスの位置

ver 3.21 : A40/A40-A1.flv
0:00

speaker	start	end	subtitles
L	2.140	3.599	めがね掛けへんで大丈夫なん
R	4.459	5.405	掛けたらね、<へへ>@W
R	5.405	6.460	@S
R	6.460	6.579	<木>
L	5.459	7.209	うん、え、それ、だて(じゃ)ないよね(え)
R	7.170	7.829	<違>うんですよ
R	8.579	9.032	今ね
R	9.175	10.859	二週間コンタクトが切れてね(え)
L	10.810	11.710	<う>ん、<ふふ>@W

図 3: リスト再生表示の画面

に縦線が表示され、その位置に対応した書き起こしがバーチャート下部に表示される。音声再生中は、再生位置に合わせて書き起こしが表示されるようになっている。音声の再生・停止はバーチャートの下に表示されているコントロールパネルを操作することによって行う。

2) リスト表示

『リスト表示』の状態では、書き起こしデータを一覧表で見ることができる（図 3 参照）。表内でマウスを動かしてクリックすると、『音声再生表示』の状態に切り替わり、マウスが指している行のスタート時間から音声再生される。画面右上の『Return』というボタンをクリックすると、『音声再生表示』に戻る。

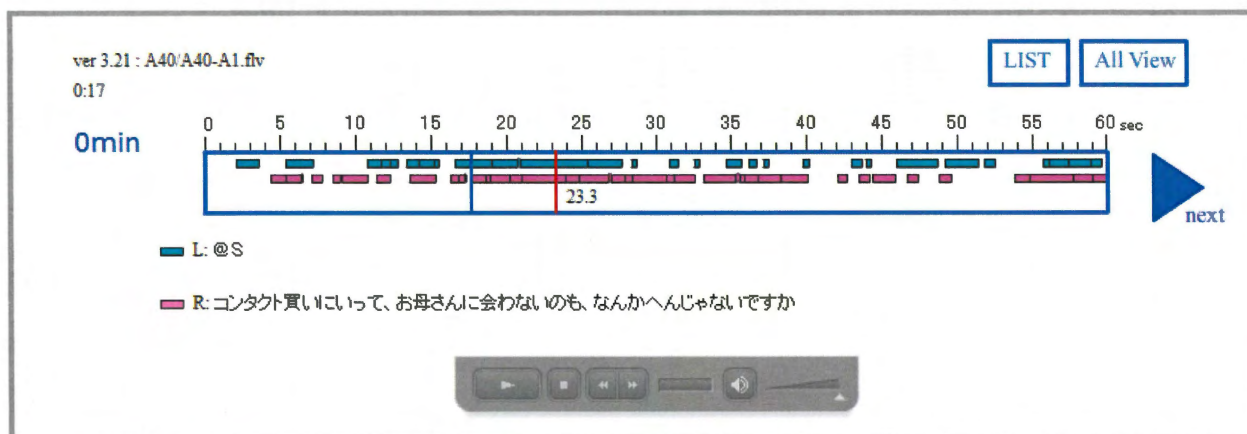


図 2: 音声再生表示の画面

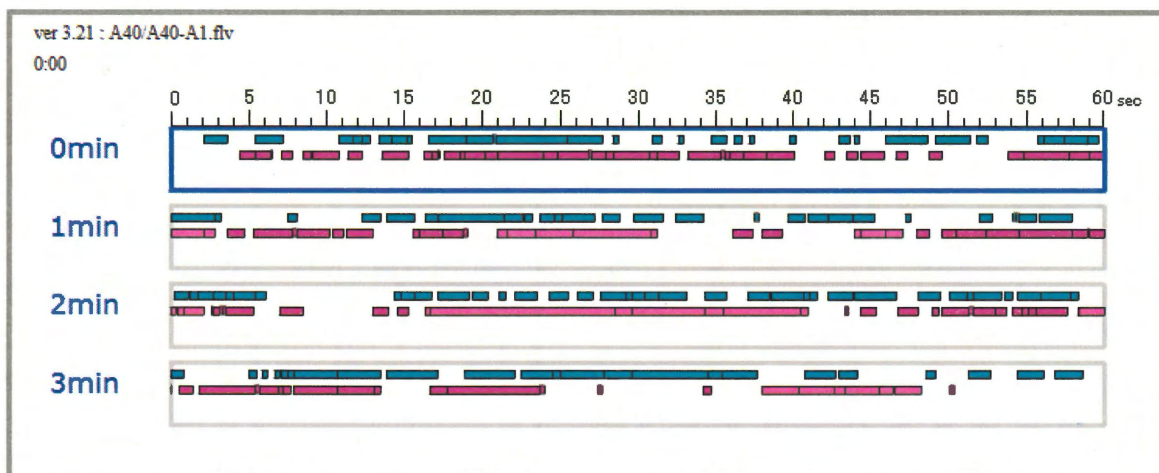


図 4: 全チャート表示の画面

3) 全チャート表示

『全チャート表示』の状態では 1 分一段のバーチャートを複数段一度に表示する（図 4 参照）。マウスを動かすことによって、聞きたい箇所を選択することができ、その状態でクリックすると『音声再生表示』に戻り、選択された段のバーチャートが表示される。

<語彙検索>

トップページからは特定の語彙を検索できる「語彙検索ページ」に移動することができる。検索したい語彙を入力し「SEARCH」ボタンをクリックすると（図 5 参照）、その語彙を含む発話の一覧が表示される（図 6 参照）。ここでは「ありがとう」と入力し、その検索結果として「ありがとう」を含むデータベース内の全ての発話が表示されている。

「PLAY」ボタンをクリックすると、該当箇所の音声再生される。コントロールパネルの

「巻き戻しボタン」「早送りボタン」を操作することによって、前後の発話を聞くことも可能である。また正規表現を用いて「^ありがとう」とすれば「ありがとう」で始まる発話が、「ありがとう\$」とすれば「ありがとう」で終わる発話が、そして「^ありがとう\$」入力すれば「ありがとう」のみの発話が表示されるようになっている。

語彙検索をする際に特定の音声ファイルのみにチェックを入れると、そのファイルのみが検索対象となる。図 5 では全てのファイル（A01～A63）にチェックが入っている。

<その他>

この WEB ページでは、2 節で説明した書き起こしルールおよび各データの話者・収録環境に関する情報も閲覧できるようになっている。話者については性別・収録当時の年齢・使用方言について書かれている。

KOBE Search
Search Word:
It may take a long time to show the result. Please be patient.

☒ A01	☒ A11	☒ A24	☒ A35	☒ A41	☒ A51	☒ A61
☒ A02	☒ A12	☒ A25	☒ A36	☒ A42	☒ A53	☒ A62
☒ A04	☒ A15	☒ A26	☒ A37	☒ A44	☒ A54	☒ A63
☒ A05	☒ A16	☒ A27	☒ A38	☒ A45	☒ A57	
☒ A07	☒ A17	☒ A28	☒ A39	☒ A46	☒ A58	
☒ A09	☒ A19	☒ A29	☒ A40	☒ A47	☒ A59	
☒ A10				☒ A48		
				☒ A49		
				☒ A50		

図 5: 語彙検索の画面

start	end	filename	speaker	PLAY	
48.091	48.687	A02-A1	L	PLAY	ありがとう
159.161	160.713	A02-A3	L	PLAY	ありがとうございます
20.767	22.147	A10-A1	R	PLAY	あ、ありがとうございます
403.339	405.544	A12-A0	L	PLAY	ありがとうと思ったけど、{xxxx}だったんですね
1821.510	1823.587	A12-A0	R	PLAY	ありがとう、ありがとうございますね、^^^@W
298.913	300.590	A16-A14	R	PLAY	なるほど、ありがとうございます
180.831	181.823	A16-A15	R	PLAY	ありがとうございます
379.840	381.218	A16-A3	L	PLAY	いや、ありがとうございますって言って
229.126	230.634	A16-A7	R	PLAY	うーん、ありがとう
7.080	8.553	A17-A1	L	PLAY	ありがとうございます
7.487	10.207	A17-A1	R	PLAY	はい、おねがいます、ありがとうございます@Wます
43.140	44.634	A19-A1	L	PLAY	はい、あ、ありがとうございました
117.959	118.339	A19-A11	L	PLAY	ありがとう
117.319	118.967	A19-A11	R	PLAY	どうも、ありがとう、ア
235.488	237.801	A25-A10	R	PLAY	一応終わりに、シ、どうもありがとうございました、もう

図 6: 語彙検索の結果を表示する画面

3. ESP コーパスのその他のデータ

ここではデータ整理が済み、すでに研究利用されている ESP コーパスのうち 2 つのサブセットについて、その内容と主な研究成果を紹介する。

3. 1. 日常会話データ

＜データ内容＞

プロジェクト開始時 32 歳であった日本語話者である女性 FAN の 2000 年から 2005 年までの 5 年間に渡る日常会話音声収録したもの。ヘッドフォン型の小型マイクとミニディスクプレーヤーを用い、話者が可能な限り日常的に機器を装着して収録を行った。この間に妊娠した話者は、2004 年 11 月 11 日の出産当日も平常と変わらず収録を続けた。

全て対話形式で行われたものであるが、対話相手の音声は収録されていない。対話相手は合計 112 人であり、家族・友人・他人・子供に大きく分類することができる。さらに話者による独り言発話も他の話者との対話中に頻繁に行われていたため、それらも収録されている。発話場は自宅や友人宅などであり、会話収録のために実験室や特別な場所に出向くことはなく、いずれも話者が日常生活を送っている場所で収録されたものである。FAN と対話相手は同じ場所で会話していることもあれば、電話で行われた会話を収録したものもある。このデータの総時間数は 600 時間であり、全てが自発的な自然発話である。

＜研究成果＞

1 話者の発話様式が、対話相手に応じていかに変化するかを計量的に示すことに成功した。

図 7 はこのデータの話者が発した 100 時間の

対話音声の声の固さと高さを示している。これらの声質の変化は、対話相手に対する親しさや発話の丁寧さと関連していることが分かっている。図 7 左図からは例えば「子供」「他人」と話すときは声が高くそして柔らかくなるのに対して、「家族」「友達」と話すときは低く固くなっていることが分かる。図 7 右図は、左図に示された家族に対する発話を、（家族の）メンバーごとに分析した結果である。ここでも例えば「子供」に話すときは声が柔らかくなるのに対して、「夫」と話すときは固くなるなど、2 つのパラメータの振る舞いから、話者と家族の各メンバーとの人間関係が伺える。詳しくは Campbell&Mokhtari.P. (2003) を参照されたい。

3. 2. 電話対話データ

＜データ内容＞

人材派遣会社を通して選ばれた日本語話者 6 人と非日本語話者 4 人が、週 1 回 30 分ずつ、10 回に渡って行った電話対話を収録したもの。総時間数は 105 時間であり、会話は全て日本語で行われた。収録開始時、10 人の話者は初対面であった。6 人の日本語話者のうち、男女 1 人ずつ計 2 人の話者は、自らの家族との会話も収録している。

＜研究成果＞

対話相手の違いに応じたものだけでなく、収録回数を重ねるごとに変化する話者間の距離に応じた発話様式の変化を観察することができた。また非日本語話者との会話を収録することにより、相手が日本語話者であるときの会話と比べて、どのように発話様式が変化するかも捉えられている。

ここでは日本語話者である女性 JFA の全収録データに高頻度で出現していた 6 発話（「ああ (a,a-)」, 「あの (ano)」, 「でも

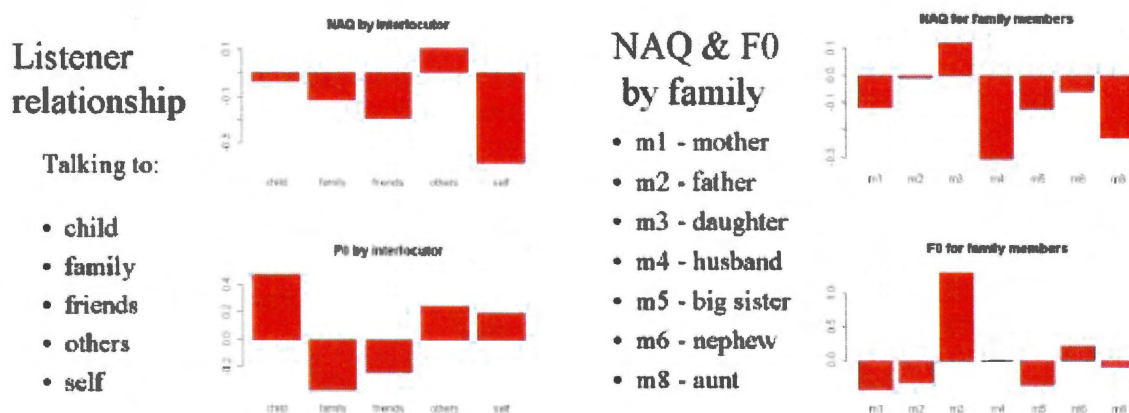


図 7: 対話相手による発話様式の違い

対話相手のグループ（左図）および家族構成員（右図）ごとの NAQ（声の固さ，上図）および F0（声の高さ，下図）。

表 2: 話者 JFA の対話相手に応じて変化する高頻度で現れる発話

JFA:	CFA	CMA	EFA	EMA	JFB	JMA
a, a-	143	145	88	89	138	170
ano	224	277	221	176	209	266
demo	41	24	31	17	89	134
e-	48	51	37	25	74	94
hai	2932	2234	2181	3239	72	33
un, un	1029	546	585	1190	909	1037

(demo)」、「えー (e-)」、「はい (hai)」、「うん、うん (unun)」を洗い出し、これらの発話の対話相手に応じた使用頻度の違いを、とりわけ日本語話者と非日本語話者を相手にしたときの違いをまとめたものを表 2 に示した。6 人の対話相手のうち、CFA・CMA・EFA・EMA の 4 人は非日本語話者であった。表から、例えば「でも」という発話が非日本語話者との対話データにはあまり多く出現していなかったが、日本語話者との対話データには多く出現していたことが分かる。それとは逆に、「はい」は非日本語話者との対話データには多く出現していたが、日本語話者との対話データにはあまり多く出現していないことが分かる。詳しくは Campbell (2007) を参照されたい。

4. まとめ

本稿では ESP プロジェクトが構築した表現豊かな大規模自然発話音声コーパスのうち、現在公開作業を進めているデータを紹介するとともに、すでに研究利用されている他のデータについてもその内容および研究成果の一部を説明した。これまで自然な音声データの収集と、音響分析に耐え得る高音質のデータの収集は一者揃いの課題のように扱われてきた。しかし音声機器の発達や収録のノウハウが確立されるにつれて、このようなジレンマも緩和されつつあるように思われる。ESP プロジェクトでも、3 節で紹介したように実験室的環境ではない環境にて収録された音声データを用いて、音響分析を行うことに成功している。今後、音響分析を必要とする研究においても、自然な環境で収録されたデータの需要が高まることが予想される。

冒頭でも触れたように、音声コーパスを用いた研究は音声工学や言語研究など様々な分野において盛んに行われている。しかし外国語教育、とりわけ日本語教育におけるコーパスの利用は遅れており、その取り組みが今まさに始まったばかりだと言われている (砂川: 2009)。その取り組みとして、日本語学習者が遭遇する様々な場面を設定し、それらに必要な語彙・表現が含まれるようなコーパスをデザインすることは大変重要である。しかし本稿で紹介したような「普通の日本人同士の、普通の会話」を収録し

たコーパスも、外国語教育に大いに貢献するだろうと思われる。多くの外国語学習者が、母語話者同士の自然な会話を聞き、全く理解できずに愕然とした経験があるだろう。映画やできるだけ自然な会話を題材とした教材も増加しつつあるが、それらが依然として不特定多数の聞き手が理解できるように話されている発話であり、「普通の人の普通の会話」に比べて聞き取りやすく作られていることは否めないだろう。今後外国語教育の分野でも本稿で紹介したような一般の話者による自然な発話コーパスが利用されることが期待される。

[付記] 本稿は科学技術振興機構 (JST) による戦略的創造研究推進事業 (CREST) 「表現豊かな発話音声のコンピュータ処理システム」(研究代表者: ニック・キャンベル)、日本学術振興会の科学研究費助成金による基礎研究 (A) 「人物像に応じた音声文法」(課題番号: 19202013, 研究代表者: 定延利之) の成果の一部である。

参考文献

- [1] Campbell, N. & Mokhtari, P.: Voice quality; the 4th prosodic dimension, *Proceedings of the 15th International Congress of Phonetic Sciences*, pp.2417-2420, 2003.
- [2] Campbell, N.: How speech encodes affect and discourse information: Conversational Gestures, *NATO Security through Science*, Vol.18, pp.103-114, 2007.
- [3] Cowie, R.: Describing the emotional states expressed in speech. *Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*. 2000.
- [4] Erickson, D., Yoshida, K., Mochida, T., & Shibuya, Y.: Acoustic and articulatory analysis of sad Japanese speech, 第 18 回日本音声学会全国大会予稿集, pp. 113-118, 2004.
- [5] Iida, A., Campbell, N., Higuchi, F., & Yasumura, M.: A corpus-based speech synthesis system with emotion. *Speech communication*, Vol.40, No.1, pp.161-187, 2003.
- [6] Maekawa, K., Koiso, H., Furui, S., & Isahara, H.: Spontaneous speech corpus of Japanese, *Proceedings of LREC 2000*, pp.947-952, 2000.
- [7] 砂川有里子: コーパスを活用した日本語教育研究, *人工知能学会誌*, 24 巻, 5 号, pp.656-664. 2009.
- [8] Williams, C. E., & Stevens, K. N.: Emotions and Speech: Some Acoustical Correlates. *The Journal of the Acoustical Society of America*, Vol. 52, No.4. pp. 1238-1250, 1972.
- [9] <http://www.speech-data.jp/tabakobedata/>

国立民族学博物館のデジタル・アーカイブズ

久保正敏

国立民族学博物館・文化資源研究センター

国立民族学博物館に収蔵されている様々な民族誌資料を文化資源ととらえ直し、文化資源のアーカイブズ整備の現状と課題を紹介する。文化人類学・民族学の観点から言えば、文化資源を扱う場合に西欧的な知的財産権だけでなく文化固有の知的財産権や個人・集団に対するリスペクトが必要である。特にデジタル化、ウェブ公開する場合に十分な注意が求められる。また、国立民族学博物館には、1974年の創設以来自ら制作してきた映像も含む様々な映像・音響・写真資料も収蔵されている。これらは無形の文化資源を記録し可視化するうえで重要な資料であるが、アナログ・デミアとデジタル・メディアの保存と活用のバランスという大きな問題を提起している。こうした問題点も含め、現状を紹介する。

キーワード：文化資源、文化固有の知的財産権、資料の保存と活用、データ・マイグレーション

Digital Archives of the National Museum of Ethnology

Masatoshi Kubo

Research Center for Cultural Resources, National Museum of Ethnology

This manuscript describes the current status of digital archives for managing cultural resources in the National Museum of Ethnology, together with the discussion about problems relating intellectual properties for each culture, data migration, etc.

Keywords: cultural resources, cultural intellectual properties, conservation and utilization of cultural resources, data migration

1. 民族誌から文化資源へ

国立民族学博物館（民博）は、人類社会の多様性の発見とその知的遺産の継承をめざして文化人類学・民族学の研究センターであると同時に、同分野および関連諸分野に関わる資料・情報センターの役割も果たしてきた。これまで収集されてきた資料の形態・種類は多岐にわたっている。民博所蔵資料は、以下の四つに大別される。文献図書資料、標本資料（モノ＝生業用具、生活用具、儀礼用具、などの人工物）、映像音響資料（写真、映像記録、映像・音響記録）、研究アーカイブズ資料（フィールドノートやメモ、雑多なモノなどフィールドワーク途上および研究途上での様々な記録や収集物）。「表現メディア」の形態（文字、画像・映像、音声・音響）や情報の回数も様々である。

現在、文献図書資料として日本語及び外国語図書 610,000 冊、16,000 タイトルの雑誌、標本資料は 275,000 点余り、映像音響資料は 70,000 点余りを所蔵するほか、様々な民族学研究アーカイブズを所蔵している。これら資料に対する書誌的事項を中心とするデータベースが研究者の協力の下で民博の職務著作物として作成されている他に、研究者の研究資料やそれに基づいた研究成果としての研究データベースも作成・公開されている。民博のウェブサイトを参照されたい (<http://www.minpaku.ac.jp/>)。

こうした資料類は、従来から総じて「民族誌資料」と呼ばれてきた。民族に関する記述や記述されたものを意味するが、これは、外部から記述する側からの視点にのみ立った言い方ではないだろうか。

1980 年代以降、文化人類学の分野では、従来のような資料収集や記述に対する見直しが生まれ、J. クリフォードのように「文化を語ること」「文化を翻訳すること」とは何かを突き詰めて考える動きがある[1]。すなわち、(a) 従来の民族誌は、書き手とテキストを権威付けて客観性を謳うためのレトリックに満ちているが、そもそも客観的なテキストはあり得るのかという問題、(b) 書き手と対象文化出身者の間には、政治的・経済的・文化的権力がもたらす非対称性（しばしば植民側が被植民側を記述する）が厳然と存在する点、(c) 西欧言語という強い言語で口頭文化を文字化することによる固定化や権威化の問題、(d) 収集された情報や資料は、記述する側に占有されたままで現地に還元されたかの問題（現地から宗主国等への資料返還＝repatriation 要求はその反映である）、などを問いかける。ポスト構造主義と呼応し、既存の権威や、自己と他者の区分について脱構築を提唱するものと言える。

こうした議論を踏まえれば、記述する側の視点のみを含意した「民族誌」資料ではなく、「文化資源」と呼び換えることが適切であろう。

なぜなら、こうすることによって、それらが本来は現地の資産であることを積極的に示し、各文化に対する外から視点だけでなく内からの理解にも役立つ資料であり、文化の継承・復興、あるいは創成にも寄与する資源と捉え直すことができること、そしてその集積が人類の知恵の宝庫と見なすことができるからである。

さらに、民族誌を文化資源と捉え直すならば、前述したような有形の資源、あるいは外在化が容易な資源だけではなく、より広い対象を資源に含めることができる。すなわち、個人が持つ技術、知識、記憶、ノウハウなども、その文化成員に「身体化された文化資源」と見なすことができる。さらには、個人が持つ様々な人的ネットワークや組織、制度、概念なども、「制度化された文化資源」と言える。例えば、知的財産権の概念について見れば、オーストラリア・アボリジニ文化においては、個人の顔を具象化したイメージや写真は、当人の死後、外部にさらすことが禁じられている。これは西欧流の「肖像権」とは異なる、宗教的な背景を持つ「文化人格権」と見なすことができよう。そこで我々は、文化資源として、図1に示すような3つのカテゴリーを設定しようと考えている。

もっとも、第2、第3のカテゴリーの文化資源も、最終的には第1カテゴリーの外在化された資源に変換されることによって、理解や共有が容易となるのであり、第1カテゴリーの文化資源の重要性は言うまでもない。

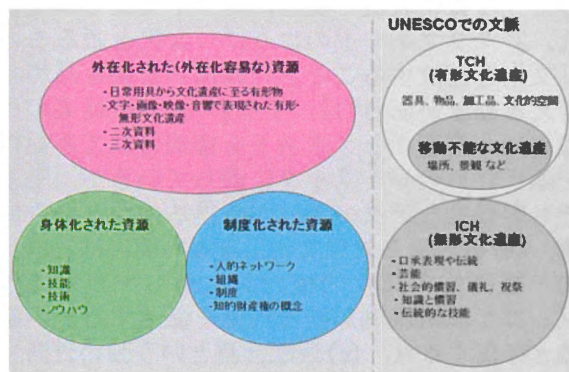


図1 文化資源の概念

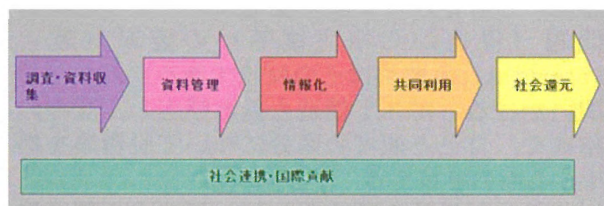


図2 開発から社会還元までのフロー

特に後述するように、現地に所有権を確保しつつ、共有・活用の両者のバランスを図る最良の方法がデジタル化であろう。ここに、客体化された文化資源のデジタル・アーカイブズを構築する意義がある。

図2には、現地における文化資源の調査・収集から、資源の保存管理、整理と情報化、研究での利用、社会還元と、より広い活用の流れをフローとして表現したものを示す。

2. フォーラム型のデジタル・アーカイブズ

前述のポスト構造主義に対応した見直しは、資料収集や取材、写真撮影についても同じ事が言える。特に前記(b)に関しては、資料の持ち去りや非現地語のみによる記述は、現地からは「文化の剥奪」と見なされても仕方がない状況である。こうした反省に立って、今後の資料収集や情報化の指針を打ち立てることが必要であろう。また、集積された資料数や情報量が膨大化した事態に対処するドキュメンテーションの方法論も確立せねばならない。

前記(a)、(b)の問題の解決方法の一つは、「共同性」の徹底である。記述される側、記述する側、記述を利用する側、三者の間による議論の場、フォーラムを保証することであろう。(a)に対しては、「一次収集、一次記述とその後の参照・活用との間でのフォーラム」の推進である。記述や収集はいずれも現実の切り取りの一つに過ぎず客観性を謳うことは無理だと認め、その代わりに、どの時点でも、収集や記述の責任は署名によって明確化されるべきことを前提とする共同性である。それに伴って、知的所有権保護や認証制度を充実させる必要がある。

(b)に関しては、「現地と非現地との間のフォーラム性」を追求することが解決法の一つである。例えば、資料は現地に残し、それに関する記述を共同で行い、その結果の共有を推進する、といった方向性である。現地からの資料返還要求についても、共同性を追求する中から解決法が見いだせるかも知れない(所有権を現地に返還するが資料保管は条件の整った先進国が担当する virtual repatriation の考え方もその一つであろう)。こうした共同性を保証する情報化の運営体制とシステム構成、「フォーラム型デジタル・アーカイブズ」が、今後、文化資源アーカイブズを構築する上で推進すべき一つの方方向性である。ウェブ 2.0 とも重なるこの考え方で次のような恩恵が得られるだろう。

(1) 専門家占有から共有・共創へ

民族学研究者のみが専門性を持つという傲慢

が否定される今、現地研究者・現地関係者・異分野研究者・非専門家が集い合うフォーラム型共同作業と知の共有の中にこそ、互恵的な成果が得られる。

(2) 研究倫理の転換

研究者が一方的に現地から文化剥奪を行ってきた従来の手法から、現地の知的所有権を保証する研究手法へ転換し、一方的な剥奪から、現地との共有・共同型の情報収集と人類共通の知的財産形成へ、という流れを確立する。

(3) 共有による情報の高精度化

現地関係者との共同による現地情報処理方式、いわば「ポイント・オブ・フィールド方式」情報化によって、情報収集・処理が高品質化される。利用者の serendipity（発見能力）を発揮させる手法である。さらに、精度の低い過去の民族誌記録に対しても、それを現地にフィードバックすることにより、現地から情報が追加されて情報の質が高まる。

(4) 現地への文化的還元

過去に集積された情報の現地との共有は、失われた現地文化の復元・伝承・復興に寄与できる。特に、グローバル化の進行と共に、既存文化の急激な変化が進む一方で、自文化表象・主張の手段として、失われつつある文化の復興に努める集団も増加している現在、現地への研究成果還元の大きなチャンネルとなり得る。

(5) 文化資源の集積から人類智発見へ

実世界の情報の宝庫である文化資源に、データマイニング手法などを適用すれば、人類智を表現する人類智オントロジーが構築できる可能性もある。

(6) 文理融合の効果

もっぱら民族学研究資料としてのみ捉えられてきた文化資源は、環境・開発・平和など世界的な問題解決の糸口となる情報を含んでいる。自然科学系も含む他分野の人々とも共有するなかから、人文社会学系と自然科学系との研究融合への架け橋となる。

3. デジタル・アーカイブズ構築のための課題

デジタル・アーカイブズの構築によって文化資源の共有を図るには、有形資源の蓄積とデジタルな無形資源の流通とを有機的に結びつけるための、次のような諸課題を明らかにせねばならない。

(1) 領域専門アーキビスト

資源化の流れの最上流の段階から考えてみると、まず、生（なま）の資料や情報を、共有に耐えるレベルにまで整理・分類などの処理を施

し、文化資源化するには、原資料や情報の領域ごとの特性に応じた適切な処理が必要である。そのためには、各フィールドに関する専門的知識、および、資料・情報の組織化に関する専門的知識、の両者が必要となる。これらを合わせ持ち、各領域ごとの専門性を備えた「領域専門アーキビスト」の養成が、今後ますます重要性を増すであろう。当然このアーキビストには、デジタル化の光と影に対する深い認識が求められる。しかし残念ながら、日本ではアーキビストの重要性に対する認識が低く、専門性の社会的位置づけも未だ低い。これを改めるには、いわば「アーキビスト市民権」を確立するための、制度的、法的な整備が必要となろう。

(2) 文化資源リテラシー

他方、民族学調査の場合のように、研究者自らがフィールドワーク途上で様々なメディアを用いて情報収集を行う際には、その後のアーカイブズ形成を見越した権利処理や情報処理を施しておくことが望ましい。こうした資源化のためのノウハウやスキルおよび様々な資源利用のスキルの両者、すなわち「文化資源リテラシー」が、今後の人文社会科学系の研究者には求められる。このリテラシー向上を研究者に奨励するためには、資源化の作業が、正当な研究活動の一環として過不足なく評価される仕組み、例えば、研究映像の撮影制作、資料整理とデータベース作成、研究成果の展示などが研究業績として論文などと同等に評価される制度導入も検討されるべきである。

(3) 文化資源の長期保存と利用の両立を図る手法としてのデジタル化

いったん形成された文化資源の長期保存に対する認識が低いことも大きな問題である。標本（モノ）資料など有形資源のみならず、映像音響などアナログの資料についても、長期保存体制の整備が必要である。ただし、デジタル化にのみ焦点が当たり、原資料の保存が軽視される事態に陥ってはならない。デジタル化は、保存と活用を両立させる手法の一つと考えるべきであろう。特に貴重な資料については、劣化の進行を遅らせる保存管理の体制を保証するような、法的・予算的な制度の確立が求められる。

アナログ・メディアの保存に関しても、メディア自体が持続可能かの問題がある。例えば、光学フィルム、印画紙、音響テープの製造中止が発表されたために、関連部材や再生機の入手が困難になっていく。さらに近年、フィルムアーカイブズなど映像資料の保存管理機関に共通し、民博も抱える悩みは、光学フィルムや磁気フィルムのベース（支持体）の劣化問題である。

1990年代から主流となったPET（ポリエチレン・テレフタレート）ベースは寿命が500年とされる。しかし、それ以前に主流だったTAC（セルロース・トリアセテート）ベースは1954年以来その接着容易性からも広く使われてきたものの、保存環境にもよるがアセテート成分が加水分解して酢酸ガスを発生し、それが更に自身と周囲の加水分解を促進する現象で、最後にはフィルム自体が雲散霧消する。酢酸ガスの臭いから「ヴィネガー・シンドローム」と呼ばれるこの現象はごく最近（1991年）に発見されたばかりで、対処法も標準化されている訳ではない。劣化の進んだフィルムを他から隔離し低温で保管することが基本だが、所蔵量の規模により隔離施設のための経済的負担は大きい。TAC以前の1950年代まで広く使われたNC（セルロース・ナイトレート）ベースは、可視光線の透過率が良いが、無酸素で燃焼する危険物で、加水分解による劣化とともに発火点も低下する。一定重量以上所蔵すれば消防法の対象となる。

対象のフィルムがこれら3種類のどれに相当するかの判定も困難で、映像編集過程で別種のフィルムが接着され1本の完成品となっていることが多いのも困難さを増す要因である。

デジタル・メディア化された資源の長期保存における問題点は、記録媒体が長期保存に耐えるかどうかともさることながら、デジタル化方式、記録媒体自体の頻繁な変化にある。例えば写真について見れば、技術進展に合わせた種々のデジタル化方式の栄枯盛衰であり、映像について見れば、ベンダー企業がオリンピックや万博などビッグイベントに合わせてデジタル化の方式と媒体を競って変化させてきたために、問題はより深刻である。これらの変化にあわせたデータの移行（data migration）を確実に行える体制や費用の持続的確保が必須となる。

すなわち、アナログ、デジタル両者の資源についての「持続可能な保存管理」という視点が必須である。時にはメディアの廃棄も念頭に置かねばならないが、唯一メディアの持つ「アウラ性」や、デジタル・メディアへの不信への配慮もまた、モノを扱う博物館として忘れてはならない[2]。

(4) 各種財産権の保護と共有のトレードオフ

文化資源の公開や流通は、所有権および知的財産権の保全と切り離すことはできない。原資料の所有者との間で、所有権と知的財産権それぞれの帰属に関する手続きなど十分な制度設計が必要である。文化資源として貴重な資料については、例えば、寄託制度や補償金制度などを確立すべきだろう。こうした財産権の処理を

個々の機関が行うことにも限界がある。知的財産権処理センター的な組織の整備も必要である。

しかし、現在の大勢である著作権保護を強化するプロパテントの流れは、資源共有とは相容れない部分の多いことも事実である。共有を促進するためには、Public Domain、あるいはクリエイティブ・コモンズ[3]の考え方に基づいた制度の整備も、考慮すべきであろう。例えば、研究資源のPublic Domainとして公的に認定された機関は、著作権の一部、公衆送信権、貴重資料に対するデジタル複製権などを免責される制度などもあり得る。実際に日本では、著作権法の改正案などで現在進行形のようなのである。

文化固有の知的財産権についても、十分な配慮が欠かせない。先述した「文化人格権」の他にも、オーストラリア・アボリジニ社会では、知的財産権を個人に帰するのではなく集団的な財産権（collective rights）と見なすことが伝統であった。部族の神話に関わる絵画デザインは部族の成員に共有されているのである。他方では、女性や子どもは見ることもすら許されない、儀礼に関わる聖物や秘物もある。これらを収集し博物館に所蔵すること、写真撮影しデジタル化することは、現地文化に対するリスペクトの観点からの検討が必要である。

アボリジニ文化を例にとれば、固有の知的財産権の概念も1960年代以降変化しつつある。西欧型社会にデザインを無頓着に利用されて来たことへの権利回復や、文化的アイデンティティ主張の一環として、知的財産権を強化するとともに個人に帰属させ、作家名を明示する美術工芸作品も登場するようになった。西欧型の知的財産権概念の影響とも言えるが、皮肉なことと同じ頃から西欧型社会では、public domainやcreative commonsなど、アンチ知的財産権の動きが現れるようになった。図3に示すように、非西欧型社会は、あたかもトラックを半周遅れで西欧型社会を追っているかのようである。

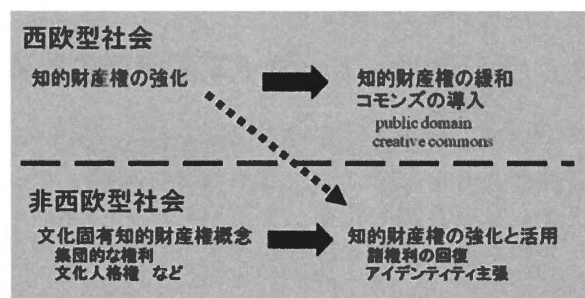


図3 知的財産権概念の相違と変化

(5) データベースの個別多様性を担保する横断検索の手法

デジタル化された文化資源の流通に関連して、しばしば文字コードや検索手法等の標準化が語られる。しかしながら標準とは、あくまでもその時点での大勢に過ぎず、未来永劫続くことは保証されていないものであり、相互利用のための手段の一つである、と割り切る必要がある。個別データを標準に統一することはむしろ危険であり、個別領域でのデータの多様性を許容すること、および、複数の標準の間での interoperability を確保すること、の二点を原則とするべきだろう。

多くの機関のデータベースを横断検索する手法に関する技術開発の動きとしては、データベース項目の共通化のためのメタデータとしてダブリン・コア、検索手続きとして Z39.50、など標準仕様を用いる手法に関する共同研究と試作開発が、1996 年以降、国文学研究資料館の安永尚志教授・原正一郎助教授（当時）を中心として進んできており[4][5]、民博もそこに参加してきた。その過程で問題になったのは、各機関が作成してきた、文学作品、美術作品や絵画から博物館の所蔵資料、あるいは自然科学的な情報まで幅広くカバーする個々のデータベースの特徴を生かした索引項目、検索項目を、如何にして共通化し横断検索を可能とするか、という古くて新しい課題であった。基本的には、図 4 に示すように、各データベース項目を、段階を経て共通のダブリン・コアにマッピングする手法が模索された。

この研究に基づいて、より大きな枠組みで実用化が図られたシステムが、次に概説する人間文化研究機構・研究資源共有化事業である。

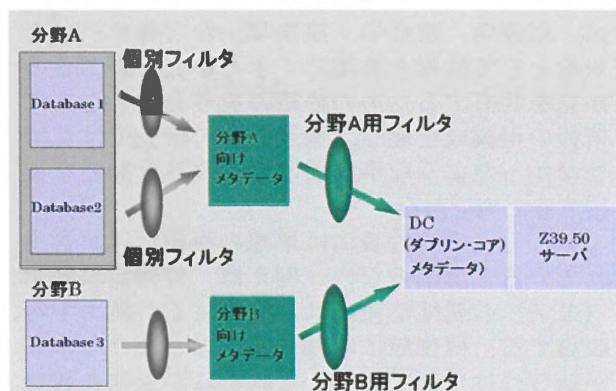


図4 分野間の相違を吸収するメタデータのマッピング（原正一郎「Z39.50 とメタデータによる研究機関間連携」『情報処理学会誌』Vol. 43 No. 9 : 968-974, 2002 年を改変）

4. 人間文化研究機構の研究資源共有化事業

2002 年 2 月、人間文化研究機構発足を控えて「人間総合科学の創出に関する調査研究委員会」（委員長：石井米雄）が発足した。そこに設置された 3 専門部会の一つ「研究資源専門部会」の提言に基づき、2004 年度の機構発足に伴い、石井米雄初代機構長の強い希望の下、研究資源共有化事業が推進されることとなった。前述の国文学研究資料館の共同研究成果などを踏まえて、2006 年度から、集中型システム、分散型システム、時空間システムの 3 機能に分けて、システム開発が開始された。2007 年度にはこれらが統合型システムとして一元化され、2008 年 4 月から一般公開された。図 5 の概念図に示すように、現在のところ、人間文化研究機構傘下の 5 機関が提供する 100 を越えるデータベースを横断検索できる「統合検索システム」と、研究者参加型のデータベースを管理する「nihuONE システム」が公開され、次のアドレスから利用できる。今後は機構外の研究機関との連携も順次図られる予定である。

(<http://www.nihu.jp/kyoyuka/database.html>)

具体的には、各機関に置かれた FES (Front End System) が原データベースの項目をマッピングしたメタデータ・データベースを作成・保持し、検索要求に応える検索サービス機能も持つ。

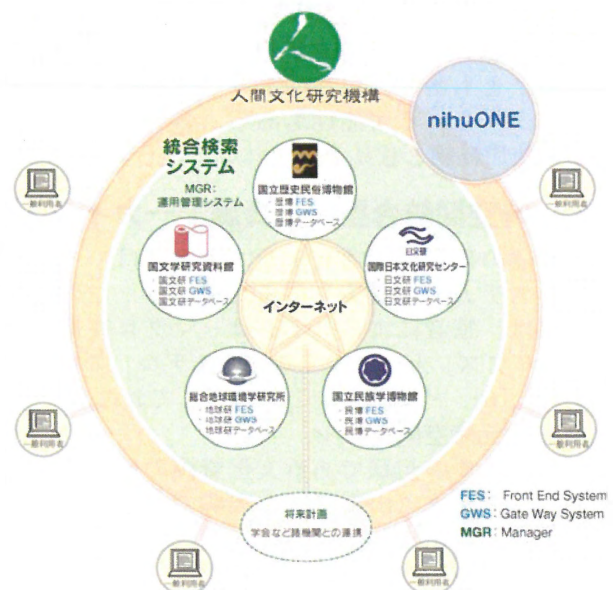


図5 人間文化研究機構・研究資源共有化システム概念図（2009 年発行のパンフレットより）

これに対応したクライアント機能は各機関に置かれた GWS (Gate Way System) が担い、ウェブからの検索要求と表示要求に答えて必要な処理を FES に渡す。各機関に GWS が置かれているので、ユーザはどの機関にアクセスしても同等の結果を得られる。ユーザからの要求にはまずメタデータ・データベースが返答して一覧リストなどを表示し、詳細データは、各機関の原データベースからユーザに提供される。

こうして横断検索システムの運用が始まり、今後、利用者の意見などから効果と問題点が見極められていくことになる。しかし、多様な分野のデータ項目を力づくでダブリン・コアにマッピングした感は否めず、異分野間の横断検索の難しさもあらためて浮き彫りになってきた。

情報検索の目的は、「ハンティング」と「ブラウジング」に大別されるが、あまりにも多様なデータベースから冗長な検索結果が得られる場合、ハンティングを目指すユーザにとって負担の増える恐れもある。そこで、研究分野を、例えば歴史・考古系、民俗系、民族系、美術系、文学系などに大別し、こうした大分類領域毎の個性・特殊性を確保したメタデータを作成することも考え得る。図4で言えば、各分野で異なるマッピングを行うことに相当する。その場合には、各大分類領域に対応した機構外のデータベースとの連携の核あるいはハブ機能を機構のシステムが担うことも方向性の一つだろう。他方では、ブラウジングを目的とする場合には、むしろ意外な検索結果が新たな発見のヒントとなることもあり、現在のような統合検索機能の効果が得られる可能性もある。多くの利用者からの使用経験と意見が待たれるところである。

5. 時空間統合型文化資源アーカイブズ

前述の研究資源共有化システムには、「時空間システム」の機能も含まれている。この出発点には、筆者による「ミクローマクロ往還型アーカイブズ」と「3次元時空間モデル」の提案があった[6]。その概要をここで示しておこう。

(1) 3次元時空間モデル

文化事象の記録である文化資源の解析によって、文化事象（出来事）を理解できるが、従来、社会学・政治学・地理学などマクロな解析と文化人類学のようなミクロな解析は、必ずしも相補的に行われてきた訳ではない。文化事象は、ミクローマクロ様々なレベルの事象が相互に関係し合うダイナミズムの中でこそ理解できる。ダイナミズムのベースとなるのが時空間情報であり、すべての文化資源は時空間情報を伴う事象として扱われるのが望ましい。

自然科学分野でも、気象など環境モニタリング・データ、地層の解析から推定される地震、洪水、津波の災害記録、など、様々な時空間情報を伴った事象が存在する。これら数値を記録文書など非数値情報と突き合わせれば、生態学的分析や地域史と気候変動史との相関関係など、大きな枠組みの中で数値の意味を明らかにできる。

このように、人文社会科学的な文化資源情報と自然科学的な情報の両者を、時空間情報をベースとする事象として捉えてアーカイブズ化し、その解析ツールを整備すれば、事象間の関係性を発見し、地域や領域の動的な関係を総合的に理解できるだろう。

例えば、筆者のフィールドであるオーストラリア・アボリジニの社会・文化を見ても、19世紀後半から対アボリジニ保護隔離政策が適用され始めた要因の一つには、雨量減少のため水場が減少した結果、白人入植者とアボリジニの接触と衝突の機会が増えたことがある。また同時期の本国英国での人道主義の流れや、当時脚光を浴びていた生物進化論が、保護隔離政策の一環として20世紀に入ってから混血児引き離し政策につながった。このような、個々の事象が相互に関係し合うダイナミズムは、個別のコミュニティ調査だけでは捉え難く、それを取り巻く諸要素の系を通時的・共時的に分析して総合し、要素間の因果関係など、自明ではない種々の相関関係を発見しない限り、理解は難しい。時空間情報を導入する目的はそこにある。

時空間情報に加えて、例えば、環境破壊、農業、自然災害、など、その事象を説明あるいは関係する複数のテーマやキーワードを付加すれば、テーマに基づく検索や解析も容易になるだろう。ここでテーマ値を取り入れるのは、例えば、気象学、社会学、民族学、など各テーマを対象として情報を蓄積してきた研究分野ごとの知見を共有するための枠組みを与えるとともに情報の再編成を促し、個別分野の研究では見出せなかったような事象間関係の発見と記述を支援するためである。

ただし、元の情報には単独の事象だけが含まれているとは限らない。例えば、博物館収蔵物（モノ）の属性情報には、製作する、使用する、収集する、博物館に収蔵する、など様々な事象が結びつけられていて、それぞれが意味の異なる時間値、空間値を持つ。文書資料にも複数の事象が記述され、それぞれ意味の異なる時空間値を持っている。従って、事象についても何らかの型を与え、時間値、空間値の意味を記述できる枠組みを用意する必要がある。

さらに、分野ごとの知識体系の差異、および、時間値・空間値の表現の差異を越えるためのメタ知識、あるいはオントロジーが必要である。このためには、研究分野ごとに異なるマイクロからマクロに至る視界のレベルや用語のレベルを越えるための「マクローマイクロ往還」の仕組み、すなわち、テーマ、時間、空間、それぞれにおけるズームイン、ズームアウトの機能も持たせる必要がある。

具体的には、時間軸の伸張に連動した様々なレベルでの時間表現、暦法の相違を吸収できる表示や検索の機能、空間軸においても、地図のズームイン、ズームアウトに連動した様々なレベルの地名での表示や検索の機能である。

そのために、時間値については、文化ごとに異なる様々な暦法や時代名、「～頃」「中世前半」などあいまいな表現を標準的な表記に統一する仕組み、空間値については、地域ごとの地名辞書を集積し、地名が指す空間的領域と標準的な緯度・経度との対応表などの仕組みが必要となる。また、地名やそれが指す領域が時間的に変化することにも対応しなければならない。

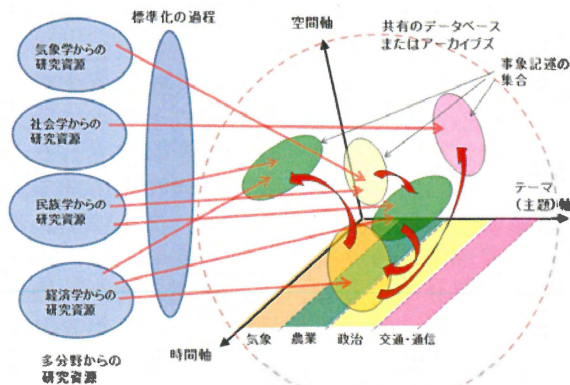


図6 研究資源共有プラットフォームとしてのマイクロマクロ往還型3次元事象空間

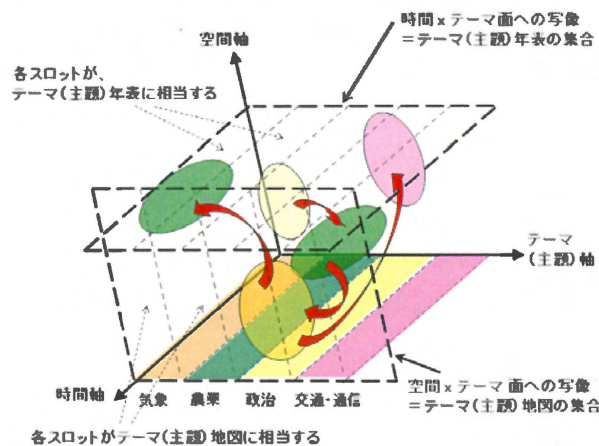


図7 3次元時空間モデルの可視化

テーマ値については意味のシソーラスやオントロジーの整備が必要となる。階層構造だけでなく網構造も前提とした意味のオントロジー整備は古くて新しい問題であり、難しい課題だが、ある研究者が作成したオントロジーを他分野の研究者も共有・追加できる仕組みを導入し、相補的にオントロジーを構築していくのが、現実的な対処であろう。

このような標準化の過程を経てアーカイブ化された事象群を、3つ組を軸とする3次元の事象スペースに再配置し、2次元表現として可視化する仕組みを用意すれば、事象間の関係の発見と表現が容易となることが期待される。これが、「3次元時空間モデル」と呼ぶものである。

これは、図6に示す通り、各研究分野からの文化資源が、標準化を経て、3次元空間内に事象記述の共有アーカイブズとして集積され、これに基づいて事象やその集合間の時空間的な関係性、すなわち図の矢印のようなダイナミズムを見出そうとするものである。この矢印は、降水量減少が農業生産減少をもたらし、栽培作物の転換指導など農業政策の変更につながり、それが農業や生態環境の変化とともに作物流通形態の変化に至る、といった事象の連鎖を想定している。共有型の文化資源アーカイブズによってこうしたダイナミズムの発見が期待できよう。

この過程を支援するためには、図7に示すように、3次元空間に展開された事象記述集合を「時間 X テーマ」面に写像した年表のような表現、「空間 X テーマ」面に写像した地理情報システム (GIS) のような表現が得られる可視化ツールが必要となる。テーマ軸との関係で言えば、テーマ軸に沿って年表上や地図上に写像されたものが、両ツール上ではレイヤーとして表示され、それぞれテーマ年表、テーマ地図に相当する。両ツールに、複数のレイヤーに対する一覧・検索・種々の演算処理機能を持たせることで、蓄積された事象間の共有のみならず事象間の関係性発見の支援が容易になる可能性がある。

現在、人間文化機構の統合検索システムに検索ツールの一つとしてウェブ・プラグインの形で組み込まれている「時空間検索機能」は、こうした可視化ツールを検索と結果表示に生かしたものである。しかし、現在の統合検索システムは、メタデータに基づく所在情報から原データベースへのアクセス・パスを提供することが主眼である。アーカイブズの一次資料を対象とした、より深い分析支援機能のためには、現在の統合検索システムを拡張するよりはむしろスタンドアロン型のシステムで展開する方向が、現時点では適切であろう。

この考えに基づき、年表のような可視化ツールを HuTime, GIS のような可視化ツールを HuMap と名付け、スタンドアロン・システムとしての開発が別途進められており、これには、筆者も参加する、京都大学東南アジア研究所・京都大学地域研究統合情報センターを核とする研究グループ（柴山守・原正一郎ほか）「H-GIS 研究会（Humanities GIS Research Group）」のメンバーによる共同研究がベースとなっている[7][8]。近々、第一バージョンが希望者に配布されて、実証実験を開始する予定である。

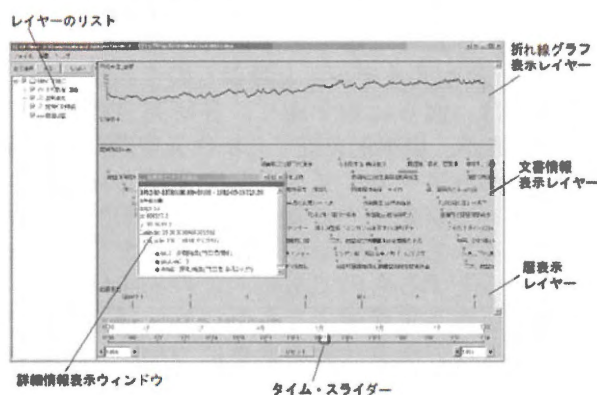


図8 HuTime の典型的な表示画面

図8に示す HuTime の典型的な表示画面例では、数値情報が折れ線グラフとして、文書情報の見出しがその下のレイヤーに表示されていて、文書の詳細は詳細表示用のポップアップ・ウィンドウに表示される。タイム・スライダーを操作することにより、時間軸上のズームイン、ズームアウトが可能なので、天地開闢から現在に至る長期間から地震のような短期間の事象までを取り扱うことが出来る。また、地域や文化に固有の暦を表示することもできるため、元号や閏月を扱うことも可能である。我々はこのために、時間値の標準値として「ユリウス通日」を採用し、すべての事象に開始時刻と終了時刻を設定し、それぞれの時間値をユリウス通日に変換することによって、共通の時間軸上での表示や解析を可能としている。

HuTime には、時間軸に基づく解析機能も組み込まれている。現在のところ、文書や数値の検索結果からマスク（特定の時間範囲）を設定し、元の情報をマスクで抽出したものをレイヤーとして生成し、それらをさらに組み合わせる、というマスク演算の連鎖、によって視覚的に情報の抽出と関係性の発見につなげる機能が実装されている。このように HuTime は、文化資源の時間に関わる情報を可視化し文化的事象の相互関係解析につなげるツールとして極めて新しいも

のである[9][10]。

(2) 事象間の相関関係の記述

事象間に隠れた相関関係の発見が文化資源アーカイブズ形成の目的の一つならば、研究者たちが仮説として立てた相関関係を記述する仕組みを、アーカイブズの枠組みに導入できないだろうか。

例えば、シソーラスの中で関係語（RT）として掲げられる、行為と道具、行為と結果、行為と受動者、起源、因果関係、事物・行為と対抗者、などの相関関係を属性として持つリンクを、事象（群）の間に設定し保存する仕組みである。他の研究者がこれを参照できれば、仮説の検証や、それを土台とした深化など、フォーラムとしてのアーカイブズの可能性が広がる。同様なアイデアは、例えば慶応義塾大学で開発された「歴象データモデル」でも実現されて歴史研究や教育へ応用されており[11]、文化のダイナミズム解明にも有効と考えられる。

(3) 「サイクロニクル」のアイデアの導入

乾期や雨期を持つモンスーン地域などの事象には季節性を反映したものが多く、周期性の観点からの解析が必要となる場面が多い。これは、一方向に進む時間軸で形成された年表を「折り畳み」、周期的な暦の集合として捉え直すことに相当する。ヘブライズムの成立とともに一方向の不可逆な流れという時間観念が確立したと言われるが[12]、元来、人間に親和性のあるのは循環する時間観念であるから、文化に関わる事象を理解するうえでも、周期性の概念は重要である。図9のような、年表を周期性から操作する考え方を、筆者は周期性を持った年表という意味で「サイクロニクル（Cychronicle = Cyclic Chronicle）」と名付け提案した[13]。

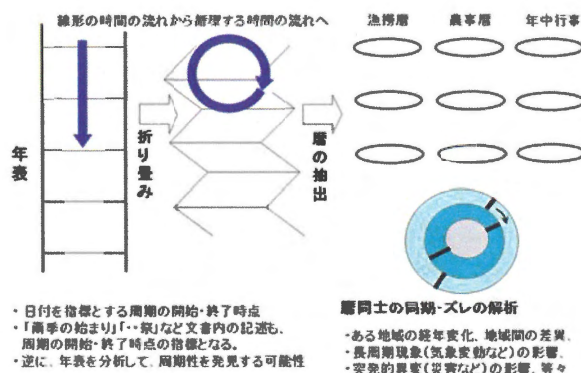


図9 サイクロニクルの概念

図9に示すように、年表から周期の契機となるような事象に着目し、例えば周期的な「漁撈

暦」「農事暦」を構成する事象群を抽出する。これらの事象群を各暦毎に積み上げて比較することで、ある地域内での時間的な暦のずれ、あるいは地域間での空間的な暦のずれなどを発見し、そこから、突発的事件、雨期など気象変化や経済・社会情勢の変化と生活文化との関係を解析することができるかも知れない。相対的な時間軸を処理・表示できるように HuTime を拡張し、マスク演算と組み合わせれば、こうした解析が可能となるだろう。

これをさらに展開すれば、様々な条件でアーカイブズを検索したり演算処理を行う中から、気象関係の数値情報から長期変動パターンを探るなど、事象群の中に隠された周期性を発見できる可能性もある。これも今後の課題の一つである。

6. 民博の民族学研究アーカイブズ

本稿の冒頭で触れたように、民博の所蔵資料の第四のカテゴリーとして、研究アーカイブズ資料がある。これらは、音声、画像、映像を含む様々な非図書資料であるが、次の2種類に大別できる。

(1) 団体や機関による調査・研究の過程で収集・作成した資料類

民博設立時に国文学研究資料館等から移管された日本民族学協会や日本民族学会に由来する資料類、あるいは京都大学アフリカ学術調査隊の記録類など団体や科研による調査隊関連資料類である。参考書類、手紙類、実際の調査過程で得られた写真、映像、音声テープ、メモ、フィールドノートなどから成る。

いずれも、第二次大戦後の日本における民族学研究史を物語る貴重なものだが、所有権が不明確なもの、知的財産権の所属が個人か団体か不明なものが数多い。

(2) 個人の研究室や書斎に保管されていたものが寄贈されて個人名の冠されたコレクションとして収蔵されているもの。ここには、調査・研究に伴う資料類、論文・著書や展示など自身の研究成果発表のバックデータとなった資料類や参考資料のコピー類、私信などが含まれる。中には複製権に問題がある書籍の全体コピー、個人情報やプライバシーの問題をはらむ私信などがある。

民博の発足当初から「民族学史資料」の重要性は認識されていたが、総体的な整理・利用ポリシー不在のまま、個別部署での保管が30年続いてきた。しかし、2004年の法人化とともに、資料・情報の共同利用と社会還元の要請、および知的財産としての明確な取り扱いの要請が強

まったのを機会に、次のような基本方針を立て、整理とその結果の書誌的情報のウェブ公開、利用受付などを開始した。

(a) 将来の利用ではなく、現時点での共同利用・社会還元を目指す。

(b) アーカイブズ資料には、現地での聞き取りや家系調査データなど、現地の第一次情報を含むものも多く、現地文化へのリスペクト、公表することによる現地社会・文化を攪乱する恐れ、研究者の倫理、など微妙な問題を数多く含む点で、公文書や史料と異なる性格を持つ。そこで「民族学研究アーカイブズ」と命名する。

(c) 提供者の意向に沿い、様々な記録メディアから成るアーカイブズ資料を一体として扱い、「コレクション名」を付けてアクセスできるように整備する。

(d) 資料のメディア（紙媒体資料、映像・音響・電子媒体資料、標本資料）ごとに、最適な保管場所・利用場所を用意する。

(e) 保管場所、利用場所が異なっても、ヴァーチャルなレベルでは一体として管理できるようにする。そのため、目録・データベース構造などを工夫するとともに、それぞれの管理部署間で調整しつつ事務処理を行う。

(f) 受入後、直ちに研究目的での閲覧利用ができることを受入の最低限の条件とし、非公開の期限がついたものは原則として受け入れない。

(g) 研究目的で閲覧利用した者に対して、利用後の成果公開の際には、資料に記述された現地文化へのリスペクト、個人情報への配慮などを特に求める。

これらの方針に基づいて、新規受入のための手続き・ガイドラインが制定された。すなわち、仮受入、寄贈受入、の二段階制を採ることとし、仮受入後は民博に一旦寄託して精査、資料目録作成、廃棄・返却分などの確定を行う。次に寄贈受入後に、書誌的な検分を通して作成した資料リストをウェブ公開して研究者による資料研究を促し、より深い資料調査報告・解題などの作成を目指すものである。一方では、これまで手つかずであった既存アーカイブズ資料の実態調査も着手した。その資料が民博に所蔵されるに至った経緯を解明し、合わせて所有権や知的財産権の帰属も明らかにせねばならない。これら調査と並行して整理が進められ、資料リストや画像データベースのウェブ公開も行っている。

(<http://nsearch.minpaku.ac.jp/index.html>)

現在は簡易なリスト公開にとどまるが、研究者による資料調査・研究が深まれば、アーカイブズ資料に対する標準 MARC に相当する EAD (Encoded Archival Description) をベースとし

たシステム開発なども視野に入ってくる。

さらに言えば、前記 3「デジタル・アーカイブズ構築のための課題」でも述べたような、メディアの保存と利用のバランスの問題に深く関わる資料の多くが、実はこのアーカイブズ資料である。長期保存媒体と利用向け媒体の選定と維持管理、劣化した資料に対しては費用／効果の観点からの保存の臨界点の見極め、など多くの課題を抱えており、悩みの多いところであるが、ウェブ公開の効果が現れたものか、外部研究者の利用が徐々に増加していることを、民族学研究アーカイブズを立ち上げた者の一人として報告しておきたい。

7. おわりに

本稿では、民博がこの 30 数年にわたって蓄積してきた資料・情報・経験に基づいて、今後の人文社会系デジタル・アーカイブズ構築にあたって考慮すべき事項などを、筆者なりの視点で紹介した。特に文化人類学・民族学の分野では、研究倫理と現地文化へのリスペクトが最も重要な要素であり、それを踏まえた、共有・共同・共創が基本的なキーワードとなる。そしてこのことは分野を越えた学際的研究にとっても同様であるのは、言うまでもないだろう。

参考文献

- [1] クリフォード, J., マーカス, G. 編『文化を書く』紀伊國屋書店, 1996.
- [2] 久保正敏「写真とアウラ」『月刊みんぱく』30(8), pp. 7-8, 国立民族学博物館, 2006.
(<http://www.minpaku.ac.jp/publication/gekkan/200608txt.html#02>)
- [3] クリエイティブ・コモンズ・ジャパン
(<http://www.creativecommons.jp/>)
- [4] 安永尚志・研究代表者, 国文学研究資料館 共同研究「文化情報資源の共有化システムに関する研究」2004 年度～2006 年度.
- [5] 原正一郎「失われゆく情報の復元・保存技術-人文科学における情報処理-(文献学・データベース共有・史料編纂): 5.Z39.50 とメタデータによる研究機関間連携」『情報処理学会誌』Vol. 43 No. 9, pp. 968-974, 2002.
- [6] 久保正敏「モノと情報班の活動: 時空間統合型データベースの構築を目指して」『総合地球環境学研究所研究 プロジェクト 4-2 2005 年度報告書 アジア・熱帯モンスーン地域における地域生態史の統合的研究: 1945-2005』pp. 338-347, 総合地球環境学研究所, 2006.
- [7] Humanities GIS Research Group :
(<http://www.h-gis.org>)
- [8] [柴山・原・貴志 08] 柴山守・原正一郎・貴志俊彦 (共編)『アジア遊学: 特集 地域情報学の創出』No. 113, 勉誠出版, 2008.
- [9] 関野樹「時間に基づいた情報解析ツール」『アジア遊学: 特集 地域情報学の創出』No. 113, pp. 140-151, 勉誠出版, 2008.
- [10] Sekino, T., "Tools to Realize Spatiotemporal Analysis in the Humanities" in *GIS in the Humanities and Social Sciences International Conference*, Oct. 7-9, 2009, Academia Sinica, Taiwan.
- [11] 慶應義塾大学 FCRONOS プロジェクト室.
"KEIO-GSEC Project on Frontier CRONOS: Research on Risk Communication and Management Based on Cronos Authoring Tool".
(<http://www.fcronos.gsec.keio.ac.jp/>)
- [12] 真木悠介『時間の比較社会学』岩波書店, 1997.
- [13] 久保正敏「時空間統合アーカイブズ構築の構想-ミクローマクロ往還, Cychronicle」『文化情報資源の共有化システムに関する研究 研究成果報告書』pp. 51-54, 国文学研究資料館, 2007.

和歌解析用 MeCab 辞書の開発 —八代集解析済みコーパスによる学習—

山元 啓史

東京工業大学

本稿では八代集 (905 年頃～1205 年) 用の品詞解析済みコーパスと品詞タグつき辞書を用いて、CRF (Conditional Random Field) 法による接続コスト計算を実施し、二十一代集 (905 年頃～1439 年) に対応した和歌用形態素解析辞書の開発について述べる。八代集テキストの CRF による接続コスト学習の結果、90.3%の接続が正しく解析できた。この辞書をもとに今後徐々に二十一代集テキストを増やししながら、辞書を育て、最終的には二十一代集対応の辞書に仕上げていく。

キーワード：和歌、辞書編集、形態素解析、八代集、二十一代集、接続コスト

Development of the MeCab Dictionary for Classical Japanese Poems Based on the *Hachidaishū* Corpus

Hilofumi Yamamoto

Tokyo Institute of Technology

This paper addresses the development of the dictionary of the *Nijūichidaishū* (ca. 905–1439) for MeCab, morphological analysis parser, based on the dictionary of the *Hachidaishū* (ca. 905–1205). The CRF (Conditional Random Fields) method is used to calculate connection rules weights which indicate the coherence of any two words (bigrams) in corpus/sentences. As a result of the compilation of a dictionary, 90.3% of the parsing accuracy is obtained.

Keywords: classical Japanese poetry dictionary compilation morphological parser, the *Hachidaishū*, the *Nijūichidaishū*, connection rules weights

1 はじめに

山元 [7] は和歌用の形態素解析辞書および形態素解析システム kh を開発した。その対象は八代集 (905 年頃～1205 年) に限定されていた。本研究の大きな目的の一つは、その辞書を八代集から二十一代集 (905 年～1439 年) に解析可能対象を拡大することである。ところが、この八代集辞書には接続規則情報がなく、未知のフレーズはうまく解析できず、入出力を逐一確認し、適切な解析が得られるように辞書を育てなければならず、膨大な作業を要して

いた。そこで、本研究では八代集辞書を初期基本辞書として用い、二十一代集の処理を通して用語の接続規則情報を計算処理によって学習し、二十一代集用解析辞書に仕上げることを試みる。

1.1 語彙の体系を目で見たい

さて、語と語は互いに結びつきあって、どんな意味のまとまりを作っているのだろうか。和歌の場合なら「梅と鶯」、「桜と時鳥」、「吉野と桜」、「龍田と紅葉」のように和歌ならではのコンビネーションが思い浮かぶが、このようなコンビネーションはい



図 1: 鶯と時鳥の合成グラフモデル: 鶯と時鳥の 2 語について、古今集の和歌とその現代語を比較し、グラフで描いた。網掛けは共有ノード。楕円は和歌の語、矩形は現代語訳にのみ現れた語。エッジの数字は共出現の頻度。山元 [6] より。

くつぐらい、どんな種類が存在し、どんな意味合いで、どの時代から使われはじめ、それらは互いにどれぐらいの強さで結びついているのだろうか。

和歌研究者の直観や経験だけでは即答にくいコンビネーションを実際し和歌データから獲得する企みとして、図1に示す可視化モデルを作成し、和歌用語の体系について論考を重ねてきた。たとえば、地名の例でいうなら「龍田」は紅葉彩る秋の風景、「吉野」は桜をとりまく春の花模様として有名であるが、可視化モデルを通して見るとそれだけでなく「龍田」は「神の地」、「吉野」は「人間世界／世俗の地」というまとまりをも観察することができた[6]。その後、2009年までに八代集用語について辞書とシソーラスを整備し[7, 8]、八代集限定ではあるが、和歌用語の可視化モデルを完成させた。本研究はこれを基礎にして、八代集(905年頃～1205年、9440首)の300年間だけでなく、二十一代集(905年頃～1439年、25,648首)、534年間の大きな古典の知識を蓄積し、体系化を進めるのが究極的な目的ではあ

るが、いくつかの基本的な問題点があるので、ここではその問題と解決について多少説明を加えたい。

1.2 辞書開発の必要性

八代集から二十一代集に処理対象を拡張すると、分析できる時代が500年間に広がるだけでなく、歌の数も25,000首以上になる。データが多くなればなるほど、語と語の組合せ頻度もある程度得られ、語相互の接続規則を統計的に推定するには都合がよい。しかし、そのためには単位分析（歌を単語に分割し、品詞名を各単語に付ける作業）が必要であるが、25,000首のすべてに対し、手作業で行うには限界がある。たとえば、単語の分割は一通りではなく、長く切る場合（例「うらふきかへす」）もあれば、短く切る場合（「うら/ふき/かへす」）もあり、切り方を統一しておかなければ、語彙一覧表に見られる単語の種類や頻度が異なり、結果的に出現頻度計算が無意味になってしまう。さらに、これを人手によって行うとなると切り方の判断に揺れが生じ、

不統一なデータができてしまう。均一な処理を何度も繰り返して実施するには、量の多少に関わらず計算機で行うべきである。

従来にも和歌を計算機処理する試みはいくつか存在する。たとえば、近藤ら [1, 2] による N グラム統計による方法や竹田ら [4] の LCS (Longest Common Subsequence) 法などである。いずれも辞書を用いずに文字列のみを操作して、その目的に応じた研究成果をあげている。しかしながら、古典知識を蓄積するためには、活用語を基本語に変換、表記を 1 つ (あるいは意味コード) に統一して、文法や意味の構造も柔軟かつ汎用的に取り扱いたい。文字列をそのまま扱う (あるいは一般の古語辞典を転用する) 方法では、異なる表記の同語 (異形同語: たとえば、京都の地名「音羽川」はその形が「おとは／をと／音羽」と「かは／がは／川／河」の組合せ数存在する)、同じ表記の異語 (同形異語: たとえば、ワ行下二段動詞「植う」の未然形あるいは連用形と一般名詞「上」は共に「うへ」である) の判別、宛て字 (たとえば「立覧 [たつらむ]」、「契剣 [ちぎりけん]」、「思ふ蝶 [おもふてふ]」など) における単語の分割と品詞の特定、基本単語 (特に目、手のような身体語)、助詞・助動詞に多く見られる一音節単語の特定は難しい。

従来、機械学習によって辞書の接続情報を得るには大量の処理済みコーパスが必要とされてきた。しかしながら、和歌の電子テキストはあるものの、機械学習によって接続規則を得るほどの量の【処理済みコーパス】はなく、ひとつひとつ手作業によって解析済みコーパスを作成せざるを得なかった。

1.3 和歌特有の問題点

和歌を計算処理するためには和歌特有の問題点があり、十分注意して取り扱わなければならない。たとえば、和歌の表記に①二句切れ、三句切れのように歌の途中で意味上終るものがあるが、句点などなく明示的に文の終りを示す手がかりがない。歌の途中で文が終っているのか、次の句を修飾しているのかかわからないため、連体形なのか終止形なのか判別できない。②和歌大系本やデータベースの中には、あらかじめ「／」のような句の切れ目を示す記号を入れたものがある場合とない場合がある。③「ゝ」や「／」などの踊り字がある場合とない場合がある。特に「ゝ」の場合、前の語尾に同じ音が語頭に続く場合、たとえ単語をまたいでいても「ゝ」が当

てられている場合があり、分割・集計した後、それがどの語であったのか分からなくなってしまう。④仮名文には清濁の明示はないが、大系本には読者への便宜を図り、清濁や漢字を適当に施したものがある。しかし、清濁を明示しないことで掛詞を示す場合は清濁をつけないこともある。

1.4 八代集辞書の問題点

山元 [7] の八代集辞書には接続規則に関わる情報がない。その代わりに八代集に見られる単語の連鎖パターンをすべて登録しており、適切な出力が得られるようにしてある。古文には単音節の単語が複数接続して、ひとまとまりの意味を示すことが多い。たとえば、「ながめせしに」は「ながめ」「せ」「し」「ま」「に」に分解され、「ながめ」以外は一音節のすべて品詞の異なる単語が接続している。山元 [7] は、このような複合した語群を単純に辞書に登録して、見出語と解析済み品詞列を入れ換えるだけの、最長一致法による品詞タグづけシステムを開発した¹。

しかしながら、上記方式では未知のパターンはうまく解析できず、新しい連鎖パターンが出てくるたびに辞書を逐一育てなければならず、膨大な作業を要していた。現代語の形態素解析では、すでに大量の解析済みデータと現代語辞書、そして標準的な日本語の表記があるため、機械学習による接続規則の学習が可能であり、その規則を有する辞書を用いて新たな現代文をほぼ完全に近い形で解析することができている。一方、古語には大量の解析済みデータもなく、接続規則情報付きの辞書もなく、そして標準的な表記もない。

そこで、本研究では今までに蓄積してきた八代集辞書と八代集解析済みデータを元手にして、徐々にテキスト量を増やし、最終的に二十一代集処理用に仕上げることを計画する。まず計算処理により、八代集までの接続規則を学習させ、その未熟な辞書を用いて二十一代集テキストを少しずつ処理させていく。始めから二十一代集すべてを処理させたのでは未知語が多く、正しい方向性を持った接続コストの計算が保証されない。少しずつ処理させ、未知語や誤解析を修正し、それを辞書に反映させ、徐々にテキスト量を増やしていき、解析精度も高め、最終的に二十一代集対応の辞書に仕上げるのである。

¹言い換えれば、八代集 (約 9,500 首) に存在する接続パターンを異なる表記も含めてすべて人手で辞書に登録したにすぎない。しかし、この作業がなければ、揺れのない解析結果は得られなかった。

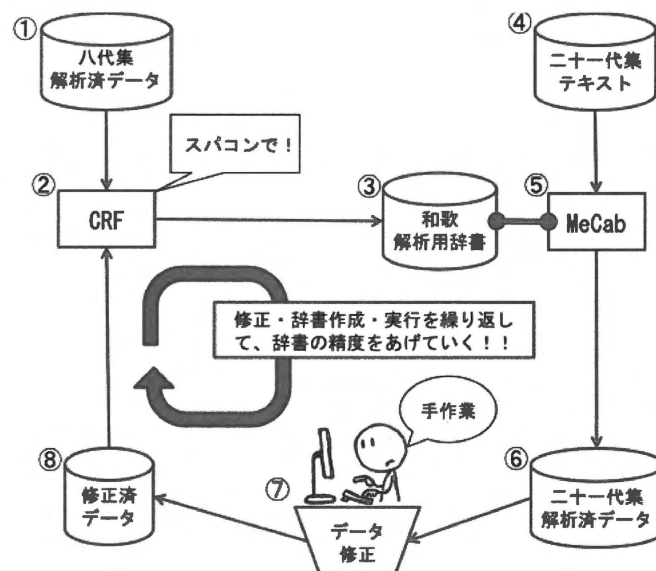


図 2: 辞書開発と接続規則獲得の手順: 山元 [7] 開発の①八代集用辞書を②CRF (連接パラメタ推定プログラム) で処理し③二十一代集用の初期辞書を作成。④二十一代集テキストを準備し⑤MeCab (形態素解析器) と③で⑥二十一代集を解析。誤りや未知語は手作業で⑦修正。⑧修正済データを再び②CRF で処理し、③辞書を作成。②～⑧を繰り返し、徐々に精度の高い辞書を得る。

2 方法

二十一代集のための辞書開発 研究方法は図 2 に沿って説明する。まず、山元 (2007) [7] 開発の①八代集用の解析済みデータと②CRF (Conditional Random Fields) 法²を用いて仮の解析辞書を作成する。CRF は語と語のつながりの程度 (コスト) を統計的に推定するプログラム (連接パラメタの推定) で、辞書の主要な部分を生成する。

多種多様な表現形式の収集 つぎに④二十一代集テキストの準備である。これは国文学研究資料館の二十一代集データベースを用いる³ほか、古典文学大系本その他をスキャンし、表記情報 (漢字仮名混じり、送り仮名などの異なりや揺れを調査したものを追加し、多種多様な表記に対応したテキストデータを作成する。この作業にはドキュメントスキャナを用いて、電子テキスト化し、コンピュータプログラムによって表記の異なりや揺れを一括して収集整理できる状態にしておく。しかしながら、手作業によるところも多い。

MeCab で形態素解析 仮の辞書と二十一代集テキストが準備できれば、⑤MeCab で形態素解析を行う。MeCab を本研究に採用した理由は、これが既成の品詞体系に依存しない設計になっており、現代語のみならず古代語であっても独自の品詞体系で形態素解析器が自作できるからである。たとえば、守岡 [5] は MeCab のこの特徴に注目し、古典中国語を形態素解析するための辞書を開発している。本研究の場合でも、和歌にありがちな独自の品詞体系も十分に設定できるものと判断した。MeCab 辞書の作り方は、MeCab 「オリジナル辞書/コーパスからのパラメタ推定」 (<http://mecab.sourceforge.net/learn.html>) の手順に従った。

MeCab で処理した後、⑥二十一代集の解析済みデータを得られるが、このデータには誤りや未知語があるので、それを手作業で⑦修正・追加し、⑧修正済みデータを作成する。

辞書の精度を上げていく ⑧修正済みデータを用いて、再度②CRF で前回よりも精度のよい③辞書を作成する。ただし、この CRF による辞書作成には大量のメモリと計算速度が必要なので、東京工業大学のスパコン TSUBAME を用いて、効率的に行う。二回目以降の修正作業では、単に辞書の追加や修正

²<http://mecab.sourceforge.net/>

³既に同館知的財産委員会より利用許諾は得ている。

だけでなく、品詞体系の見直しや新たな接続規則の導入も試みる。これには辞書やテキストを実際に目で追いかけて、接続規則を分析的に眺める作業が不可欠である。この点が本研究における本質的かつ忍耐力が必要な部分である。以上、②～⑧までを何度も繰り返し、徐々に精度の高い辞書を作成して行く。

時代別・歌集別の辞書の検討 二十一代集テキストは一度にすべてを処理せず、歌集ごとに辞書に項目を追加しながら辞書を育てて行く方法にしておけば、類似の誤解析を減らすことができるだけでなく、その育てていく過程において、時代別・歌集別に辞書を分割しておいた法が望ましいかどうかを検討していく。時代別・歌集別に分割した方が効率が良ければ、辞書開発を通して、それぞれの特徴が抽出できることも考えられる。

以下では、第一段階として、従来の八代集辞書と八代集処理済みコーパスを用いて、八代集における用語接続コストの学習と、MeCabで八代集テキストを処理してみた結果、どの程度の再現率が得られたのかについて報告する。

3 材料：八代集辞書の収録内容

初期辞書として利用される八代集辞書には、新編国歌大観 CD-ROM 版の二十一代集に相当するデータ [9]、国文学研究資料館編集二十一代集データベース [3]、新日本古典文学大系本二十一代集に相当する書籍その他、新潮日本古典集成の新古今集、ヴァージニア大学日本語テキストイニシアティブ (<http://etext.lib.virginia.edu/japanese/>) 監修の二十一代集データから、それらにすべてに見られる用語がそれぞれの表記で登録されている。MeCab による処理実験に用いるテキストは国文学研究資料館二十一代集データベースの中の八代集のすべての和歌を用いる。

4 解析結果

八代集辞書を MeCab 用辞書に変換して、再び八代集テキストを解析し、評価を実施した。また、八代集以外のテキスト（新後撰和歌集）についても解析実験を行った。CRF による接続コスト学習は、所有の自作パソコン、Linux Kernel version 2.6.27.15 (gcc version 4.2.4)、Intel(R)Core(TM)2 Duo CPU E7200 2.53GHz (cache size:3072 KB)、全記憶容量

2GB、スワップサイズ 20GB で計算した。すべての和歌テキストを接続コストの学習に用いたかったが、メモリを使い尽したため、テキストを国文学研究資料館二十一代集データベースを中心にメモリ容量極限の 11,119 行に限定し、接続コストの計算をやり直した。

表 2 は八代集収録以外の歌、新後撰和歌集の 4 番歌を新編国歌大観 CD-ROM と国文学研究資料館データベースの 2 種類のテキストで処理したものである。「まで／まで」のように清濁の有無は問題なく解析できているが、「みよし野の／みよしのゝ」は踊字「ゝ」がうまく解析できていない。EOS の前の「かな」はいずれにおいても解析できている。

どの程度正しく解析できたかを評価するために、MeCab パッケージ標準添付の mecab-system-eval プログラムを用いて、MeCab の結果とすでに山元 [7] において処理した結果との差異 (precision / recall) を計算した。その結果、1 番目の素性 (主に品詞名のみ) の特定については、99.7%、すべての素性 (活用形などに代表される品詞の下位分類) の特定については、90.3%の解析再現が確認できた。

5 おわりに

本稿では八代集用の辞書を用いて、接続コストを CRF により学習し、二十一代集を処理するための初期基本辞書の試作を行った。現行では手持ちのパソコンの計算速度ならびに記憶容量の制約ですべての八代集テキストを用いた試作実験が行えなかった。今後はスパコン上に処理環境を作り、まず今回行えなかった八代集の全テキストで初期辞書を作成していく。その上で、二十一代集を処理しつつ、徐々に精度をあげて、二十一代集すべての和歌が柔軟に処理できる辞書に仕上げる予定である。さらに、この作業を通して得られる接続情報をもとに、和歌の接続規則の理論化も進めていきたいと考えている。

参考文献

- [1] 近藤みゆき：n-gram 統計による語形の抽出と複合語——平安時代語の分析から——, 日本語学, Vol. 20, pp. 79-89 (2001).
- [2] 近藤泰弘, 近藤みゆき：平安時代古典語古典文学研究のための N-gram を用いた解析手法, 言語処理学会第 7 回年次大会発表論文集, 第 7 巻, pp. 209-212 言語処理学会 (2001).
- [3] 中村康夫, 立川美彦, 杉田まゆ子：国文学研究資料館データベース 古典コレクション『二十一代集』(正保版) CD-ROM, 岩波書店, 東京 (1999).

表 1: MeCab 用に整理しなおした八代集辞書 (Seed 辞書の例)

うつつ,0,0,0, 名詞, 一般,****, 現, うつつ,**
うつぶしぞめ,0,0,0, 名詞, 一般,****, 空五倍子染め, うつぶしぞめ,**
うつぶしぞめ,0,0,0, 名詞, 一般,****, 空五倍子染め, うつぶしぞめ,**
うつぶしぞめ,0,0,0, 名詞, 一般,****, 空五倍子染め, うつぶしぞめ,**
うつま,0,0,0, 動詞,****, 四段・マ行, 未, 埋む, うづむ, 埋ま, うづま
うつまぎ,0,0,0, 名詞, 一般,****, 渦巻, うづまぎ,**
うつむ,0,0,0, 動詞,****, 下二段・マ行, 終, 埋む, うづむ, 埋む, うづむ
うつもる,0,0,0, 動詞,****, 下二段・ラ行, 体, 埋もる, うづもる, 埋もるる, うづもるる
うつもれ,0,0,0, 動詞,****, 下二段・ラ行, 未, 埋もる, うづもる, 埋もれ, うづもれ
うつもれ,0,0,0, 動詞,****, 下二段・ラ行, 未用, 埋もる, うづもる, 埋もれ, うづもれ
うつもれ,0,0,0, 動詞,****, 下二段・ラ行, 用, 埋もる, うづもる, 埋もれ, うづもれ
うつら,0,0,0, 動詞,****, 四段・ラ行, 未, 移る, うつる, 移ら, うつら
うつら,0,0,0, 動詞,****, 四段・ラ行, 未, 映る, うつる, 映ら, うつら
うつら,0,0,0, 名詞, 一般,****, 鶉, うづら,**
うつり,0,0,0, 動詞,****, 四段・ラ行, 用, 移る, うつる, 移り, うつり
うつりか,0,0,0, 名詞, 一般,****, 移り香, うつりが,**
うつりが,0,0,0, 名詞, 一般,****, 移り香, うつりが,**
うつり香,0,0,0, 名詞, 一般,****, 移り香, うつりが,**

表 2: MeCab による解析結果 (例): 国文学研究資料館のデータベースには「／」が含まれているが、取り除いた上で解析した。

新後撰集 (新撰国歌大観版) 4 番歌	
000004	昨日までふる郷ちかくみよし野の山もはるかにかすむ春かな
000004	記号, 一般,****,*
昨日	名詞, 一般,****, 昨日, きのふ,**
まで	助詞, 一般,****, まで, まで,**
ふる郷	名詞, 一般,****, 故郷, ふるさと,**
ちかく	形容詞, ク,****, 用, 近し, ちかし, 近く, ちかく
みよし野	名詞, 地名,****, み吉野, みよしの,**
の	助詞, 格助詞,****, の, の,**
山	名詞, 一般,****, 山, やま,**
も	助詞, 係助詞,****, も, も,**
はるかに	形容動詞, ナリ,****, 用, 遥かなり, はるかなり, 遥かに, はるかに
かすむ	動詞,****, 四段・マ行, 終体, 霞む, かすむ, 霞む, かすむ
春	名詞, 一般,****, 春, はる,**
かな	助詞, 終助詞, 詠嘆,****, 哉, かな,**
EOS	
新後撰集 (国文学研究資料館データベース版) 4 番歌	
000004	昨日まで／ふる里ちかく／みよしのゝ／山もはるかに／かすむ春かな
000004	記号, 一般,****,*
昨日	名詞, 一般,****, 昨日, きのふ,**
まで	助詞, 一般,****, まで, まで,**
ふる里	名詞, 一般,****, 旧里, ふるさと,**
ちかく	形容詞, ク,****, 用, 近し, ちかし, 近く, ちかく
みよしの	名詞, 地名,****, み吉野, みよしの,**
ゝ	助詞, 格助詞,****, と, と,**
山	名詞, 一般,****, 山, やま,**
も	助詞, 係助詞,****, も, も,**
はるかに	形容動詞, ナリ,****, 用, 遥かなり, はるかなり, 遥かに, はるかに
かすむ	動詞,****, 四段・マ行, 終体, 霞む, かすむ, 霞む, かすむ
春	名詞, 一般,****, 春, はる,**
かな	助詞, 終助詞, 詠嘆,****, 哉, かな,**
EOS	

- [4] 竹田正幸, 福田智子, 南里一郎: 歌集間における表現特徴の自動抽出—部分文字列の生起頻度に見る—, 情報処理学会研究報告 00-CH-47, Vol. 47, pp. 39–46 (2000).
- [5] 守岡知彦: MeCab を用いた古典中国語の形態素解析の試み (セッション 1), 情報処理学会研究報告. 人文科学とコンピュータ研究会報告, Vol. 2008, No. 73, pp. 17–22 (2008).
- [6] 山元啓史: コンピュータによる歌枕の分析, イタリア日本語教育協会, 第 3 回シンポジウム論文集, pp. 373–382, イタリア日本語・日本語教育学会 (2006).

- [7] 山元啓史: 和歌のための品詞タグづけシステム, 日本語の研究, Vol. 3, No. 3, pp. 33–39 (2007).
- [8] 山元啓史: 分類コードつき八代集用語のシソーラス, 日本語の研究, Vol. 5, No. 1 (2009).
- [9] 新編国歌大観編集委員会 (編): CDROM 版新編国歌大観, 角川書店 (1996).

作文指導に向けた表現データベースの構築と 表現評価指標の検討

安藤 秀明 秋吉 信吾 竹内 和広
大阪電気通信大学 工学研究科 情報工学専攻

高校・大学生の作文の表現指導は評価者の主観が強く影響するため、一定の基準で作文指導を行う上では指導者間で共通する評価指標を設定することが課題となる。我々は現在、標準的な表現で記述された文章データに対して書き換え処理を行い、誤り表現の評価基準を設定するための誤り表現データベースの構築を試みている。今回はそのようなデータベースを用いて高校・大学生の作文表現を客観的に評価する数値指標の比較、検討を報告する。

キーワード：教育支援情報処理，作文自動評価，学習者コーパス

A Comparison of Statistical Error Detection Methods for Objective Evaluation of Students' Composition

Hideaki Ando Shingo Akiyoshi Kazuhiro Takeuchi
Graduate School of Engineering
Osaka Electro-Communication University

In this paper, we propose a set of pseudo-error rewriting rules based on an analysis on students' errors in their compositions and introduce a pseudo-error database that is created from regular texts by applying the rules. In order to establish an objective evaluation method which adopts various viewpoints for students' composition, we compare some statistical error detection methods on the database.

Keywords: Computer Aided Education, Automatic Error Detection, Learners' Corpora

1. はじめに

人間が書く作文には本質的に多様性がある。そのような多様性は文学作品に作家特有の文体が存在するように、必ずしも悪いものでばかりではない。しかし、高校や大学での作文教育の場面では、作文の不適切な部分を指導する基準があいまいであると問題を招くこともある。そこで、学生が書いた作文の不適切さを多様な観点から基準設定することが求められる。しかし、不適切な表現の基準を策定するためには、不適切な事例を大量に収集、蓄積することが最良であるが、そのようなデータ収集は簡単には実現できないことが問題となる。本稿ではこの問題に対して、適切な表現を使っていると考えられる標準的な文章を書き換えて疑似的な不適切表現として利用し、多様な作文評価を可能とする評価指標設計の資源として利用することを検討したい。

2. 不適切表現のデータベース化

我々は、既に、作文指導のために学生の作文を教員が添削したデータを収集し、データベ

ス化を行っている[1]。このデータベース化作業を進めていく過程で、学生が作文中に用いる表現の多様性に改めて驚くと同時に、教員が特定の表現を不適切表現ととらえる観点も多様であることに気付いた。このような教員の経験則である観点はそれぞれに妥当性があり、どの観点が優れているとは言い難い。例えば、ある表現が、文体の統一性から見て不適切なのか、表現が含まれる文そのものを見た場合でも不適切なのかは観点の相違であり、その観点は指導の目的によって変化する性質をもつはずである。前者の文体や論理構造に関わる問題は文章レベルの特徴付けが必要であり、その観点に関しても既に我々は検討を進めているが[2]、本稿では後者の文内のレベルの不適切表現をデータベース化することを目的とした研究を報告する。

文脈を文内に限定した表現レベルでも、特定表現を不適切と見る観点には多様性がある。また、このレベルの不適切表現を整理する上での難しさは、不適切さの理由はある程度説明することができるにも関わらず、それが文内レベルのどのような文脈においても成立するわけではなく、不適切となる文脈を限定することが難しい点である。

不適切表現を整理・限定するための文中文脈を統語解析といった言語解析モジュールの解析結果に基づいて行うためには、もう一つの問題が存在する。不適切表現を含んだ学生の作文では、その不適切表現が問題となって係り受け解析や形態素解析が精度よく機能することが期待できないという問題である。そこで、自然言語処理モジュールを使わず、基本的な言語統計量を用いて不適切表現の成立する文脈を特定する方法を検討したい。具体的には、主に OCR の分野で開発されてきた手法を参考に、作文の文を単位に不適切性を議論したい。

3. 疑似的な不適切表現の生成

学生の作文を添削したデータベースを参考に、学生が不適切な表現を行う理由を以下の典型的な 5 種類に大別し、疑似的な不適切表現のデータベースを作成した。

DB1: 1 文字単位での誤字脱字

DB2: 格助詞の不適切な使用

DB3: 連体修飾をとともう不適切な名詞句の形成

DB4: 名詞の表記が不適切なもの

DB5: 文節要素の順が不適切なもの

この 5 つの誤り類型に対して具体的には表 1 のような 11 の疑似的な不適切表現への書き換え規則を作成し、Wikipedia¹ のデータに対して表現書き換え規則を適用することで表中の各行ごとに 1000 文からなる疑似的な不適切表現データベースを作成した。すなわち 11 ファイルから成る不適切表現データベースが完成した。なお、書き換え処理規則に自然言語処理モジュールの解析結果を利用するが、利用したモジュールは形態素解析器 ChaSen[5]と係り受け解析器 CaboCha[6]である。

表 1 のそれぞれのファイルについて、大別類型の DB1 から順番に詳細を説明する。

DB1 では正データの文字列に対して、文中の適当な文字に対して下記の 3 種類の書き換え規則を設計した。なお、それぞれの規則は、1 文中に 1 文字程度を書き換えるように設定している。

- ひらがな文字を挿入する (c_ins)
- 文字を削除する (c_del)
- 文字をひらがな文字に変換する (c_henkan)

DB2 では書き換え対象文にまず形態素解析を

かけ、その結果に対して、文中の品詞情報が助詞-格助詞-一般となる語に以下の 2 種類の書き換え規則を用意した。

- 格助詞をランダムに変換する (case_case)
- 格助詞を削除する (case_del)

各規則では、1 文中の 1 つの格助詞がランダムに選ばれて書き換えられるように設定している。

表 1. 疑似不適切表現データベースの内容

	DB	略称	不適切の種類
1	DB1	c_ins	文字挿入
2	DB1	c_del	文字削除
3	DB1	c_henk	文字変換
4	DB2	case_case	格助詞変換
5	DB2	case_del	格助詞削除
6	DB3	rentai_col	連体化部分変換
7	DB3	rentai_del	連体化部分削除
8	DB4	noun_hira	ひらがなに変換
9	DB5	rule_a	格要素の位置交換
10	DB5	rule_b	副詞的文節後方移動
11	DB5	rule_c	副詞的文節前方移動

DB3 は書き換え対象文の形態素解析の結果から、連体化が名詞節を同定することを前提とした疑似的な不適切表現生成規則を以下のように作成した。

- 連体修飾節の接続を別の形にランダムに変換する (rentai_col)
- 連体修飾節中の助詞を 1 つ削除する (rentai_del)

以上も DB2 と同様に連体修飾節がある文の中で 1 箇所のみ書き換え規則が適用されるようになっている。

DB4 では正データの形態素解析で同定された形態素の品詞情報が名詞-一般かつ形態素の字種がカタカナもしくは漢字のみで構成されている場合、形態素の読みをひらがなとして変換する (noun_hira) という書き換え規則を用意した。

DB5 の作成には、係り受け解析で得られた情報を用いた疑似不適切表現の書き換え規則を用いる。具体的には下記の 3 つの種類がある。

- 同じ係り受け先をもつ格要素を入れ替える (rule_a, 図 1 中の a)
- 副詞で終わっている副詞的な節を 1 つ後の文節と入れ替える (rule_b, 図 1 中の b)

¹ <http://ja.wikipedia.org/>

- 副詞で終わっている副詞的な節を1つ前の文節と入れ替える (rule_c, 図1中のb)

これらの規則も適用されるのは1文中、1箇所限定するように調整した。

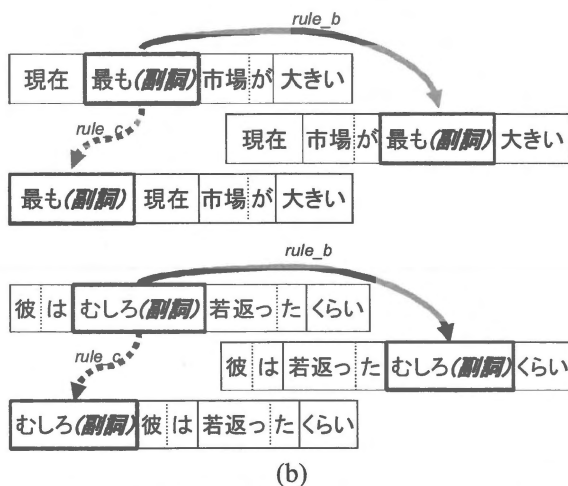
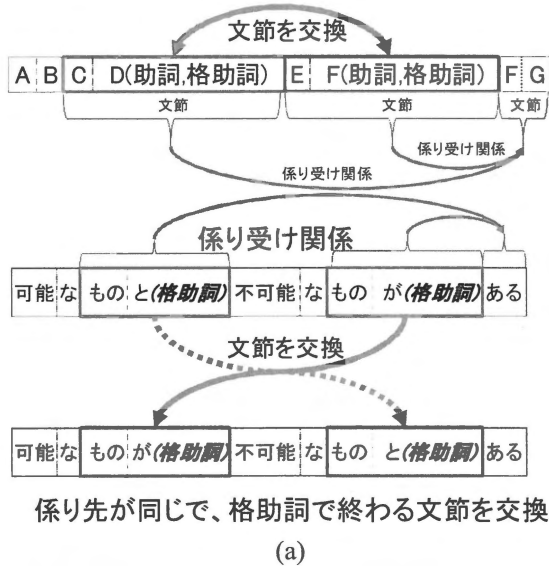


図1. 係り受け情報を用いた不適切表現への書き換え規則

4. 不適切表現の検出指標

本稿では、適切な表現が使われていることが想定される標準的な文書集合（正の例）を用意しておき、正の例から採集した統計情報を用いて前節で作成した疑似的な不適切表現のデータベースから本当に不適切表現となる例を検出することを検討する。

このように文中の表現の不適切性を評価する手法としては、図2に示したように文中の表現（灰色で示す）に対して基本的な2つの方向性がある。1つは文内に検査用の区切りを設け、

区切りの文頭側と文末側の文字列の結合性（つながりやすさ）を基準にする方向性である。他方は、文中の連続文字列が基準となる正の例においてどの程度存在するか（正の例での出現度合）を基準とする方向性である（図2中の手法2）。

手法1：区切りの左右での文字列結合性で評価

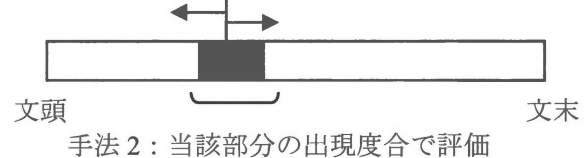


図2. 文中の表現の不適切性を検査する方法

手法1の方向性に着目して表記誤りを検出する研究には、村田手法[3]がある。村田手法では、「表現xが実際の正の例には出現していない場合」に表現xが負の例である度合 $P(x)$ を定義し、この値を使って不適切表現の検出を行う。具体的には $P(x)$ は式(1)によって求める。

$$P(x) = 1 - (1 - (\frac{O(a)}{N} \times \frac{O(b)}{N})^N) \quad (1)$$

式(1)で用いる統計量は $O(a)$ 、 $O(b)$ 、 N であるが、ここで、 $O(a)$ と $O(b)$ は表現x中に分割点を用意し、分割点の文頭側の文字列a、文末側の文字列をbとした場合の正の例中の出現数である。Nは正の例中に出現する全ての文字の出現数である。

実際の不適切表現の抽出には、作文中の文の文頭から文末まで1文字ずつ順に分割点を移動させ、分割点を基準に文頭側、文末側の文字列の生起数を比較する。

それぞれの値の具体例を表2に示す（[3]中の表を説明のために簡略化して引用）。この例は「負の事令を検出する。」という文の「事令」の間に分割点を置き、左右の文字列a、bをそれぞれ3文字から1文字まで考えた場合の統計値を示した表である。

表2. $P(ab)$ を計算した実例

A	O(a)	B	O(b)	P(ab)
負の事	0	令の検	2	0
...		...		
の事	20837	令の	648	0.0324
...		...		
事	384718	令	16227	0.9999

表2は分割点を基準に文頭側文末側3文字以内のパターンを整理しているため、実際の表の行

数は 9 パターンにわたる。分割点を基準にどれだけの文字数を表現の単位と考えるかは指標の使用者側で決定する必要がある。例えば、分割点の前後 5 文字までを見るならば、25 パターンの統計値を計算することとなる。また、 $P(x)$ を認定する際に不適切表現である度合い $P(x)$ にどのような閾値を用意すればよいかも決める必要がある。以上の村田手法は、特定の表現を二項関係(a,b)からなるものと仮定し、二項関係(a,b)の独立性を調べることを発展させた手法と考えることができる。本稿では村田論文を参考にし、実装した手法を以後手法 1 と呼ぶ。

次に図 2 中の手法 2 の方向性の例として、荒木手法[4]について述べる。これは文中の連続文字列が正の例においてどれだけ出現するかを基準とする方法であるが、手法 1 と同様に 1 から無限長の長さの文字列についての出現検査をするわけではない。具体的には、 m 重マルコフモデルを使い、前方 m 文字に続く文字 x_i の連鎖確率 $P(x_i|x_{i-m},x_{i-m+1},\dots,x_{i-1})$ を評価する。

一般的に文字 x_i の連鎖確率は式(2)のように表され、文字 x_i の連鎖確率 $P(x_i|x_{i-m},x_{i-m+1},\dots,x_{i-1})$ は、文字 x_i を含む長さ $m+1$ の文字列 $x_{i-m}x_{i-m+1}\dots x_{i-1}x_i$ の出現確率と文字 x_i の前方の長さ m の文字列 $x_{i-m}x_{i-m+1}\dots x_{i-1}$ の出現確率の比として求められる。

$$P(x_i | x_{i-m}, x_{i-m+1}, \dots, x_{i-1}) \\ \equiv \frac{P(x_{i-m}, x_{i-m+1}, \dots, x_{i-1}, x_i)}{P(x_{i-m}, x_{i-m+1}, \dots, x_{i-1})} \quad (2)$$

実際にはこの m として大きな値を使うのではなく、 $m=2$ といった値を用いる。すなわち、図 3 のように検査対象文の文頭から特定文字数（図の場合は 3 文字、丸が 1 つの文字を示す）ずつ文字列をとり、その正の例での出現数を(2)の確率値として計算する。図 3 ではそれぞれの確率値を縦軸として示してある。

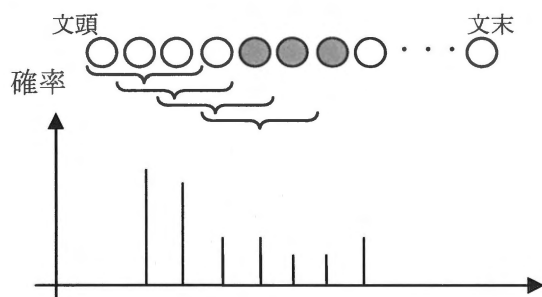


図 3. 文中の 3 文字単位ごとの確率値の計算

例えば図 3 中の不適切な表現が灰色の丸とすれば、その確率値は通常の表現よりも低くなることが考えられるので、この値が閾値より小

さい（落ち込む）部分を不適切表現と判定する手法である。

さらに荒木はこの連鎖確率の落ち込み回数と位置によって、

- (1) 誤りの開始位置が検出できる
- (2) (i)誤字誤りかまたは誤挿入誤り、(ii)脱落誤りの 2 種類が区別できる
- (3) また(i)の場合、誤り文字数が決定できる

としている。

本稿では荒木論文を参考に実装した手法を手法 2 と呼ぶ。

5. 実験

5.1 作文および未知の適切文に対する実験

前節で紹介した指標（手法 1,2）を使って、擬似的な不適切表現から、確度高い不適切表現を抽出・蓄積するために、いくつかの文章集合に対し適応し、それぞれの誤り傾向をみた。対象とした文章集合は、以下の 3 つの文章集合である。

- 大学生が書いた不適切性の高い文（以下「作文」と呼ぶ、390 文）
- インターネット上の掲示板 2 ちゃんねる¹(以下「2ch」と呼ぶ、610 文)
- 青空文庫²より収集した文学作品(以下「青空文庫」と呼ぶ、1000 文)

上記の文集合はすべて正の例には含まれない文から成り、かつ、作文、2ch は青空文庫よりも不適切な表現が多いことが期待できる。このことを利用して、手法 1,2 を使い、下記の適切率を用いてそれぞれの文集合の特質を区別できるかを実験した。

$$\text{適切率} = \frac{\text{指標が適切と判定した文数}}{\text{対象とした文数}}$$

まず、作文と 2ch の検出率を比較したグラフを図 4 に示す。次に、学生の作文と 2ch の書き込みを一つの文集合として、青空文庫との不適切率を比較し、その結果を図 5 として示す。

¹ <http://www.2ch.net/>

² <http://www.aozora.gr.jp/>

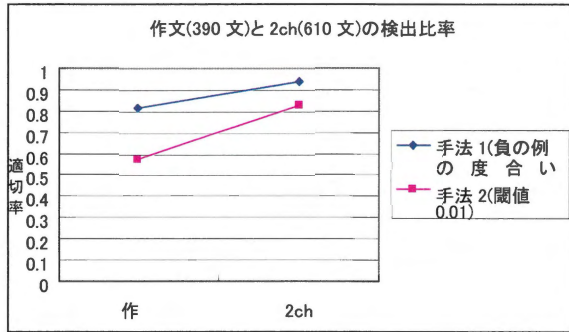


図 4. 作文(390 文)と 2ch(610 文)の比較

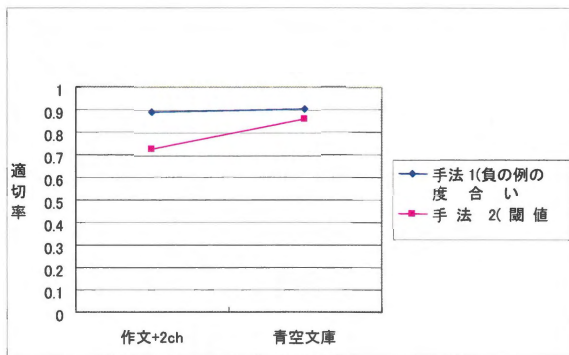


図 5.(作文+2ch)と(青空文庫)の比較

以上の図 4, 図 5 から, 手法 2 によって, 特に文集合の不適切な違いを分割することが可能となると考えられる. また, 我々が収集した学生の作文 360 文が 2ch の例と比較しても手法 1, 2 どどちらを用いても適切率が低いと判定された.

5.2 表 1 のデータベースを用いた実験

5.1 節で確認した手法 1, 2 の不適切表現検出の傾向をさらに調査する. 検討手法として 3 節で作成した各不適切表現ファイルに対する適切率をみることで検討を行う. 適切率は 3 節で作成した各ファイルの文数が各 1000 文であることから以下のように定義する. 5.1 節の文集合とは異なり, これらのファイルに含まれる文は表 1 の書き換え規則を用いて 1 文に必ず何らかの擬似的な不適切表現があるように書き換えられている.

$$\text{適切率} = \frac{\text{手法が適切とした文数}}{1000}$$

手法の閾値や比較文字列長を 5.1 節の実験と同様の設定で, 各ファイルに対する適切率を出したグラフが図 6 となる.

図 6 から不適切性の原因によって手法 1, 2 が不適切性を検出する度合いが変化することがわかる. すなわち, 特定の表現が不適切となる文

脈を区別する評価指標が設計可能ではないかと考える.

例えば, 5.1 節の実験結果である図 5 の作文+2ch と青空文庫に比較では, 手法 1 による適切率にはあまり差がなかったものの, 手法 2 では明確な差があった理由は, 作文+2ch の文集合に手法 2 が感度良く検出するタイプの不適切表現が多く含まれている可能性が考えられる.

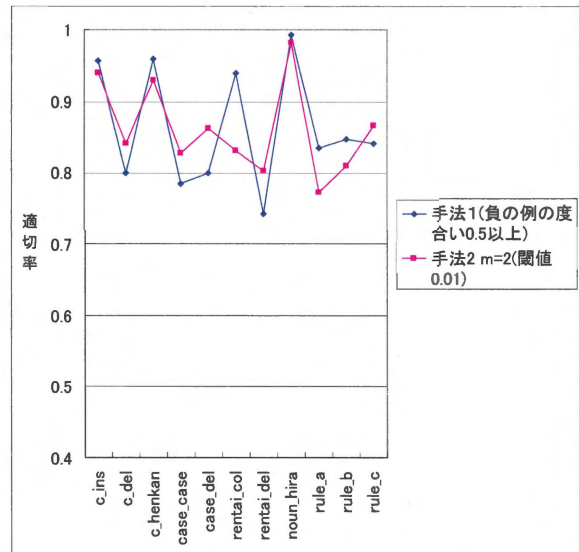


図 6. 擬似不適切表現への手法 1, 2 の適用

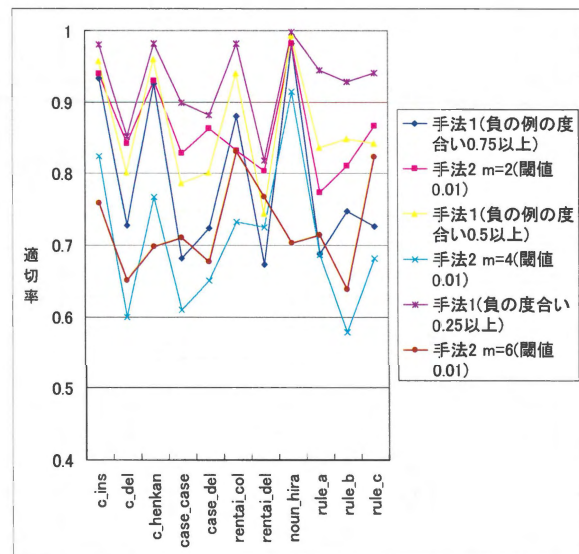


図 7. 手法 1 の負の度合いと手法 2 の m の次数を変えた状態の再現率

さらに, 手法 1 と手法 2 の設定を変え, 細かく傾向をみる実験も行った. 主に手法 1 では不適切性を検出する際の不適切性と適切性を分ける閾値, 手法 2 では誤りを特定する上での走査文字列長が感度調整に役立つ. 実験結果を図 7

に示す。図 7 が示すように、これらの設定が、図 6 よりもさらに、指標間の不適切性検知の感度が調整可能であることが期待できる。

6. まとめ

本稿では、人間が書いた不適切な表現を大量に収集、蓄積する代わりに、適切な表現を使っていると考えられる標準的な文章を書き換えて疑似的な不適切表現を作成する研究を報告した。疑似的な不適切表現はそのままでは工学的応用には用いることができず、評価指標開発と組みになってはじめて機能する。本稿では、この点に関して既存の代表的な文表現適切性を評価する指標を用いて、このような目的に使用可能かを評価した。本稿の内容は、それら手法の疑似的な不適切表現に対する総合的な評価のみにとどまった。今後はさらに評価指標の精緻な分析をすすめると同時に、評価指標と人間の不適切性を検知する観点との相関を調査していきたい。

謝辞

本研究は科学研究費補助金（基盤研究(C)課題番号 21500113）の補助を受けて遂行されている研究の一部である。ここに記して謝意を示す。

参考文献

- [1] 竹内和広：作文指導のための作文添削データベースの構築, 第 12 回「人文科学とデータベース」シンポジウム予稿, pp.55-60, 2006.
- [2] 橋本喜代太 他：複数コーパスからの特徴表現抽出を利用した文体統一性の評価とその活用について, 言語処理学会第 14 回年次大会, pp.384-387, 2008.
- [3] Masaki Murata, Hitoshi Isahara : Automatic Detection of Mis-Spelled Japanese Expressions Using a New Method for Automatic Extraction of Negative Examples Based on Positive Examples, IEICE Transactions on Information and Systems, VOL.E85-D, No.9, pp.1416-1424, 2002.
- [4] 荒木哲郎, 池原悟, 塚原信幸, 小松康則, 田川崇史, 橋本憲久：m 重マルコフ連鎖モデルを用いた日本語の誤字・脱落・誤挿入誤り文字列の検出と訂正法, 信学論(D-II), vol.J83-D-II, no.6, pp.1516-1528, 2000.
- [5] 形態素解析器 ChaSen:
<http://chasen-legacy.sourceforge.jp/>
- [6] 係り受け解析器 CaboCha:
<http://chasen.org/~taku/software/cabocha/>

『古事記』学術支援データベースの構築 —系図表示システムにおけるデータ構造の検討—

生田 敦 司*¹ 杉 山 正 治*²
柴 田 み ゆ き*¹ 宮 下 晴 輝*¹

大谷大学文学部*¹ 立命館大学情報理工学部*²

概要：現在開発中の『古事記』学術支援データベースでは、系図を快適に表示・作成することを目指している。これまでに、インターフェイスに重きを置いたシステム MaSSRiDGe を実装した。現在は、大容量情報をストレスなく扱う手法と、紙媒体上の系図表現に馴染んだデータ入力・閲覧・変更などが可能な作図手法の実現を目指している。本研究では、系図表示システムの諸問題を検証し、上記課題を解決する手法を考察した。

キーワード：系譜史料, MaSSRiDGe, 系図表示, 線分交叉

KOJIKI Knowledge Assistant Database System for Academic Usage —A Study of Data Structures for Displaying Genealogy—

ATSUSHI IKUTA*¹, SEIJI SUGIYAMA*²,
MIYUKI SHIBATA*¹ and SEIKI MIYASHITA*¹

Department of Humane Informatics, Otani University*¹
College of Information Science and Engineering, Ritsumeikan University*²

Abstract: In this paper, we showed a method to display family trees with segment intersections. *KOJIKI Knowledge Assistant Database System for Academic Usage* aims to display and make genealogy system comfortably. Family trees are drawn using segments with vertical, horizontal, and sometimes intersects on paper media. Existing family tree display systems, however, do not have the function of displaying segment intersections. To display segment intersections in our system, we researched existing family tree display methods and showed display element patterns.

Keywords: Genealogy in historical materials, MaSSRiDGe, Displaying family trees, Segment intersection

1. はじめに

『古事記』学術支援データベースは、紙媒体の索引をテキストベースで補完する史料検索ツールである。2006年に最初のプロトタイプを発表し、現在その拡張に向けた基礎研究を行っている[1]。将来的には、他の諸史料においても快適に入力・編集・表示ができるような汎用性の獲得を目指している。

本稿は、この過程で判明した、コンピュータ上での系図表示の問題について取りあげる。

情報技術を用いた人文科学に益する系図表示の研究は、すでに複数の有益な成果を挙げることができ

る[2]-[5]。こうした関心に対する共通の課題として、いかにして直観的な理解を得られるデータの提示ができるかという点が挙げられる。それは各個性間の情報が複雑に連携するときに解決困難な問題となる。

紙媒体に系図を表記するルールには明確な定義があるわけではない。しかし、系図表示をコンピュータ上で実現し、直観的な理解を促すには、紙媒体上で行われる経験則を、数式化できる定義へ変換する必要がある。

『古事記』学術支援データベースでは、これまでに、系譜情報全体像の俯瞰的な一覧性と部分的な詳細表示とを両立し、一画面でシームレスに切り替えるための

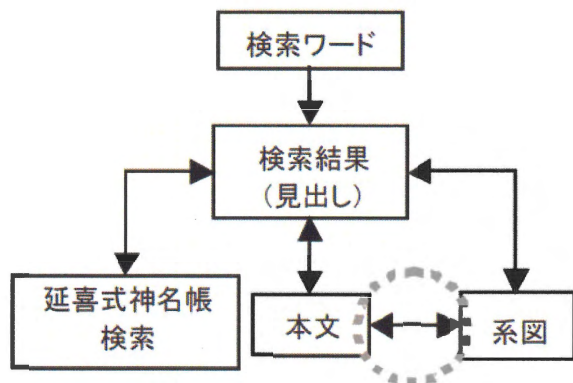


図 1: 『古事記』学術支援データベースの概念図

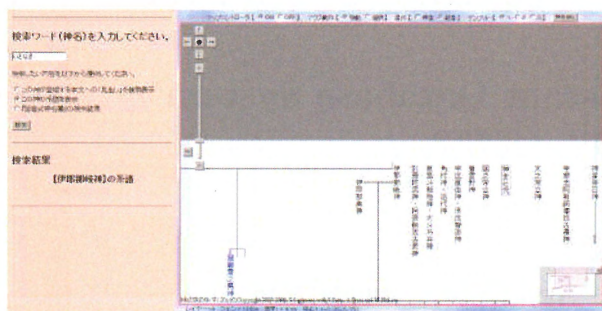


図 2: MaSSRiDGe による系図表示画面

プロトタイプシステム Magnifying And Simplifying System for Retrive and Display GENEalogy (略称, MaSSRiDGe)」を作成した (図 1, 図 2) [6]。

現在, 系図表示の問題をさらに深化させ, 紙媒体による系図表記の利便性を検討し, その実現を目指している。紙媒体では, 個性を配置して, 各個性が相互に多くの個性と関係性を有するとき, 線分が交叉してでもその場所に円弧を用いて表現することがしばしば行われる (図 3)。紙媒体上での系図表現をコンピュータ上で実現させる場合には, この点も克服しなければならない。

本研究では, 上記の問題を克服するために必要な情報を調査・検討した。具体的には, 日本史料や慣習的な紙媒体上での系図表記・描法の変遷の整理 [7], 既存の系図表示ソフトウェアがどのようなアプローチで表示を試みているかの調査 [8], 系図の線分が交叉する際の基礎的な情報整理 [9] を試みた。最後にこれらの結果を踏まえて, 現状コンピュータでの系図表示の問題点を, 問題打開の可能性を考察する。

2. 系譜・系図史料の変遷

本章では, 従来紙面上に表現されてきた日本史料の系譜・系図の変遷をたどり, コンピュータ表示に必要なパターンを考察する。

2.1. 口承・文章系譜

系譜伝達の初期形態は口承もしくは文章系譜である。古いものでは, 5 世紀後半頃とみられる稲荷山古墳出土鉄剣銘の系譜がある。7 世紀後半から 8 世紀はじめにかけての『上宮記』『一云』系譜・山ノ上碑などは一系系譜を婚姻で結びつけるなどの連結形である [10]。『天寿国曼陀羅繡帳銘』や『古事記』『日本書紀』では, 皇統譜・氏族などの系譜が, 同族婚や異世代婚の形で複雑に叙述されている [10]-[12]。

後述するように, 既存の系図作成ソフトには, 婚姻を前提とする描き方に特化したものが見られる。史料の内容を過不足なく表現するためには, 様々な様式に柔軟に対応し, 各様式を自由に連結させるシステムが不可欠である。

2.2. 縦系図

平安時代の初期には, 系譜内容を線分でつないで系図化することがはじめられ, 紙面を上から下へ描き継ぐ「縦系図」が現れる。その最初期の史料として知られるものに, 『和気系図』(『円珍俗姓系図』) や, 丹後国籠神社宮司家海部氏の『籠名神社祝部氏系図』(以下『海部氏本系図』と称する) が挙げられる。

『和気系図』では, 水平・垂直線分の長さが不揃いである。また紙面空間の制約を回避して, 垂直線分の迂回や斜め方向への延伸などがみられる。兄弟関係の表記は, 横一列に同世代を配しながら, 水平線分の下に文字列のみが列挙されている。『海部氏本系図』では, 垂直線分が人名を貫いて引かれている。また, 線分途中で水平線分が引かれ, 付帯情報が記されている。

情報整理の観点でいえば, 世代間の違いは垂直線分, 同世代 (兄弟関係等) や婚姻は原則水平線分に統一されればよい。

2.3. 横系図

中世ごろから, 縦系図で描かれた内容が卷子や冊子にまとめられ, 横方向に読み進める「横系図」が一般化する。系図を参照する研究では, 『群書類従』系譜部・『尊卑分脈』・『寛永諸家系図伝』・『寛政重修諸家譜』などが多用される。

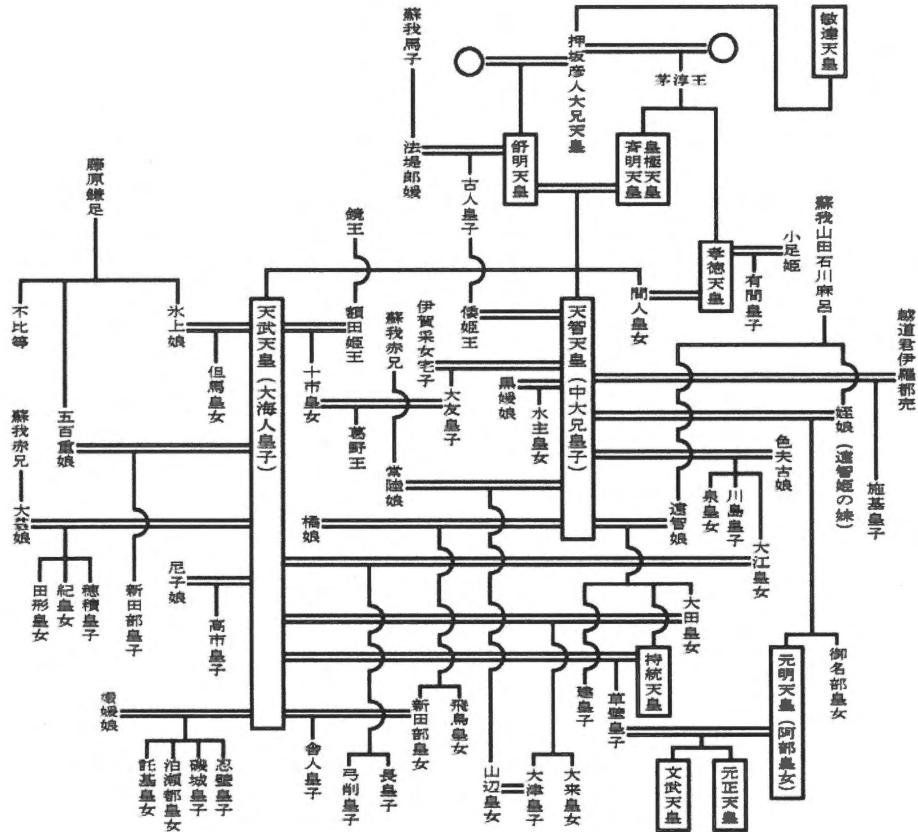


図 3: 線分交叉を用いた系図表示の例 (『口語訳 古事記 [完全版]』[16] 掲載図より作成)

横系図は史料形態の制約上、縦方向に伸びる世代の連続性が鉤形に屈曲するため、横系図を参照するのに不慣れな者にとっては視認性を悪くしている側面がある。

コンピュータは理論上、紙面形態や面積の制約とは無関係であるので、前節で述べた水平・垂直線分の基本原則を踏襲することができる。ただし『尊卑文脈』は現存する諸本とも、概ね人物情報の配置や線分の描法に大きな異同がなく [13]、史料成立時の系図作成上の意図を考慮する必要もあるだろう。

2.4. 慣習の描法

前節までに述べたような性質をもつ史料から、系譜・系図閲覧者は各自の関心に基づいて必要な情報を抽出し、自らの手で系図化しようとする場合がある。その際には、系図の線分を円弧を用いて交叉させながら、1つの個性を1箇所1度記すことが慣習的に行われている(図3)。この描法より、1つの個性は1回の目視ですべての情報を認知される。

人文研究者をはじめとする系図作成者・閲覧者の利便性を考慮すると、慣習的に行われてきたこのよう

な手描きによる系図描法を，コンピュータで復元する必要がある．

3. 既往ソフトウェアの検討

一般に系図表示ソフトウェアでは、系図化される個性の属性とそれに付随する関係性がデータベースに記録・管理される。そして、その系譜情報から自動的に系図化する際には、一定の描画ルールに則ったアルゴリズムによって描画される。多くの場合は、個性同士を直接結び付けて関係性を表現する。

この方法では、ある個性が他の多くの個性と関係性を有したり、近親的な間柄で複雑な関係性を有したりすると、プログラム上の制約から、個性の配置や線分描画が困難になる。

ここでは、これまでに開発され、普及度が高いと判断できる系図表示ソフトウェアをサンプルとして、これらがどのような描画手法を用いているかを調査し、上述までの要求に応えるにはどのような観点が必要であるかを考察する。

3.1. ソフトウェアの選定

系図表示ソフトウェアは、市販品や無料ソフトウェアなど、多数存在する。調査対象とする系図表示ソフトウェアの前提としては、調査の簡便性と第三者による追試の容易性、さらに利用言語、OS、完成度、リリース状況、拡張機能、経済的制約、関連ソフトウェアの要・不要を考慮した。そこで次の箇条書きのように、

- 利用言語が日本語または英語である
- 利用 OS が Windows XP
- 正式にリリースされている (Ver.1 以降)
- 機能面において複数婚に対応し他家系と連携する
- フリーソフトウェアあるいはシェアウェアである
- 他の有料ソフトウェアが不要である

という 6 点の条件を設定し、これらの条件を満たすソフトウェアとして、

- ancestry[14]
- myHeritage[15]
- アライアンス [5]

の 3 本を検討対象とすることに決定した。

3.2. 調査結果

選定された 3 本の系図表示ソフトウェアを利用して、複数婚が発生するデータが入力された時に、それぞれどのような表示がなされるかを調査した。

ancestry では、(1) 画面上に一度に表示できるデータが限られているため、家系を構成する個性の全体像が把握しづらい。そして (2) 複数婚の存在を見せない画面構成であるため、婚姻関係等の実態が完全には把握できない。さらに、(3) 同一家系内で複数の婚姻関係が結ばれたような情報が入力された場合には、同一個性が複数箇所に表示される。(3) の時、データベースで各個性に割り振られる単一 ID が表示されないため、各個性の同定が直感的には難しい。

myHeritage では、同一家系内で複数の婚姻が結ばれるような情報入力が行われると、ancestry と同じ問題を生じる (3)。また、(4) 他の家系と複数の婚姻関係を結んだ場合に、それぞれの相手を切り替えて表示する手法を採用している。このため、一回の視認では関係の全体像を把握することができない。

アライアンスは各個性に付される単一 ID が画面上の個性の肩に表示する。その上で複数ある婚姻関係の相手を複数箇所表示する。これにより、ID を頼りに複数箇所表示された個性の同定は可能となる。しかし、同じ個性を複数箇所表示する系図に不慣れなものにとっては、直感的な理解がやはり難しいといえる。

3.3. 問題点の整理

以上 3 本の系図ソフトウェアを調査した結果、複数の個性と婚姻関係を結んだり、同族家系の中で婚姻を結んだりする系譜情報に対しては、同一画面での表示を行わなかったり、同一個性を複数箇所に表示したりする処理を行っていることが指摘できる。

同一の個性を複数箇所に表示させることは、コンピュータ上のグラフィカルな問題、すなわち、線分が交叉してでも個性を 1 箇所に 1 度配置するという紙媒体で行われてきた慣習的描法を回避し、できるだけ単純なアルゴリズムでデータ処理を行うことが常であったことを意味している。

コンピュータ上で線分が交叉しても 1 回の目視で関係性を把握できる系図表示には、ひとつの方法として 3 次元表示を求めることができる。しかしこれにはコンピュータ上の技術的側面や [5]、人文研究者はじめ、系図を閲覧する者が「系図」として視認するにはまだ馴染まない現状も想定しなければならない。

以上の諸点から、コンピュータによる従来からの情報認知方法の問題も 2 次元で克服しておく必要があると考えられる。

4. 線分交叉表示の前提

公開されている既存の系図表示ソフトの多くは線分を交叉させず、同一個性が 2 箇所以上に配置されるので、関係性の直感的把握が困難であることを指摘した。コンピュータ上で処理される系図にも紙媒体のように線分を交叉させながら表示できれば、従来人文研究者らが行ってきた紙媒体の系図表示に近づけることができる。

本章ではその第一歩として、系図の線分が交叉する場合について考察を行う。

4.1. 図像化規則

視認性の良い系図表示を実現するには、簡素な図形表現が求められる。また、処理が重くならないようにするため、単純なアルゴリズムが必要である。これらの点を考慮し、本研究では以下のように図像化規則を定義した上で、線分交叉の組み合わせを調査する。

- (1) 線分や記号の種類については文化人類学の慣例を踏襲する [17]。
- (2) 紙媒体と同様に二次元平面で表示する。
- (3) 関係線の表示には水平・垂直線分のみを使用し、複雑な迂回や斜め方向への延伸等は行わない。

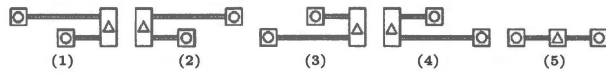


図 4: 婚姻相手 2 人の接続型

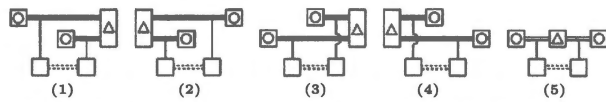


図 5: 婚姻相手 2 人の子の表示 (異母系親族婚含む)

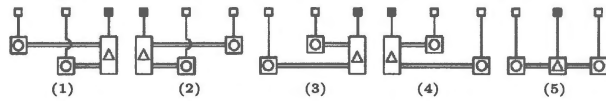


図 6: 婚姻相手 2 人の家系または親を表示

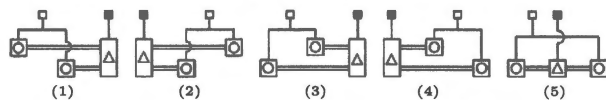


図 7: 婚姻相手 2 人が姉妹の場合

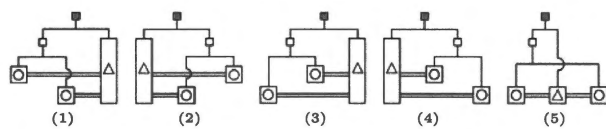


図 8: 婚姻相手 2 人が姪姉妹の場合 (異世代婚)

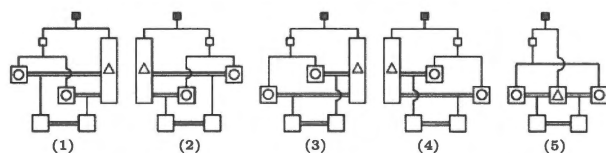


図 9: 異世代婚・異母系親族婚などを同時に表示

- (4) 交叉の表示には半円弧を使用する。
- (5) 1 つの個性を 2 カ所以上に表示しない。
- (6) 複数の個性を格子状に密集配置する。
- (7) 個性が増えるに従い、密集配置が不可能な場合にのみ開離配置とし、関係線の線分を伸ばす。

4.2. 線分交叉の最小単位

系図の線分交叉で考える組み合わせの最小単位として、ある個性が別の 2 つの個性と婚姻した場合と、兄弟姉妹関係にある 2 つの個性がそれぞれ婚姻した場合について議論する。また、本稿では便宜上、男系中心による婚姻関係を用いて記述する。

4.3. 2 つの個性との婚姻

婚姻相手が 2 人の場合、婚姻順や年齢順を考えなければ、図 4 の (1)~(5) のいずれかの図形で表現できる。

婚姻相手 1 人につき子が 1 人いる場合は、図 5 のように表現できる。ここで、(3)(4) の時に 1 回の交叉が発生するが、(1)(2)(5) では交叉は発生しない。また、二重点線で示した異母系親族婚を表示する場合についても交叉回数は同一である。

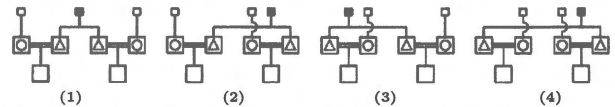


図 10: 2 人の兄弟それぞれに婚姻相手 1 人の接続型

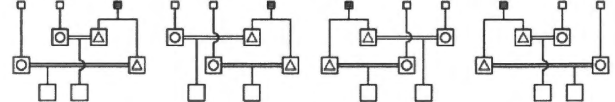


図 11: 2 人の兄弟が 2 人の姉妹と婚姻 (カリエラ型)

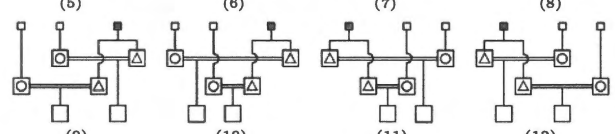


図 12: 終端ノードが交叉している場合

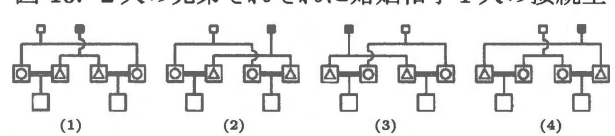


図 12: 終端ノードが交叉している場合

婚姻相手の家系または親を示す場合は図 6 のように表現できる¹。ここで、(1)(2) の時に 1 回の交叉が発生するが、(3)~(5) では交叉は発生しない。

婚姻相手 2 人が姉妹の場合は、図 7 のように表現できる。ここで、(1)(2)(5) の時に 1 回の交叉が発生するが、(3)(4) では交叉は発生しない。

婚姻相手 2 人が姪姉妹、即ち異世代婚となる場合は、図 8 のように表現できる。この場合の交叉回数は図 7 の交叉回数と同一である。

異世代婚、異母系親族婚などを同時に表示する場合は、図 9 のように表現できる。図 5~図 8 では交叉しない接続型が存在したが、図 9 では全て 1 回の交叉が発生する。

4.4. 2 つの兄弟姉妹関係の個性の婚姻

2 人の兄弟がそれぞれ婚姻相手を 1 人ずつ持つ場合は、図 10 のように表現できる。ここで、(1)~(4) は左右横並びに、(5)~(12) は上下左右並びに、それぞれ接続したものである。(2)(3)(5)~(9)(12) では 1 回の交叉が、(4)(10)(11) では 2 回の交叉が、それぞれ発生する。

2 人の兄弟が 2 人の姉妹と婚姻、文化人類学で「カリエラ型」と呼ばれる婚姻形態となる場合は図 11 のように表現できる。交叉回数はいずれも 1 回である。

¹ ■は男系親世代、□は婚姻相手の女系親世代を表す。

4.5. 考察

図 11 を図 10 (1)~(4) と比較すると、図 11 (4) については 2 回の交叉が予想されるが、個性の上下左右配置関係に違いがなく、かつ交叉する箇所が移動できる場合には、交叉回数を減らすことができる。

また、図 12 に示すように、末端の個性に接続された線分（終端ノード）の交叉は解消することができる。この設定を適用すれば、図 5 (3)(4) で点線の婚姻が無い場合、図 6 (1)(2)、図 7 (5)、図 10 (2)~(8)、図 11 (1)(4) では、交叉が発生しないように配置転換することが可能となる。更に、図 10 (10)(11) の交叉は 1 回で済む。

本稿で述べた接続型は、単に図形の配置を表しているにすぎない。実際、どのように個性を配置して系図化するかについては、個別の婚姻関係や上位下位世代との関係、年齢順等に従って、自由に作図することが望ましい。そこで、系図作成者により指定された個性の位置関係や終端ノードの条件等をもとに、それに見合った接続型と交叉の有無とを判定すれば、交叉表示を行うことができる。

5. 議論

本研究では、『古事記』学術支援データベースの拡張を視野に置きながら、主にコンピュータ上で日本史料の性質や、それらを慣習的に扱う系図作成者・閲覧者のインターフェイスに根ざした表示方法を考察した。

こうした系図を表示させるには、個性同士を結ぶ線分が交叉しても 1 つの個性を 1 箇所に 1 度表示させるプランが必要であることが確認できた。我々が開発した MaSSRiDGe を含め、これまで開発、リリースされてきた系図表示システムでは、上記の要求に応えることのできるアルゴリズムでは設計されていない。

したがって今後は、2 次元上で線分を交叉させてもストレスのないプログラムの構築と、それに合わせた系図データベースを再考する必要がある。また、人文系研究者をはじめとして、系譜情報を系図化したい者がわかりやすいインターフェイスで入力し、編集することも本データベースの大きな主眼の 1 つである。

これらを総合的に勘案すると、システム全体のプログラム設計がオブジェクト指向であることが必要である。これに伴い、付属する系図表示システムもオブジェクト指向データベースを用いた開発に切り替える必要があると考えられる。

6. おわりに

以上、コンピュータを用いた従来の系図表示について、その問題点と今後の可能性を論じた。現在、前節に述べた課題を実現するために、アルゴリズムの定式化をはじめ、目的の系図表示システムにむけたデータベースの検討を行っている。順次、システムの試作を重ね、『古事記』学術支援データベースに搭載し、動作実験を継続する予定である。

参考文献

- [1] 生田敦司, 齋藤晋, 柴田みゆき, “『古事記』学術支援データベースの構築—基本機能の検討—”, 情報処理学会研究報告, 人文科学とデータベース, 第 12 回公開シンポジウム-5, pp.47-54, 2006.12.23.
- [2] 相田満, “日本古典系図データベースの構築”, 情報処理学会研究報告, 2001-CH-51(8), pp.15-22, 2001.
- [3] 田中猛彦, 富金原賢次, 宇都宮啓吾, 中川優, “平安・鎌倉時代を対象とした僧侶データベースシステム”, 情報知識学会誌, 13(2), pp.18-31, 2003.
- [4] 朴明哲, 森本雅史, 立花純児, 村川猛彦, 宇都宮啓吾, 中川優, “人文研究を支援するデータベースシステム: 聖教検索および系図表示”, 情報知識学会誌, Vol.17, No.2, pp.105-110, 2007.
- [5] 杉藤重信, “人類学調査支援ツール、親族データベース「アラリアンス」について”, オセアニア学会ニュースレター, no.86, pp.10-37, 2006.
- [6] 生田敦司, 齋藤晋, 杉山正治, 柴田みゆき, 宮下晴輝, “『古事記』学術支援データベースの構築—系譜の図像化とインターフェイスの検討—”, 情報処理学会研究報告, 人文科学とデータベース, 第 13 回公開シンポジウム-2, pp.9-16, 2007.12.22.
- [7] 生田敦司, 柴田みゆき, 齋藤晋, 杉山正治, 宮下晴輝, “線分交叉を伴う系図表示の基礎的研究—人文研究が求める表現—”, 情報処理学会第 71 回全国大会, pp.4, 385-386, 2009.
- [8] 柴田みゆき, 杉山正治, 齋藤晋, 生田敦司, 宮下晴輝, “線分交叉を伴う系図表示の基礎的研究—既存の系図表示アプリケーションの現状と課題—”, 情報処理学会第 71 回全国大会, pp.4, 387-388, 2009.
- [9] 杉山正治, 柴田みゆき, 生田敦司, 齋藤晋, 宮下晴輝, “線分交叉を伴う系図表示の基礎的研究—線分交叉の前提と定式化に関する考察—”, 情報処理学会第 71 回全国大会, pp.4, 389-390, 2009.
- [10] 生田敦司, “記紀を遡る系譜史料の基礎的考察”, 龍谷史壇, 115, pp.21-52, 2001.
- [11] 義江明子, “天寿国續帳銘の一考察—出自論と王権論の接点—”, 日本史研究, 325, 1989. のち義江明子, “日本古代系譜様式論”, 吉川弘文館, 2000 所収.
- [12] 義江明子, “日本古代の氏の構造”, 吉川弘文館, 1986.
- [13] 黒板勝美, 国史大系編修会, “尊卑文脈第一篇 凡例”, “国史大系尊卑分脈”, 吉川弘文館, 1957.
- [14] The Generations Network, “Genealogy, Family Trees and Family History Records online”, <http://ancestry.com>, 2008.11.
- [15] MyHeritage “MyHeritage”, <http://myheritage.jp>, 2008.11.
- [16] 三浦佑之, “口語訳 古事記 [完全版]”, 文芸春秋, 2002.
- [17] 蒲生正男, 山田隆治, 村武清一 (編), “文化人類学を学ぶ”, 有斐閣, 1979.

古事記崩年干支についての疑念

Uncertainty due to the sexagenary-cycle year description used in *Kojiki*, an ancient book of Japanese history

小沢一雅（大阪電気通信大学）

Kazumasa Ozawa

Osaka Electro-Communication University

あらまし 古事記・日本書紀における年代表記の基本は、60年周期の干支年である。日本書紀は編年体で記述されているため年代情報の欠損はないが、古事記は多くの天皇の崩御年（崩年）の干支年が欠落していることもあり、絶対年への換算が唯一にできないという問題がある。とくに、日本古代におけるキーパーソンである崇神天皇の崩年を、古事記は戊寅年としているが、これを絶対年に換算するに際して、318年説と258年説の2つの説がある。本稿では、両説ともに大きな矛盾をはらんでいることを明らかにする。

Summary In ancient books of Japanese history, the sexagenary-cycle year description (SYD) is used to date historical events. The oldest book *Kojiki* also presents some years of death of emperors using SYD. The tenth Emperor *Sujin* is the most important person who plays a key-role in clearing the mystery of ancient Japan in the late 3rd century. A problem is that since *Kojiki* is missing many SYD years of death of emperors, given *Sujin*'s SYD year can not automatically be converted to a unique year in terms of non-cyclic year description, e.g. a Christian year. Specifically his SYD year has sometimes been addressed either to 258 AD or to 318 AD, depending on opinions of historians or archaeologists. This paper presents a quantitative discussion on whether each of 258 AD and 318 AD is reasonable or not in the context of *Kojiki*.

キーワード：古事記、崇神天皇、崩年、数理的分析

Keywords: *Kojiki*, Emperor *Sujin*, Year of death, Quantitative analysis

1. はじめに

古事記では、崇神天皇の崩年を戊寅（つちのえ・とら）年とししている。これが「正しい」記録であるという前提で、崇神天皇崩年を西暦換算で318年、あるいは258年とする見解が流布されてきた。近年、奈良県桜井市の箸墓古墳を卑弥呼の墓とする邪馬台国畿内説がマスコミを中心に喧伝されているが、この説は戊寅年＝258年とする仮説を土台にして組み立てられている。一方の、戊寅年＝318年説は、古代史・考古学の世界で長年にわたって定説的な見解として“市民権”をもちつづけている。これに依拠した学

説は、邪馬台国畿内説・九州説を問わず、枚挙にいとまがない。

崇神天皇の古事記崩年としての戊寅年に対して、ただ1つの西暦年が決まらず、このように多義性をもつのは、古事記における年代情報の記述が不完全なことに起因する。つまり、古事記では、多くの天皇について崩年（干支）の記述が欠落しているのである。完全な編年体をとっている日本書紀とくらべると大きな差異がある。

本稿では、古事記崩年干支に関する問題を数理的な視点からくわしく分析し、戊寅年＝318年説、および258年説ともに古事記の文脈上重大な矛盾をはらんでいることを明らかにする。

2. 古事記・日本書紀の天皇崩年干支

表1 古事記・日本書紀の天皇崩年一覧表 [1]

		古事記		日本書紀	
天皇諡号	代	崩年干支	西暦年	崩年干支	西暦年
神武天皇	1			丙子	BC585
綏靖天皇	2			壬子	BC549
安寧天皇	3			庚寅	BC511
懿徳天皇	4			甲子	BC477
孝昭天皇	5			戊子	BC393
孝安天皇	6			庚午	BC291
孝霊天皇	7			丙戌	BC215
孝元天皇	8			癸未	BC158
開化天皇	9			癸未	BC98
崇神天皇	10	戊寅	318	辛卯	BC30
垂仁天皇	11			庚午	70
景行天皇	12			庚午	130
成務天皇	13	乙卯	355	庚午	190
仲哀天皇	14	壬戌	362	庚辰	200
神功皇后				己丑	269
応神天皇	15	甲午	394	庚午	310
仁徳天皇	16	丁卯	427	己亥	399
履中天皇	17	壬申	432	乙巳	405
反正天皇	18	丁丑	437	庚戌	410
允恭天皇	19	甲午	454	癸巳	453
安康天皇	20			丙申	456
雄略天皇	21	己巳	489	己未	479
清寧天皇	22			甲子	484
顕宗天皇	23			丁卯	487
仁賢天皇	24			戊寅	498
武烈天皇	25			丙戌	506
継体天皇	26	丁未	527	辛亥	531
安閑天皇	27	乙卯	535	乙卯	535
宣化天皇	28			己未	539
欽明天皇	29			辛卯	571
敏達天皇	30	甲辰	584	乙巳	585
用明天皇	31	丁未	587	丁未	587
崇峻天皇	32	壬子	592	壬子	592
推古天皇	33	戊子	628	戊子	628

古事記は神武天皇から推古天皇までの33代、日本書紀は持統天皇までの41代の天皇について記載している。日本書紀は、先行する史書である『漢書』などの体裁にならって編年体で書かれている。つまり、天皇の即位年を起点として、在位中の事蹟を年代順に整然と記載していく記述形式である。一方の古事記は、編年体をとっていない上に、年代情報の記述そのものが不完全であって、崩年が記載されていない天皇も多い。

表1は、記紀それぞれに記載されている推古天皇までの天皇崩年の一覧表である。

同表に記入されている西暦年は干支年から

の換算であるが、日本書紀については干支年から一義的に西暦年が導かれるが、古事記の場合は上述のように、多義性をもっている。表1の古事記崩年干支に対応する西暦年の数値は、これまで“定説”的に採用されてきた換算例であるが、これを一括して導いたのは、筆者の知るかぎり、末松保和氏であって、昭和初期の頃であった[2]。末松氏の手続きは、つぎのようなものであった。

推古天皇崩年の西暦年は、628年であって日本書紀とも一致する。そこで、これを正しい起点として、そこから古事記崩年干支が記載されている天皇について順次遡上し、1運(60年)以内で対応する西暦年をあてはめていったのである。たとえば、継体天皇(丁未年=527年)のつぎは雄略天皇(己巳年)であるが、己巳年を継体天皇の丁未年から60年以内で算定すれば、38年前になる。これをもとに、雄略天皇の西暦崩年を、

$$527年 - 38年 = 489年$$

と算定するわけである。つまり、両者の途中にある清寧・顕宗・仁賢・武烈という4天皇を無視する。末松氏は、このように崩年干支の記載のない天皇を無視する理由についてくわしくは述べていない。推測であるが、実在性を認めていないのかもしれない。

定説的に採用されてきた崇神天皇崩年318年とは、推古天皇からこのように順次算定されて到達した成務天皇崩年355年から、同じ手続きで得られる最後の結果なのである。

末松氏の西暦年換算法を、ここでは末松換算とよぶことにする。末松換算から導かれた崇神崩年318年説とは、たとえば仲哀天皇や成務天皇の実在を想定しないかぎり成立しえないものであることがわかる。逆に、318年説に立ちながら、仲哀天皇や成務天皇の非実在を説くことがあるとすれば、自己矛盾のきわみということになる。

3. 崇神天皇崩年の意味するもの

崇神紀には、有名な三輪山伝承が記載されている。その中で、崇神天皇の在位中に、祖父の姉にあたる倭迹迹日百襲姫(ヤマトトトヒモソヒメ)を葬る大市墓(箸墓古墳)を築造した、という記事が重視されてきた。その理由は、箸墓古墳が(1)女性墓であるとする伝承であること、および(2)崇神天皇の在位中の築造とされていることである。

この背景には、箸墓古墳を最古級の前方後円墳とみなす考古学的な年代観がある。すなわち、前方後円墳の時代、つまり古墳時代が、箸墓古墳の築造を嚆矢としてはじまったと考える年代観である。

魏志倭人伝からは、邪馬台国の卑弥呼が魏に遣使したのは239年、卑弥呼が没したのは248年頃という年代が確定している。

崇神天皇の古事記崩年・戊寅年は、末松換算によれば318年となるが、卑弥呼の没年にかかなり近い年代である。一方の258年説をとれば、卑弥呼の没年にさらに近接した年代となる。箸墓古墳と卑弥呼を、まさに間接あるいは直接に関連づける年代になるわけで、そうした“想い”が崇神紀の伝承を重視する根底にあるとおもわれる。

以下では、崇神崩年の戊寅年=258年説と、318年説について、古事記の文脈上か

ら読みとれる数理的な疑念について考察する。

4. 崇神崩年258年説と318年説がはらむ問題点

258年説 —笠井新也氏の説—

近年、箸墓古墳が卑弥呼の墓だという、笠井新也氏が大正時代に唱えた畿内説[3]が一部で再評価されている。笠井氏は、魏志倭人伝の文献学的検討、および大和地方の遺跡に関する考古学的知見を総合して、畿内説をみちびいた。畿内説となれば、卑弥呼の墓は当然奈良につくられたことになる。笠井氏は、箸墓古墳こそそれだと主張したのである。

魏志倭人伝には、卑弥呼が死んだあと「大いに冢(つか)を作る 径百余歩 徇葬する者 奴婢百余人」とある。箸墓古墳の後円部の直径は、筆者の計測によれば161mである。1歩を適当なメートルに設定すれば、笠井氏が主張するように、「径百余歩」という記述と大きな矛盾はきたさない。ただし、倭人伝の「径百余歩」という描写からごくすなおにイメージされる墳墓の形状は、あきらかに円形であって、「前方後円形」ではないことは、笠井説の難点の1つであろう。また、1歩の長さを適当に解釈すれば、後円部だけをみて「径百余歩」になる前方後円墳は、箸墓以外にも奈良にはたくさんある。これらの中から箸墓古墳一基にしぼりこむには、さらにもうひとつの論拠が必要になる。

私見だが、この「もうひとつの論拠」こそ、じつは笠井説の真の核心であり、着想のそもその原点ではないかと推測している。笠井氏は、崇神天皇崩年・戊寅年=258年とした。卑弥呼が没したのは248頃である。崇神天皇が258年に没したとすれば、卑弥呼の死は、まさに崇神天皇の在位中か、即位直前の出来事になる。

崇神紀が伝える、神がかりで大規模な箸墓築造の物語は、まさに卑弥呼の墓の築造にふさわしい。笠井氏はそう考えて、卑弥呼=倭迹迹日百襲姫という等式を着想した。わたしはそう推測している。卑弥呼が「鬼道を事とし、能く衆を惑わす」と魏志倭人伝が伝えていることと、崇神紀にある倭迹迹日百襲姫の呪的な伝承との対比などについて、氏はいろいろ言を重ねている。が、笠井説の根元は、なんといっても崇神天皇の古事記崩年干支・戊寅年を258年と解釈した一点につきるとおもう。年代という数値が、笠井説の成立に決定的な役割をはたしているのは、まずまちがいない。

それでは、この年代観は妥当といえるであろうか。検証をおこなうことにしよう。

検証の方法であるが、258年は古事記にある崩年干支を根拠としている点を重視した方法でなければならないだろう。つまり、古事記崩年干支という世界の中で、戊寅年=258年という仮説の妥当性を検証するのである。すなわち、表1の成務天皇崩年355年を基準とした検証をおこなう。

いま、ここに問題がある。成務天皇の実在性である。井上光貞氏は、成務天皇の実在をみとめていない[4]。ただし、井上氏の見解は、成務天皇を別系統の人物におき替えるということであって、皇統からただ削除するという、単純消去法ではない。したがって、その代替者の没年を想定して検証すればよいことになり、これから行う検証の有効性にまった

く影響しない。

そこで、成務天皇崩年355年を定点としよう。記紀が伝える皇統は、崇神、垂仁、景行、成務の順である。古事記は、表1にあるように、垂仁天皇と景行天皇の崩年干支を記載していないが、崇神天皇の崩年がわかれば、垂仁・景行・成務三天皇の三代にわたる在位年数は計算できることになる。

笠井氏は、崇神天皇の崩年を258年としたわけだ。そうすると、

$$355年 - 258年 = 97年$$

であるから、垂仁・景行・成務三天皇の三代在位年数は97年ということになる。

ここで検証しようとしているのは、この97年が妥当な数値なのか、あるいはそうでないのか、という一点である。検証の手順を説明しよう。まず、記紀の崩年が信用できる古代天皇（允恭天皇）から昭和天皇にいたるまでの、すべての天皇の崩年（厳密には退位年）を基礎データとし、連続三代の在位年数をくまなく算出する。その結果をグラフ化すれば、天皇の三代在位年数の実態が歴然とする。図1がそれである。

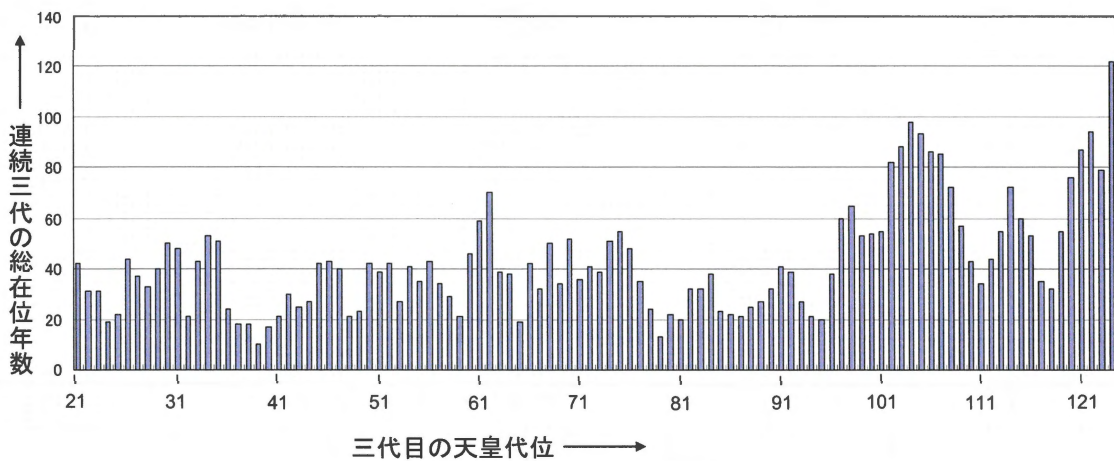


図1 連続三代の在位年数集計

この棒グラフの見方であるが、たとえば、明治・大正・昭和の連続三代を例にとると、この三代の在位年数122年を昭和天皇（124代目）が代表する年数として「棒」の高さを決めている。1代くり上げた、孝明・明治・大正の三代の在位年数79年は、大正天皇（123代目）が代表する年数とするのである。もっとも古代には、雄略天皇（21代）が代表する、允恭・安康・雄略三代の在位年数42年が、最左端の「棒」の高さになっている。

このグラフをみれば、問題は一目瞭然であろう。たしかに、室町時代末期～安土桃山時代、および江戸時代末期以降には、三代在位年数が、100年近い年数になる事例（明治・大正・昭和の三代は100年を超える）もあるが、すくなくとも、醍醐天皇（60代）以前の古代にはまったくない。長くても、40年前後だということは一見してわかる。

そこで、笠井氏の年代観から導かれる、成務天皇（13代）が代表する垂仁・景行・成

務三代の在位年数97年はどうだろうか。答えは明白だ。ありえない年数といわねばならないだろう。すなわち、笠井氏が採用した戊寅年=258年という年代観は、図1に示めされる三代在位年数集計結果から推してまず成立しえないことが判明する。

318年説

戊寅年=318年説は、三角縁神獣鏡の同範鏡分有仮説にもとづく邪馬台国畿内説で有名な小林行雄氏をはじめ、多くの考古学者や古代史家が長年にわたり採用してきた、いわば「定説」に近い西暦換算である。これについても、数理的な検討をくわえてみる。

まず、結論からいうと、古事記の編者は、本来崇神天皇の崩年を318年よりはるかむかしに想定していたはずである。これは、以下のように、古事記の文脈を数理的に点検することによって明らかになるのである。

古事記によれば、崇神天皇は168歳で崩御とある。つづく垂仁天皇は153歳、景行天皇は137歳、成務天皇は95歳である。一方、表1から、成務天皇の古事記崩年は355年である。天皇の系譜から、崇神天皇の曾孫である成務天皇が、95歳で崩御したとすると、崇神天皇の崩年318年には、すでに58歳だったことになる。

崇神天皇が崩御してから成務天皇が崩御するまでの期間は、表1から計算すると37年間である。この“わずかな”期間内に、それぞれ153歳、137歳、95歳で崩御した垂仁、景行、成務の3天皇が、つぎつぎ即位し在位したことになる。すくなくとも、垂仁天皇と景行天皇は、100歳を超えてから即位した計算になるだろう。

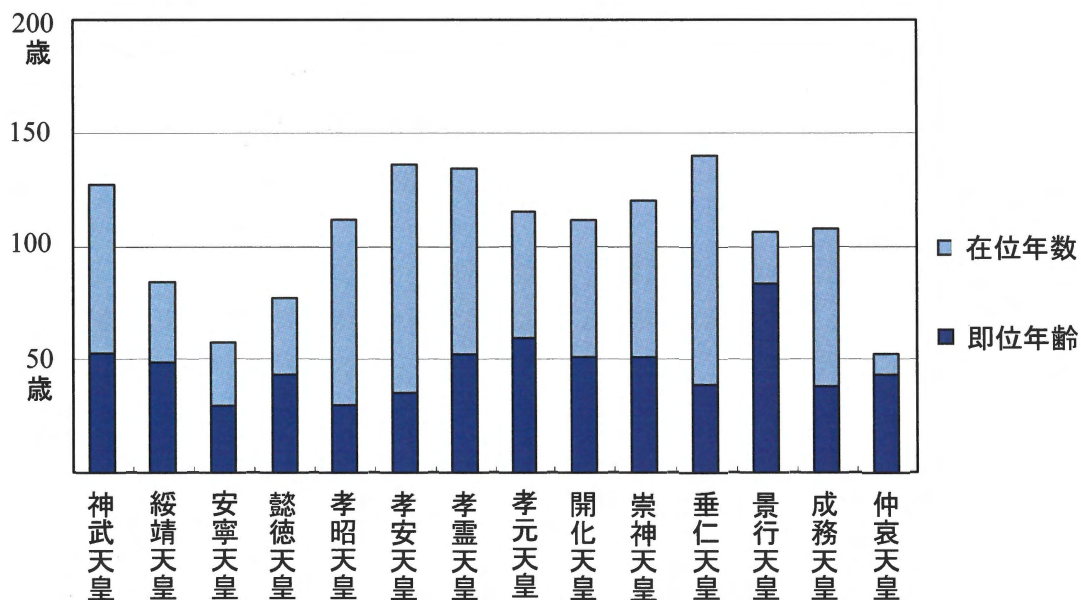


図2 日本書紀における古代天皇の年歴一覧

153歳といった長い寿齢は、いまでいえば装古操作だが、編纂時には「大むかしの天皇は長寿だった」という設定が異常とは認識されなかったからである。しかし、すくなく

とも2人の天皇が引き続いて100歳を超えてから即位するなどという事態は、あきらかに異常である。

日本書紀では、初期の天皇にあっても20歳前後で皇太子になり、30～60歳ぐらいで即位したと記述しているケースがふつうである。図2に、日本書紀に記載されている古代天皇の即位年齢と在位年数を表示した年歴グラフ（棒の高さは天皇の寿齢になる）を掲げている。もちろん、100歳を超えて即位するような年歴の組み立ては行われていない。

古代天皇の年歴（いつ即位し、何歳まで在位したか）は、古事記・日本書紀ともに、編者がいわば「創作的に」組み立てているものとおもわれるので、古事記編者にしても、2人の天皇が100歳を超えてから即位するというような“非常識”なシナリオを想定していたとは、とうてい考えにくい。

153歳というような長い寿齢を設定している理由は、長い在位年数を確保したいという編者の構想のあらわれであって、決して即位年齢を異常なまでに高齢にしようと意図したものではないはずである。100歳を超えるような高齢で即位し、ごく短期間しか在位しなかった、というようなシナリオを編者は描かないだろう。

こうした矛盾は、そもそも、戊寅年＝318年とする換算から発生するわけであって、古事記の年代記述が不完全であるため、一見、数値的に歴然とした矛盾にはみえないが、編者の本来の認識とはあきらかにちがっているはずである。もし、垂仁天皇と景行天皇の2人が、日本書紀が描いている古代天皇たちの年歴の「シナリオ」（図2）から類推して、50歳程度で即位したとしてみれば、戊寅年を318年よりも、すくなくとも120年、あるいは180年くり上げて換算しなければならない計算になる。

つまり、もし、古事記の年代記述が完全ならば（編者の構想が明確に記述されているならば）、おそらく崇神天皇崩年は2世紀中の戊寅年になったはずである。戊寅年＝318年説は、このように古事記の文脈上、重大な矛盾をはらんでいることがわかる。

表1における、末松換算によって与えられた崇神天皇以下数代の天皇の古事記崩年（西暦年）についても、文脈を無視して、崩年干支だけをピンポイントでとりあげ、近年の年代観にあうように西暦年をあてはめただけの産物とみなさざるをえない。古事記編者の意図とは、あきらかにちがっているはずだ。これは、日本書紀の崇神天皇崩年干支・辛卯年に対して、文脈を無視して、たとえば4世紀中の331年にあてることとまったくおなじである。

5. むすび

末松換算は、さきにしたように、信頼できる推古天皇崩年・戊子年＝628年を起点として、順次遡上しながら古事記崩年干支に西暦年を割りあてていったものであることから、推古天皇に近い天皇については、ある程度の信頼性はあると考えられる。実際、表1で、日本書紀崩年（西暦年）と比較してみると、多少の誤差をゆるせば、允恭天皇あたりまで両書の崩年は数値的にほぼ一致している。ひとつの判断として、この辺までは、末松

換算による西暦崩年がほぼ信頼できるとみなすことができよう（どちらかといえば、日本書紀崩年の方が信頼性が高いと考えている）。

一方、崇神天皇の崩年については、表1の古事記崩年318年と日本書紀のそれを比較してみると、日本書紀があまりにも荒唐無稽な数値（紀元前30年）であるため、対照的に318年に妥当性があるようにみえる。

たしかに、318年は、数値としてみれば、日本書紀崩年より“まし”であるが、これを真値と断定できる根拠はまったくないばかりか、それを崇神天皇の崩年とすれば、さき書いたように、古事記の文脈上、あきらかな矛盾をひきおこすことになるのである。

すなわち、笠井新也氏が採用した崇神天皇崩年・戊寅年＝258年が、三代在位年数からみて、とうてい成立しえないことを前節であきらかにした。

258年説は、このように現代の合理的な視点からみ成立しえないし、一方の318年説は、前述のように古事記の文脈上あきらかな矛盾をひきおこしている。深刻なディレンマである。この解決不能のディレンマの原因をつきつめると、結局、崇神天皇崩年を「戊寅年」とする前提にあるとしか考えられない。井上光貞氏は、崇神崩年干支・戊寅年を「信用できない」と述べている[4]。その根拠を明らかにしていないが、老練な歴史学者の直感力がはたらいた結果と推測している。

筆者の見解としては、こうした払拭できそうにない疑念のある古事記崩年干支に頼らず、まったくちがった角度から、日本古代のキーパーソンである崇神天皇の崩年を決定していく必要があると考えている。その具体的方法として、すでに崩年モデルによる推定法を提案している[5]。さらに、崩年モデルを用いた天皇崩年の詳細な数理的検討を通じて日本古代の年代軸を再構成し、新たな視点からの邪馬台国論を試みている[6]。

【参考文献】

- [1] 和田萃『大系日本の歴史2—古墳時代』、小学館、1987.
- [2] 末松保和『日本上代史管見』、(自己出版)、1963.
- [3] 笠井新也「卑弥呼即ち倭迹迹日百襲姫命(1)」、『考古学雑誌』、14巻7号、1924.
- [4] 井上光貞『神話から歴史へ(日本の歴史1)』、中央公論社、1965.
- [5] 小沢一雅「崇神天皇の崩年はいつごろか—崩年モデルによる数理的検討—」、第13回公開シンポジウム「人文科学とデータベース」論文集、pp.113-120、2007.
- [6] 小澤一雅『卑弥呼は前方後円墳に葬られたか』、雄山閣、2009.

第15回公開シンポジウム
「人文科学とデータベース」

2009年11月28日(土)
神戸大学国際文化学部F301教室

プログラム

- 10:00～ 受付
- 10:30～10:40 開会挨拶 実行委員長：森下 淳也（神戸大学）
- 10:40～11:10 【自由論題報告】
田中 順子，森下 淳也（神戸大学），中島 和子（トロント大学）
「継承語教育文献データベースの構築」
- 11:10～11:40 永田 良茂
「日本語の起源，解明方法と数値評価」
- 11:40～12:10 モクタリ 明子（神戸大学），ニック・キャンベル（ダブリン大学），
田畑 安希子（神戸大学）
「表現豊かな自然発話コーパスのアクセスについて」
- 12:10～13:30 昼休み
- 13:30～14:30 【招待講演】
久保 正敏（国立民族学博物館教授）
「国立民族学博物館におけるデジタルアーカイブズ」
座長：岡田 浩樹（神戸大学）
- 14:30～14:50 コーヒーブレイク
- 14:50～15:20 【自由論題報告】
山元 啓史（東京工業大学）
「和歌解析用 MeCab 辞書の開発
—八代集解析済みコーパスによる学習—」
- 15:20～15:50 安藤 秀明，秋吉 信吾，竹内 和広（大阪電気通信大学）
「作文指導に向けた表現データベースの構築と表現評価指標の検討」
- 15:50～16:20 生田敦司（大谷大学），杉山正治（立命館大学），柴田みゆき（大谷大学），
宮下晴輝（大谷大学）
「『古事記』学術支援データベースの構築
—系図表示システムにおけるデータ構造の検討—」
- 16:20～16:50 小沢一雅（大阪電気通信大学）
「古事記崩年干支についての疑念」
- 16:50～16:55 閉会挨拶
- 17:20～19:20 懇親会

第 15 回公開シンポジウム「人文科学とデータベース」発表論文集

発行日 2009 年 11 月 28 日

編集・発行

第 15 回公開シンポジウム「人文科学とデータベース」実行委員会

事務局 〒657-8501 神戸市灘区鶴甲 1-2-1

神戸大学大学院国際文化学研究科

清光 英成 研究室内