古辞書データベース構築の過程
—院政期の国語辞書『色葉字類抄』を例に—
Construction of a Database for Japanese Old Dictionaries:
The Case of the *Iroha-Jiruishō*, Compiled in the Late Heian Period

藤本 灯[†1], 志村 誠[†2], 津村 昌祐[†3], 北﨑 勇帆[†4]

Akari Fujimoto, Makoto Shimura, Masahiro Tsumura, Yuho Kitazaki

[†1]国立国語研究所, 東京都立川市緑町 10—2

[†2]株式会社ドワンゴ [†3]富士通株式会社 [†4]東京大学大学院人文社会系研究科

[†1]National Institute for Japanese Language and Linguistics,

10-2 Midori-cho, Tachikawa, Tokyo

[†2]DWANGO Co., Ltd. [†3]Fujitsu Ltd.

[†4]Graduate School of Humanities and Sociology, University of Tokyo

あらまし:本発表においては, 院政期の国語辞書『色葉字類抄』(30811 字語を収録)を対象として発表者らが構築したデータベースについて述べる. 本書が国語辞書として広く利用可能となるよう, Web サービスとして構築を行ったが, その際に問題となった, 1) 検索結果のフィルタ方法 2)異体字データの表現方法 3) 文字コードへの対応方法 の 3 点についての具体的な解決法と, 今後の国語辞書系古辞書のデータベース化の展開についての議論を行う.

Summary : In this article, we discuss an old Japanese database published by us. This database contains 30,811 words from the *Iroha-Jiruisho*, a Japanese dictionary compiled in the late Heian period. First, we describe the features of the *Iroha-Jiruisho*, such as its historical background, the kind of words it contains, its composition, and typical features. Second, we explain the database guidelines that specify how to search a particular character. Users can search using words like *kanji*, *modern kana*, and *traditional kana* and see their new character forms and syllabary spelling. We also discuss three major difficulties that surfaced when we developed a web service interface for the database: (1) how to filter the result; (2) how to express *itaiji*, a variant form of a *kanji*; and (3) encoding of the database web service. Third, we explain how this database was developed and how we resolved these three difficulties. We used PHP, bootstrap, and MySQL to develop a database that effectively uses multibyte characters. Finally, we reviewed the difficulties and subsequently considered the possibility of developing the old Japanese database.

キーワード:色葉字類抄, データベース, 日本の古辞書

Keywords : *Iroha-Jiruisho*, database, Japanese Old Dictionaries

1. はじめに

『色葉字類抄』は, 平安時代末期(12 世紀半ば頃)に橘忠兼(伝未詳)によって編纂された国語辞書である. これまでの研究により, 本書は, 男性貴族が文章を記す際に, 語の漢字表記を調べる用途のために編纂されたものと結論付けられている. 現代においては, 反対に, 中古・中世の漢字文献における漢字の読み方を推定する根拠としても用いられることが多いが, いずれにせよ, 日本語学, 日本文学, 日本史学において, 本書は, 古代日本語を読み解くための必携の古辞書の一つとして位置付けられてきた. しかし,『色葉字類抄』に収録された語を調べることは, (全語を対象とした索引が, 部首引きで手書きのものしかないなどの理由から)必ずしも容易ではなかった面がある.

こうした背景から, 発表者らは, 本辞書を対象としたデータベース『三巻本「色葉字類抄」収録語彙データ

ベース』を作成した[1]. 本発表ではデータベース構築の過程を述べ, 古辞書をデータベース化する際の問題について論じる. 以下, 本発表の構成を述べる.

まず, 第2節では対象とした『色葉字類抄』の概要と, 各項目の持つ情報について概説する. 第3節ではデータベースの実装, 第2節で述べた各情報に対する検索方法と, 構築過程で生じた「検索結果の絞り込み」「異体字の表示方法」「文字コードの対応」といった問題について述べる. 第4節で今後の課題と展望として, 複数の辞書に対応した辞書横断データベースについて述べる.

2 色葉字類抄の概要

2.1 底本

『色葉字類抄』には数々の異本や伝本が存在するが, 編者の自筆本は残っていない. そこで我々は, 現在最もよく利用されている三巻本『色葉字類抄』のうち, 成立後まもなく書写された前田家尊敬閣文庫蔵本(=前田本, 中巻と下巻の一部欠)[2]と, それを江戸時代に忠実に写した黒川家本(=黒川本, 現在国立国会図書館古典籍資料室蔵, 完本, 図1)[3]を底本としてデータベース構築を行うこととした.

図1 黒川本色葉字類抄 (現国立国会図書館蔵)

2.2 本書の構成

『色葉字類抄』では, 見出しとなる漢字語を, その音読みないし訓読みの頭音にあたる音節によりいろは順に並べ, さらに各篇の中を形や意義で21部[4]に分類し, 排列する. これはこの後江戸時代に至るまで, 国語辞書の主流となった分類排列法である. すなわち『色葉字類抄』は図2のような階層構造を持ち, 各項目は, 篇(イ〜ス)・部(天象〜名字)・所在(頁等)の情報を持つこととなる.

図2 『色葉字類抄』の構造

2.3 各項目の持つ情報

前項に述べた各項目は, 読みや語義といった情報を持つ. 以下, 各要素について概説する.

2.3.1 見出し語

見出し語は漢字二字〜六字から成り, 多くは一, 二字である. 字体には旧字や異体字を多く含む.

2.3.2 読み

各見出し語は, 音読み(もしくは, 漢字の音読みを利用した読み)か訓読み(もしくは, 漢字の訓読みを利用した読み), あるいはその両者を読みとして持つ. ただし両者を持つ場合, 見出し語に対する主たる読み(=頭音が所属篇と一致)は, 原則として見出し語の下に配置される. 図3の「雷」の右に「ライ」, 下に「イカツチ」, 図4の「戸」に「コ」「へ」とあるようなものであるが,

1 本データベースは http://jiruisho.l.u-tokyo.ac.jp/ において一部試験公開中である.

2 公刊されているものに, 前田育徳会尊経閣文庫編 (1999)『色葉字類抄』八木書店 など.

3 公刊されているものに, 中田祝夫・峰岸明共編 (1977)『色葉字類抄研究並びに総合索引 (黒川本・影印篇)』風間書房 など.

4 天象・地儀・植物・動物・人倫・人体・人事・飲食・雑物・光彩・方角・員数・辞字・重点・畳字・諸社・諸寺・国郡・官職・姓氏・名字から成る.

図5のように「タヽシタヽス」「タンシヤウ」とあるようなものでも，主たる読みは「タンシヤウ」であることが分かる．ただし図6のように読み方が示されないものも稀にある．

図3 「雷」

図4「戸」

図5「端正」

図6「歎息」

2.3.3 注文

語の注釈・解説を持つ項目がある．図3の「雷」の項目には「又乍靐」（見出し語の雷の異体字に靐があることを示す字体注記，乍は作の略字），図4には「民―」（「民戸」という熟語形があることを示す），図5には「云形美也」（形が美しいことであるという語釈），図7には「芋根也」（言い換えによる語義説明）などとあり，これらをまとめて「注文」と称するが，注文を持たない項目も多くある．

なお，国郡部に収録された国名などでは，

伊勢 （以下割注）大東海 桑名〈クハナ〉 員辨〈ヰナベ〉 朝明〈アサケ〉 三重〈ミヘ〉／河曲〈カハワ〉 鈴鹿〈スヽカ〉〈国府〉 奄藝〈アムギ〉 安濃〈アノ〉 壹志〈イキシ〉 飯高〈イヒタカ〉／多気〈タケ〉 飯野〈イヒノ〉／度會〈ワタライ〉〈大神宮／在此郡〉／（朱）田万九千廿四丁 行程上四日下二日

のように，項目の下に非常に長い文章が連なるなど，注文の長さは語によってまちまちである．

図7「魁」とその声点

2.3.4 声点

アクセントや音の清濁を示す符号で，約六種類（とそれぞれの清濁）がある．漢字の四隅，もしくは四隅を頂点とした四角形の辺上に朱点で付される．図3・図7では漢字の左下に点（平声の声点）が見える．

図8「生」

2.3.5 合点

同訓異字が羅列される場合，より一般的に使用する漢字に付く符号とされる．項目の右肩に朱で付される．例えば，図8「生」（イク，生きるの意）の項目の後には，同じ「イク」を表す字として「活」「存」「居」「穀」「蘇」「穌」の字が掲出されるが，このうち，「活」「存」には右肩に合点が付されている．

2.4 まとめ

以上，『色葉字類抄』全体の構造と，各項目が持つ要素について述べた．これまでに挙げた見出し語を例として各項目に含まれる情報を一覧すれば，次表1のようである．

表1 項目情報一覧

見出し語	雷	戸	活
所属篇	イ	ヘ	イ
所属部	天象	地儀	人事
所在（前田本）	上2オ	上50オ	上6オ
所在（黒川本）	上2オ	上39ウ	上5オ
音読み	ライ	コ	クワツ
訓読み	イカツチ	ヘ	（イク）
注文	又乍靐	民―	－
声点	平	－	－
合点	－	－	有

3 データベースの構築

3.1 システム構成

色葉字類抄をデータベースとして公開するにあたり，我々は検索キーワードを入力すると，それに関する検索結果を表示するWebアプリケーションとして実装した．これはWeb上に公開されている多くの辞書と同じ形式であり，色葉字類抄も辞書の一種であるため，同様の手法が有効だと考えた．構築したWebアプリケーションの外観を図9に示す．

図9　データベースの外観

具体的なシステム構成としては，PHPのWebアプリケーションフレームワークCakePHP1.3を用い，バックエンドのデータベースとしてMySQLを使用した．ここでMySQLについては，4バイト対応のUTF-8文字を取り扱うことが可能な，MySQL5.5.3以降のバージョンを使用している．これは『色葉字類抄』に掲載されている漢字の中に，従来の3バイト対応のUTF-8では表示できないものが数多く含まれているためである．またTwitter社が公開しているcssフレームワークであるBootstrap[5]を用いることで，レスポンシブデザインに対応し，小さなディスプレイやスマートフォンからでもレイアウトを崩さずに表示が可能となっている．

3.2 実装

本節では2.3.で述べた各項目について，具体的な実装方法を説明していく．実装においてキーとなる箇所は，利用者が入力したキーワードに対する検索処理部と，マッチした結果の表示方法の 2 つに大別される．以下に，詳細を述べる．

[5] Bootstrap http://getbootstrap.com/

3.2.1 検索処理

検索処理部について，主となる検索フローを図10に示す．まずユーザーにより入力されたキーワードに対して，正規化処理を行う．具体的には，(1)SQLインジェクションを防ぐためキーワードのサニタイズ処理を行い，(2)全角かなと半角カナを全角カナに変換，(3)「A」「B」「【」「(」など，本データベース内で特殊な意味を持つ文字を除去，といった処理が含まれる．

続いて正規化されたキーワードを用いて，データベースへの問い合わせを行う．本システムでは，読み方と見出し語が検索キーワードによる探索範囲となる．その上で，完全一致による検索結果と，部分一致による検索結果をそれぞれ得る．なお，ここでの部分一致による検索結果は．完全一致による検索結果と照らし合わせて，完全一致検索には含まれないが部分一致検索には含まれる結果のみを取り出すものとする．

また読み方の検索においては，原本の仮名遣い(および和訓の場合，それを修正した歴史的仮名遣い)と，現代仮名遣いのいずれによっても検索可能とした．具体的には，データベースのテーブル上に原本の仮名遣いカラムとともに，現代仮名遣いのカラムも用意した上で，両方のカラムに対して同時に検索を行っている．

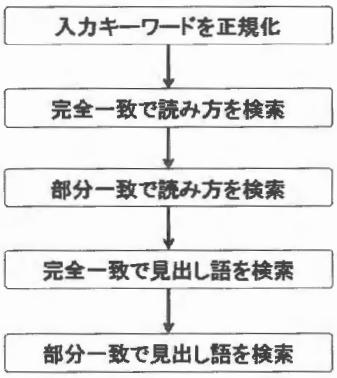

図10　検索処理のフロー

3.2.2 結果表示

続いて，得られた検索結果を表示する部分の実装について述べる．結果の提示順としては，図10に示した検索順に，得られた結果を上から並べて表示する形を取った．これにより，主となる読み方の検索結果について，完全一致→部分一致と並び，その下に従となる見出し語が続くこととなり，ユーザーの求めている結果にマッチしやすいものから順に並ぶと考えられる．

検索結果として並ぶ，各単語の詳細情報の表示例を図11に示す．見出し語を一番上に載せ，各項目を順に表示する形を取っている．UTF-8で表示不可能

な文字については，該当カラムに対して「A」「B」といった文字を仮に置き，これの近似字体を画像データで表示するか，解字情報を示すかの対応を行った．

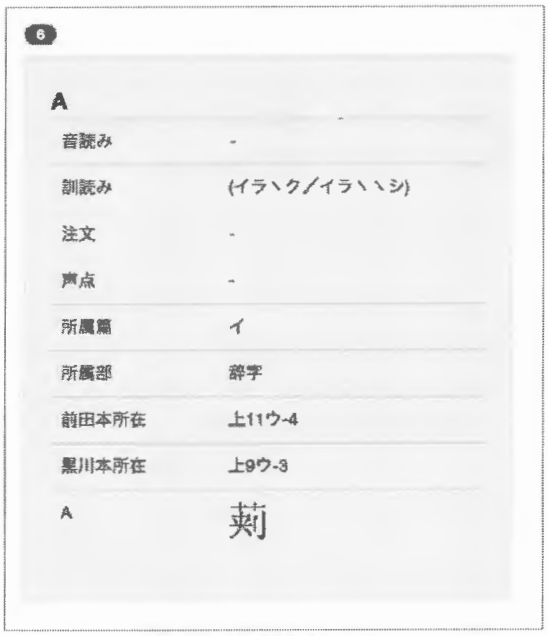

図 11　検索結果の表示

3.3　実装時の問題点と対処法

3.3.1　検索結果のフィルタ方法

ここまで説明してきた実装方法には，検索結果の表示について大きな問題がある．それは例えば「イ」のような非常にありふれたキーワードで検索を行うと，膨大な単語が該当してしまい，実用上の意味をなさない点である．これに対処するため，我々は検索結果に対するフィルタを実装した．具体的には所属篇，所属部，および見出し語の漢字数の 3 種類について，プルダウンメニューから指定することで，検索結果を動的に絞り込めるようになっている(図 12)．

3.3.2　異体字データの表現方法

古代日本語のデータベースを Web 上に表示するにあたり，大きな問題となるのが既存の文字コードで表示不可能なデータをどう扱うか，についてである．色葉字類抄においても，UTF-8 では表示不可能な文字が多数存在する．本データベースにおいては，表示不可能な異体字を「A」「B」といった文字で仮に表現し，別途設けた「A」「B」欄で，文字鏡研究会により提供されている今昔文字鏡の画像データ[6]を表示するか，解字情報を示すかの対応をとった．

3.3.3　文字コードへの対応方法

前節とも関連するが，古代日本語を含むデータベースにおいては，そもそもどのような文字コードを用いるべきか，という問題がある．色葉字類抄に含まれる文字を取り扱うためには，JIS X 2013 で定められている第 3 水準および第 4 水準の漢字を用いる必要があった．そこでこれらの漢字に対応しているutf8mb4を利用可能なデータベースとして，バージョン 5.5.3 以降の MySQL を採用した．

4　課題と展望

以上，平安時代院政期に成立した三巻本『色葉字類抄』のデータベース構築過程を通して，各項目への処理の方針および実装方法を明らかにした．

発表者らは今後，フィルタの条件を増やすなどの改変を予定しており，その他のシステムの改善にも意欲的である．また，本データベースは現在試験運用中で

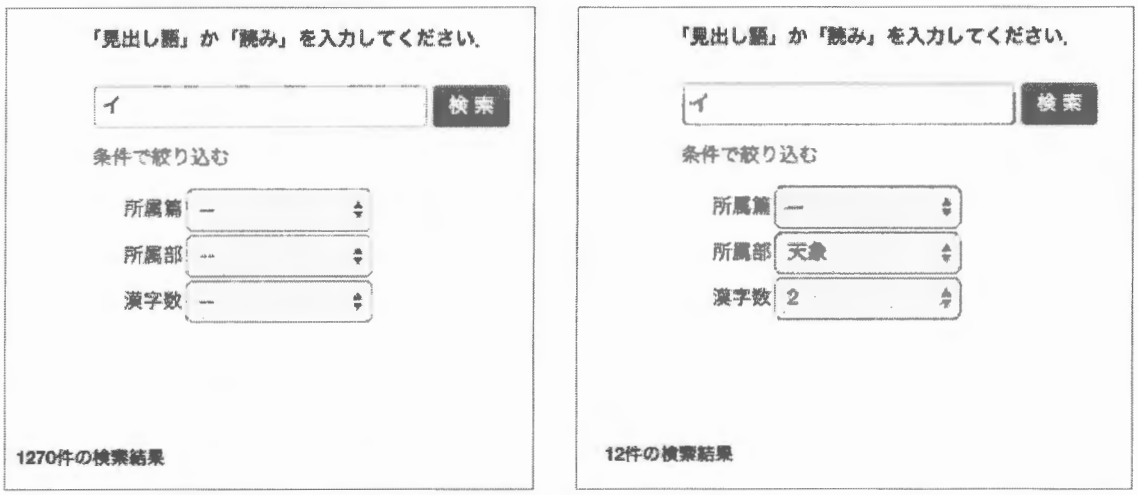

図 12　検索結果に対するフィルタ

[6] 今昔文字鏡 http://www.mojikyo.com/

あるが，今後，利用方法(利用可能性)についての分析を行い，利用者側の利便性に注目した設計の方針も明らかにしていきたいと考えている．

さらに，『色葉字類抄』と同じく，漢字語とその情報を並べるタイプの古辞書(ここでは近代初期までのものを含めてそう呼ぶこととする)は，本データベースに類する仕様の検索システムで覆うことが可能となる．

図 13　明応五年本節用集

図 14　易林本節用集

図 15　合類節用集

例えば図 13～15 のような，室町・江戸期の節用集類などは，『色葉字類抄』をより単純化した構造を持っており，本システムの流用が充分可能であると考えられる．母体となるデータの作成や入手を以て，検索可能な辞書や語を増やすことは，本データベースを拡大していく第一の方向となるであろうが，複数の辞書を横断検索する[7]システム構築については，特に表示方法の面において今後の課題と言えよう．

参考

[口頭発表]藤本灯「色葉字類抄データベースの構築と展望」(訓点語学会第 113 回研究発表会，於東京大学山上会館，2015 年 11 月 8 日)

7 国語辞書に収録される語が時代によって変化することは現在と同様であり，平安期から明治期にいたる 1000 年間の，辞書に登録された語の歴史すなわち「語史」を一目で追えるようになれば，データベースの利用可能性も格段に拡がるであろう．