色葉字類抄データベースの構築と展望

国立国語研究所論集 (NINJAL Research Papers) 11: 1–9 (2016)
ISSN: 2186-134X print/2186-1358 online

色葉字類抄データベースの構築と展望

藤本　灯

国立国語研究所 理論・構造研究系

要旨

本稿は，三巻本『色葉字類抄』収録語彙データベース（http://jiruisho.l.u-tokyo.ac.jp/）の部分公開開始に伴い，その目的と構築過程，延いては古辞書データベース構築における課題と展望を示すものである。内容は，「古辞書データベース作成上の問題点」「本データベースの基本方針」「本データベースのデータ項目」「課題」「展望」から成る。指針の概要については以下の通りである。

【全体の指針】利用者は，いずれの見出し語についても，少なくとも漢字・字音・和訓のいずれかより検索可能とする。また，新字・現代仮名遣いでの検索を可能とする。

【指針① 漢字の処理】新字での検索に対応する。検索結果には原本に近い字形を持つ漢字を表示させるが，厳密に再現することはしない。

【指針② 仮名の処理】原本の仮名遣い・現代仮名遣いのいずれによっても検索可能とする。また和語については歴史的仮名遣いでも検索可能とする。

【指針③ 画像と全文テキスト】底本使用の権利上，画像と全体翻字テキストの提供・公開は見送り，当面は検索システムを介して個々の語を表示させることとする。

本データベース作成の契機は，『色葉字類抄』中の漢字の検索方法に困難があったことにあるが，これが完成すればその点が克服され，また古代語研究の効率化が期待される *。

キーワード：色葉字類抄，データベース，日本の古辞書

1. はじめに

『色葉字類抄』は，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，日本史学の分野において，本書は，古代日本語を読み解くための必携の古辞書[1]の一つとして活用されてきた。しかし，『色葉字類抄』に収録された語を調べることは，必ずしも容易ではなかった面[2]がある。こうした背景から，筆者は，本辞書を対象とした検索システム「三巻本『色葉字類抄』収録語彙データベース」を作成することとした。

本稿では，本システムの部分公開（http://jiruisho.l.u-tokyo.ac.jp/）開始に伴い，その目的と構築過程，延いては古辞書データベース構築における課題と展望を示すこととする。

* 本稿は，訓点語学会第 113 回研究発表会での発表内容に加筆修正を加えたものである。また，本データベース並びに研究は，平成 26-27 年度科学研究費補助金（研究課題番号 26884013，研究活動スタート支援，『色葉字類抄』を中心とする国語辞書史研究，代表者：藤本灯）の成果の一部である。なお，本データベースのシステム構築は，志村誠氏・津村昌祐氏に，本データベースのデータ作成は北崎勇帆氏に支援を受けた。

[1] 本稿で述べるところの「古辞書」の範囲は，前近代に成立したものとする。

[2] 島田友啓氏の漢字索引（1966 ～ 1970）が存在し，有用であるが，手書きによって文字が特定し難く，また大部であり，更に検索方法や所在注記が初心者には分かり辛いという難点があった。

2. 古辞書データベース作成上の問題点

従来，古文献・古字書データベース作成の難しさについては，特に文字コードの面から種々に述べられ，解決策が模索されてきた（當山日出夫氏，池田証寿氏らに御論考がある，當山 1987・1995・1996・2013，池田 1999 〜 2002，池田氏のウェブサイト[3] 等）。漢字や仮名の字体を如何に処理するかは，字書（本稿では漢字字書系の文献を指す）・辞書（本稿では国語辞書系の文献を指す）に限らず，日本語文献のデータベースを構築する上で，必ず方針を定めなければならない点である。その他，原本の画像を付すかとか，プレーンテキストを提供するかとかいった点も，そのデータベースを特徴付ける点となり得る。これらの方針は，データベース構築者にとっての要不要のみならず，データベースの規模や利用者の便，サーバの性能等，様々な視点と条件との摺り合わせによって決定されていくことが普通であろうが，無論，最終的な判断は構築者に委ねられる。

いま改めて古字書・古辞書データベース化にあたって障壁となる点，方針策定にあたっての困難さを挙げれば，次のように集約出来ようか。

①【漢字の処理にかかる難しさ】
  a. 漢字の特定
  b. 字形・字体の再現（誤字の処理を含む）
  c. 検索条件の設定
②【仮名の処理にかかる難しさ】
  a. 字体や字母の再現（万葉仮名の処理を含む）
  b. 仮名遣いの揺れや乱れ，欠落部分の補訂
  c. 検索条件の設定
③【原本画像／プレーンテキストデータの提供にかかる難しさ】
④（①②③に関連して）【提供する情報の種類や量の設定にかかる難しさ】

例えば①b に関して，原本に「己・已・巳」というような字形の差や「斎・齊」等の揺れがある場合，これらの情報を残しつつ利用者の便に沿うようにデータベースを構築するためには，少なくとも

「原本にある字形」
「当時の然るべき字体」
「現行の字体」

の 3 種のデータが必要となる。熟語となればその乗算分の情報が必要となるだろう。しかるに，「原本にある字形」は，たとえ近似の字体を準備しても利用者の環境で検索文字として使用出来なければ意味がないし（検索結果に表示させるためだけの目的でその字形を用意することは，原本や影印の閲覧が一般に困難な場合を除いては，コスト過多で現実的ではない），後二者はデー

[3] 池田証寿氏「漢字字書データベースの作成とその利用」
http://rose.hucc.hokudai.ac.jp/~o16404/shikeda/jallc15p.html（1994 年公開。2016 年 1 月 31 日参照）

タ作成者の（多くの場合，高度な）解釈が多分に関わるために，常にこの 3 種を揃えることは容易な作業ではないと言える。

①ｃは①ｂにも関連するが，新字体や通行字体での検索を許容するか，完全一致や部分一致（前方一致等），熟語を構成する漢字数によるフィルタ（検索条件）を設けるか，外字の検索を如何に可能にするかといった問題が考えられる。

②ｂは，和訓や字音，その他の注記が仮名や漢字仮名交じりで表記されている場合を想定したものである。漢字と同様に，

「原本の仮名遣い」

「歴史的仮名遣い」

「現代仮名遣い」

の 3 種の情報並びにそれぞれの「字種（万葉仮名／片仮名／平仮名）」「字体／字母」について，処理の方針を定める必要がある。

③は，原本の所蔵機関の画像公開方法やリンク・翻字公開の可否，データの容量等に大きく左右される問題ではあるが，画像とテキスト，検索システムの三者全てがオープンアクセスにて提供されることが利用者にとって理想であることは言うまでもない。

④は①～③に挙げた困難さを益々複雑にする要因である。古辞書に慣れない者は，原本に書いてある情報が全て翻字された上で，当該辞書特有の性格や使い方，符号や注記の解説も付してあることを願うであろうし，中級者以上が，最低限，語の有無のみが分かれば良いと考えるのも自然である。要は，④は利用者層の設定に関わる問題ということになろうが，更に，原本の再現性・データの正確性・提供される情報量，検索方法等において，構築者の理想と利用者の需要とが一致することは――どこにターゲットを設定したとしても――まずないと言ってよい。

そもそも古字書・古辞書が，字や語をある目的のもとに集め，更にその多くについては一定の規則・基準のもとに排列した書物である以上，他の典籍に比して，データベース化の需要が劣る点は事実である（すなわち原則として，本来の用途と異なる検索を可能にするために，データベースの需要があると言える）。その一方で，現代辞書のように凡例の完備されない古字書・古辞書の読み解き方は，前述の如く初心者にとっては分明でない場合が少なくない。そうなれば，これを構築するのは，適性という点においても熱意という点においても自然と古辞書研究者自身かその周辺分野の研究者であることとなり，大部の辞書をデータベース化することには相当の労力と時間を費やすこととなる。もし多くの時間と私財を費やしてデータベースを作成したとすれば，それを躊躇なく無償で公開することも心情的に容易ではないかもしれない。また既にデータを所有し，公開の意志があったとしても，継続してデータベースを管理し続けられる環境にあるかどうか。このような点も現実的には問題点と成り得るであろうが，今は別次元の問題として置いておく。次節では，実際に筆者が公開を開始した，三巻本『色葉字類抄』収録語彙データベースについて，上に示したような問題点と照らし合わせながら，方針と現状について述べることとする。

3. 本データベースの基本方針

前節で述べた①～③の問題点につき，三巻本『色葉字類抄』のデータベース化にあたっては，当面，次のような指針を設けて進めることとした。

【全体の指針】いずれの見出し語についても，少なくとも漢字・字音・和訓のいずれかより検索可能とする。また，新字・現代仮名遣いでの検索を可能とする。

＊例えば漢字での検索が難しく，かつ原本に読み仮名のないような語については，便宜的な読み仮名で検索可能となるようにした。

【指針①　漢字の処理】新字での検索に対応する。検索結果には原本に近い字形を持つ漢字を表示させるが，厳密に再現することはしない。

〈見出し漢字の検索〉新字で検索可能とした。

＊『色葉字類抄』の原表記が旧字に近いものであれば，新旧両字で検索可能となる。（図1）

〈見出し漢字の表示〉見出し語の漢字の字体は，UTF-8 の範囲内で，近似の字形を持つ通行字に置き換え，旧字・異体字には，新字も添えて見出し語欄に表示した。UTF-8 で表示不可能なものについては「A, B, C…」の如く表示し，別途設けた「A, B, C…」欄に，各字に近似の字体を「今昔文字鏡」フォントや解字情報により表示した。（図 2）

＊ 1 つの見出し語とその注記を 1 レコード（=1 件のデータ）とした場合，レコード毎に必要に応じて「A, B, C…」と振り直した。

榮(栄)

音読み	-
訓読み	(イハフ)
注文	-
声点	-
所属篇	イ
所属部	辞字
前田本所在	上10オ-3
黒川本所在	上8オ-7

図 1　旧字（原本の字形に近い字）・新字

A(爨)

音読み	-
訓読み	イヒカシク
注文	-
声点	-
所属篇	イ
所属部	飲食
前田本所在	上8ウ-1
黒川本所在	-
A	Bの灬が火
B	爨

図 2　字体表示方法

【指針②　仮名の処理】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）・現代仮名遣いのいずれによっても検索可能とする（図 3）。また平仮名・片仮名，促音・（合）拗音の大小，清音・濁音はいずれによっても検索可能とする（図 4）。

窟

音読み	コツ
訓読み	イハヤ
注文	石ー
声点	入
所属篇	イ
所属部	地儀
前田本所在	上3オ-1
黒川本所在	上2ウ-4

図 3 「いわや」で検索

牽牛

音読み	ケンキウ
訓読み	イヌカヒホシ／又ヒコホシ
注文	-
声点	平平
所属篇	イ
所属部	天象
前田本所在	上2オ-4
黒川本所在	上2オ-2

図 4 「けんぎゅう」「ケンキユウ」で検索

【指針③　画像と全文テキスト】底本使用の権利上，画像と全体翻字テキストの提供・公開は見送り，当面は検索システムを介して個々の語を表示させることとする。

4. 本データベースのデータ項目

本書の語彙掲出形式例を図 5 に，本データベースの構築に使用した入力用テーブルを表 1 に示

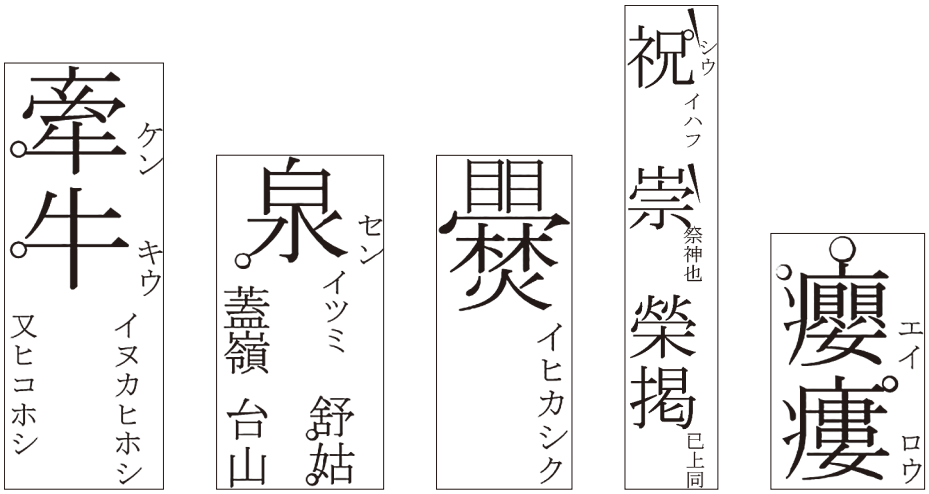

図 5 『色葉字類抄』語彙の原本における掲出形式

す。なおシステムについては別稿に論ずる予定である[4]。

表 1　入力用テーブル[5]

本文の情報	見出し語	1	見出し語	**牽牛**	**泉**	**A（爨）**	**癭痩**
	見出し語の漢字の情報	2	異体字 A			B の灬が火	
		3	異体字 A_ 漢字番号				
		4	異体字 B				
		5	異体字 B_ 漢字番号			73695	
		6	異体字 C				
		7	異体字 C_ 漢字番号				
		8	声点	**平平**	**平**	**ー**	**上去**
		9	文字数	2	1	1	2
	字音・和訓の情報	10	音訓	訓	訓	訓	音
		11	音	**ケンキウ**	**セン**	**ー**	**エイロウ**
		12	音完全一致	ケンキウ	セン		エイロウ
		13	音（現代仮名遣い）	ケンギュウ			
		14	訓	**イヌカヒホシ／又ヒコホシ**	**イツミ**	**イヒカシク**	**ー**
		15	訓完全一致	イヌカヒホシ／ヒコホシ	イツミ	イヒカシク	
		16	訓（現代仮名遣い）	イヌカイボシ	イズミ	イイカシグ	
	注記の情報	17	注文		**舒姑【平平】／蓋嶺台山**		
所在		18	所属篇	**イ**	**イ**	**イ**	**口**
		19	所属部	**天象**	**地儀**	**飲食**	**人体**
		20	前田本所在	**上 2 オ -4**	**上 2 ウ -3**	**上 8 ウ -1**	**上 17 ウ -5**
		21	黒川本所在	**上 2 オ -2**	**上 2 オ -8**	**ー**	**上 14 オ -6**
作成者注		22	作成者注				**口篇（ママ）**

■表 1 解説

1　本文に大字で示される見出し語。漢字 1 ～ 6 字による。漢字の表示は，「α」ないし「α（β）」の形式で示した。後者は，α が旧字や非通行字を含む場合や，原本の字体に錯誤がある場合に，新字や然るべき字を（β）として示したものであり，いずれも検索対象となる。よって結果的に，「榮（栄）」等の如く，新旧両字で検索可能となる語句が存在することとなる。

2・4・6　解字情報。字形が複雑な場合，3・5・7 欄を優先して入力してある。

3・5・7　今昔文字鏡番号。検索結果画面では，対応する（A, B, C 等の）欄に，今昔文字鏡フォントによる字形を表示する。

8　見出し語に付された声点。二字熟語の上字にのみ加点される場合は「平－」等と表示する。

[4] システム構築につき，「第 21 回公開シンポジウム 人文科学とデータベース」（2016 年 2 月 27 日，於同志社大学寒梅館）にて口頭発表（藤本灯・志村誠・津村昌祐・北﨑勇帆「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―」）を行った。

[5] 1 ～ 22 の番号およびそれより左欄は，今便宜的に設けた欄である。また，実際には MySQL を用いたデータベース上で管理を行っているが，今は表示の都合上，列と行や排列順を入れ替え，管理用の各種通し番号等の欄は省略してある。なお，太枠で示したゴシック体の部分は，検索結果画面に表示される情報である。

9　見出し語の漢字数。検索条件に用いる。

10　該当語が音読み・訓読みのいずれで掲出されたかを示す（音読み・訓読みは必ずしも字音語と和語に対応せず，漢字の音訓いずれを用いた語であるかを便宜的に入力してある。11 〜16 欄もこれに準じる）。

11　原本に示された音読みとそれに準じる情報（12 参照）。10 欄が「音」でありながら原本に音読みが示されない場合，想定される音読みを（　）に括って示した。また音読みが「同」等とある場合,「同（　　）」「（　　）」として想定される音読みを示した。なお, 片仮名／平仮名，清音／濁音，ア行／ヤ行／促音ツの大小字についてはいずれでも検索可能とした（ワ行小字は不可）。

12　本文の情報から「又・俗・已上・同・【(声点)】」等の注記を除き，音読みの情報のみとした欄。将来的に検索の絞り込み用キーとするか。

13　12 欄の音読みを現代仮名遣いに改めたもの。ただし現在，いわゆる字音仮名遣いへの訂正等は行っていない。また現行の一般的なあり方に従って適宜濁点を付与してある（前述の如く, 検索の際は清濁いずれによっても可）。12 欄が現代仮名遣いと相違ない場合は空欄とする。

14・15　原本に示された訓読みについての情報。11・12 欄に準じる。

16　15 欄の訓読みを現代仮名遣いに改めたもの。13 欄に準じる。

17　本文にある「見出し語・声点・見出し語を示す音訓」以外の情報全て。注文に付された片仮名は〈　〉で，声点は【　】で示した。

18　当該語の所属する篇（イ・ロ・ハ〜ヱ・ヒ・モ・セ・スの 47 篇のうち）。検索条件に用いる。

19　当該語の所属する部（天象〜名字の 21 部のうち）。検索条件に用いる。

20・21　前田本・黒川本における所在。巻（上中下）・丁数表裏・行数による。

22　本文の誤脱等に関する，作成者（筆者）による注記。現在，合点の有無も本欄に記述。

5. 課題

本データベースは，現在，日本語・日本文学研究者一般向けに構築中である。そのため，可能な限りシンプルな検索様式を模索し，一旦，上記のような仕様に落ち着いたところである。その方針内において，今後の課題となり得るのは以下の点である。

・漢字情報の充実…大漢和辞典コードの付与や，原本画像の表示。
・音訓での絞り込みに対応…「入力用テーブル」10 欄を用いて，検索キーに指定することが可能であるが，音訓の区別が困難な語や，音訓混ぜ読みの語の扱いに問題が残る。
・同訓異字中の掲出順の付与…『色葉字類抄』特有の事情として，同訓異字の掲出順がその語の表記の当時の一般性とある程度比例するという特性があるため，何らかの形で表示されれば研究上有用であるが, 例えば「3/13」（同訓異字が 13 字並ぶ中で上から数えて 3 番目に位置する）と表示することが妥当であるとしても，専門性が高く，一見して意味するところが分かり辛いという難点がある。

・声調や所在での絞り込み機能の付加。
・注文内文字列の検索対象化。

◆機能拡張例（□部は現在使用可能のもの・下線部は現在使用不可のもの）

・検索名目キー　【部分一致／前後一致／完全一致／文字コード／注文／所在】
・絞り込みキー　【所属篇／所属部／文字数／音訓／声調／合点有無】
・検索結果表示画面　【一語ごとの結果／画像】
・他　【全文テキスト（黒川本を底本とする）の提供】

6. 展望

さて，本稿では，筆者が構築中の三巻本『色葉字類抄』収録語彙データベースの現状について述べた。

そもそも『色葉字類抄』が，現在では漢字の読み方や語の収録の有無を知るために用いられることが多いにもかかわらず，漢字の検索方法に困難があったことが本データベース作成の契機であるが，本データベースが完成すれば，第一にはその点が克服されること，またそれによる，古代語研究の効率化が期待される。データベース拡張の方向性としては，(「いろは字類抄」諸本データベースとして，『世俗字類抄』等の）異本情報の付与，(「国語辞書」通時データベースとして）異なる時代の辞書情報の追加等が考えられ，現段階では未定であるが，将来的には，古辞書の大型データベース化に寄与すべく，他の研究者との連携を試みたいと考えている。

なお，本データベースは個人ベースで運営しており，今後もシステムの改善には意欲的に取り組む予定である[6]。

参照文献

池田証寿（編）(1999 ～ 2002)『古辞書と JIS 漢字　第 1 ～ 5 号』私家版.
島田友啓（1966 ～ 1970）『色葉字類抄漢字索引』私家版.
當山日出夫（1987）「コンピュータではあつかえない漢字―「和漢朗詠集」の場合―」『汲古』12: 87–93.
當山日出夫（1995）「古典籍と JIS 漢字―テキストの本文校訂との関係において―」『公開シンポジウム　人文科学とデータベース　vol. 1』29–36.
當山日出夫（1996）「白氏文集の本文校訂と JIS 漢字―文字集合間の整合性を中心として―」『シンポジウム　人文科学における数量的分析』15–22.
當山日出夫 (2013)「古典籍と JIS 漢字についての再考察―何が変わったか, 変わらないでいるか―」高田智和・小助川貞次（編）『訓点資料の構造化記述　成果報告書』（国立国語研究所共同研究報告 12-08）23–34.

[6] 本稿で述べる本データベースの仕様は，平成 28 年 2 月末時点のものである。

Construction and Prospects of the *Iroha-Jiruishō* Database

FUJIMOTO Akari

Department of Linguistic Theory and Structure, NINJAL

Abstract

This paper presents the purpose, process, and difficulty of constructing the *Iroha-Jiruishō* Database (DB), which is now partly published, and describes the general prospects of the DB, which consists of words from old Japanese dictionaries. The following are the DB guidelines.

【Overall guidelines】 Users can search for words in the *Iroha-Jiruishō* DB with Chinese characters or kana. Both traditional and new forms are available, as are the traditional and modern kana.

【guideline i】 New character forms, which are invisible in the manuscript and search results, are available to use for searching. Further, the forms similar to them in the manuscript will be displayed in the search results.

【guideline ii】 Users can search for words using the syllabary spelling of the manuscript, modern kana, and, for native Japanese words, traditional kana.

【guideline iii】 In consideration of the owner of the manuscripts, the search results show individual words in the DB, instead of the whole transliteration, for the time being.

Key words: *Iroha-Jiruishō*, database, old Japanese dictionary