夢梅本『倭玉篇』全文テキストデータベースの構築

著者	高橋 大希, 劉 冠偉, 池田 証壽
雑誌名	言語資源活用ワークショップ発表論文集
巻	1
ページ	49-56
発行年	2017
URL	http://doi.org/10.15084/00001457

夢梅本『倭玉篇』全文テキストデータベースの構築

高橋 大希（北海道大学文学研究科修士課程）
劉 冠偉（北海道大学文学研究科博士課程）
池田 証壽（北海道大学文学研究科）

Construction of a Mubaibon Wagokuhen Full-text Database

Daiki Takahashi (Graduate School of Letters, Hokkaido University)
Guanwei Liu (Graduate School of Letters, Hokkaido University)
Shoju Ikeda(Graduate School of Letters, Hokkaido University)

要旨

本発表は,夢梅本『倭玉篇』の全文テキストデータベースの構築と,その利用について述べるものである。『倭玉篇』は中世に生まれ近世まで広く用いられた漢和字書である。多種の写刊本が現存し,それらについて研究が行われてきたが,その多くは部首配列や特定の部を対象にした部分的なものであった。そこで『倭玉篇』の,特に和訓に関する全体的な研究を目的として,慶長 10 年（1605)刊行の夢梅本『倭玉篇』の全文テキストデータベースを構築した。この字書は『大広益会玉篇』を中心とした中国辞書を編纂基盤としており,すでに構築されているデータを用いて効率的に入力作業を進めることができる。構築したデータベースは約 22,000 字を収録する掲出字テーブルと,約 24,000 の和訓を収録する和訓テーブルからなり,中世末期の字訓対応の資料としても価値がある。また,このデータは「平安時代漢字字書総合データベース」の和訓データの整備にも使用される予定である。

１．はじめに

『倭玉篇』は室町時代初期に成立したとされる部首分類体の字書である。鈴木（2014）によると,慶長年間までのものでも 50 種以上の写刊本が現存しており,現存最古は『延徳三年写本』（1491）である。『倭玉篇』は室町時代における漢字と訓の対応を示す資料として価値のあるものであり,いくつかの影印本や索引が出版されて活用されている。

『倭玉篇』に関する研究には,諸本の分類・系統に関する研究が多く,部首配列や少数の部をサンプルとした掲出字配列・和訓についての調査が行われてきた。一方で,それぞれの本の全巻を通した調査に基づいてその特徴を明らかにしようとした研究は少ない。この背景としては,『倭玉篇』諸本の掲出字数の多さ,文字同定の難しさから,全体的な調査を容易にするテキストデータの作成が困難であったということが考えられる。

日本古辞書のテキストデータベースとしては,すでに「平安時代漢字字書総合データベース（HDIC)」（URL : http://hdic.jp/）が一部公開されている。このデータベースは,日本字書の『篆隷万象名義』,『新撰字鏡』,図書寮本『類聚名義抄』,観智院本『類聚名義抄』と,中国字書の『玉篇』,『大広益会玉篇』,『大宋重修広韻』,『龍龕手鑑』からなる。池田（2014）が述べるように,このデータベースの構築に際しては,日本字書の解読を効率的かつ正確に行うために,関連する中国字書のデータを同時に構築し,それを参照するという方法をとっている。

以上のような状況を踏まえ,『倭玉篇』のうちの一本である夢梅本『倭玉篇』（以下,『夢梅本』と略す）の掲出字,注文のすべてをテキスト化した全文テキストデータベースを構築した。これは初の中世漢字字書のデータベースである。本データベースには,約 22,000 字の掲

出字と,約 24,000 項目の仮名書きの義注が収録されている。

『夢梅本』は慶長 10 年（1605）に刊行された,『倭玉篇』の中ではごく初期の版本である。書誌的な情報については,岡井（1933）,中田・北（1976）に詳しい。『夢梅本』の各部首内の掲出字は宋本『大広益会玉篇』を基盤としており,配列もほぼ同じであるため,HDIC でとられた方法と同様に『大広益会玉篇』のデータを用いて文字同定と入力を効率的に行うことができた。また,『大広益会玉篇』の収録字に日本漢字音と和訓を付したような『夢梅本』のデータは,他の『倭玉篇』諸本や,『大広益会玉篇』の影響を受けて成立した日本字書の電子テキスト化にも活用できる可能性がある。

本データベースは,辞書史研究のみならず,大量の漢字と訓の対応を示す言語資源として語彙史の研究にも資することを目的に構築された。『夢梅本』の国語資料としての価値は,その注文構造の特殊性にある。『倭玉篇』の多くは掲出字に対して仮名で音と和訓が付されるのみで,ほとんど漢字注を持たないのに対し,この『夢梅本』は掲出字のほとんどに漢字注が付いている。漢字注を付すことには,和訓の意味を分かりやすくすることや,当該の和訓が付されている根拠を明示するという目的があったとみられる。漢字注を持たない字書においては,漢字と和訓の対応関係は,掲出字とそれに付された和訓という構図でしかとらえることができないのに対し,ほとんどの掲出字に漢字注がある『夢梅本』では,ある和訓が掲出字自体に結びついた和訓なのか,漢字注に結びついた和訓なのかを分析することが可能となる。

本発表では,構築したデータベースの構造と入力方針を説明し,利用の一例として『大字典』和訓データベースとの比較結果を示す。また,今後のインターネット公開や,他の辞書データベースとの連携に向けた課題について述べる。

２．データベースの構造

底本には,中田・北（1976）の無窮会神習文庫蔵本の影印版を用いた。データを入力するソフトウェアには,Excel を使用した。

データベースの範囲は夢梅本の掲出字と注文のみで,目録,付録,刊記は含まない。

一般に字書データベースに求められる機能としては漢字検索と,仮名検索とがある。これに対応するため,（A）掲出字テーブルと（B）仮名注テーブルの二つを設けた。

（A）掲出字テーブル

掲出字テーブルは,掲出字一字を単位としたテーブルで,漢字検索に対応する。掲出字テーブルは①ID,②部首番号,③部首,④掲出字,⑤漢字注,⑥仮名音注,⑦仮名注,⑧備考の八つのフィールドから成る。

①掲出字 ID

掲出字の所在位置を表す ID を入力する。ID は「冊_頁_行_段」であらわす。

「冊」は,五分冊されているうちの所在を表す。数字の範囲は 1 から 5 である。

「頁」は,所在する頁が,第一冊から文字が印刷されている頁を全巻通して数えて何番目であるかを表す。この頁数は中田・北（1976）の影印本に書かれているものである。数の範囲は 001 から 594 である。

「行」は半丁七行のうち,右側から何行目にあたるかを表す。例えば,ある部首 A が一行目で終わり,一行空白があって次の行から部首 B が始まる場合,部首 B の字は三行目から始まっているとする。数の範囲は 1 から 7 である。

「段」は,ある行の中で,その字が上から何番目の掲出字であるかを表す。数の範囲は 1 から 9 で,10 字目,11 字目を表すためにそれぞれ a,b を用いる。

例：5 巻 445 ページ 4 行 11 段　→　5_445_4_b

②部首番号

当該の部首の出現順の番号を入力する。

③部首

部首名を入力する。

④掲出字

掲出字を入力する。

⑤漢字注

本文中に漢字で書かれた注と,項目の一部に付された振り仮名を入力する。「打也」「亦作赴」のように注文の意味でまとまった単位を項目と呼ぶこととする。同一のセル内に複数の項目が入る場合には「/」で区切りを示す。

⑥仮名音注

掲出字の周辺に置かれる片仮名で書かれた音注。

⑦仮名注

掲出字の下に置かれる片仮名で書かれた注文。漢字注と同様に,複数の項目の間は「/」で区切る。

⑧備考

本文と合わせて使用する上での注意点。

次の表 1 は掲出字テーブルの入力例である。上から順に,図 1～3 が対応する。引用の都合上,画像には本文の写しを使用した（以下も同様)。

表 1　掲出字テーブル入力例

①	②	③	④	⑤	⑥	⑦	⑧
1_033_6_3	4	言	討	治也	タウ	ヲサム/ウツ	
2_188_5_6	27	牛	犅	養牛羊也/今作劬	ス	ウシヒツシヲヤシナフ＊/クサ	訓点
5_548_6_2	165	血	衅	牲〈セイ〉血〈ケツ〉塗（ヌリテ）器ニ祭（マツル）也/亦作釁	キン	チヌル	訓点

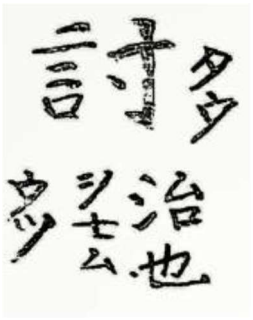

図 1　言部「討」

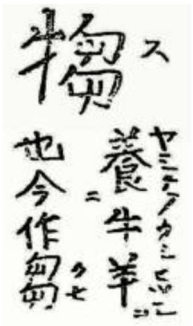

図 2　牛部「犅」

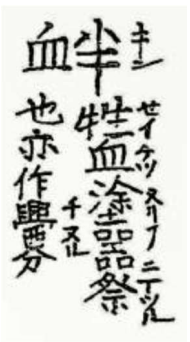

図 3　血部「衅」

（B）仮名注テーブル

仮名注テーブルは,仮名注の一項目を単位としたものであり,仮名検索に対応する。『夢梅本』では掲出字下に仮名書きされる注文には和語と漢語のどちらも含まれているため,語種を区別せずに仮名注と呼ぶこととする。

仮名注テーブルは①掲出字 ID,②仮名注 ID,③仮名注,④仮名修正,⑤掲出字,⑥漢字注,⑦部首,⑧部首番号,⑨備考の九つのフィールドから成る。

①掲出字 ID

掲出字の所在位置を表す ID を入力する。

②仮名注 ID

掲出字 ID の末尾に「出現順」を付したものを入力する。

「出現順」は仮名注を区別するために便宜的に付した値である。基本的には,掲出字下のスペースの中で,右上から右下,左上から左下に数えて何番目に出現するかを示している。

③仮名注

掲出字の下に置かれる片仮名で書かれた注文を入力する。

④仮名修正

夢梅本では,仮名遣いの乱れ,活用形の不統一,濁点の有無によって,仮名注をそのまま入力するだけでは検索に不便が生じる。そのため,仮名注の形式を修正したものを「仮名修正」に入力する。修正に際しては,影印本の索引,『日本国語大辞典』,『時代別国語辞典室町時代篇』を参考にした。また,踊り字と合字は開き,漢字は仮名に直した。和訓の検索や他の古辞書との連携にはこの列を用いる。同音異義語を区別するため,その語に対応する代表的な漢字を丸括弧に入れて末尾に付す。

⑤掲出字

掲出字を入力する。

⑥漢字注

本文中に漢字で書かれた注を入力する。

⑦部首

本文の各部首冒頭に掲げられた部首を入力する。

⑧部首番号

当該の部首の出現順の番号を入力する。

⑨備考

本文と合わせて使用する上での注意点を入力する。

次の表 2 は仮名注テーブルの入力例である。上から順に図 4～6 と対応する。

表 2　仮名注テーブル入力例

①	②	③	④	⑤	⑥	⑦	⑧	⑨
4_346_1_5	4_346_1_5_2	アニ	アニ（兄）	兄	昆也/男子先生爲一（兄）	兄	70	
5_535_6_3	5_535_6_3_1	アニ	アニ（豈）	豈	安也/焉也	喜	151	
3_290_4_1	3_290_4_1_3	サイワヒ	サイハヒ	履	皮曰一（履）/又踐也/禄也	尸	52	

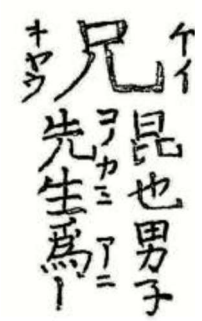

図 4　兄部「兄」

図 5　喜部「豈」

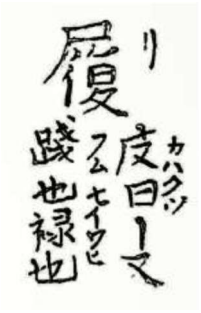

図 6　尸部「履」

３．入力方針

３．１　漢字

漢字は,Unicode の範囲内で可能な限り原本に近い字体を用いて符号化した。符号化できない字の場合,IDS(Ideographic Description Sequence)による方法で表す。たとえば,「⿰扌丁」の場合,「打」の字体を表す。IDS の部品となっている符号は「{}」で括る。IDS を使っても符号化が難しい文字に関しては対応する康熙字典体にギリシャ文字（β,γ）を付して表す。対応する康熙字典体もない場合は「■」で表す。

熟語や異体字が連続して掲出される場合は,そのそれぞれに注が付いているものとして考え,一字ずつに分ける。熟語や異体字が連続して掲出されている字には,「備考」にそれぞれ「熟字」「連字」と記す。

注文の中で掲出字を表す「一」は,その次に掲出字を丸括弧に入れて補って示す。

３．２　仮名

仮名は現在使われている字体に直す。

合字「メ（シテ）」「ㄱ（コト）」はそのまま合字を入力する。

二字連続の踊り字（くの字点）については「ヽヽ」で表す。

音読符「ー」は,漢字注のフィールドでは音読として扱うため角括弧で括って示す。仮名注のフィールドに入る場合は,括弧に入れずに示し,後ろに推定される漢字音を角括弧に入れて示す。例えば,次の図 7 の場合は部分的にしか仮名がついていないため,漢字注のフィールドになり,「ト〈ー〉メ問フヲ吉凶ヲ曰フー（敷）ト」となる。図 8 の場合は,仮名注のフィールドに入れるため,「カイコー<クワ>スル」となる。

図 7　又部「敷」

図 8　卵部「孵」

３．３　記号類

欠損などで判別不可能な字については「□」で表し,推測ができるものについては「□(ル)」のように丸括弧の中に補って示す。

明らかな誤字は,原文通りの表記の後に正しい文字を丸括弧に入れて表す。誤字がある場合,「備考」に「誤字」と記す。

注文の意味でまとまった単位を項目と呼ぶこととする。漢字注であれば「打也」「亦作赴」,仮名注であれば「アタヽカナリ」「ウツコトハナハタシ」などが項目にあたる。同一のセル内に複数の項目が入る場合には「/」で区切りを示す。

異本注記が禾部「秅」の注文に一箇所見られた。漢字注の「四百秉爲一」の中の「秉」に対して「束イ」と書かれていたため,この注記を鉤括弧に入れて「四百秉「束イ」爲一（秅）ト」のように表した。

４　符号化率

本データベースの掲出字符号化率は次の表 3 の通り 99.23%であった。符号化できなかった掲出字は 176 字で全体の 0.76%であった。この中には IDS 方式で表したものと,符号化できずに「■」を入力した字が含まれている。

表 3 掲出字符号化率

所属	字数	パーセント
CJK	13,079	57.75%
拡張 A	3,157	13.94%
拡張 B	6,186	27.32%
拡張 C	1	0.00%
互換漢字	47	0.21%
総計	22,423	99.23%

同じように,注文内の文字についての符号化率を示したのが次の表 4,5 である。符号化できなかった文字は異なりで 112 字（0.14%）,延べで 112 字（1.50%）であった。

表 4　注文異なり符号化率

所属	字数	パーセント
CJK	6,058	81.38%
拡張 A	501	6.73%
拡張 B	754	10.13%
拡張 C	1	0.01%
拡張 D	2	0.03%
互換漢字	16	0.21%
総計	7,332	98.51%

表 5　注文延べ符号化率

所属	字数	パーセント
CJK	77,863	97.72%
拡張 A	612	0.77%
拡張 B	853	1.07%
拡張 C	2	0.00%
拡張 D	2	0.00%
互換漢字	232	0.29
総計	79,564	99.86%

いずれにおいてもほとんどの文字の符号化ができており,十分に使用できる水準であると考えられる。

５.『大字典』和訓データベースとの比較

他のデータベースと連携させた利用の一例として,本データベースと同じように和訓データを持つ『大字典』和訓データベースとの比較を行った。

『大字典』（初版 1917 年）は国語学者の上田萬年によって編纂された漢和字書である。約 18,000 の掲出字を『康熙字典』と同じ部首画数順に配列している。重要視する和訓は太字で示され,品詞情報が付されている。

『大字典』和訓データベースは,『大字典』の重要和訓と品詞情報を収録するデータベースである。重要和訓が付された約 6,100 字の掲出字を収める和訓付き掲出字テキストテーブルと,約 10,500 個の重要和訓と品詞情報を収める和訓テキストテーブルからなる。
比較を行ったのは掲出字と和訓の二項目である。

まず両データベースの掲出字について共通字数を調べた。掲出字次の表 6 の「大字典」の列の値は『大字典』和訓データベースに収録されている掲出字数,「夢梅本」の列の値は夢梅本の掲出字数,「共通字数」は両本に共通して掲出されている掲出字の数を表す。共通字数は 5,388 字あり,これは『大字典』和訓データベースに収録されている掲出字のおよそ 88%にあたる。

表 6　掲出字比較

大字典	夢梅本	共通字数
6,106	22,646	5,388

不一致となったものの中には,「昂」と「昂」,「内」と「内」のようにそれぞれ微妙に異なる形で符号化しているものが含まれていた。このような符号化方針の違うデータベース間での掲出字の比較方法は今後の課題となる。

次に,『大字典』和訓データベースの和訓と,夢梅本『倭玉篇』全文テキストデータベースの仮名注で,同じ掲出字に同じ和訓がついている項目の数を調べた。夢梅本『倭玉篇』全文

テキストデータベース側で使用したのは「仮名修正」のデータである。結果が次の表 7 である。

表 7　和訓比較

大字典	夢梅本	一致数
10,519	24,446	3,111

一致する項目の数は 3,111 項目であり,これは『大字典』の和訓の約 3 割にあたる。これらの漢字と訓の対応は,比較的安定性の高いものとみることができる。

６．おわりに

本発表では,夢梅本『倭玉篇』全文テキストデータベースの構築について述べ,その利用例を示した。

最後に今後の展望と課題について述べる。

本データベースは掲出字に仮名書きの字音,和訓が付いたデータを収録しているため,同様の形式を持った他の辞書のデータ構築にも利用できる。本データベースの構築にあたってデータを利用した HDIC は,日本字書の字音・字訓が未整備の状態であるが,本データベースを用いることでデータ整備の効率化が期待できる。

また,本データベースは,広く国語資源として使用できるようインターネットでの公開を計画している。公開にあたっての使用許諾,公開形式については今後の課題としたい。

謝辞

本研究におけるデータベース構築は,JSPS 科研費 16H03422 による成果の一部である。
また成果の公表に関しては,北海道大学大学院文学研究科「共生の人文学」プロジェクトの助成を受けた。

文献

池田証壽（2014）.「平安時代漢字字書総合データベースの構築」北海道大学文学研究科紀要 142 号,pp. 79-90.
岡井慎吾（1933）.『玉篇の研究』,東洋文庫.
鈴木功眞（2014）.「字鏡集と和玉篇の境界と継承について」『国語語彙史の研究三十三』pp.147-162,国語語彙史研究会編.
中田祝夫・北恭昭（1976）.『倭玉篇夢梅本篇目次第研究並びに総合索引』,勉誠社.

関連 URL

平安時代漢字字書総合データベース（HDIC）　http://hdic.jp/