『大字典』データベースをつくる

高田智和

北海道大学大学院

概要

本データベースは、『大字典』の漢字集合としての規模と質を検討するために作成され、『大漢和辞典』番号、JIS 漢字（JIS X 0208:1997）の区点番号、UCS（JIS X 0221:2001）の 16 進コード番号を付している。作成にあたって直面した重出字の処理や、包摂規準による掲出字入力の問題点を述べるとともに、漢字調査における 15000 字程度の漢字集合の有効性について考える。

Making of "Daijiten" database

Takada Tomokazu

Hokkaido Graduate School of Letters

abstract

This database is created in order to examine the scale and quality as a kanji character set of "Daijiten", and it attaches the "Daikanwajiten" number, the code number of JIS X 0208:1997, and the code number of JIS X 0221:2001. I describe the problem of Jushutsuji (the same characters arranged two radicals) and inputting by the unification rules. So I consider that the validity of about 15000 kanji characters when we investigate of kanji characters.

1 はじめに

明治以降、『和玉編』や『康熙字典』の流れをくむ辞典が数多く編まれ、親字に熟語を付し、字源解釈を取り入れるなどして、次第に現在の漢和辞典の形態に近づくに至った。大正年間には、『詳解漢和大字典』（服部宇之吉・小柳司気太著、1916）、『大字典』（上田万年・岡田正之・飯島忠夫・栄田猛猪・飯田伝一編、1917）、『字源』（簡野道明著、1923）など、著名な漢和辞典を輩出した。特に『大字典』は版を重ね、世に広く行われたことは、金田一京助博士による普及版の序文からも察せられる。また、彌吉光長「『大漢和辞典』を引く」（紀田順一郎編『『大漢和辞典』を読む』大修館書店、1986 所収）は、「規模において親字二万、熟語十万、内容においては漢音と国語の訓を一つにまとめ、親字に通し番号を付し、本文の欄外に見出し語を載せるなどの新機軸を編み出した」と、漢和辞典史上の『大字典』の意義を述べている。

築島裕『大般若経音義の研究索引編』（勉誠社、1983）所載の「漢字索引」は、『大字典』の配列順に掲出字を記し、このような『大字典』の国語学における利用も散見する。また、北海道大学文学部の石塚晴通教授の演習では、長年にわたって漢字字体の調査が行われ、画像データベースとして蓄積されており、文字の整理と配列は『大字典』によってなされている。

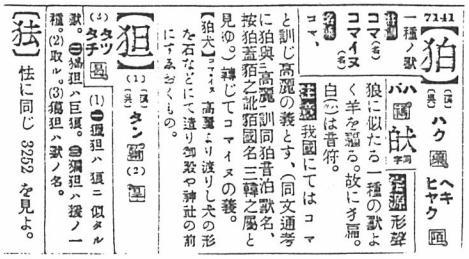
2　本データベース作成の目的

本データベース作成の目的は、およそ15000字にのぼる『大字典』掲出字の、漢字集合としての規模と質を吟味することにある。明治以降の漢字字典・漢和辞典を相互に比較し、文字種の異同と増減を明らかにすることは、JIS漢字やユニコードなどにつながる文字の標準化の系譜をたどるにほかならない。また、各書における掲出字の漢字字体の異同、あるいは、異体字の扱いを捉えることは、近年に至るまでの異体字の体系を明らかにし、電子化テキストにおける異体字の処理・漢字シソーラスなどに寄与する面もあろう。さらに、電子化テキストを用いた漢字使用頻度調査によって、辞書掲載字と実際に用いられている漢字とを比較し、漢字使用の実態と辞書掲載字との関わり、漢和辞典の漢字集合としての有効性を実証的に解明することができるものと考える。そのための第一歩として、『大字典』のデータベース化を試みるのである。

3　『大字典』の掲出字

『大字典』の掲出字は、図1に示すように、次の4通りがある。

図1『大字典』掲出字（1）



1.【 】で括られた番号つき掲出字
2.【 】で括られた番号なし掲出字
3.〔 〕で括られた番号なし掲出字
4.【 】で括られた掲出字の項目に、「略字」「俗字」などの字体注記を付してあげられた掲出字

将来的には、これら4通りの掲出字すべてを入力して完成版とするつもりであるが、現在は、【 】で括られた番号つき掲出字と【 】で括られた番号なし掲出字の入力までが完了している。

また、『大字典』の掲出字は、掲出字に対する記述の内容から大きく2通りにわけられる。

図2『大字典』掲出字（2）

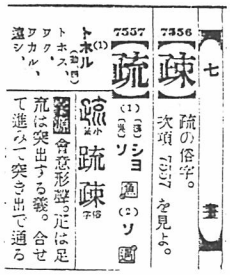


図2のように、掲出字に対して音義などを記述しているものと、参照情報のみをあげているものとがある。本稿では便宜的に、前者を「正字」、後者を「異体字」と呼ぶことにする。

なお、底本とした『大字典』は、1966年9月1日発行の普及版第60版（講談社）である。

4　データ項目

『大字典』データベースは次の 9 つのフィールドからなる。

1. 文字番号
   『大字典』の文字番号をそのまま利用する。番号なし掲出字については、直前にある掲出字の文字番号の末尾に x をつけて表す。番号なし掲出字が連続する場合は、出現順に「…x1」「…x2」とする。
2. 掲出字
   JIS 漢字（JIS X 0208:1997）包摂字体を入力する。
3. 部首番号
   『大字典』の部首番号を示す。1 から 214 まであり、配列は『康熙字典』に同じ。
4. 部首内画数
   掲出字の『大字典』による部首内画数を示す。
5. 参照文字番号
   掲出字が「異体字」である場合、参照先の文字番号を示す。
6. 大漢和辞典番号
   諸橋轍次著『大漢和辞典』（修訂版、1984-86、大修館書店）の検字番号を示す。
7. JIS X 0208 の区点番号
   JIS 漢字（JIS X 0208:1997）第 1 水準及び第 2 水準の区点番号を示す。
8. UCS の 16 進コード番号
   UCS（JIS X 0221:2001）の CJK 統合漢字の 16 進コード番号を示す。
9. 備考
   重出情報や、包摂を行った場合に包摂規準の連番などを記述する。

データベースの一部を以下に示す。

729, 刀,18,0,,1845,37-65,5200,
730,,18,0,729,1847,,5202,
731,,18,0,,1846,,5201,
732, 刃,18,1,,1848,31-47,5203, 包摂 170
733,,18,2,,1852,,5205,
734, 分,18,2,,1853,42-12,5206,
735, 切,18,2,,1858,32-58,5207,
736, 刈,18,2,,1859,20-02,5208,

5　『大字典』の文字種—作成上の諸問題と『大字典』の特色—

5.1　重出字

入力作業によって、【 】で括られた番号つき掲出字を 14923 字、【 】で括られた番号なし掲出字を 547 字、合計 15470 字を抽出した。

ところが『大字典』には、同一の文字が複数の部首に重複して採録されている重出字がある。例えば、「旦」は一部 4 画（21）と日部 1 画（4447）とに 2 字重出、「舎」は人部 6 画（307）、口部 5 画（1271）、舌部 2 画（9637）に 3 字重出となっている。重出字は 55 組 113 字（2 字重出が 52 組 104 字、3 字重出が 3 組 9 字）である。以下に全用例を示す。

「旦」21, 一部 4 画／ 4447, 日部 1 画　「両」31, 一部 5 画／ 631, 冂部 5 画

「厤」43, 一部 12 画／ 1087, 厂部 11 画　「𠂉」83, 丿部 2 画／ 700, 几部 1 画
「匆」89, 丿部 3 画／ 921, 勹部 2 画　「𠆢」400, 人部 8 画／ 9444, 肉部 6 画
「吳」604, 八部 5 画／ 1206, 口部 4 画　「釜」614, 八部 8 画／ 12384, 金部 2 画
「黃」615, 八部 9 画／ 14694, 黃部 0 画　「囘」627, 冂部 3 画／ 1604, 口部 2 画
「冄」633, 冂部 5 画／ 2817, 巾部 5 画　「冑」635, 冂部 7 画／ 9387x, 肉部 5 画
「尽」668, 𠂇部 4 画／ 2475, 尸部 3 画　「憑」695, 𠂇部 13 画／ 3483, 心部 11 画
「𠙵」719, 凵部 2 画／ 2519, 屮部 1 画　「鹵」726, 凵部 6 画／ 8637, 米部 2 画
「齒」728, 凵部 8 画／ 5668, 止部 11 画　「飡」943, 匕部 8 画／ 13504, 食部 0 画
「衷」1011, 十部 7 画／ 10653, 衣部 4 画　「亶」1014, 十部 8 画／ 1995, 大部 7 画
「肅」1017, 十部 10 画／ 8661, 米部 5 画　「翰」1018, 十部 15 画／ 9237, 羽部 11 画
「厠」1064, 厂部 6 画／ 4475 日部 4 画　「对」1120, 又部 3 画／ 2414, 寸部 2 画
「叫」1226, 口部 4 画／ 4341, 斗部 3 画　「虽」1308, 口部 6 画／ 10300, 虫部 3 画
「埜」1804, 土部画／ 7532, 田部 7 画　「墱」1805, 土部 9 画／ 1911, 士部 9 画
「梦」1941, 夕部 8 画／ 4991, 木部 7 画　「𠆢」1978, 大部 5 画／ 3104, 彡部 5 画
「爽」2003, 大部 8 画／ 7041, 爻部 7 画　「娄」2107, 女部 6 画／ 8646x, 米部 3 画
「斃」2296, 子部 17 画／ 4315, 子部 16 画　「冗」2303, 冖部 2 画／ 8292, 穴部 0 画
「对」2416, 寸部 3 画／ 4324, 文部 3 画　「爭」2417, 寸部 4 画／ 7025x2, 爪部 3 画
「尉」2432, 寸部 10 画／ 6878, 火部 8 画　「常」2450, 小部 7 画／ 2840, 巾部 7 画
「𤔔」2520, 屮部 2 画／ 7462, 生部 1 画　「歸」2858, 巾部 9 画／ 5660, 止部 8 画
「応」2929, 广部 4 画／ 3198, 心部 3 画　「鹰」2970, 广部 10 画／ 6945, 火部 10 画
「鼻」3028, 廾部 11 画／ 14810, 鼻部 0 画　「擧」4116, 手部 14 画／ 9631, 臼部 11 画
「舉」4117, 手部 14 画／ 9632, 臼部 11 画　「攵」4223, 攴部 0 画／ 4322, 文部 0 画
「變」4319, 攴部 19 画／ 11215, 言部 16 画　「者」4494, 日部 4 画／ 9254, 老部 4 画
「爨」6944, 火部 10 画／ 14711, 黒部 0 画　「羴」9193, 羊部 14 画／ 11194, 言部 13 画
「隲」12934, 阜部 14 画／ 13777, 馬部 7 画　「鱻」14354, 魚部 15 画／ 14922, 補遺
「奇」162, 一部 7 画／ 1307, 口部 6 画／ 8358, 立部 4 画
「𠂤」225, 人部 4 画／ 4457, 日部 2 画／ 4652, 曰部 2 画
「舍」307, 人部 6 画／ 1271, 口部 5 画／ 9637, 舌部 2 画

重出字の多くは、参照情報のみが記された「異体字」であるものが多い。そこでこれらの文字を『康熙字典』で検してみると、『康熙字典』に採録されているもの 27 文字、採録されていないもの 28 文字であった。『康熙字典』非掲載字を以下に示す。

両　厤　𠂉　匆　奇　舍　𠆢　吳　釜　黃　鹵　齒　亶　翰
厠　对　叫　虽　飡　爽　娄　对　常　応　鹰　鼻　者　隲

『大字典』の凡例に、「文字は康熙字典を準據としたれども、素より之を取捨選擇したる上に、更に通常世間に慣用せる俗字・略字・新製文字等を新聞・雜誌其他諸記録中より捃摭し、普く之を網羅せり。」とあることからも、これらの文字は『大字典』編者が採取した「俗字・略字・新製文字」の類であろう。重出となっているのは確たる典拠が乏しいために、部首の配属先が未だ定まっていないということが考えられる。『大漢和辞典』では、「𠂉、匆、奇」の 3 文字が重出字、「𠆢、亶、翰、常」の 4 文字が非掲載字である以外は、残りの 21 文字は部首配属先が一箇所になっている。

また、重出字には現在でもよく見る字、よく使う字が含まれていることから、検索の便宜を図る目的があったかと思われる。

さて入力作業では、重出字の一方を空見出しとし、備考欄に参照先を示す等の処置を行った。重出字を修正した結果、『大字典』掲出字の総数は、【 】で括られた番号つき掲出字 14868 字、【 】で括られた番号なし掲出字 544 字、合計 15412 字となった。

5.2　欠画字

『大字典』には『康熙字典』の欠画字（皇帝の諱を避けて最終画を落とした字形）をそのまま踏襲したと考えられるものが 17 字ある。

傛（453）　弦（3055）　慉（3440）　搐（3943）　昡（4501）　泫（6092）
滀（6426）　炫（6815）　珎（7260）　畜（7513）　痃（7596）　眩（7832）
稸（8255）　絃（8796）　舷（9673）　蓄（10035）　鉉（12429）

欠画字であることは明らかなので、欠画を補った字形を想定して扱うことにする。入力作業では、備考欄に「欠画字」と記した。

5.3　包摂規準による掲出字の入力

掲出字の入力は、JIS X 0208:1997 の包摂規準にしたがって行った。JIS X 0208:1997 には 185 の包摂規準と、過去の規格との互換性を維持するための包摂規準が 29 立てられている。

例えば、一部 2 画の「与（11）」は包摂規準連番 39（与｜与）によって包摂し、「与（45-31）」で表す。また、人部 7 画の「俠（344）」は過去の規格との互換性を維持するための包摂規準（侠｜俠）によって包摂し、「侠（22-02）」で表す。このような場合には、備考欄に包摂規準の連番、あるいは互換規準字である旨を記した。これは、豊島（1999）が説くような「暗黙の包摂」を極力排除するためである。

ところが、包摂規準にしたがって掲出字を符号化した結果、2 字を 1 区点で表すような事態に陥った。例えば、亠部 6 画の「享（156）」と亠部 7 画の「享（159）」とは、包摂規準連番 145 によって、同じ「享（21-93）」で符号化される。このようなものが 93 組 187 字（「即（1037）」「卽（1048）」「卽（1049）」は「即（34-08）」で 3 字包摂）ある。これらについては、『大字典』の方が包摂の範囲が狭く、JIS X 0208:1997 の方が包摂の範囲が広いということになろう。

では、UCS（JIS X 0221:2001）を用いた場合、JIS X 0208:1997 で 2 字包摂されたものがどのようになるか見てみよう。人部 6 画の「併（269）」と人部 8 画の「倂（362）」は、JIS X 0208:1997 を用いると、包摂規準連番 97（并｜幷）によって包摂されていたが、UCS を用いると、「併（269）」は「併（4F75）」、「倂（362）」は「倂（5002）」でそれぞれ表現可能となる。このように、2 字包摂を分離できるのは 41 組 82 字（「即（1037）」と「卽（1049）」は別コード、「卽（1048）」は「卽（1049）」に包摂か）である。

しかし、先にあげた「享（156）・享（159）」などは UCS を用いても分離することができない。「享（4EAB）」で包摂して符号化するか、「享（159）」を符号化不能とするかどちらかである。大字典が両字体を掲出字としていることを尊重するならば、「今昔文字鏡」を用いるほかあるまい。入力作業では包摂扱いとした。

さて、JIS 漢字で表現したい、すなわち、包摂したいが包摂規準のないものがある。

まず、人部 10 画の「舒（460）」、水部 5 画「沿（6057）」、臼部 11 画の「舉（9631）」、艸部 10 画の「蓆（10037）」、衣部 17 画の「襷（10811）」は、JIS X 0213:2000 で追加された包摂規準によって、それぞれ「舒（48-16、連番 194 適用）」「沿（17-72、連番 189 適用）」「擧（58-09、連番 188 適用）」「蓆（72-78、連番 197 適用）」「襷（75-07、連番 188 適用）」に包摂する。入力作業では、備考欄に JIS X 0213:2000 の追加包摂規準の連番を記した。

次に、以下にあげる 44 字は JIS 漢字に包摂する。

「亂（121）」―「亂（48-12）」　「儼（555）」―「儼（49-23）」
「剳（835）」―「箚（68-20）」　「勒（885）」―「勒（80-53）」
「嚴（1581）」―「嚴（51-78）」　「囃（1590）」―「囃（51-82）」

「巖（2757）」—「巌（54-62）」　「戚（3409）」—「慼（20-22）」
「慌（3444）」—「慌（25-18）」　「扞（3705）」—「扞（57-19）」
「換（3907）」—「換（20-25）」　「敢（4277）」—「敢（20-26）」
「曄（4619）」—「曄（59-01）」　「棔（5223）」—「棔（60-43）」
「槍（5237）」—「槍（33-68）」　「槍（5413）」—「槍（59-56）」
「歿（5692）」—「歿（61-39）」　「毒（5839）」—「毒（38-39）」
「涵（6223）」—「涵（62-30）」　「滄（6431）」—「滄（62-75）」
「滿（6476）」—「滿（62-64）」　「濟（6641）」—「濟（63-27）」
「燈（6971）」—「燈（37-85）」　「猾（7208）」—「猾（64-49）」
「獪（7231）」—「獪（64-54）」　「獵（7244）」—「獵（64-58）」
「疊（7552）」—「畳（65-40）」　「睥（7864）」—「睥（66-46）」
「紀（8735）」—「紀（21-10）」　「網（8885）」—「網（44-54）」
「蘇（9078）」—「蘇（69-91）」　「荒（9851）」—「荒（25-51）」
「蕚（9960）」—「萼（72-53）」　「虔（10269）」—「虔（73-42）」
「虞（10275）」—「虞（22-83）」　「誤（11020）」—「誤（24-77）」
「譫（11114）」—「譫（75-66）」　「譜（11174）」—「譜（41-72）」
「跋（11515）」—「跋（76-77）」　「雅（12962）」—「雅（18-77）」
「鬮（14034）」—「鬮（82-13）」　「鵄（14486）」—「鵄（83-11）」
「鷗（14533）」—「鴎（83-19）」　「鼬（14797）」—「鼬（83-76）」

これらは、文化庁文化部国語課（2000）の23種の活字総数見本帳に見える。包摂の根拠はこの点にあるが、中には包摂してしまうには抵抗が大きいものもあろう。ともあれ、『大字典』の掲出字は、明治以降の活字の実態を反映しているのである。入力作業では、備考欄に「望包摂」と記し、続けて文化庁文化部国語課（2000）の所在（ページ数のみ）を示した。

このような手続きを経て、JIS X 0208:1997で表現したものは6072字（39.4%）である。一方UCSでは13351字（86.6%）を処理できる。約6300字のJIS X 0208:1997の漢字集合は、約15000字の『大字典』の漢字集合に大部分が含まれるものであり、約27000字のUCSの漢字集合は、『大字典』の集合とおよそ2000字の隔たりがある。それでも、9割近くの掲出字を処理できることは、十分に魅力的である。

5.4　『大字典』シソーラス

ここでは『大字典』の「異体字」について述べる。【 】で括られた掲出字のうち、「正字」は12394字（80.4%）、「異体字」は3018字（19.6%）である。『大字典』掲出字の5分の1が「異体字」群である。「異体字」を参照先の「正字」に結び付け、複数字体をまとめると、複数字体の組合せが2294組できる。その一部を以下に示す。左側の1列目が「正字」、2列目以下が「異体字」である。

一（1）－弋（3033）
万（5）－卍（999）
丈（6）－丈（19）
三（7）－弍（3036）
丑（17）－丒（16）
丕（23）－丞（1000）
世（25）－丗（24）－丗（995）－丗（996）
丘（26）－丠（33）－坵（1702）
丸（68）－丮（67）
久（85）－乆（84）

「異体字」を『大漢和辞典』に検してみると、『大漢和辞典』に採録されているもの 2654 字、採録されていないもの 364 字であった。『大字典』掲出字全体で、『大漢和辞典』非掲載字は 384 字であるから、大部分を「異体字」が占めていることがわかる。これらの典拠を求めることで、『大字典』の異体字の体系を明らかにすることができようが、これは今後の課題としたい。

6　文字種の有効性

『大字典』15000 字の集合がどれだけ有効性を発揮するのか、最後に少しだけ検証してみようと思う。青空文庫の Web ページに「文学作品に現れた JIS X 0208 にない文字」(http://sumomo.sakura.ne.jp/~aozora/gaiji/gaiji0208/mokuji.html) が掲載されているので、これを用いて JIS 外字のカバー率を調べる。

「文学作品に現れた JIS X 0208 にない文字」は、1999 年 3 月末までの青空文庫収録作品と、CD-ROM 版の『明治の文豪』『大正の文豪』『新潮文庫の 100 冊』(いずれも新潮社) から JIS 外字を採取しており、近代・現代文学の主要な作品を扱っていると評価できる。JIS 外字の総数は 725 字である。

なお比較の対象として、掲出字およそ 10000 字の辞書である『角川新字源改訂版』(小川環樹・西田太一郎・赤塚忠編、1994、角川書店) と、掲出字およそ 50000 字の辞書である『大漢和辞典』とで、同様に JIS 外字のカバー率を求める。

『角川新字源』　421 (58.1%)
『大字典』　616 (85.0%)
『大漢和辞典』　690 (95.2%)

『大漢和辞典』のカバー率が突出しているのは当然の結果として、『大字典』もかなり健闘している。『大字典』に上乗せすること 35000 字で、カバー率の伸びは約 10%であり、対して『新字源』に上乗せすること 5000 字で、カバー率の伸びは約 20%となる。15000 字から 50000 字よりも、10000 字から 15000 字の方が密度が高い。

外字のカバー率が 8 割をこえる点は大いに評価でき、『大字典』15000 字の集合は、漢字調査の基礎母体として一定の有効性を持っていると判断できよう。

7　おわりに

「大正新脩大藏經テキストデータベース」(http://www.l.u-tokyo.ac.jp/~sat/kanji/index.html をはじめとして、『大漢和辞典』をコードブックとして利用する方法は広く行われているようである。「今昔文字鏡」を利用すると、『大漢和辞典』掲載字であれば Web 上での表示も可能となり、『大漢和辞典』を媒介とした漢字の使用はこれからも揺るぎないもののように思われる。

しかし、『大漢和辞典』に拠らないデータも少なからず存在するのである。『大字典』のデータベースは、『大字典』に依拠したデータを他とつなぎ、将来にわたって継承していくことに貢献できるだろう。『大漢和辞典』の検字番号と対照したのは、電子の海で孤島にならないための予防策でもある。

最後に本稿は、漢字調査の基礎母体として、15000 字程度の集合をひとつの目安として設定すること主張する。今後はこの集合の規模と質の検証をさらに推し進めていくことが課題である。

参考文献

石塚晴通 (1984)『圖書寮本日本書紀研究篇』汲古書院
田嶋一夫 (1984)「漢字シソーラスの構想と課題」『日本語学』3-3

當山日出夫（1999）「コンピュータの文字に対する意識について―錯綜するJIS漢字論の根底にあるもの―」『国語と国文学』76-5
豊島正之（1999）「書評　横山詔一・笹原宏之・野崎浩成・エリク＝ロング『新聞電子メディアの漢字―朝日新聞CD-ROMによる漢字頻度表―』」『日本語科学』6
日本規格協会（1997）『7ビット及び8ビットの2バイト情報交換用符号化漢字集合　JIS X 0208:1997』
日本規格協会（2000）『7ビット及び8ビットの2バイト情報交換用符号化拡張漢字集合　JIS X 0213:2000』
日本規格協会（2001）『国際符号化文字集合（UCS）―第1部:体系及び基本多言語面　JIS X 0221:2001』
文化庁文化部国語課（2000）『明朝体活字字形一覧―1820年～1946年―』大蔵省印刷局