日本語漢字の異形字シソーラス

小熊善之，永森光晴，阪口哲男，杉本重雄，田畑孝一
図書館情報大学
E-mail: {oguma, nagamori, saka, sugimoto, tabata}@ulis.ac.jp

概要

漢字の中には、「国」と「國」「竜」と「龍」などのように、意味・読み・用法が全く同じでありながら姿形だけが違う漢字“異形字”が存在する。これらは文中で互いに入れ替えたとしても、文意が変化することはない。また、読解上区別することもない。

一方、近年急速に普及したコンピュータ上での漢字処理は、文字コードに基づいている。文字コード上で別々のコードポイントが与えられた漢字は、「国／國」のように同音・同義であっても別の文字であると処理される。このことが、日本語情報処理の中で、小さくない障碍となってきている。

人間が知識と経験に基づいてごく自然に行っている同一視作業を、コンピュータ上で行おうとすると、どの漢字とどの漢字が同一視できるのかという情報を用意してやらなければならない。本研究では、JIS X 0208:1997 と JIS X 0212-1990 に含まれる 12,156 字を対象に、漢字同一視のためのシソーラスを構築した。

キーワード

異形字, シソーラス

A Thesaurus of Japanese Kanji Variants

Yoshiyuki OGUMA, Mitsuharu NAGAMORI, Tetsuo SAKAGUCHI, Shigeo SUGIMOTO, Koichi TABATA
University of Library and Information Science
E-mail: {oguma, nagamori, saka, sugimoto, tabata}@ulis.ac.jp

Keywords

kanji variant, thesaurus

1. はじめに

一般的に日本語は、漢字仮名混じり文で記述され、漢字と縁を切ることはできない。この漢字という文字体系は、古代中国に端を発する超大規模文字体系であり、その総数は10万とも15万とも言われている。近年、これまでJIS X 0208（JIS第1第2水準）の6千字強だった漢字環境が、Unicodeなどの浸透とともに1万字を越える漢字が使用できる環境へと移行しつつある。さらに現在、10万字規模の文字集合の策定が試みられている。このような状況の影で、小さくない問題として浮かび上がってきているのが、"異形字"である。

一例を挙げれば「国」と「國」がそうである。この両字は相互に同じ字であると解釈され、通常どちらで書かれていても区別されることはない。しかし、文字コード上では両字には別々のコードポイントが与えられており、処理上は別字として扱われている。このことがもたらす最も典型的な問題を、本学のULIS OPACに見出すことができる［図1,2］。検索キーとして「柳田國男」を与えた場合と「柳田国男」を与えた場合では、得られる検索結果が全く違う。利用者側の立場からすると、「柳田國男」でも「柳田国男」でも同じ結果が返って来るように、「国」と「國」は同一視して欲しい。しかし文字コード上からだけでは、これらの異形字を同一視するための情報を得ることは不可能である。これらの同一視のためには、異形字情報を別途用意する必要がある。

結果一覧

検索条件式

- [標準]INDEX-F(INDEX-2):eq('柳田國男')

件数： 12件

項番	内　　容
1	海上の道 / 柳田國男著.-- 筑摩書房, 1967.-- (筑摩叢書 ; 85).
2	海南小記 / 柳田國男著 . 山の人生 / 柳田國男著 . 北の人 / 金田一京助著 . 東奥異聞 / 佐々木喜善著 . 猪・,鹿・,狸 / 早川孝太郎著.-- 平凡社, 1961.-- (世界教養全集 ; 21).
3	居住習俗語彙 / 柳田國男,山口貞夫共編.-- 国書刊行会, 1975.
4	禁忌習俗語彙 / 柳田國男著.-- 国書刊行会, 1975.
5	婚姻習俗語彙 / 柳田國男, 大間知篤三共著.-- 国書刊行会, 1975.
6	歳時習俗語彙 / 柳田國男編.-- 国書刊行会, 1975.
7	退讀書歷 / 柳田國男著.-- 書物展望社, 1933.
8	服装習俗語彙 / 柳田國男編.-- 国書刊行会, 1975.
9	明治大正史 : 世相篇 / 柳田國男著.-- 平凡社, 1967.-- (東洋文庫 ; 105).
10	柳田国男生誕百年記念民俗調査報告書.-- 柳田国男生誕百年記念会, 1976.
11	柳田国男生誕百年記念国際シンポジウム・民俗調査報告書 / 柳田国男生誕百年記念会, 日本民俗学会編.-- 柳田国男生誕百年記念会, 1976.
12	柳田國男 ; 折口信夫 ; 萩原朔太郎 ; 宮澤賢治 ; 高村光太郎 ; 斎藤茂吉 ; 高浜虚子 ; 久保田万太郎 ; 幸田露伴 / 柳田國男[ほか]著.-- 小学館, 1989.-- (昭和文学全集 / 井上靖[ほか]編 ; 4).

図1

結果一覧

検索条件式

- [標準]INDEX-F(INDEX-2):eq('柳田国男')

件数： 75件

項番	内　　容
1	遠野物語 / 柳田国男著.-- 大和書房, 1972.
2	火の昔 : 少年少女のための文化の話 / 柳田国男著 ; 笠原正夫挿画.-- 海鳴社, 1991.-- (柳田国男児童読み物集 / 庄司和晃監修).
3	海上の道 / 柳田国男著.-- 筑摩書房, 1967.-- (筑摩叢書 ; 85).
4	山島民譚集 / 柳田国男著 ; 関敬吾,大藤時彦編.-- 増補.-- 平凡社, 1969.-- (東洋文庫 ; 137).
5	少年と国語 / 柳田国男著 ; 笠原正夫挿画.-- 海鳴社, 1992.-- (柳田国男児童読み物集 / 庄司和晃監修).
6	昔話と文学 / 柳田国男著 ; 大藤時彦解説.-- 白凰社, 1971.-- (白凰社名著選).
7	村と学童 / 柳田国男著 ; 笠原正夫挿画 ; 槇垣一彦解説.-- 海鳴社, 1990.-- (柳田国男児童読み物集 / 庄司和晃監修).
8	退讀書歷 / 柳田國男著.-- 書物展望社, 1933.
9	定本柳田国男集 第10巻 / 柳田国男著.-- 筑摩書房, 1962.
10	定本柳田国男集 第11巻 / 柳田国男著.-- 筑摩書房, 1963.
11	定本柳田国男集 第12巻 / 柳田国男著.-- 筑摩書房, 1963.
12	定本柳田国男集 第13巻 / 柳田国男著.-- 筑摩書房, 1963.
13	定本柳田国男集 第14巻 / 柳田国男著.-- 筑摩書房, 1962.
14	定本柳田国男集 第15巻 / 柳田国男著.-- 筑摩書房, 1963.
15	定本柳田国男集 第16巻 / 柳田国男著.-- 筑摩書房, 1962.
16	定本柳田国男集 第17巻 / 柳田国男著.-- 筑摩書房, 1962.
17	定本柳田国男集 第18巻 / 柳田国男著.-- 筑摩書房, 1963.
18	定本柳田国男集 第19巻 / 柳田国男著.-- 筑摩書房, 1963.
19	定本柳田国男集 第20巻 / 柳田国男著.-- 筑摩書房, 1962.
20	定本柳田国男集 第21巻 / 柳田国男著.-- 筑摩書房, 1962.
21	定本柳田国男集 第22巻 / 柳田国男著.-- 筑摩書房, 1962.
22	定本柳田国男集 第23巻 / 柳田国男著.-- 筑摩書房, 1964.
23	定本柳田国男集 第24巻 / 柳田国男著.-- 筑摩書房, 1963.

図 2

本研究では、JIS X 0208:1997 と JIS X 0212-1990 の計 12,156 字を対象として、異形字関係を記載したシソーラスを構築した。

2. 異形字

“異形字”とは、情報処理学会文字コード体系専門委員会で提示された新しい用語で、「“字体”が異なるが、意味や読みなどが同じ漢字で、文中でそれらを入れ替えても善意の第三者にとっては同じ意味の文になる漢字」と定義されている [1]。これまで「異体字」「古字」「籀文」「略字」「俗字」「譌字」「新字・旧字」などと辞書・論者によって呼ばれてきたものを総称する。異形字は相互の関係を示す言葉であり、どちらが正しい字であるといった区別はない。通常、これらの字は相互に字音と字義を共有している。

異形字の発生過程は多種多様であり一概に分類することは難しく、またその総数も判然としない。我々に最も身近な異形字としては「常用漢字体」と「旧字体」が挙げられるが、常用漢字表 1,945 字の中ですら 355 字について 357 字の異形字が示されている。しかもこれは草冠やしんにょうなどの差異を除いた数

である。また、この表に記載されていない異形字も存在する。特殊なものではあるが無視できない例としては、人名・地名などの固有名詞に使われる異形字が挙げられる。「齊」や「邉」などはその典型例であるが、戸籍上の字とは別に、日常は略字や新字などの異形字を用いているという人も少なくない。日常的によく使用する異形字もあれば、辞書を引かないとまず分からないものまで、異形字には実に多くの種類が存在する。

2.1 文字コードと異形字

文字の電算処理を考えるとき、最も一般的に用いられるのが文字コードである。文字コードは、文章記述に必要な文字を文字セットとして集め整理し、これにコードポイントを付与したものである。

最も利用されている日本語文字コード規格・JIS X 0208:1997 は、漢字だけで 6,355 字を規格化している。その符号化は字体を拠り所としており、基本的に同一視される異形字であったとしても、字の形が違えば別字として別のコードポイントを付与している。この結果、JIS X 0208:1997 の中には、相互に異形字となる関係を持った漢字が、少なくとも 600 対 1,300 字以上別字として収録されている。

たとえ同字と看做される異形字であったとしても、規格上で別のコードポイントが与えられてしまえば、処理上は別字として扱われる。また、異形字相互の句点番号になんらの規則性もないので、単純に文字コードだけを見て異形字関係を判別することもできない。そのため多くの処理系では、異形字の同一視といった処理は行われず、末端利用者に押し付ける形になってしまっている。現行文字コードでは、異形字の同一視処理を行うためには、文字コードとは別に異形字情報が必要とされる。

2.2 図書館情報システムと異形字

文字コードだけでは異形字の同一視ができないという問題は、その文字コードを用いた図書館情報システムにもにそのまま引き継がれた。本稿の最初に例示した本学 OPAC をはじめ、NACSIS Webcat などでも同様の問題を見出すことができる。この問題によって、利用者が異形字の存在に気を配っていないと、検索要求に対し充分な回答を引き出せない状況が発生してしまっている。

この問題に比較的早くから対応を試みていたのが国立国会図書館である。JAPAN MARC の編纂に際し「漢字等の字種採用の基準」[2] を打ち立て、異形字によって発生する発生する問題の回避を試みた。具体的な方法としては、JIS C 6226-1978（現、JIS X 0208）の中に定められた漢字の中に、使わない文字を設定したのである。なんらかの異形字を規格内に持つ漢字については、その相互に異形字関係を持つ漢字の中から代表一字のみを統制字として用い、他の字は用いないようにする。規格外字については、諸橋大漢和番号を付与するというものである。しかしこの方法でも、末端利用者が異形字を全く考慮しないで済むわけではなく、数ある異形字の中で、どの字が統制字として採用されているかを意識していなければならない。日常良く使う異形字ならばともかく、異形字であることを知らなかった場合、全く検索結果を得られないこともあり得る。

現在進行している電子図書館計画に於いては、より一層、問題となることが予測できる。目録だけではなく、全文を電子化して蓄積する電子図書館に於いて、全文検索を実行したときに、その中に現れる異形字は膨大な数になる。人文科学分野に於いては特に、歴史文献からの引用など、異形字を用いる機会が多い。しかし必ずしも全ての局面で異形字を入力できるとは限らないし、どの異形字が用いられていたかを正確に記憶しているわけでもない。引用時に常用漢字体に置き換えてしまう場合もある。

こういった文字コードだけでは対処できない状況を打開するためにも、異形字情報の整備が望まれるのである。

3. 異形字シソーラス

今後の各種情報処理を考えたとき、これまで読み手が無意識的に行ってきた異形字の同一視／区別を、電算機上でも行えることが望ましい。しかし前述したように文字コードそのものは、異形字に関する情報を持っていない。そのため、異形字同定のためには、別途異形字情報を用意しなければならない。異形字同定ツールを開発するにも、この異形字情報の整備が最も重要であることは論を待たない。

異形字情報を記載したものとしては、これまで漢和辞典、漢字字書が通用してきている。しかし、これらの辞典・字書は異形字情報を知るのが第一義ではない。異形字の情報は記載されてはいるが、相互の異形字関係を総覧するのにはあまり向いていない。また、電子化も余り進行しておらず、電子化されているものも、専用アプリケーションを介して利用するものが殆どであり、その情報を二次利用するには制限がある。

電子的な異形字情報に求められる情報は、ある漢字について、異形字関係にある漢字を網羅し、相互に参照を可能とするものである。そして二次利用が容易であることが望まれる。

そういった要求への一つの回答として、シソーラスという形での情報提供が行われてきた。

3.1 先行研究

異形字シソーラスとしての性格を持つ最初の研究としては、1980年の国立国文学研究資料館「計算機による日本語文字システムの実用的処理」研究班による『データ処理システムのための漢字シソーラス〔試作版〕』[3]が挙げられる。

この研究の中心的人物であった田嶋一夫氏（現、いわき明星大学教授）らの手によるこのシソーラス（以下“田嶋シソーラス”と呼称）は、見出し字として当時のJIS C 6226-1978を立て、関連字の調査範囲は大漢和、新字源に及んでいる。

しかし残念なことに、現在一般に目にすることができる田嶋シソーラスはこの〔試作版〕のみであり、フォントの都合からか随所に欠字が見られ、実際の利用に際しては、新字源と大漢和辞典を用意する必要がある。一般に公開されているのが紙媒体のみであるという点も、利用を困難にしている。

またこれは田嶋シソーラス自体の問題ではないが、田嶋シソーラス編纂後、JIS C 6226は83年、90年、97年と改定されており、その間に字体の変更（29字）、文字の入れ替え（26組）、文字の追加（6字）が行われている。そのため、当時の規格と現行規格の間に小さくない差異があり、そのことが更に田嶋シソーラスの利用を困難にしている。

3.2 “あやのふひと”

本研究では異形字シソーラスとして、新たに“あやのふひと”を編纂した。この“あやのふひと”には以下のような特徴がある。

- 本体はCSVファイルとして提供され、電子的な利用が容易である。
- JIS X 0208:1997とJIS X 0212-1990を含んでいる。
- 二次利用に制限がない。

特に第二点については、フリーのものとしては国内では唯一である。

シソーラスを実際に編纂するに当たっては、このシソーラスで異形字と看做す文字の基準を以下のように立てた。

1) 常用漢字表に掲げられている、所謂“旧字”
2) JIS X 0208:1997 に於いて参考として示されている参照字
3) JIS X 0212-1990 に於いて参考として示されている同義漢字
4) 上記以外で、字書に於いて異形字関係が指摘されているもの

1)~3) については現行の法令・規格を参照した。4) については、法令ならびに規格票には記載されていないものの中で、明らかに異形字と看做せるものや、異形字と看做した方が適切であるものについて、追加を行った。

3.2.1 “あやのふひと”の形式

“あやのふひと”はレコードとして見出し字、JIS 区点番号、Unicode 番号、検字番号、部首番号、部首内画数、総画数と、異形字集合あるいは参照先を持っている。

JIS 区点番号は JIS 規格によって振られた番号で、Unicode 番号は Unicode 1.1 に基づくコードポイントである。

見出し番号は見出し字に対して振られる通し番号である。

部首番号は、清朝に編纂された『康熙字典』で立てられた 214 部首に、1 から 214 まで順につけられた番号である。

部首内画数は、その部首の中での画数を表す。部首番号と部首内画数は、主として排列のために用いる。

総画数は言うまでもなく、漢字の総画数である。

そして見出し字がディスクリプタの場合は異形字集合の要素が、非ディスクリプタの場合は対応するディスクリプタの見出し字と見出し番号が参照先として示される。ディスクリプタは、第一に常用漢字体とし、集合内に常用漢字体がない場合は、句点番号が一番若いものとする。

3.2.2 使用ツール・作成環境

主に二つの環境で、“あやのふひと”を構築した。

```
〈環境 1〉
OS:     Microsoft Window95
Tools:  Gnu Awk (gawk) 2.15, patchlevel 6 + multi-byte extension 1.04
        GNU sed version 1.18 + multi-byte extension 1.03
〈環境 2〉
OS:     Personal Media Corporation B-right/V R.2.010 （超漢字）
Tools:  Microscript
        MicroCard
```

Windows 環境は主に JIS X 0208 領域の作業をしていたときに、超漢字環境は JIS X 0212 領域での作業に使用した。

また、各種漢字情報については、以下の二つを主に利用した。

a) JIS X0208-1983「情報交換漢字用符号系」の為の漢字字書 [4]

b) 最新 JIS 漢字辞典 [5]

なお、上記の情報だけでは不足と思われた場合は、適宜、漢和辞典を当たって勘案した。

最終的に完成したデータは、TRONコードによるCSV形式のものと、それをeuc-jpに変換したものの二つである。

4. 考察

4.1 異形字の同定基準について

今回のシソーラス構築に於いては、最も根源的な問題である。

「一」と「壱」、「二」と「弐」は、一般的には異形字として認識されているが、今回のシソーラスではそれぞれ異形字とはされていない。これは、典拠にした資料類において別字であるとされており、常用漢字表にも別々に記載されていることに鑑みた。しかし実用上は異形字として扱う局面が多い。

また国字については、異形字としての記述にかなりばらつきがあった。ある資料では異形字として記載されていても、別の資料では別字となっていたりした。

元異形字と言うべきものも、存在していた。かつては同じ意味の字であったにもかかわらず、歴史の中で片方の字の意味が変わってしまったものである。これは、対象資料によって異形字とするか否か、判断を分けねばなるまい。

こういった同定基準の問題は、思想的な部分を少なからず含むので、突き詰め始めると水掛け論に陥りやすい。よって、シソーラスの利用者が利用状況を勘案して、適宜異形字を追加、あるいは削除するのが望ましい。

4.2 包摂について

ある句点番号を振られた文字に対して、どこまでの字体の揺らぎを許容するかが、本来の包摂基準の意義である。例えば、草冠は別れていても繋がっていても、同じ句点番号に包摂し、実際に表示される字形はどちらであっても構わないとされる。しんにょうや示偏も、このように規格上の包摂が行われている。

これが異形字のみを対象に包摂が行われていれば問題はなかったのだが、JIS X 0208に於いては、別字に対しての包摂が規格上で行われている。有名な所では柿とこけら（図3）の問題が挙げられる。JIS X 0208ではこの柿とこけらについては、同字と看做している。しかし、本来この二つの漢字は、姿形こそ似ているが別字であり、またJIS X 0212には独立した一字として句点番号が与えられているのである。JIS X 0208とJIS X 0212の間の包摂基準の不一致が、混乱を生んでいる。

これらの文字については、今回はJIS X 0208の包摂を敢えて無視し、JIS X 0212の字を別字としている。しかし、用法によっては、このような“同形異義字”をシソーラスで吸収する必要が生まれるかもしれない。

4.3 より柔軟な用途に向けて

今回のような単漢字シソーラスでは吸収できない、同義語の問題もある。

これはより一般的なシソーラスの守備範囲となるべきものなのだが、略語字とでも言うべき字が、存在する。「図書館」という単語に対して図4という字が、「貝多羅」という単語に対して図5がそれぞれ存在し、略語字として利用されている。将来的に異形字シソーラスを拡張し、漢語シソーラスとして整備することがあれば、外せない項目であろう。

かき　　こけら

柿柿

図3

圖

図4

逆に、今回のような単漢字シソーラスで吸収できるであろう問題としては、「置き換え」語が挙げられる。「置き換え」語は、當用漢字制定後、国内各方面で法令用語や学術用語などが當用漢字表内字で置き換えられたものである。一般に異形字で置き換えられたものは大きな問題はないが、音借で書き換えられたものは注意が必要である。例を挙げれば「雇傭」は「雇用」に、「国際聯合」は「国際連合」に、一律に置き換えられた。これらは確かに同じ語として用いられているが、置き換えられた文字同士が同じ意味だったことを示してはいない。しかしこれらを敢えて同一視したい場合は、各置き換え字をも異形字に含めるようにシソーラスに追加することで、対処が可能である。

また、漢字以外では、変体仮名や平仮名／片仮名の同一視、アルファベットの大小の同一視などが挙げられる。これらについても、シソーラスの収録基準を変更し、追加登録することによって実現が可能である。

椢

図5

5. おわりに

本研究では、ここまで述べた通り、日本語漢字における異形字をまとめて、1万2156件のシソーラスとした。この規模の異形字シソーラスでは、これまで電子的なものが一般公開されたことがないが、"あやのふひと"は広く一般に公開する予定である。公開に伴って、誤謬・不備の指摘などが寄せられると思うが、できるだけ真摯に受け止め、修正に務める所存である。

この研究の為に利用した、様々な字書、辞書、ツール類を編纂執筆作成した先人達に深く感謝の念を捧げる。この研究は到底私一人の力で成し得るものではなく、先達の研究の成果がなかりせば、開始直後に挫折していた事だろう。改めて、畏敬の念を深くした。とりわけ、『今昔文字鏡』と文字鏡研究会には、公私共にお世話になった。また、国文学研究資料館の原正一郎先生には完成前のシソーラスを見て頂き、貴重な意見を頂いた。

弥縫したい所や、今後への課題はいくつか遺したものの、12,156字中、1,255組2,760字の異形字を集成し、一つの区切りとしたい。

今後後進には、このシソーラスを用いて、より良い日本語情報環境を獲得することを期待する。

参考文献

[1] 情報処理学会文字コード標準体系検討専門委員会会議資料

[2] 国立国会図書館. 漢字等の字種採用の基準, 1992, http://www.ndl.go.jp/librarian/wtnews/kanzi.html
国立国会図書館. 文字種の取り扱い基準の変更について, 1998,
http://www.ndl.go.jp/librarian/wtnews/mojishu.html

[3] 山中光一, 田嶋一夫. データ処理システムの為の漢字シソーラス〔試作版〕.「計算機による日本語文字システムの実用的処理」班, 1980

[4] 豊島正之, 金水敏, 古田啓.JIS X0208-1983「情報交換漢字用符号系」の為の漢字字書, ftp://fan.shinshu-u.ac.jp/pub/kanjidic/ydic.lzh

[5] 田嶋一夫 監修　日本規格協会 編集協力. 最新JIS漢字辞典. 東京, 講談社, 1990, (ISBN4-06-123264-9)