2005－CH－67（8）
2005／7／29

階層的素性名を用いた異体字記述の試み

秋山　陽一郎 †　守 岡 知 彦 †　浦 田 衣 里†

CHISE (CHaracter Information Service Environment) で本格的な字義データベースを構築するための準備として、『大漢和辞典』(修訂第二版、第 9〜11 巻) における異体字記述を我々が提案する『Chaon モデル』および『階層的素性名方式』に基づき機械可読形式でデータ化することを試みた。この試みはいわゆる『代表字形』に相対する『別字形』という 1 対 1 もしくは 1 対 n の関係ではなく、文字間関係の種類と方向性を重視するという点において画期的であるといえる。本報告はそのデータ化の中で浮かび上がって来た諸問題とその解法について述べる。

Description method for relations of character variants based on the Hierarchical Feature Name method

Yoichiro AKIYAMA,† MORIOKA Tomohiko† and Eri URATA†

This report's task is to suggest description of Kanji character variants by Hierarchical Feature Name. This will be one of the tasks based on Chaon model advocated by the CHISE (CHaracter Information Service Environment) Project. Characters variants in coded character sets and current character databases are often associated in 1 to 1 or 1 to n model, and they are just associated as one of the "variant character" of the target character. The Hierarchical Feature Name method suggested in this report, will associate Kanji character variants by categories of variants and directions of association. Additionally, the refered sources can be described for each character association object, so that any association indicated by optional sources or individuals are totally free from any single fixed semantic. The source used in this report is *Daikanwa dictionary* (2nd edition, Vol.9 - 11) edited by Tetsuji Morohashi (Taishukan Press, 1999).

1. はじめに

CHISE (CHaracter Information Service Environment) で本格的な字義データベースを構築するための準備として、伝統的な漢字辞典における異体字記述を我々が提案する『Chaon モデル』および『階層的素性名方式』に基づき機械可読形式でデータ化することを試みた。

データ化の対象としたのは大漢和辞典で、実際に 3 巻分 (9 〜 11) のデータ化を試みた。ここでは、この試みの中で浮かび上がって来た諸問題とその解法について述べたい。

1.1 異体字関係とその表現

ここで述べる異体字記述法における『異体字』は、基本字に対応する『異体字』という従来の辞書的な意義も含むが（一例を挙げれば『本字』・『正字』に対する『俗字』という概念など)、文脈自由に置き換えが可能な等価な文字など、必ずしも一方的な代表字・別字関係を前提とはしないため、本稿では混同を避けるためにこれを『類字』と呼ぶことにする。たとえば『大漢和辞典』の「無」(M-19113) には禁止の助辞として「毋 (M-16721)・勿 (M-02501)・莫 (M-31078) に同じ」とあるが、参照先の項目を確認してみても、「毋」(M-16721) では「無に通ず」とあり、「勿」(M-02501) にも「無、莫に通ず」とあることなどから、これらの文字間関係が必ずしも 1 対 1 もしくは 1 対 n ではないことがわかる。

異体字 (類字) 関係を表現する手法としては、「符号化文字基本集合 (Basic Subset of Coded Character Sets)」(以下、BUCS)[1] のように『代表字形 (字体)』(親字) に対して『別字形 (字体)』(異体字) を挙げるという方法や、特に代表字形 (字体) を定めずに対応関係が認められる文字間の関係を異体字シソーラスとして表現する手法などがある。

前者の手法は個々の文字の使用頻度を踏まえた実装を行う上ではある程度有効だが、文字間の関係は必ずしも『正字』(もしくは『本字』) に対する『俗字』、あ

† 京都大学人文科学研究所
Institute for Research in Humanities, Kyoto University

るいは基本字に対する異体字といったようなものに収斂されるとは限らず、このモデルでは文字間の関係を扱う上での幅を狭めてしまう。後者の手法はこのような恣意性を避けることができ、また、検索などでは有効であるが、『旧字』(繁体字) と常用漢字や簡体字の変換や異体字の正規化処理などを実現する上で問題がある。

こうしたことを鑑み、我々は『代表字形 (字体)』(親字) を定めることなく、異体字 (類字) の種類を分類し記述することで、両者の利点を併せ持った異体字 (類字) データベースの実現を目指している。『同字』『通仮字』『正字』『俗字』といったような文字間の関係の種類を記述することは文字間の関係がいかなる要素 (字形・字音・字義・字種 etc..) に起因・関係しているかを認識する上で重要である。即ち、これは文字間の関係をどのようにとらえるかを示す情報であるといえる。しかしながら、このことは文字間の関係の記述に一層の恣意性を持ち込み、『正字』『俗字』といったような政治的視点・偏見を導入するものである。しかし、ここで、その関係の種類に対し、その出典を明示し、また、その関係の種類が有効となる範囲を明示することによって、その関係の種類が持つ恣意性や政治性・偏見を認識する主体を明示したり、その視点が流通し得る限界を示すことが可能であると考えられる。我々はそのための仕組みとして『階層的素性名方式』を提案するとともに、その方式の有効性を実証するために大漢和辞典に現れる異体字関係に対して、この方式を適用することを試みた。

1.2 なぜ大漢和辞典か

今回、大漢和辞典を中心ソースとして利用しているのは、以下の２点の理由による。

(1) さまざまな先行ソースが参照されている
(2) 公知で誰もが比較的容易に参照できるソースで、かつ情報量も比較的豊富である

特に重要なのは前者の要因である。

類字関係が記述されている先行する字書や工具書は、字形をベースにモデル化された『説文解字』・『玉篇』・『康熙字典』といった字書類、『切韻』・『廣韻』・『集韻』・『一切經音義』など字音をベースにモデル化された韻書類、字義によってカテゴライズされている『爾雅』・『廣雅』・『釋名』といった辞書類をはじめとして枚挙に暇がない。

いずれも編纂された時代や背景事情・目的・思想などが異なっており、細かい類字の取り扱い方は一様ではない。

このうちのどれか一点を採用すれば、モデルの精度は向上するが、我々が目指しているのはあくまで大規模汎用文字データベースの実現であって、たとえば『説文解字』単体の類字 (もしくは異体字) データベースではない。

そこで、上に挙げたような各種字書や工具書ソースを可能な限り網羅的に参照していて、かつ公知で需要の高いソースとして『大漢和辞典』を用いることにした。

時代や背景事情・目的・思想などが異なる多数のソースを元にしていることで、あるいは類字関係に齟齬が生じる可能性は充分にありうるが、そうした齟齬や矛盾も柔軟に包摂できるようにしている (そもそも CHISE では多様な文字間関係の学説を許容できるようにしており、むしろ矛盾や齟齬が生じてもそれらを共存させうることが大規模文字データベースにおいては重要な要素であるといえる)。また同時に出典情報も入力しておくことで『大漢和辞典』内における文字間関係をソースごとに抽出できるようにしている。こうして構築される文字間関係データベースのモデルは、もちろん個別の字書や工具書の類字データベースを独自に制作・実装するような際にもほとんどそのまま転用できると考えられる。

2. CHISE

CHISE (CHaracter Information Service Environment) というのは CHISE Project で開発している次世代文字処理環境で、文字符号に制約されることなく文字や文字の性質を自由に表現・処理するための環境として開発を進めているものである。

符号化文字モデルでは文字に関する知識は文字符号の定義の中に存在し、利用者が勝手に変更することはできない (利用者が勝手に変更すれば情報交換できなくなる) し、そうした知識は計算機の中に存在しない。それに対し、CHISE では文字に関する知識は計算機の中に機械可読な形で (文字オントロジとして) 保持される。よって、文字オントロジを変更することにより、文字のセマンティクスを変更することができる。

2.1 Chaon モデル

CHISE では『Chaon モデル』と呼ぶ方法によって文字を表現するようになっている。これは汎用符号化文字集合に依存することなく自由に文字を表現するために我々が提案しているもので、表現したい文字に関する知識 (文字の性質の集合) の機械可読な表現によっ

て文字を表現し操作する方法である。

Chaon モデルでは、文字を説明するための要素（文字の性質や用例など）を『文字素性』(character feature) と呼ぶ。文字素性としては、部首、画数、部品の組合せ方に関する情報（漢字構造情報）、発音、意味、用例、その他文字処理で必要となる各種情報などが考えられる。

Chaon モデルにおける文字表現は文字素性の集合であり、文字に関するさまざまな要素を文字素性として記述可能である。また、対象とする文字を 1 つの文字素性だけで記述することも可能であるし、多数の文字素性を用いて非常に詳細に記述することも可能である。なお、Chaon モデルでは文字と文字の集合は本質的に区別されない。符号化文字モデルでは、通常、文字符号の仕様（規格）が想定する抽象文字を文字とし、抽象文字はある範囲の字体を包摂するものとなっている。ここで、文字か字体かが必然的に判別可能なものであるなら問題はないのであるが、実際にはこれは文化的なものに依存しており、社会的に共有される規範や文字観念はあるにせよ、必ずしもかっちりと線を引けるものではないといえる。そして、これらは歴史的・地域的に変化することがあり、用途や文脈にも依存し得る。よって、Chaon モデルでは、技術的な枠組としては、抽象文字、字体、字形のような文字の抽象度に関する固定的な線引きを行わない。ある文字素性の集合を字形を表したものと思うか、字形の集合である字体を表したものと思うか、字体の集合である文字を表したものと思うか、はたまた文字の集合であると思うかは、その文字素性の集合の解釈に依存し、Chaon モデルはその解釈を規定しない。それは利用者の自由である。

2.2 文字間の関係の記述

Chaon モデルは文字素性の集合として文字（の集合）を表現（指示）する手法であり、そのモデルそのものには文字間の関係で構成されるネットワーク構造を表現する仕組みは存在しない。しかしながら、文字間の関係を表す文字素性を導入し、その値として文字（の集合）（の集合）を取るようにすれば、文字間の関係で構成されるネットワーク構造を記述することができる。

2.3 階層的素性名方式

汎用的な文字データベースを作る場合、用途や立場・学説などによって、文字素性に値に複数の選択肢を設けたくなる場合がある。この場合、しばしば、単純に複数の値が記述できるだけでなく、それらの値の出典情報などのメタデータも付加したい場合がある。こうした場合、文字素性の値か名前のどちらかを構造化する必要がある。『階層的素性名方式』というのは後者の方法の一種である。

階層的素性名方式は構造化の対象となる文字素性の名前（文字素性基底名）に値を選択するための識別子(『ドメイン識別子』と呼ぶ）を付けた文字列を生成し、それを名前（文字素性具象名）として用いたり、同様にメタデータ識別子を付けた文字列を生成しそれを名前（文字素性メタデータ名）として用いる方法である。この名前は次のような規則で生成される：

文字素性具象名
:= 文字素性基底名 @ ドメイン識別子
| 文字素性具象名 / ドメイン識別子

文字素性メタデータ名
:= 文字素性具象名 * メタデータ識別子

例えば、総画数を表す文字素性名を `total-strokes` とし、ドメイン識別子として `ucs` を用いる時、文字素性具象名は `total-strokes@ucs` となる。また、出典情報を表すメタデータ識別子を `sources` とする時、`total-strokes@ucs` の出典情報は

```
total-strokes@ucs*sources
```

で表される。

部首と部首内画数のように異なる種類の文字素性の値が対応関係を持っている場合、ドメイン識別子を用いてその対応関係を表すことができる。例えば、部首を `ideographic-radical`，部首内画数を `ideographic-strokes` で表す時、

```
ideographic-radical@ucs
ideographic-strokes@ucs
```

の両者は対応する。

値を構造化する手法と名前を構造化する手法を比べた場合、前者はドメイン識別子を必要としないという利点を持っているものの、値が構造データとなるので C のような単純な記憶管理機構しかない環境では不便である。また、高速化を要するような単純な処理の場合、大抵、複数の値やメタデータを必要としないといえる。また、Chaon モデル的には文字が文字素性の単純な集合になっている方が自然であり便利であるが、値を構造化すると複数の値同士の集合演算を要し、処理が複雑になる。このようなことを鑑み、現在の CHISE では共有文字データベース内では原則として値ではなく名前を構造化する方針を採っている。

3. 大漢和辞典における類字

3.1 大漢和辞典における字義階層

大漢和辞典の字義解説には必要に応じて幾つかの階層が設定されている。この大漢和辞典の字義階層モデルについて第 1 巻の「凡例」には以下のように説明されている。

> 「一字に二音以上ある場合には、[一][二][三] 等の記号を用ひて区別を明らかにした。…字義の解説は…音によって意義の異なる時は、それぞれ字音の [一][二][三] の記号の下にこれを分説した。」
>
> 「一語を二つ以上の項目に分けて説明する場合は、㊀㊁㊂ 等の記号を用ひてこれを区別し、一項の中に於て更に細説を要する場合は、適宜（イ）（ロ）（ハ）（甲）（乙）（丙）等の記号を使用した。」

まず最上階層は字音によって切り分けられている (以下、これを大漢和辞典の字音層と呼ぶ)。これは字音によって意味が変わることがあるのに対応している。たとえば「說」(M-35556) は、[一] セツ・セチ、[二] エツ・エチ、[三] [慣] ゼイ・ネイ、[四] セイ、[五] タツ・タチといった字音があるが、考え（㊃）を説いたり（㊀）、道理（㊆）を説き明かしたり（㊂）する場合にはセツ（[一]）、悦ぶ（㊀）とか楽しむ（㊂）とかいう場合にはエツ（[二]）と音読するなど、漢字の字音はしばしば字義によって変わることがある。これは漢字の字音の違いが、往々そのまま言葉としての由来の違いになっていることがあるために起こる。

ここで「字音の違い＝言葉の由来の違い」であると言い切っていないのは、もちろんそうでないケースもあるためである。つまり、同じ字義が複数の字音で共有されているケースが実際に多数存在するのだ。このような場合、大漢和辞典では同じ字義を複数の字音中に併記している。たとえば上述の「說」(M-35556) なら、「よろこぶ」という字義が [二]-㊀と [三]- ㊀、つまり「エツ（エチ）」と「ゼイ（ネイ）」という字音で共有されている。極端な場合、複数の字音でまったく同質同量の字義が共有されているケースもある。

最上層である字音層の下には細かい字義ごとに仕切られた字義層がある。複数の字義がひとつの字音中に併在する例は大漢和辞典中に常見するが、中には派生義として分化した字義や、元来持っている字義の幅が広く、さらに一層字義を細かく分類することができるものある。このような場合は、字義層をさらに広義（㊀・㊁・㊂…）と狭義（イ・ロ・ハ）とに細分化する。（これよりさらに細分化できる場合は、最大でもう 1 階層深く掘り下げて、甲・乙・丙…で区別する。）

以上のように大漢和辞典の字義階層モデルは、字音＞字義（広義）＞字義（狭義）という階層を持っているが、字義にバリエーションがない字については、上記のような記号は原則として省略されている。

3.2 文字間関係の表記ゆれ

大漢和辞典の字義階層には上記のようなモデルが設定されているが、文字間関係については凡例をみても大漢和辞典内での一貫したモデルは設定されていない。紀元前からある「作ル」・「通ズ」・「同ジ」といった関係表記を、ほとんど依拠した原典そのままに利用しているためである。ところが、実際には大漢和辞典内で出典として引用されている原典を細かく検討してみると、必ずしも原典の関係表記にしたがっていなかったり、複数ある関係表記をひとつに集約してしまっていたりしていて一貫性がない。大漢和辞典における文字間関係の記述法には概ね以下のようなパターンがある。

A 大漢和辞典の文字間関係表記は原則として依拠した原典の関係表記をそのまま踏襲している。

B 依拠した原典が複数ある場合には、そのいずれかの関係表記を採用している。

C 依拠した原典の数量に関わりなく、原典表記に依らずに大漢和辞典が独自に関係を提示しているケースがある。

D そもそも出典が明示されていないものがある。

上記の特徴を要するに、大漢和辞典が依拠した原典に忠実に従っているものと、大漢和が特定の原典に従わずに独自に（あるいは恣意的に）文字間関係を定義しているものとに区別されよう。言い換えれば、大漢和辞典が関係定義を放棄しているものと、独自に関係を定義しているものとに区別されるとも言える。

4. 類字データベースの実現

4.1 形　　式

CHISE では文字データベースを記述するのに通常『define-char 形式』と呼ぶ Chaon モデルに基づく S 式 (Lisp) 表現を用いることになっている。この形式は入れ子状のデータが表現できるなど表現力が高い半面、S 式 (Lisp) に不慣れな者には扱いにくい面がある。このため、入力作業には define-char 形式とは別の行指向の形式を定義し、これを用いた。

この行指向の形式は具体的には次のような形式で表

される：

見出し字, 大漢和辞典字義階層, 参照先の類字, 文字素性名, 大漢和辞典における出典

実例は 1 の通り。たとえば「罘」(M-28224) の [一]（「フウ」・「ブ」という字音）[二]（「ヒ」・「ビ」という字音）- ㊀（「あみ」という広義の字義）-（イ）（「うさぎあみ」という狭義の字義）に、「(M-28275) に同じ」として『集韻』・『史記』司馬相如列傳・その注（裴駰『史記集解』）などが出典として引用されている。これを上記形式で記述すると、

M-28224,[1][2]-1-イ,M-28275,¡-same,jiyun shiji shiji-zhu

となる。見出し字の大漢和番号につづいて、大漢和の字義階層、参照先の類字の大漢和番号、さらに「〜に同じ（同字）」という文字間関係の種類（”same”）を記述し、出典リスト（ピンインベースの書名表記）をスペース区切りでデータ化している。同様のことを罘 (M-28224) と罦 (M-28279)、罦 (M-28279) と罟 (M-28275)、罦 (M-28279) と罥 (M-28256)、罟 (M-28275) と罾 (M-28282) などに対しても行うことになるが、このようにして 1 対 1、1 対 n の枠には収まらない類字関係を記述している。また罦 (M-28279) と罥 (M-28256)、罟 (M-28275) と罾 (M-28282) は「もと〜に作る（本字）」(本報告の試みでは ”original” としている）とあって文字間関係の種類が異なる。このように類字（異体字）をカテゴライズして、単なる「異体字」や「類字」ではなくパリエーションを持たせるように記述している。

なお、この形式は define-char 形式に機械的に変換可能である。

4.2 字義階層の記述

CHISE では大漢和の字義階層を [1]-6-イ のように順に記述して大漢和辞典の字音区分・字義区分を保持したまま取り扱っている。

例 M-39076,[1]-6-イ, M-06664,<–synonyms, shuowen-tongxundingsheng guangya lushichun-qiu lushichunqiu-zhu

なお、このほか見出し字に対して特に字形・字音・字義に関する疑義・異説が存在する場合には「参考」(データ上では “note”）という欄が上記の字義階層の外に別途設けられる。

4.2.1 ドメイン：大漢和辞典独自の解釈

CHISE では特定ドメインにおける特異な説を柔軟に扱えるようにしているが、これはこれまでの大規模文字データベースにはなかった画期的な点である。

文字間の関係を記述する際には、その関係が普遍的に通用するものなのか、特定のドメイン内でのみ規定されているものなのかという点が区別されている必要がある。これについては 2.3 節で述べた階層的素性名方式を用いて以下のように記述している。

文字間関係@ドメイン

たとえば「諸韻書皆誤って蚌（M-33084）に作る。今、正す」(M-33218）とある事例ように、大漢和辞典が依拠した原典の説に反して、ある文字と別のある文字との関係を『誤字』であると異義を提起している場合は、wrong@daikanwa （誤字@大漢和辞典） のように記述する。

こうすることによって「大漢和辞典では誤字だと主張している」という、特定ドメインに限定的に主張されている文字間関係を、普遍的なものと区別して取り扱うことができるようになる。

この手法は大漢和辞典内（多くは「参考」というオプション的な項目中）で紹介されている、大漢和辞典以前の固有の学説にも適用しうる。

たとえば「段玉裁は、小篆、趨（M-37333）を趨 (M-37351）に改める」のように表記されている個所も、wrong@duanzhuben（誤字@段玉裁）のように記述できる。もちろん、このケースは段玉裁（『説文解字注』）の学説と、それ以外（特に『説文解字注』以外の各種『説文解字』校注本）の従前説を区別するためのものである。

4.2.2 出典依存の情報

先述したように、大漢和辞典の文字間関係は依拠した原典の表記に依存している。つまり、それぞれの関係の種類（〜に作る、〜に通ず、〜に同じ、本字、古字など）のモデルも依拠している原典に依存することになり、少なくとも大漢和内では個別の文字間関係について一貫した定義はなされていない。そこで、文字間関係の後にその大漢和辞典が依拠している出典を挙げることにした。こうすることによって、データが揃った後で、文字間関係表記のモデルを大漢和辞典が依拠した個別の出典ごとに抽出・検討することができるようになる。出典名は原則として声調なしピンイン表記としている。出典は複数挙げることも可能で、同一文献の特定版本が限定的に指定されている場合は、前項のドメイン表記の手法を採用している。たとえば「大徐本説文」(北宋の徐鉉という人物が中心となって校定した『説文解字』。徐鉉の弟 徐鍇の『説文解字繫傳』を「小徐本」と呼ぶのに対応して一般に

「大徐本」と呼んでいる）とある場合（『説文解字』という文献の中でも「大徐本」と呼ばれるテキストでしか確認できない場合）は shuowen@daxuben（説文@大徐本）のようになる。またすでに散佚した古典籍の佚文も同様で、李善『文選注』所引『蒼頡篇』などは cangjiepian@wenxuan-lishan-zhu（蒼頡篇@文選-李善-注）のように表記する。

4.2.3 文字間関係：原典における関係表記の保持

大漢和辞典の文字間関係表記は原則として依拠した原典の関係表記をそのまま踏襲しているため、1 対 1 の文字間関係であっても、その文字間の関係がまったく異なるモデルを持った字書に依拠していることがある。したがってたとえば一方の字が他方の字の『俗字』であると書かれていたとしても、必ずしもそれがその『俗字』の『本字』もしくは『正字』と定められているとは限らないし、『俗字』『本字』『正字』とは一見 縁のない文字間関係が書かれていることもある。

このような一見 破綻ないし矛盾しているかのように見える多様な文字間の相互関係をも取り扱えた上で、原典における関係表記をそのまま保持するという要求をこれまで満たした大規模文字知識ベースはなかった。

この文字間関係については以下のようにデータ化している。

見出し字（文字 A）,, 参照先の字（文字 B）
<–文字間関係の種類,
見出し字（文字 B）,, 参照先の字（文字 A）–>
文字間関係の種類,

たとえば賓（M-36790）という見出し字に対して「M-36788 の古字」であるという字解がある場合は、

例）M-36790,,M-36788,<–ancient,

のように記述する。これは見出し字の賓（M-36790）が、参照先の文字「賓」（M-36788）の古字（ancient）であることを意味している。逆に「裕」（M-34305）という項目に「古、衮（M-34306）に作る」という解説がある場合には、

例）M-34305,,M-34306,–>ancient,

のように記述する。こちらは見出し字「裕」（M-34305）の古字（ancient）が参照先の文字「衮」（M-34306）になるため、前の事例とは関係の方向性（<–|–>）が逆になる。以上のように記述することで、単なる異体字シソーラスにとどまらず、文字間関係の方向性まで扱えるようになる。

ちなみに同時に 2 件以上の文字間関係があるケースにも対応している（以下のように参照先の諸橋番号を併記する）。

例） 【賓】(M-36788) [一]-⑮：古、賚（M-36790）・賮（M-07350）・賓（M-25542）・贐（M-07088）に作る。

M-36788,[1]-15,M-36790,M-07350,M-25542,M-07088,–>ancient,

また 1 つの項目に 2 つ以上の文字間関係が指摘されているようなやや複雑なケースの記述も行うことができる（この場合は参照先の諸橋番号と関係の種類と方向性をセットにして併記する）。

例)【車】(M-38172)[一][二]-⑦：籀文、轗（M-38586） に作り、古文、𨐈（M-02763） に作る。

M-38172,[1][2]-7,M-38586,–>zhouwen,M-02763,–>ancient,

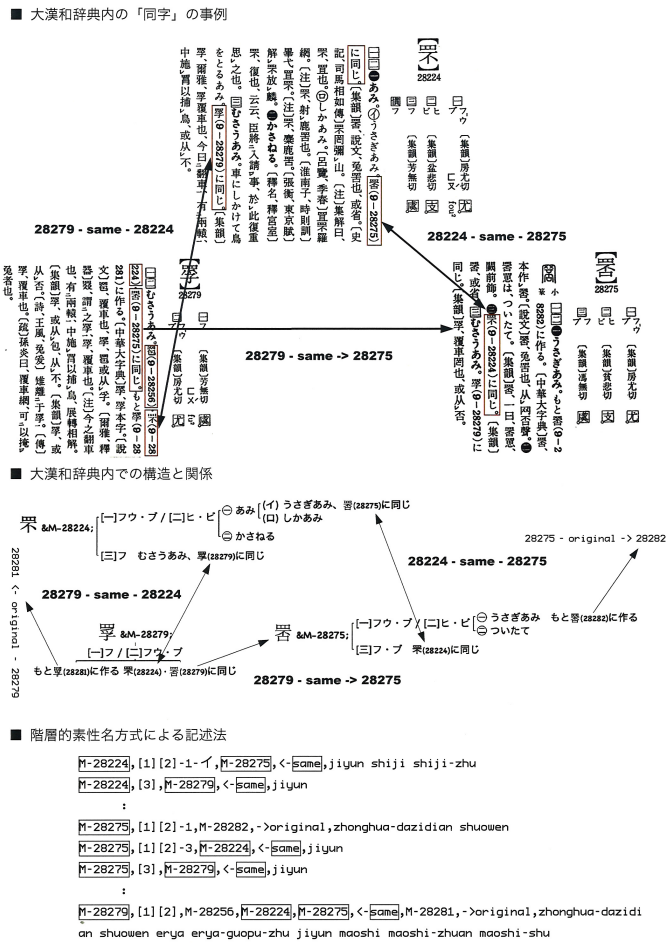

図 1　文字間の関係の例

なお文字間関係の種類と方向性については現在のと

ころ、以下の表のようにしている。

表 1　類字関係等

<–synonyms	〜に通ず；〜の類字
–>synonyms	〜に通ず
–>formed	〜に作る
<–same	〜に同じ
<–simplified	〜の略字
–>simplified	略して〜に作る
<–original	〜の本字；〜の正字
–>original	もと〜に作る
<–vulgar	〜の俗字
–>vulgar	俗に〜に作る
<–wrong	〜の譌字、誤字 (特例/〜に誤り通ず)
–>wrong	誤って〜に作る
<–different	〜の別字
–>different	〜は別字

表 2　字種対応等

<–ancient	〜の古字
–>ancient	古、〜に作る
<–zhouwen	〜の籀文
–>zhouwen	籀文は〜に作る
<–Small-Seal	〜の小篆（篆文）
<–Large-Seal	〜の大篆
<–liwen	〜の隷文、隷書
–>liwen	隷文、隷書は〜に作る

5. おわりに

CHISE で本格的な字義データベースを構築するための準備として、伝統的な漢字辞典における類字記述を『Chaon モデル』および『階層的素性名方式』に基づく試みとして、大漢和辞典のデータ化の中で出てきた諸問題とその解法について述べてきた。まだ 3 巻分 (9 〜 11) のデータにとどまることもあってデータ的にも偏りがあり、今後本稿において論じていない問題も出てくるかもしれないが、ひとまず以上をもって現時点での報告としておく。

参 考 文 献

1) 情報処理学会. 符号化文字基本集合 Basic Subset of Coded Character Set, 2002. 情報処理学会 試行標準 IPSJ-TS 0005:2002