Vol.2022-CH-128 No.12
2022/2/19

圏論に基づく漢字構造記述のモデル化の試み

守岡 知彦[1]

概要：RDF や IPLD のような有向非循環グラフに基づくデータモデルは人文学で扱われる多様な対象を表現する上で有用であり人文情報学分野のさまざまなデータセットで用いられている。構造や関係を扱うための数学である圏論はこうしたデータの構造を扱う上で有用な道具であるといえるが、白須裕之氏の先駆的な仕事などのいくつかの事例はあるものの現状では人文情報学での応用例は少ないといえる。著者は漢字の構造記述のためのモデルとして『多粒度漢字構造モデル』を提案しているが、これは文字と部品間の関係とそれぞれの漢字構造の対応関係の対応に基づくモデルであり自然変換の一種とみなすことができる。ただし、現実にはこうした綺麗な関係がほつれる場合もあり、大部分相似であるが似て非なる多量なデータを扱う必要が生じているのが現状である。しかしながら、自然変換の観点からモデル化することにより、何が対応し何が例外なのかをはっきりさせることにより例外に注力した簡潔な記述が可能になるかもしれない。本項ではこうした圏論の利用の可能性について議論したい。

1. はじめに

人文学では構造と構造の間の関係に関るような対象を扱うことが少なくない。文献学における引用や翻訳の問題は互いに対応関係が認められるが表現形は異なるようなテキスト間の関係を扱っているといえる。そしてこうした対応関係は表現形から概念や意味のレベルまでさまざまなレベルで議論され得る。

著者が扱っている漢字の知識表現の問題においても、形音義のそれぞれの側面で、文字全体と部品、あるいは、形態素とそれを指示する異なる時代・地域（の異なる字体標準）の字体といったものは差異をはらんだ緩やかな対応関係を描くものと考えることができる。

計算機上に漢字の知識表現を載せる場合、現実にあった差異を無視できない以上、実用例ベースでデータ化せざるを得ないが、それは多量の実用例を採取しないといけないことを意味する上、それでも実際にありえた（また、これからありうる）全ての現象を網羅できない。また、書記言語も音声言語と同様に時代・地域によって変化することを鑑みれば、実用例の集合をどうカテゴライズするかという問題も生じ、また、そのカテゴライズが実現できたとして、任意の時代・地域を網羅することの困難さに直面することになる。それゆえになんらかのモデルが必要となるといえる。そして、そのモデルは漢字と部品、あるいは、異なる時代・地域間の字体の対応関係に着目し帰納的に構成しつつも演繹的に利用可能なものでなければならない。

一般に文字は多様な字形で書かれ得るものであるが、漢字は文字の種類が多く、長い期間・広い地域で使われてきたため、任意の２字形を取り出した時、それらの文字が同一かどうかを判定するのが容易ではないことが少なくない。「大」と「犬」のように、点の有無のような小さな差異が字音・字義の異なる別字になることもある一方、「丈」と「丈」や「類」と「類」のように、こうした差異が捨象されることも少なくない。こうしたことから漢字の符号化には『字体の包摂』という概念と『包摂除外』という例外的運用則が用いられる。前者は似た形の部品を同一視するための規則であり、後者はその例外である。前者は演繹的に利用可能なものであり、後者は字の弁別や字源、ソースとなった文字集合等の実例に基づくものといえる。

また、『字体の包摂』という概念は理論的には楷書体の歴史に関する漢字字体史研究に基づくものと考えることができる。特に、藤枝晃らによる敦煌写本の研究 [3][4] に基づき石塚晴通らが作成した「漢字字体規範史データベース」(HNG DB)[6] は藤枝晃らの経験則・仮説を実際のデータによって実証した点で画期的なものといえる。[5] また、著者は項書き換え系に基づいた包摂規準のモデル化を行っており [7][*1]、試験的な実装も行っている。このように、楷書体の字体史研究の歴史において、人文学的研究に基づき帰納的に構成された演繹的にも利用可能な計算可能なモデルは一応の実現を見たと考えられるが、その数学的な基礎付け

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

[*1] 項書き換え系の理論に基づくより妥当なモデル化については [11] を参照のこと。

にはまだ課題が多く残されていると思われる。

藤枝晃・石塚晴通らによる唐代の楷書体に初唐標準字体(『書写体の楷書』)と開成石經規範(『字典体の楷書』)という２つの字体標準／規範があったとする仮説・モデルは資料に関するコディコロジー的な側面と字体に関する側面の対応関係の分析に基づいているが、同様に、人文学研究においては同じ対象を異なる側面で切り取ったものや異なる対象、ある文献に関連する異なる写本・版本の集合、人間関係といった、さまざまな対象間の比較や対応関係の分析を行うことがある。こうした対象は多かれ少なかれ構造を持っているが、その構造が必ずしも斉一であるとはいえず、また、本当は綺麗な規則性を持っていても研究段階ではそれが発見されておらずデータをうまく構造化できないことも少なくない。

そうした半構造データを記述する上で XML のような木構造のマークアップ形式、あるいは、RDF や IPLD のような有向非循環グラフに基づくデータモデルは有用であり、XML や RDF は実際に人文情報学分野のさまざまなデータセットで用いられている。こうしたデータの構造を扱う上で圏論は有望な道具といえるが、白須裕之氏の先駆的な一連の仕事 [9][10] などはあるものの人文情報学での応用例は少ないといえる。

著者は漢字の構造記述のためのモデルとして『多粒度漢字構造モデル』を提案しているが、これは文字と部品間の関係とそれぞれの漢字構造の対応関係の対応に基づくモデルであり自然変換の一種とみなすことができる。ただし、前述のように、現実にはこうした綺麗な関係がほつれる場合もあり、大部分相似であるが似て非なる多量なデータを扱う必要が生じているのが現状である。しかしながら、自然変換の観点からモデル化することにより、何が対応し何が例外なのかをはっきりさせることにより例外に注力した簡潔な記述が可能になるかもしれない。本項ではこうした圏論の利用の可能性について議論したい。

2. 漢字と漢字構造の自然変換

ここでは、文字と対応する UCS 符号位置を集合、漢字構造記述とそれを UCS で符号化した数列をリストの集合で表現することを想定し、[2] の第 5.3.1 節の例を流用して議論を進める。

> 集合 X に対し、$\mathrm{List}(X)$ は成分が X の要素であるようなリスト全てからなる集合である。この時、X を $\mathrm{List}(X)$ に移す関手 $\mathrm{List}:\mathbf{Set}\to\mathbf{Set}$ がある。射 $f:X\to Y$ が与えられたとき、リストのそれぞれの成分に f を適用することによって、成分が X に属するリストを成分が Y に属するリストに変換できる。この処理をリストの翻訳と呼ぶ。

例えば、文字オブジェクトの集合 X があった時、その各要素の UCS の符号位置の集合 Y を考え、X の各要素に対して UCS の抽象文字の符号位置を返す関数 $f:X\to Y$ があるとする。

例えば、

$$X=[花,花,⺾,艹,化,化,亻,匕,匕,⿰,⿱]$$

ならば、

$$\begin{aligned}Y=[&\#\mathtt{x82B1},\#\mathtt{x82B1},\\&\#\mathtt{x8279},\#\mathtt{x8279},\#\mathtt{x5316},\#\mathtt{x5316},\\&\#\mathtt{x4EBB},\#\mathtt{x5315},\#\mathtt{x5315},\\&\#\mathtt{x2FF0},\#\mathtt{x2FF1}]\end{aligned}$$

となる。

また、集合 X を X のリストのリストの集合に移す関手 $\mathrm{List}\circ\mathrm{List}:\mathbf{Set}\to\mathbf{Set}$ を考えることができる。例えば、

$$X=[花,花,⺾,艹,化,化,亻,匕,匕,⿰,⿱]$$

ならば、

$$[[花,⿰],[花,花,花,化,⺾],[化]]$$

や

$$[[]]$$

や

$$[[⿱],[],[花,花,花],[亻,匕]]$$

のような要素が $\mathrm{List}\circ\mathrm{List}:\mathbf{Set}\to\mathbf{Set}$ に属する。

これを使って、漢字に対応する漢字構造記述を表現することを考える。これも X のリストのリストの集合に移す関手であり、但し、その要素の要素は [] か IDS に限定されるものとする。即ち、

$$[[⿱,艹,花],[⿰,艹,化],[⿱,艹,⺾]]$$

や

$$[[],[⿱,艹,化],[⿰,亻,花],[⿱,⺾,化]]$$

や

$$[[],[]]$$

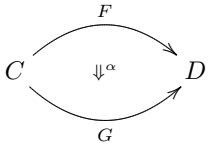
のような要素がこれに属する。

この漢字から漢字構造記述への変換は、下記の例のように、文字をまず漢字構造記述に変換してから UCS の抽象文字の符号位置に翻訳しても、文字をまず UCS の抽象文字の符号位置に翻訳してから漢字構造記述に変換しても同じ結果が得られる。これをこの漢字から漢字構造記述への変換が翻訳に対して自然であるという。

$$\begin{array}{ccc}[花,花] & \xmapsto{\ \mu x\ } & [[⿱,艹,化],[⿱,艹,化]]\\ \Big\downarrow{\scriptstyle \mathrm{List}(f)} & \checkmark & \Big\downarrow{\scriptstyle \mathrm{List}(f)\circ\mathrm{List}(f)}\\ [\#\mathtt{x82B1},\#\mathtt{x82B1}] & \xmapsto[\mu y]{} & [[\#\mathtt{x2FF1},\#\mathtt{x8279},\#\mathtt{x5316}],\ldots]\end{array}$$

定義：自然変換

C, D を圏とし、$F : C \to D$ と $G : C \to D$ を関手とする。$\alpha : F \to G$ と表記される F から G への自然変換 (natural transformation) α は



と図示し、

コンポーネント　それぞれの対象 $X \in \mathrm{Ob}(C)$ に対する D の射 $\alpha x : F(X) \to G(X)$. これを X における α のコンポーネント (component) と呼ぶ。

という構成要素が次の規則

自然性の規則　(natural transformation law)

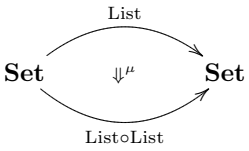
C の全ての射 $f : X \to Y$ に対して、f の自然性 (naturality) の図式と呼ばれる次の図式

$$\begin{array}{ccc} F(X) & \xrightarrow{\alpha x} & G(X) \\ {\scriptstyle F(f)}\downarrow & \swarrow & \downarrow{\scriptstyle G(f)} \\ F(Y) & \xrightarrow[\alpha y]{} & G(Y) \end{array}$$

が可換でなければならない。

を満たすものとする。

前述の漢字から漢字構造記述への変換の例がこの自然変換の定義に合致するかを確認する。圏 C, D はともに **Set** であり、関手 $F : C \to D$ は List で関手 $G : C \to D$ は List $\circ$ List である。自然変換は $\mu : \mathrm{List} \circ \mathrm{List} \to \mathrm{List}$ である。これは



と図示できる。

3. 字体から抽象文字への自然変換

この自然変換の定義に基づき、字体粒度の文字オブジェクトと抽象文字粒度の文字オブジェクトの漢字構造の対応について考える。

例えば、字体粒度のオブジェクトの集合 J_z があった時、その各要素に対応する抽象文字粒度の文字オブジェクトの集合 J_a を考え、J_z の各要素に対して対応する抽象文字粒度の文字オブジェクトを返す関数 $f_a : J_z \to J_a$ があるとする。

例えば、

$$J_z = [\text{花}, \text{花}, \text{艹}, \text{艹}, \text{化}, \text{化}, \text{亻}, \text{匕}, \text{匕}, \text{⿴}, \text{⿱}]$$

ならば、

$$J_a = [\langle\text{花}\rangle, \langle\text{花}\rangle, \langle\text{艹}\rangle, \langle\text{艹}\rangle, \langle\text{化}\rangle, \langle\text{化}\rangle, \langle\text{亻}\rangle, \langle\text{匕}\rangle, \langle\text{匕}\rangle, \langle\text{⿴}\rangle, \langle\text{⿱}\rangle]]$$

となる。

また、文字から漢字構造記述への変換は 2 節のものを流用する。

この漢字から漢字構造記述への変換は、下記の例のように、文字をまず漢字構造記述に変換してから抽象文字に翻訳しても、文字をまず抽象文字に翻訳してから漢字構造記述に変換しても同じ結果が得られる。

$$\begin{array}{ccc} [\text{花}, \text{花}] & \xrightarrow{\mu_{J_z}} & [[\text{⿱}, \text{艹}, \text{化}], [\text{⿱}, \text{艹}, \text{化}]] \\ {\scriptstyle \mathrm{List}(f_a)}\downarrow & \swarrow & \downarrow{\scriptstyle \mathrm{List}(f_a)\circ\mathrm{List}(f_a)} \\ [\langle\text{花}\rangle, \langle\text{花}\rangle] & \xrightarrow[\mu_{J_a}]{} & [[\langle\text{⿱}\rangle, \langle\text{艹}\rangle, \langle\text{化}\rangle], ...] \end{array}$$

4. 多粒度漢字構造モデル再考

著者は複数の包摂粒度をサポートする漢字の構造記述のためのモデルとして『多粒度漢字構造モデル』（図 1）[1] を提案し、CHISE 文字オントロジーにおける漢字の記述においてこのモデルを適用してきた。これは文字単位の包

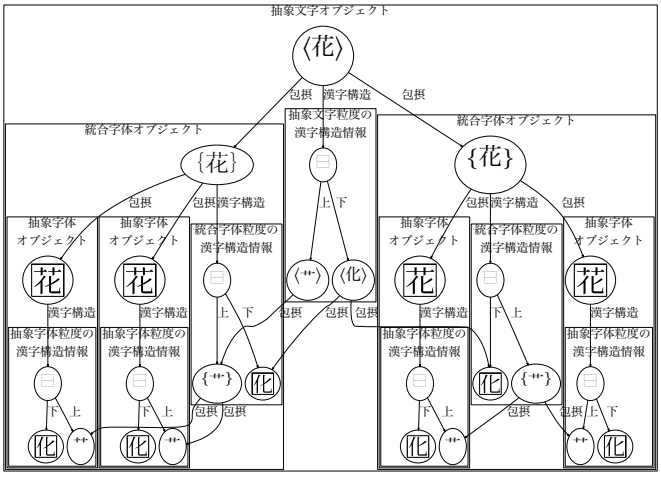


図 1　多粒度漢字構造モデルの概念図（花）

摂関係とそれらの漢字構造記述中の部品の包摂関係が対応するように整合性をとって漢字の構造記述を行うためのモデルであるが一見すると複雑に見える。しかしながら、これが含意することは抽象文字や字体、字形といった異なる包摂粒度間で漢字構造が対応する（同様である）ことと、文字単位での包摂関係と部品単位での包摂関係が対応するということを示しているに過ぎない。

つまり、3 節の例における字体粒度から抽象文字粒度への翻訳を拡張し、字形粒度から抽象字形粒度、抽象字形粒度から詳細字体粒度、詳細字体粒度から字体粒度、字体粒度から統合字体粒度、統合字体粒度から抽象文字粒度、抽象文字粒度から超抽象文字粒度とその上下に翻訳の層を付

けたものと見なすことができる。それぞれの層の翻訳は 2 節や 3 節で述べたのと同様な方法で構成でき、それぞれ自然変換をなす。

しかしながら、漢字の包摂関係やある包摂粒度の漢字と漢字構造の関係をばらばらに記述した時、このような綺麗な関係は一般には成立しない。

例えば、2 節の例の文字オブジェクトの集合 X

$$X = [花, 花, ⺾, ⺾, 化, 化, 亻, 匕, 匕, □, ⿰]$$

に対応する UCS の抽象文字の符号位置の集合 Y'

$$\begin{aligned} Y' = [&\#\texttt{x82B1}, \#\texttt{x82B1},\\ &\#\texttt{x8279}, \#\texttt{x8279}, \#\texttt{x5316}, \#\texttt{x5316},\\ &\#\texttt{x4EBB}, \#\texttt{x5315}, \#\texttt{x2090E},\\ &\#\texttt{x2FF0}, \#\texttt{x2FF1}] \end{aligned}$$

のように〈匕〉(U+5315) と〈匕〉(U+2090E) を別の抽象文字とした場合、例えば、集合 J_a のリスト [化, 化] の各要素を先に抽象文字化すると [〈化〉,〈化〉] になり、これを漢字構造記述を表現するリストのリストに変換すると

[[〈⿰〉,〈亻〉,〈匕 (U+5315)〉], [〈⿰〉,〈亻〉,〈匕 (U+5315)〉]]

になるのに対し、先に漢字構造記述を表現するリストのリストに変換すると [[⿰,亻, 匕], [⿰,亻, 匕] になりこれの各要素を抽象文字化すると

[[〈⿰〉,〈亻〉,〈匕 (U+5315)〉], [〈⿰〉,〈亻〉,〈匕 (U+2090E)〉]]

となるので自然変換が成立しない。

この問題を解決するためには、〈匕〉 と〈匕〉 を包摂する抽象部品〈匕・匕〉 を定義すれば良い。[*2]

このように、自然変換に基づくモデルを置き、そのような関係が成立するように記述することによって、抽象文字と抽象漢字構造から字体の漢字構造記述を導いたり、逆に互いに対応関係にある異なる字体の漢字構造から抽象漢字構造を導くといったことが可能になり、記述の省力化や頑健性を増すことに寄与するといえる。ただ、その一方で、現実にその例外となる事例やこうしたモデルを置いたことによって設けられた不自然な部品や関係が生ずることもある。[*3]

5. 説文小篆と現代漢字の対応

[8] では、現代漢字の漢字構造記述と現代漢字と説文小篆の対応関係から説文小篆の漢字構造記述を生成する手法を提案しているが、2 節 や 3 節の場合と同様に捉えることができる。

例えば、現代漢字（の抽象文字粒度）のオブジェクトの集合 J_m があった時、その各要素に対応する説文小篆の文字オブジェクトの集合 J_s を考え、J_m の各要素に対して対応する抽象文字粒度の文字オブジェクトを返す関数 $f_s : J_m \to J_s$ があるとする。なお、ここでは字体の包摂粒度は考えないこととするため、抽象文字粒度を示す〈 〉は省略する。

例えば、

$$J_m = [説, 言, 兑, 說, 兌, ⿰, □]$$

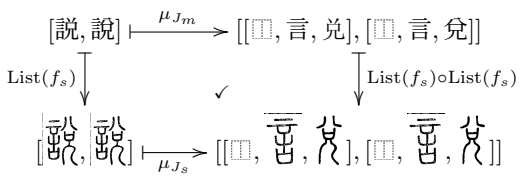
ならば、

$$J_s = [\text{  }, \text{  }, \text{  }, ⿰, □]$$

となる。

また、文字から漢字構造記述への変換は 2 節のものを流用する。

この漢字から漢字構造記述への変換は、下記の例のように、現代文字をまず漢字構造記述に変換してから説文小篆に翻訳しても、現代文字をまず説文小篆に翻訳してから漢字構造記述に変換しても同じ結果が得られる。



プラクティカルな利点としては、こうした自然変換が概ね成り立つと仮定できる場合、既知の情報から未知の情報の候補を自動生成したり、そのクロスチェックのためのワークフローを考えるための手がかりが得られると考えられる。例えば、説文小篆の漢字構造記述の場合、[8] でも議論しているように、

(1) 現代漢字から漢字構造情報記述への対応関係 (μJ_m)
(2) 現代漢字から説文小篆への対応関係 (f_s)

が既知であれば、説文小篆からその漢字構造情報記述への対応関係 (μJ_s) の候補となる写像（関数）を map 関数等を用いて具体的に構成できるということが判る。即ち、関数 f_s を用いて現代漢字の漢字構造情報記述中の各部品を説文小篆のものに置き換えれば良い訳である。現実には説文小篆に対して直接生成した漢字構造記述との間で不一致点があるはずだが、この自然変換のほつれを解消するようにデータを修正すればよい訳である。つまり、これで生成された説文小篆の漢字構造記述を説文小篆から直接生成した漢字構造記述と比較して不一致点がなくなるように関数 f_s や漢字構造記述 μ を修正すれば良いことが判る。

[8] で議論しているように、もちろん、そのためには現代漢字と説文小篆の間の対応関係を適切に記述したり、現代

[*2] この操作を『完備化』と呼ぶことにする。[7] で述べた『包摂規準の完備化』に対応する操作だと思われる。
[*3] こうした便宜上表われてしまったものを不自然に思えるのは実用例に表われない人工物だからであるが、実用例から帰納的に導かれたモデルと考えれば、説文部首以来の伝統的な産物と考えることもできるかも知れない。

漢字と説文小篆の漢字構造がちゃんと対応するように工夫する必要があり、『多粒度漢字構造モデル』の場合と同様に例外の問題や便宜上表われてしまうものをどうするかという問題が生じ得る。しかしながら、その代わりに現代漢字の漢字構造記述の曖昧性に一定の制約をかけることで記述の一貫性を保ったり、隷変時に生じたバリエーションをまとめたり、初唐標準字体から開成石經規範への変化時における楷書字体に対する説文小篆の影響を記述するなど、記述の頑健性やデータの応用において一定の意義があると考えられる。また、自然変換が綻びる部分にこそ重要であると考えれば、それを見つけるための道具としても有望であると考えられる。

6. おわりに

CHISE における現代漢字の包摂関係の記述や現代漢字と説文小篆の対応関係の記述を例に圏論、特に、自然変換の利用を試みた。しかしながら、著者の圏論に対する理解は非常に浅く誤っている部分も多々あると思われる。もし、そうした部分を見つけたら是非御指摘頂きたい。また、漢字構造記述の定式化について十分に議論することができなかった（ここではとりあえず文字のリストとして扱うことを想定しているが、より適切な定式化がありうると思われる）。

本稿を書こうと思ったのは人文学研究や人文情報学の研究のための道具としての圏論の可能性を議論することであり、上述の例はその具体例を意図したものであるが、著者の能力が不十分なため、それが十分に示せなかったことは残念である。ただ、圏論を用いることで、漠然とした対応関係に対して具体的な変換を構成することで、抽象的な視点と具体的なデータを行き来するためのツールとしての可能性がいくばくなりとも伝われば幸いである。

一方、著者が参考にした [2] では自然科学系に偏っているとはいえ、数学以外の多彩な対象において圏論を使うためのさまざまなレシピが載っており、特に、グラフやデータベースに関する説明は人文学においても非常に有益なものだと思われる。とりわけ、本書の著者の David I. Spivak 氏が提唱するデータベースやデータベーススキーマの圏論的な定式化は大変興味深く、等式論理との統合や実装[*4]の進展がどうなるかなども気になる所である。

参考文献

[1] Tomohiko Morioka. Multiple-policy character annotation based on CHISE. *Journal of the Japanese Association for Digital Humanities*, Vol. 1, No. 1, pp. 86–106, 2015 年 11 月.
[2] David I. Spivak, 川辺 治之 [訳]. みんなの圏論: 演習中心アプローチ (Category Theory for the Sciences). 共立出版, 2021 年 10 月.
[3] 藤枝晃. 敦煌写本の編年研究. 学術月報, Vol. 24, No. 12, pp. 7–11, 1972 年.
[4] 藤枝晃. 文字の文化史. 講談社, 1999 年 12 月.
[5] 池田証壽. 漢字字体史の資料と方法 : 初唐の宮廷写経と日本の古辞書. 北海道大学文学研究科紀要, Vol. 150, pp. 201–236, 2016 年 12 月.
[6] 石塚晴通, 池田証寿, 岡墻裕剛. 漢字字体規範データベースとその応用. 東洋学へのコンピューター利用 第 17 回研究セミナー, 全国文献・情報センター人文社会科学学術セミナーシリーズ、京都大学学術情報メディアセンター 第 78 回研究セミナー, pp. 53–63, 2006 年 3 月.
[7] 守岡知彦. 項書き換え系を用いた漢字字体の包摂規準の形式化の試み. 情報処理学会論文誌, Vol. 59, No. 2, pp. 332–340, 2018 年 2 月.
[8] 守岡知彦. 説文小篆に対する漢字構造記述の試み. 東洋学へのコンピューター利用 第 34 回研究セミナー, pp. 17–24, 2021 年 7 月.
[9] 白須裕之. 文字の指示概念に関する試論. じんもんこん 2008 論文集, pp. 311–318, 2008 年 12 月.
[10] 白須裕之. 古辞書のテクストアーカイブズ構築について—文字転写の理論とその応用—. じんもんこん 2011 論文集, pp. 395–402, 2011 年 12 月.
[11] 白須裕之. 漢字構造の代数的記述についての予備的考察. 東洋学へのコンピューター利用 第 30 回研究セミナー, pp. 129–138, 2019 年 3 月.

[*4] https://www.categoricaldata.net/