CHISE に基づくグリフ・オントロジーの試み

守 岡 知 彦[†1]

機械可読な（構造化された）グリフ・コーパスを実現し、書記言語としての構造に即したグリフ情報の分析を行うことはグリフに関わる問題を考える上で重要だといえる。このためには、グリフおよびその規範意識に関わるようなさまざまな要素を機械可読に表現したデータベースやオントロジーを実現することが重要である。本論文では、グリフ・オントロジーの構成法について議論するとともに、CHISE の文字オントロジーに基づいた具体的な記述法について述べる。また、グリフ・コーパスの形態素解析について議論し、文法的・意味的情報や文脈依存性にも考慮した多面的なグリフ解析のための手法を提案する。

Glyph ontology based on CHISE

MORIOKA Tomohiko[†1]

To realize machine-readable (well structured) glyph corpora and analyze glyph information based on structures of written language are important to research problems about glyphs. To realize them, we need machine-readable database and/or ontology about glyphs and various ideals, concepts and terms about glyphs. This paper describes required factors and construction of glyph ontology, and representation in CHISE. In addition, we discuss morphological analysis of glyph corpora and propose multimodal processing which can treat grammatical/semantical information seamlessly and support context dependency.

1. はじめに

常用漢字改訂に関わる議論でも明らかなように、漢字を考える上で字体・字形（ここではこの2つを総称して『グリフ』と呼ぶことにする）の規範に関わる問題は避けては通れないといえる。しかしながら、グリフおよびその規範意識に関わるようなさまざまな要素を機械可読に表現したデータベースやオントロジーの類はこれまでほとんど作られてこなかったといえる。もちろん、字形情報自体は画像データとして盛んに電子化されてきているし、そうしたデータベースは多いとはいえないものの存在していて、実証的文字研究をする上での重要なツールとなっている。しかしながら、視覚的な情報としての字形情報に関するメタデータ、例えば、どこがどう違っているのか？ どういう意識で変えたのか、どれが似ていてどれが似ていないのか(違和感を感じるのか)、といった視覚的な情報の構造やその意味に関わるような情報の機械可読化はあまり進んでいないといえる。そのため、こうした事項は人間が目で見て判断するしかないが、漢字は文字数が多くこうした作業は大変な手間である。また、グリフが置かれた文脈情報まで含めて扱うことは難しく、こうしたことから、1文字単位の頻度情報に頼ることが多くなってしまったといえる。しかしながら、書記言語もまた自然言語の一種であると考えられ、書記系としての構造（文法）を持ち、出現頻度の少ないものが重要な意味を担い得るような性格を持っていると考えられる。こうしたことから、機械可読な（構造化された）グリフ・コーパスを実現し、書記言語としての構造に即したグリフ情報の分析を行うことがグリフに関わる問題を考える上で重要だといえる。そこで、本論では構造化されたグリフ・コーパスの実現の基礎となるようなグリフ・オントロジーについて考察する。

2. 漢字の知識表現

漢字文献の電子化を考える場合、その論理構造や内容に関わる側面と同時にその視覚的な構造に関わる側面の双方を適切に扱うことが重要であるが、このためにはこの両者の関係をどのように記述するかということが問題となってくる。いわゆる異体字の問題というのはそれを文字レベルで見たものだといえるが、このようなことは語彙レベルでも生じてくるし、文書の

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

論理構造と媒体の視覚的・物理的構造の対応のようなレベルでもあり得る。ただ、その基礎となるのはやはり文字であり、文字をユニットとして文字によって構成されるさまざまな上位階層の問題を記述し得るといえる。

このように、論理構造や内容に関わる側面と同時にその視覚的な構造に関わる側面とのインターフェイスとして漢字を捉えた場合、視点や記述の荒さ／細かさ等によって、幾つかのアスペクトを考えることができる。いわゆる『形』『音』『義』というのは漢字に対する視点の一種であるといえる。こうした視点で漢字を見た時、それぞれの視点での切断面はしばしばグラデーションを作る。例えば、『形』の場合、字形（デザイン差）レベル、字体（文字の抽象的な視覚的表現）レベル、字体の包摂レベル、もっと包摂したレベル、…、を考えることができる。各レベルの境界がはっきりしていれば、各レベルを別のレイヤーとして分けて考えることができ、問題を単純化することができるのであるが、実際には、どこに境界を設けるかは一般には恣意的であるといえ、ある程度共通の規範がある部分もあれば、揺れている部分もあると考えられる。また、仮に規範があったとしても、それは時代や地域、コミュニティ等に依存し、変化するようなものと考えられる。結局、漢字には各視点（モーダル）毎に具象から抽象へのグラデーションを描く多次元空間上の場のようなものといえ、文字概念や観念、書かれた文字、発音、観念上の音、意図、解釈、運用、といったさまざまな要素や現象等はその場の中の点や領域として捉えることができる。

漢字の知識表現とはこうした場を記述するということだといえる。この際、重要なのは、解釈や規範の揺れなどで変化しにくいものを基準に記述することと、互いに矛盾するような解釈や規範等であっても両立するような枠組を用いることである。また、記述対象となる場は各アスペクトがしばしば連続的なグラデーションを描くようなものであり、あるいは、幸いにして明確な境界で切断できるようなものだとしても各要素の組合せは無数に存在し得るようなものとなり得るが、記述というものは有限（でかつ、なるべく少数）のものでなければならない。このため、内包的記述と外延的記述をうまく組み合わせることが重要である。このことは言い替えれば、比較的客観的に観測される『書かれたもの』の有限個の記述と、それらを代表（標本点）する解釈や規範、抽象文字といった無限のインスタンスを内包する概念・観念をどう関連付け、どう運用するかということである。著者が提案する Chaon モデルやそれに基づき著者らが開発している文字処理環境 CHISE[1)2)] ではこうした問題意識に基づき漢字を扱おうとしている。

3. グリフ・コーパス

グリフに関わるような問題、例えば、どういう時にどういうグリフの差異が別字として書き分けられ弁別されるのかだとか、どういう時に対応する異体字と看做されるのか、あるいは、グリフの規範意識といったものを考える場合、そのグリフが用いられた文脈を考慮する必要があるといえる。このためには、グリフのためのコーパス（これを『グリフ・コーパス』と呼ぶことにする）が有用である。

グリフ・コーパスというのはグリフを用いられたテキストと対応付けた形でデータ化したものであり、透明文字付き文字画像[3)] や画像マークアップされた文字画像[4)] というのはその一種と考えることができる。文字列としてのコーパスと対比させて考えれば、前者はプレイン・テキストに相当し、後者はマークアップ・テキストに相当すると考えられる。単なる文字画像もグリフ・コーパスの一種と考えるが、現在の古典中国語テキストに対する文字認識技術を前提にした場合、文字に分節化されていないままでは精度の点で難があるといえ、文字単位に切り出され、その切り出された各文字の位置関係が判り、また、各文字がどういうグリフであるかということを示し得るような情報が付与されていることが望ましい。こういう観点からいえば、検索のためになるべく異体字を正規化した透明文字付き文字画像はグリフを指し示す際の精度という点で難がある。（異体字を正規化した）抽象文字はグリフにとって重要な素性のひとつといえるが、同じ抽象文字に対応するグリフ間の差異を指し示すことができない。こうしたものは現状では基本的に目で見て判断するしかなく、大量のデータに対して機械的に処理することが困難である。

4. グリフ・オントロジーとは

グリフ・オントロジーとは、文字の視覚的な側面（ここでは、これを『グリフ』と呼ぶことにする）、および、それが表すセマンティクスを機械に理解させることを目的にした知識表現である。

画像データとしての字形情報はグリフを表現したものといえるが、ビットマップ画像やアウトライン・フォントでのベクトル表現等の場合、画像データそのものは人間にとっての構造や意味を機械可読に表現していないので、画像データからグリフの意味論に関わるよ

うな情報を読み取るには、画像理解や文字認識技術を用いない限り、人が見て判断するしかなく、グリフのセマンティクスを機械可読に表現したものとはいえない。また、画像理解や文字認識技術を用いるためには、グリフ（の特徴ベクトル）とそのセマンティクスの対応関係に関わるデータが必要であるといえ、結局、グリフのセマンティクスに関わる情報の機械可読な表現を定義し、データ化する必要があるといえる。

グリフ・コーパスとの関わりから考えれば、グリフ・オントロジーはグリフ・コーパスを構成するための基礎を与えたりグリフ・コーパスの情報を別の情報と関連付けるためのインターフェイスの役割を果たすような知識表現と考えることができる。

5. グリフ・オントロジーの構成要素

グリフ・オントロジーは、

(1) 字形やグリフの指示
(2) グリフの視覚的構造
(3) グリフのグループ化
(4) グリフの性質
(5) グリフに関する諸概念の記述

といった要素によって構成されると考えられる。

(1) は物や画像データ等で表現されるような字形、あるいは、抽象的（さまざまな粒度の）グリフをどのように指示するかということである。これは、URI やグリフ ID のようなものによって名前付けすることにより実現できる。

(2) はグリフの視覚的構造をどのように機械可読な形で表現するかということである。漢字の場合、どういう部品がどのように組み合わされているかという情報（漢字構造情報）が有用である。

(3) は具象的なグリフと抽象的なグリフの包摂関係をどう記述するかということである。2 節で述べたように、グリフの抽象／具象関係は字形、字体、抽象文字をはじめとしてさまざまな粒度が考えられ、無数の階層を作り得るといえる。また、漢字の場合、部品として使われた場合、その位置や結合関係によって変形することがあるが、一方で、単独の文字と共通する簡略化や異体字関係も存在する。このように、関係の記述においては用途や文脈等を考慮する必要がある。

(4) はグリフに関するその他の性質、メタデータ等である。

(5) は、直接的には、グリフ・オントロジーを記述するために必要な語彙の定義の問題であり、より一般的には、分野・出典毎での用語の対応関係をどのように記述するかという問題としてとらえることができる。

これらの 5 要素は、文字オントロジーの場合と同様に、Chaon モデル[1)2)] に基づき、素性の集合によって表現できるといえる。すなわち、

(a) 一般素性
(b) 構造素性（e.g. ideographic-structure 素性）
(c) 関係素性
(d) ID 素性
(e) 素性の素性

を用い、(1) は (c), (2) は (b), (3) は (c), (4) は (a), (5) は (e) によって記述できると考えられる。各素性は『階層的素性名』を用いることにより、用途や文脈、出典等を考慮した記述が可能である。

5.1 グリフのリソース

物として存在したり、あるいは、電子的な画像データとして表現された字形はグリフに関する一次的な資料のひとつであるといえる。

あるいは、Adobe-Japan1 のような符号化されたグリフ集合は汎用的である程度抽象化されたグリフを表しているといえ、これもまたグリフに関するリソースの一種であるといえる。

あるいは、符号化文字集合はその規格が規定する抽象文字レベルのリソースとして利用できるとともに、規格表の例示字形のリソースとしても利用できる。

このようなグリフのリソースは、電子化された情報であれ、物や場所であれ、適切に名前付けすることにより、URI 等を用いて指し示すことができるといえる。[5)]

また、Chaon モデルに基づいた場合、リソース（の種類）に対して固有の素性名を対応付けることにより、素性名と値の組として扱うことができる。

5.2 グリフ間の関係

ある抽象的なグリフ（例えば字体）がそれよりも具象的なグリフ（例えば字形）を包摂する範囲や関係を記述することを考えた場合、その両者を適切に区別できる必要がある。

例えば、JIS X 0208:1997 の任意の符号位置を考えた場合、その例示字形とその包摂範囲は別のものであり、前者は字形レベル、後者は字体を包摂したレベルである。実際にはその中間的なレベルや JIS X 0208:1997 や UCS では包摂されないようなものを包摂するような複数の抽象文字を包摂したような『超抽象文字』のレベルも考える必要があるし、字形のデザイン差の問題を考えれば、字形レベルに関しても複数の段階が必要かも知れない。

こうした異なる粒度のグリフ間の包摂関係は、グリフ間の関係として一般化して考えることができる。グ

リフ間の関係は有向グラフを用いて表現可能であり、CHISE（あるいはそれを一般化した Concord[6]）や RDF (Resource Description Framework) 等が利用可能である。

CHISE 文字オントロジーの場合、グリフ間の関係は関係素性を用いることによって表現可能である。CHISE 文字オントロジーでは、原則として、字形レベル以下の包摂関係には->subsumptive, 字体レベル以上の包摂関係には->denotational を用いることにしている。

ところで、グリフ間の関係は包摂関係以外の関係も記述可能であり、また、『階層的素性名方式』を用いることにより、用途や文脈、出典等に依存した関係を記述したりそのメタ情報を記述することも可能である。

5.3 派生グリフの名前付け

基準となるグリフ・リソースに対して、そこから関係素性を用いて、そこから派生したグリフ記述を行い、グリフ間の意味ネットワークを記述することができる。このネットワーク中のノードとなるさまざまな粒度のグリフは、Chaon モデルに基づき、その性質を表した素性の集合を付与することで、ID によらず検索することも可能である。

しかしながら、もし、基準となるグリフ・リソースに ID が付与されているならば、それを元に派生的な ID を対応付け、その ID を用いて名前解決を行った方がより効率的な検索が可能になるといえる。

こうした観点に基づき、CHISE では、あるベースとなるグリフ・リソースの名前に対して、さまざまな粒度の派生的なグリフ ID 素性名を対応付けるための命名規則を導入した。これは、次のようなものである：

あるグリフ・リソース *foo* に対し、例示字形の素性名を =*foo*, 抽象文字（字体の包摂レベル）の素性名を =>*foo*, 抽象字体レベルの素性名を =>>*foo*, 抽象字形レベルの素性名を =>>>*foo*, 以下、それよりも細かいデザイン差等に関するレベルは必要に応じて = の後の > の個数を増やすことで表すこととする。一方、抽象文字を包摂したレベルは ==>*foo*, 以下、それよりも抽象度の高いレベルは必要に応じて > の前の = の個数を増やすことで表すこととする。

このような命名規則を用いることにより、あるベースとなるグリフ・リソースの名前に対してその例示字形、（抽象）字形、（抽象）字体、抽象文字、超抽象文字等の無数のレベルを一貫した方式で指定することが可能である。

5.4 グリフの視覚的構造の表現

グリフを表現するための素性としては、文字認識で用いられるヒストグラムといった特徴量もあるが、この種のものは機械可読ではあるものの、人間にとっての可読性が低く、したがって、人間にとっての意味を必ずしも反映しないといえ、グリフを弁別し指示するための素性としてはあまり適切ではないと考えられる。

一方、IDS (Ideographic Description Sequence)[7] のような漢字の部品の組合せ方を表現したもの（これを『漢字構造表現』と呼ぶことにする）がある。これはグリフ表現という観点では必ずしも全ての漢字を表現できる訳ではないという点で問題があるものの、多くの漢字を表現することができ、また、人間にとっての理解に近いという点で優れている。本来の IDS は部品として UCS の統合漢字および漢字部品を用いることになっているが、漢字構造表現自体は必ずしもそれに限定されるものではなく、部品を指し示すことができるならばどのような部品を使っても成り立ち得るといえ、実際、CHISE では CHISE の文字オントロジーにある任意の文字オブジェクトを利用できるように拡張している。[8] この拡張された漢字構造情報では、部品として抽象文字をとることもできるし、字形レベルのものをとることができ、異なるレベルの部品を混在することもできる。この結果、グリフ間の包摂関係やどの部分の微細な差異に着目しているのかといったことなども表現できる。

6. 漢字処理の多層化

6.1 文字と形態素

古典中国語においては 1 文字からなる形態素が少なくなく、文字と形態素はさまざまな面で形式的に重なって見えるといえる。形態素辞書における各種素性は、形態素が 1 文字の場合、文字素性の一種と看做すことが可能であり、実際、字義や発音に関わるような情報は漢字辞書に記載される主要な項目のひとつであるといえる。品詞をはじめとする文法的な情報もまた同様に考えることができるだろう。

漢字処理において重要な問題のひとつである異体字の問題もまた文字と形態素の双方の領域にまたがる問題のひとつといえる。例えば、常用漢字を中心とする現代日本語表記における漢字（以下、『新字』とする）をいわゆる『康熙体』を中心とする伝統的な漢字（以下、『旧字』とする）に変換する場合、文字単位に新字を旧字に変換するのでは一意に決定できなくなったり不適切な変換をすることになる訳だが、このことは異体字の問題の幾つかは形態素や語彙の世界における表記の揺れとして捉えなければならないことを意味している。漢字の異体字関係は時代や地域、分野、テキス

ト、文脈等に依存するということが知られているが、こうした現象をきちんと捉えるためには、文字の世界だけに閉じて記述することはできず、形態素解析のための文法コーパスのように、対象となる文字／形態素をそれが出現する文脈を含んだ形で記述する必要があるといえる。[9] 形態素解析のための辞書やコーパスと文字オントロジーを連携させるということは、形態素解析にとっても有用なことであるが、文字処理の側から見ても重要なことだといえる。

2文字以上からなる形態素の場合、それを文字と看做すことはできない訳であるが、素性の集合で文字を表すという Chaon モデルの方法は、実のところ、概念一般に対する知識表現の一種に他ならなく、文字に限定されるものではないので、文字の場合と同様のやり方で形態素のオントロジーを記述することは可能である。[*1] もし、CHISE の文字オントロジーと同様な方法で形態素のオントロジーを構成すれば、文字と形態素の差異を考慮しつつ、両者をシームレスに扱うことが可能になると考えられる。

6.2 グリフ・コーパスの形態素解析

6.1 節で述べたような文字処理と形態素解析の連携やそのためのオントロジーの統合といったことは、グリフと形態素に対しても成り立つといえる。グリフに関わる規範意識やグリフ間の対応関係、どの差異に着目するかという問題はしばしば単一の文字だけで論じられるものではなく、グリフが用いられた文脈や文法的、意味的、視覚的構造なども含めて考える必要があるといえる。こうしたことを鑑みれば、文字レベルだけで考えるのではなく、形態素をはじめとする上位層の構造を含めた表現や処理が有用なのではないかと考えられる。そのための第一歩として、グリフ・コーパスの形態素解析について考えてみる。

グリフ・コーパスの形態素解析は、どこまで細かくグリフを区別するかという問題を無視すれば、透明文字付き画像を形態素レベルの文法的な情報を含む文字画像マークアップ・テキストに変換する問題という風にとらえることができる。これは基本的にはグリフの位置情報を管理しながら形態素解析を行うことで実現できるといえる。

前節で述べたように、グリフを表現する手段は複数考えられるが、これは文字オントロジーの場合と同様に、それぞれを素性として表現し、その組合せでグリフを表現することによって、これらを併用したり、ある情報から別の情報を取り出したりといったことが可能になるといえる。こうした方法によってグリフの知識表現を記述した『グリフ・オントロジー』を構成することができるが、これは実のところ、文字オントロジーの一部をなすものといえる。

6.1 節で議論したように、文字レベルのグリフと同様に、形態素レベルのグリフ (列) の知識表現を考えることができるが、これがグリフ・コーパスの形態素解析器における辞書に相当するものといえる。そして、文法的文字画像マークアップ・テキストが文法コーパスに相当するものと考えられる (これを『文法グリフ・コーパス』と呼ぶことにする)。

結局、グリフ・オントロジーと文法グリフ・コーパスを蓄積することによって、グリフ・コーパスの形態素解析が実現できると考えられる。当然のことながら、この情報はこれらがその一部として含むグリフの文字レベルでの頻度情報や文脈的情報、通常の形態素解析のためのデータ、それに関わる異体字情報等を含むといえる。

7. おわりに

本論文では、グリフ・オントロジー、および、その構成法について議論した。本論文ではグリフ・オントロジーを構成するための5要素を抽出するとともに、その Chaon モデルに基づく分類と CHISE における表現形式について述べた。また、グリフ・コーパスの形態素解析について議論し、文法的・意味的情報や文脈依存性にも考慮した多面的なグリフ解析のための手法を提案した。

グリフ・オントロジーを用いることにより、画像マークアップされたグリフ・コーパスの各出現字形を文字オントロジーと関連付けることができるといえる。これにより、テキスト処理や形態素解析等の自然言語処理とグリフ情報の解析をリンクして行うことが可能となり、より多面的な漢字の分析が可能となることが期待される。

残念ながら、現在のところ、古典中国語のための形態素解析器やグリフ・コーパスの整備は十分でなく、本論文で議論したグリフ・コーパスの形態素解析はまだきちんと実現できていない。しかしながら、一見複雑に思われるような複数の要素が絡み合った多面的な知識処理も、オントロジーを介することによって、比較的単純な要素に分解して整理することができるといえる。

---

*1 MeCab の辞書やコーパスの形式も (制限があるとはいえ) 素性の集合で表現されたものである。

参 考 文 献

1) Morioka, T.: CHISE: Character Processing based on Character Ontology, *Large-scale Knowledge Resources (LKR2008)*, LNAI, No.4938, pp.148–162 (2008).
2) 守岡知彦：文字オントロジーに基づく文字処理について, 情処研報, Vol.2006, No.112, pp.25–32 (2006). 2006-CH-72.
3) 安岡孝一：透明テキスト付き画像へのいざない, 東洋学へのコンピューター利用第 14 回研究セミナー, pp.31–42 (2003).
4) 守岡知彦：文字画像のマークアップの試み, 東洋学へのコンピューター利用第 14 回研究セミナー, pp.21–30 (2003).
5) Thompson, H. S.: Web Architecture and Naming for Knowledge Resources, *Large-scale Knowledge Resources (LKR2008)*, LNAI, No.4938, pp.334–343 (2008).
6) 守岡知彦：Concord: プロトタイプ方式のオブジェクト指向データベースの試み, Linux Conference 抄録集, Vol.4 (2006).
7) International Organization for Standardization (ISO): *Information technology — Universal Multiple-Octet Coded Character Set (UCS)* (2003). ISO/IEC 10646:2003.
8) 守岡知彦：CHISE 漢字構造情報データベース, 東洋学へのコンピューター利用第 17 回研究セミナー, pp.93–103 (2006).
9) 守岡知彦：MeCab を用いた古典中国語形態素解析器の改良, 情処研報, No.2009-CH-84 No.3, pp.1–5 (2009).