

思想史としての文字情報処理 問題提起として

師 茂樹*

2004年6月8日

1 はじめに

本シンポジウムは、副題に「過去・現在・未来」とあるように、文字情報処理の問題を、これまであったような技術的・工学的な議論（例えば Windows で使える／使えないといった問題やラウンド・トリップ・コンバージョンの問題など）もしくはある狭いコンテキストに依存した議論（例えば外字の問題や異体字の認定の問題など）に限定することなく、より広い視点から議論することで、これまで疎かになってきた「文字とは何か?」「コンピュータ上で文字を使うということはどういうことなのか?」といった一般的な問題群について考えることを大きな目的とする。

本シンポジウムの報告者に国際化や多言語情報処理、古典データベースなどの開発者が多いのは決して偶然ではない。なぜなら、このような開発の現場では、国境や時間を越えて複数のコンテキストを往還する中で、それぞれに共通する文字の本質もしくはモデルとは何か、という問いに直面し続けることになるからである。またそれと同時に、開発したものを世に問うことを通じて、開発者自身が属しているコンテキスト（現代人である、日本人である云々）を否応なしに自覚させられることになるからである。もし文字の一般モデルというものが見出されるのであれば、このような現場の人々によってではなかろうか。

本シンポジウムでは、歴史的経緯が現在に与えて

いる影響、現在の最新開発状況とそこで起こっている問題点、そして近い将来起こりうる技術的、社会的な問題への展望など、目前の技術的な話題に留まらない、人文・社会科学をも射程に入れた幅広い問題群について報告しあい、討論を行うことで、これまで以上の議論を可能にする文字（情報処理）の科学と、それに基づく新たな文字処理環境が立ち上がってくるのが期待したい。

2 Unicode はどこから来たのか

2.1 Unicode の character 概念

議論の手がかりとして、現代を代表する文字コードである Unicode ([8]) の文字モデルについて見てみよう。ここでは、character という用語が以下のように定義されている。

The Unicode Standard draws a distinction between *characters* and *glyphs*. Characters are the abstract representations of the smallest components of written language that have semantic value. (...) Characters represented by code points. (...) The Unicode Standard deals only with character codes. ([8], p. 15)

この説明からまずわかることは、Unicode で言われる character が、視覚的な表象の差異を捨象した抽象的なものとして規定されている点*¹、そして書

*¹ 同様な考え方は、日本の文字研究においても見られる。樺島忠夫氏は、「同じと判定される文字の集合について、字形の異なりを捨象して得られる文字観念を文字素とよぶ」([11])と述べる。

* 花園大学 (s-moro@hanazono.ac.jp)

記言語の中の最小構成部品とされている点である。この character にひとつのコードポイントが対応することになる。

この中、特に前者は、後に見る他の原則の前提となるものであり、Unicode の設計思想を考える上で非常に重要である。これに関連して、他所では、抽象的な文字を表す用語として abstract character が定義されており ([8], p. 64)、character の意味の一部であるとされる ([8], p. 1365)。abstract character の定義は以下の通りである。

Abstract Character: A unit of information used for the organization, control, or representation of textual data.

- When representing data, the nature of that data is generally symbolic as opposed to some other kind of data (for example, aural or visual). Examples of such symbolic data include letters, ideographs, digits, punctuation, technical symbols, and dingbats.
- An abstract character has no concrete form and should not be confused with a *glyph*.
- An abstract character does not necessarily correspond to what a user thinks of as a “character” and should not be confused with a *grapheme*.
- The abstract characters encoded by the Unicode Standard are known as Unicode abstract characters.
- Abstract characters not directly encoded by the Unicode Standard can often be represented by the use of combining character sequences.

ここでは明確に character が、glyph などの視覚的、具象的、物理的な実体*2を持つ文字ではなく、

*2 場合によっては、ディスプレイに表示されたりプリンタで印字された物理的に存在する glyph を「glyph image」とし、それ以前のデータとしての glyph と区別すること

「抽象的な文字」であるとされている。確かに、我々は視覚的に多少の差がある文字があったとしてもそれを同じ文字として読むことができるので、その意味ではこのような抽象的な文字の考え方も我々の持つ文字観と矛盾するものではない。

2.2 活字棚の末裔か、電信の子孫か

では、このような文字のモデルはどのような経緯で形成されてきたのであろうか。吉目木晴彦氏は、文字コードの比喩的な説明として、活版印刷で用いられた活字棚を用いている。

コンピュータの内部には、色々な文字の字形を集めた活字棚があります。この活字棚に並んだ字形データには、やはり一つ一つ番号が振られています。(…) コンピュータは記録された文章を画面に表示したり、紙に印刷する時、この活字棚から一つ一つ該当する番号の字形データを拾ってくるのです。([17], p. 17)

このように、文字の符号化の概念を印刷 (技術) の方面から捉えようとする見方に対して、安岡孝一氏は次のような批判を述べる。

漢字コードに関する文献を調べていると、しばしばこういう記述に出くわす。このような形で、漢字コードを活字棚にたとえるのは、たしかに世人の興味を引きやすい。しかし、この手の記述は、漢字コードの説明としては非常に危険であり、多くの誤解を生みかねない。というのも漢字コードは、そもそも印刷のために作られたものではなく、したがって本来、活字棚とは何の関係もないからである。(…) 文字を符号 (コード) に変換する手法は、通信技術とともに発展してきた。(…) 漢字コードもまた、通信技術とともに発展してきた。(…) 漢字コードはそもそも情報交換に用いるためのものであり、そこでおこなわれる符号化には、文字の抽象化という概念が不可欠である。したがって、漢字コードを

がある。

印刷という局面で論ずるのは、文字コードの本質を全く理解していない輩の愚行にすぎない。(16)

ここで言われている通信技術における「文字の抽象化という概念」は、先に見た Unicode における character の定義として結実したものであると見てよい。

本シンポジウムにおける安岡氏の発表は、この見解を踏まえて、現在の文字情報処理、テキスト処理が、如何に電信時代の技術(紙テープ)に規制されているかを指摘するものである。紙テープという一次元的な媒体についての批判的検討は、データ形式としてのプレーンテキストや XML など、プログラム言語においても一般的な文脈自由文法的な在り方(世界観?)を問い直すことにもつながり、極めて重要であると考えられる。

2.3 Unicode の音声中心主義

このようなメディア論^{*3}的議論は、現在支配的な文字やテキスト(文字列)のモデルを批判的に捉えるために有効であろう。

しかしながら、なぜ現在のようなモデルが(誤解されつつも)広く受け入れられたのかについては、メディア論的な視点からでは説明が難しい。そこで、ここではもう一本の補助線を引いておきたい。すなわち、文字コードの文字概念の背景に見て取れる音声言語中心主義的な観念である。この点について、再び Unicode の character 概念に注目してみよう。

Unicode における character の定義およびそれに関連する情報は、主に第2章の“10 Unicode Design Principles” ([8], p. 14) にまとまって見られる。

1. Universality
2. Efficiency
3. Characters, Not Glyphs
4. Semantics

^{*3} ここで言う「メディア論」は、マクルーハンらが主張する、メディアによって人間の思考や身体、感覚器官が変質し規定される意味のメディア論であり、ルーマンらの社会システム論におけるそれではない。

5. Plain Text
6. Logical Order
7. Unification
8. Dynamic Composition
9. Equivalent Sequences
10. Convertibility

この中、三番目の「Characters, Not Glyphs」の原則については見たが、この10項目は密接に連係しており、相互の連関において把握しなければならない。ここで注目したいのは、六番目の「Logical Order」である。

Unicode text is stored in *logical order* in the memory representation, roughly corresponding to the order in which text is typed in via the keyboard. In some circumstances, the order of characters differs from this logical order when the text is displayed or printed. (...) For the most part, logical order corresponds to *phonetic order*. ([8], pp. 18–29)

ここでは、Unicode の文字によって構成されるテキストが、メモリ上において「論理的な順序(logical order)」によって並べられなければならない、とされている。そしてこの「論理的な順序」とは、キーボードから入力する順序、もしくは音声言語において発話される際の順序(phonetic order)に近いものであると言う。この「論理的な順序」と、ある状況において対立するのは、画面表示や印刷などにおける視覚的な文字の順序である。アルファベット/英語や漢字/中国語の場合、両者が対立することはないが、デーヴァナーガリ/ヒンドゥー語などの場合には、子音と母音の見た目の位置が通常の文字の流れと逆になるなどの対立が発生することがある。これは、先の原則中であつた character と glyph との分離と重なる内容である。

音声言語的な順序を「logical」と見なす背景には、一般に広く受け入れられている文字言語に対する音声言語の先行性があるのであろう。これに関連

することとして、I. J. Gelb が提唱する「theory of writing」を指摘したい。Gelb は、文字の歴史を絵画的なものからアルファベットのなものへの進化と捉えた上で、音声言語に限りなく近い IPA の発音記号的なものになっていくのが「書記の理論」であるとしている ([4])。

ところで、現代思想に詳しい読者であればすぐに頭に浮かんだであろうが、このような音声言語中心主義を徹底的に批判し、それに代わる grammarology (文字学) を提唱したのが、脱構築で有名なジャック・デリダである ([3])。デリダは、西洋哲学における真理観に共通する現前性に注目し、これを「現前の形而上学」とした。この枠組においては、内的 (自分の声を聞く)、非物質的な音声言語 (parole) に対して、文字言語 (écriture) は外的、物質的性格を持つ二次的で補助的な道具に過ぎない、とされる (音声中心主義)*4。また、エクリチュールの中でも、表音文字 (アルファベット) / 表意文字という二項対立があるという。先に指摘した Gelb の「書記の理論」がデリダによって批判されているのは言うまでもない。

さらに、音声言語は透明であるが故にロゴスを忠実に再現できるとされ (ロゴス中心主義)、善/悪、自然/技術、精神/身体、男性/女性などの二項対立と、後者の前者に対する従属を支配するという。これらの概念は Unicode における character/glyph の分離、対立や、logical order という考え方と重なる

*4 文字を音声言語に対する二次的なものであるとする同様の見解は、文字を研究する学者の間でも共通して見られるものである。池上禎造氏は、「(…) 文字の研究のむつかしい第一の理由は、文字そのものの本性にあるのである。文字は言ふまでもなく、言語を写しとどめるものとして生れた二次記号である。だから常に言語に平行して、それを完全に写し得るものならば、文字だけが特に複雑な相を呈することはないわけである。しかし実はそれでも、言語が変化するものであるからといって、その変化に常に応じて文字が変ずるといふことはあり得ない。(…) 平行しないのみならず、第一完全に写すなどといふことがあり得ないのである。(…) 表音文字の場合ですら、言語を完全に写し得る文字などは無いのである。言葉調子、イントネーションは勿論のこと、アクセントや個々の音韻に対しても一対一の関係に徹することなどなかなか困難なのである」と述べ、文字の独自性を指摘しつつも「二次記号」であるとする ([10])。

るところが多く、分析装置として極めて有効ではないかと考えられる ([15])。小林龍雄氏が報告する予定の、文字コードの制定をめぐる政治的な運動の背景に、以上のような二項対立を見通すことは容易である。

同様な批判は開発者の立場からも表明されている。Haralambous 氏夫妻は、Unicode に顕著な character と glyph とを分離するモデルに対して、以下のような組版の視点からの批判を述べている。

For us, a glyph is “the image of a typographical sign.” You may object why we use the term “typographical” in our definition. Well, typography has been a first modelization of human writing. Books are based on this modelization (even if in some cultures books are still written by hand) and books are the carriers of human culture. Computers are based on this modelization. (...) Our definition of a character is: “a character is an equivalence class of glyphs, based on a simple, linguistic or logical description.” (...) Our argument is that the bipolarity character/glyph is not sufficient for modelizing text. ([5])

すなわち、組版は手書きをモデル化したものであり、本やコンピュータもまたこのモデルに依っているので、glyph や character の定義もまた組版的文脈でなされるべきであると主張するのである*5。

3 文字情報処理はどこへ行くのか

3.1 オブジェクトとしての文字

メディア論的な観点で言えば、昨今の大規模化し安価になった計算機環境においては、一次元的なデータ構造からの脱却が唱えられても不思議ではな

*5 とは言え、この論文で提示されているモデルは、“Rich Unicode Model”であり、その意味では Unicode の音声中心主義から完全に脱出しているとは言えないだろう。それは氏が Unicode を “rich” にするための典拠を、学術的にすぐれた辞書類に求めていることからわかる。

い。実際、現在コンピュータ上のモデル化の方法として注目を集めているオブジェクト指向は、しばしばマシン・セントリックなモデルからオブジェクト指向へ、という流れで位置付けられることが多い。

文字コードについてはどうだろうか。再び Unicode の 10 大原則に目を向けると、4 番目の原則「Semantics」において、

Characters have well-defined semantics. Characters property tables are provided for use in parsing, sorting, and other algorithms requiring semantic knowledge about the code points. The properties identified by the Unicode Standard include numeric, spacing, combination, and directionality properties (...). Additional properties may be defined as needed from time to time. ([8], pp. 17–18)

と述べられている通り、数字の持つ数値、文字の占める幅、組み合わせて用いる文字か否か、書記方向で変化する振る舞い方などといった文字プロパティが、各 character に対して定義されている。つまり、Unicode は文字とコードポイントとの対応を定義するだけでなく、各文字の性質や操作方法についても定義しているのである。これはすでに「オブジェクト」とよんでも差し支えないものであろう。

本シンポジウムにおける川幡太一氏や風間一洋氏による報告は、オブジェクトとしての Unicode の文字を処理するシステムを開発する立場からなされるものである。特に風間氏の報告においては、オブジェクト指向言語である Java における文字オブジェクト（文字クラス）の可能性について議論がなされる予定であり、興味深い*6。

*6 スクリプト言語 Ruby の開発者のメーリングリストにおいても、文字オブジェクトの必要性についての議論があった。<http://blade.nagaokaut.ac.jp/cgi-bin/scat.rb/ruby/ruby-dev/11450> 前後のスレッドを参照。

3.2 新しい文字のモデルは可能か

ここで注目すべきは、「semantics」が「well-defined」とされている点ではないかと思う。これはすなわち、「semantics」として適切でないもの、曖昧なものを排除していることに他ならない。言い換えればこれは文字にとって何が本質的で何が本質的でないか、ということをあらかじめ規定しているということであろう。ここから報告者（師）はかつて、Unicode の文字概念がアリストテレスの本質主義に類似するのではないかと指摘したことがある ([15])。これに関連して、Java などの（クラススペースの）オブジェクト指向言語は、しばしばアリストテレスなどの思想との類似性が論じられる ([7, 12, 14] など)。そこから、先に見たデリダ的観点からすれば、オブジェクト指向は音声中心主義、ロゴス中心主義の延長線上にあると言える。

では、デリダが考える文字 (écriture) とはどのようなものであろうか。少し長いですが、東浩紀氏による要約を引用しよう。

前期デリダはいくつかの論文で、「同じ mème」と「同一的 identique」という二つの形容詞を峻別し用いている。(...) 論文「署名コンテキスト」でデリダは、「署名の同じものの性 [mêmeté]」こそが、署名の同一性 [identité] と単独性とを変質させることで署名の封印を分割する」と記している。ここで「署名」という語は、記号一般がもつ性質を際立たせるために用いられたものだと考えてよい。記号は反復される。しかしそれは同一のものではない。それぞれの記号は、反復されるたびに異なったコンテキストに規定されるからだ。それはいわゆる「署名」が、反復されつつも、また書くたびに異なった筆跡で記されるのと類比的な現象である。(...)

「同一性」はコンテキストにより与えられる。それゆえ同じ記号でも、異なったコンテキスト内であればそれらはもはや同一ではない。しかし「同じものの性」はそれとは異なる。デリダの用語法においてはそれは、記号の「反

復可能な、繰り返し可能な、模倣可能なひとつの形態」を指示する。記号は記号である以上、つねにこの形態的な反復可能性に支えられている。そしてこれはエクリチュールの観念と等しい。なぜなら、エクリチュールが記号に与える引用可能性、つまり記号が本来のコンテキストから「断絶」する力は、異なった複数のコンテキストを貫いてひとつの記号が「同じ」であり続ける可能性により保障されるからだ。(…)

ここでひとつの隠喩を導入しよう。英語の war やドイツ語の war は、それぞれ意味＝同一性に満たされている。対してエクリチュール「war」はそうではない。「war」は諸言語のあいだを循環し、つねに「同じもの」であり続けながら異なった同一性を受け取る。七〇年代以降のデリダは、エクリチュールのこの「受容体」的特徴、「根底的な処女性」を名指すためしばしば「コーラ khôra」という隠喩を用いている。それはプラトンの対話篇『ティマイオス』で用いられたギリシア語であり、一般には場所、容器、苗床、国家などを意味する。この隠喩においては、ひとつのエクリチュールが複数のコンテキストに同時に属することは、ひとつの容器コーラに対し複数のコンテキストが同時に意味を流し込むこととして捉えられる。英語とドイツ語は「war」というひとつの容器を分かちあう。(9), p. 35-38)

デリダの指摘するような、エクリチュールの複コンテキストに対する多重所属については、日本の文字研究においても指摘されたことである。

音声言語はそもそもある一定の現実の場面に話手と聴手がいて、その間に行われる伝達の重要な手段である。その伝達は従って直接的伝達である。音声は発せられても時々刻々に雲散霧消して迹を残さない。(…) 文字言語は主として間接的伝達の役に廻る。(…) 文字が間接的伝達に役立つのは、文字という視覚記

号が聴覚記号に比べると恒久性があるからである。文字による記録は書かれた瞬間に消えるようなものではない。(…)

恐らく最初は空間的な、いわば横の連絡のためのものであったであろう。しかしそれはやがて縦の連絡、すなわち時間的距離を隔てての連絡にも役立つことになった。(…)

(…) 音声に依る言語は性質上、場面に依存する度合いが大きい。(…) これに対し、文字言語ではそうはいかない。もちろん、文字言語といえども言語である以上は、それ特有の言語的場の中で行われる。(…) しかし音声言語を支えている現実の場面が欠けており、原則として書手と読手は同一場面に参与しない。(…)

音声言語の変遷は前の世代から後の世代へと連続して継承して行く間に次第に現れてくるが、文字言語の変化は非連続的である。([13], p. 5-7)

このように、複数のコンテキストに多重に所属し、「反復可能な、繰り返し可能な、模倣可能な」écriture をコンピュータ上でモデル化することは可能であろうか。

既存のモデルの中で思い付くのは、プロトタイプベース ([6, 1]) のオブジェクト指向である。プロトタイプベースにおいては、クラスベースとは異なりクラス A のインスタンスがクラス B のインスタンスにもなり得るので、言わばクラスへの多重所属といったものが表現可能かもしれない。

本シンポジウムにおける守岡知彦氏の発表では、クラスベースではない文字オブジェクトを用いた文字のモデル化と実装について、様々なトライ&エラーの経験を交えつつ報告して頂く予定である。

4 最後に

以上、若干の論点について主に思想史的な観点から問題提起を行った。言うまでもなく、ここで尽くせなかった多くの論点が残されており、今後の検討が必要である。

また、この報告ではデリダのécriture 論に基づく文字のモデルの可能性について述べてみたが、これはこの方向性が唯一の道だと言うのではない。現在、唯一無二のモデルである文字コードのモデルを相対化するための強力な思想的装置として、デリダが有効であることを示すことが目的である。

いずれにせよ、チョムスキー派の言語研究が記述的妥当性から説明的妥当性へと歩みを進めたように、文字（情報処理）の現場で起きている事例に真摯に耳を傾けつつ、それに拘泥することなく文字（情報処理）の一般モデルを学際的に検討していくことが求められるのではないだろうか。最後に、チョムスキーの言葉を引いて、自戒としたいと思う。

何となく地に足がついているとような安心感があるのだと思います。抽象化というものに危惧を持ち、データから決して離れまいとする人がいます。... 人文科学や自然科学における研究活動を見れば、ごくわずかの例外を除いては、データに依っている度合いが非常に高いことがわかります。([2], p. 56)

参考文献

- [1] プロトタイプベース・オブジェクト指向。Online (Wiki). <http://sumim.no-ip.com:8080/wiki/493>.
- [2] Noam Chomsky. 生成文法の企て。岩波書店, Nov 2003. 福井直樹、辻子美保子訳。
- [3] Jacques Derrida. *De la Grammatologie*. Collection Critique. Minuit, Paris, 1967. 足立和浩訳。根源の彼方へ グラマトロジーについて。現代思潮社, 1972.
- [4] Ignace J. Gelb. *A Study of Writing*. University of Chicago Press, revised edition, 1963.
- [5] Tereza Haralambous and Yannis Haralambous. Characters, glyphs and beyond. 「書体・組版ワークショップ」報告書, Feb 2004.
- [6] Kamran Parsaye, et al. *Intelligent Databases: Object Oriented, Deductive Hy-*
permedia Technologies. John Wiley and Sons, 1989. 近谷英昭訳。知的データベース—オブジェクト指向・演繹・ハイパーメディア, オーム社, 1992.
- [7] Derek Rayside and Gerard T. Campbell. An Aristotelian understanding of object-oriented programming. In *Proceedings of the 15th ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications*, pp. 337–353, 2000.
- [8] The Unicode Consortium. *The Unicode Standard, Version 4.0*. Addison-Wesley, Boston, 2003.
- [9] 東浩紀. 存在論的、郵便的 ジャック・デリダについて。新潮社, 1998.
- [10] 池上禎造. 文字論のために。国語学, Vol. 23, , 1955.
- [11] 樺島忠夫. 文字の体系と構造。岩波講座日本語, Vol. 8, , 1977.
- [12] 河合昭男. オブジェクト指向と哲学。Online. <http://www1.u-netsurf.ne.jp/~Kawai/>.
- [13] 河野六郎. 文字論。三省堂, 1994.
- [14] 早川てつろう. アリストテレス的オブジェクト指向入門。Online, 2003. http://itpro.nikkeibp.co.jp/NIP/nip04_bn.jsp?BN=Y&OFFSET=1.
- [15] 師茂樹. Surface or Essence: Beyond the Coded Character Set Model. 「書体・組版ワークショップ」報告書, pp. 26–35, Feb 2004.
- [16] 安岡孝一. 漢字と漢字コードのはざまで。人文科学研究のフロンティア 京都大学人文科学研究所要覧, pp. 46–47, 2001.
- [17] 吉目木晴彦. いま、何が、なぜ、問われているのか。電腦文化と漢字のゆくえ 岐路に立つ日本語, pp. 7–55. 平凡社, 1998.