人文学のためのテキストデータ構造化のチュートリアル

第2章

デジタル翻刻における課題

永崎研宣

version 1.0

2026.3.21 作成

本資料は、文部科学省委託事業「人文学・社会科学のＤＸ化に向けた研究開発推進事業」（JPMXP1624）において、学校法人慶應義塾が、大学共同利用機関法人人間文化研究機構から再委託を受けて作成したものです。

1．はじめに

テキストデータの構築にあたっては、テキストの構造以前に、そのテキストを構成する個々の文字をどう扱うか、それらをどういう風にデジタルメディアへと移して（写して）いくか、という課題がある。写本であれば、文字の形の様々な揺れについて、どういう違いまでは同じ文字とみなすか、どういう違いなら区別するか、という判断が必要になる。テキストデータにおいては、文字が同一かどうかということは文字コードの相違という形で確実に判別できてしまうため、文字の形の揺れに対して何らかの判定を行なわねばならないのである。また、古いテキスト資料においてはコンピュータでは表示できない字形も未だに登場する場合がある。そのような場合に、どう対処すべきか、対処の際にかかるコストはどう見積るべきか。本章では、文字をデジタルメディアに転移・転写することをデジタル翻刻[1)]と呼ぶことにした上で、その際の課題について検討する。

2．文字と文字コード

デジタル環境において文字を扱うためには、文字を何らかの形で数値に対応づける必要がある。この対応関係を定めたものが文字コードである。コンピュータは文字そのものを理解しているわけではなく、あくまで数値として処理しているため、「どの数値がどの文字を表すのか」をあらかじめ決めておかなければ、文字の保存・表示・通信は成り立たない。

文字コードの歴史を振り返ると、初期のコンピュータでは、英語圏を中心とした限られた文字集合のみが想定されていた。代表的な例として、ASCII（American Standard Code for Information Interchange）がある。ASCII では、アルファベット、数字、基本的な記号などが定義されていたが、日本語や中国語、アラビア文字などは含まれていなかった。その後、各国・各地域で独自の文字コードが作られ、たとえば日本では JIS コード、中国では GB 系コード、台湾では Big5 などが用いられるようになった。

しかし、こうした状況は大きな問題を抱えていた。異なる文字コードを用いる環境同士では、テキストを正しく交換できず、同じ数値がまったく異なる文字として解釈されることもあった。研究データを国際的に共有し、長期的に保存するという観点から見れば、文字コードが国際的に乱立していたことは深刻な障害となっていた。

この問題を解決するために構想されたのが、世界中の文字を単一の体系で扱おうとする Unicode である。Unicode は、あらゆる言語・文字体系を包含することを目標とし、各文字に一意の番号を割り当てることで、文字コードの共通基盤を提供する。現在では、アルファベット、日本語、中国語、アラビア文字、チベット文字、ヒエログリフ等の歴史的文字体系から絵文字に

至るまで、15 万字を超える文字が定義されている。

Unicode を理解するうえで重要なのは、Unicode が扱っているのは「文字の見た目」ではなく、「文字という抽象的な単位」であるという点である。Unicode では、同じ意味・機能を持つ文字であれば、フォントや書風、細かな字形差があっても、原則として同一の文字として扱われる。つまり、Unicode が規定しているのは字形（glyph）ではなく、文字（character）である。この設計思想は、人文学のなかでもテキストを意味や言語構造の単位として扱う分野の基本的な立場とはよく対応している。

Unicode では、各文字にコードポイントと呼ばれる一意の番号が割り当てられている。これは 16 進数で表現されることが多いが、たとえば、「A」なら U+0041、「あ」なら U+3042、「学」U+5B66 といった形で、それぞれ異なるコードポイントを持つ。コードポイントは文字を識別するための抽象的な識別子であり、実際にどのような字形で表示されるかは、フォントや表示環境に委ねられている。この分離によって、Unicode は OS やソフトウェア、言語環境を越えたテキストの互換性を実現している。

これに関して留意すべきなのは、Unicode が単なる民間団体や業界団体によって恣意的に決められている仕様ではない、という点である。Unicode Standard の策定を担っているのは Unicode Consortium であるが、Unicode で使用されている文字コード表、すなわち文字レパートリーと符号化体系は、国際標準化機構（ISO）および国際電気標準会議（IEC）の合同技術委員会である ISO/IEC JTC1 の下に設置された ISO/IEC JTC1/SC2/WG2 によって策定・維持されている。Unicode の文字集合は、ISO/IEC 10646（Universal Coded Character Set）として国際標準として正式に承認されており、Unicode Standard と ISO/IEC 10646 は、実質的に同一の文字レパートリーを共有する関係にある。

さらに、東アジアの文字文化に深く関わる漢字については、より専門的な議論の場が設けられている。漢字（CJK 統合漢字）に関する詳細な検討は、SC2/WG2 の下に設置された Ideographic Research Group（IRG）において行われている。IRG には、日本、中国、台湾、韓国、ベトナムなどの専門家が参加し、新しい漢字の追加、既存文字の統合の可否、字形差を独立した文字として扱うかどうかといった問題について、文献学・書誌学・文字史的な根拠に基づいて慎重な審議が行われている。Unicode における漢字の扱いは、単なる技術的判断や企業主導の決定ではなく、各地域の文字文化と学術的知見を背景とした国際的合意形成の成果なのである。

とはいえ、Unicode は、文字に関するすべての課題を解決しているわけではないということには注意が必要である。人文学研究、とりわけ歴史資料や写本、古典籍を対象とする研究では、Unicode に未収録の文字や、研究上は区別すべき重要な字形差が存在することが少なくない。そのため、Unicode を基盤としつつも、異体字注記、外字管理、文字画像への参照などを併用する必要がある。この点で、Unicode は十分条件ではなく、あくまで必要条件として位置づけられるべきである。

Unicode はまた、純粋に技術的な規格というよりも、文化的・学術的な妥協の産物でもある。

どの文字を独立した文字として認め、どこまでの差異を「同一」とみなすかという判断は、常に歴史的・地域的背景を伴う。その意味で、Unicode を用いることは、標準化の恩恵を受けると同時に、その限界と前提を理解したうえで文字資料を扱うという、批判的な姿勢を研究者に要求する。

このように、文字コード一般から Unicode を位置づけてみると、Unicode は単なる文字コード表ではなく、文字文化をデジタル環境で共有するための国際的・公共的インフラであることが分かる。人文学研究において Unicode を理解することは、テキストをデータとして扱う技術的前提を知ることにとどまらず、文字とは何か、どのような制度的枠組みのもとで標準化されているのかを理解することにほかならない。その理解の上に、TEI などの構造化記述や画像・注釈を組み合わせることで、はじめて学術的に信頼できるデジタル・テキストが成立するのである。

3．元資料とテキストデータは同じではない

ボーンデジタル、つまり、最初からデジタルデータとして作成されたテキストには生じないが、デジタル以外の媒体から翻刻されたテキストデータには大きな課題がある。それは、「元の媒体上でのテキストと完全に同じではない」ということだ。

まず、字形や文字の大きさが完全に同じになることはあり得ない。これについては、この違いが内容の相違につながることにはならない場合が多いため、問題になることはあまりない。しかし、読み手側が受ける印象には少し違いが出てくる場合もあるだろう。近年の資料であれば、目が悪い人向けに少し大きな文字にしているかどうか、あるいは、読字障害者向けにユニバーサルデザイン（UD）フォントを使っているかどうか、ということも、やはり読み手を意識した時にはやや大きな違いとなるだろう。古い資料であっても、くずし字であればまったく同じ字形を再現することは困難であり、漢字であっても時代が異なれば字体も字形も様々である。テキストの内容を分析する場合でも、テキストに対して読み手がどう考えたかということが研究に含まれるのであれば、そのあたりの差異も関わってくることがあるかもしれない。さらに言えば、石に彫られたものと紙に印刷されたものとでは受ける印象は大きく異なると思われるが、石から文字起しされたプレーンなテキストデータにはそういった相違も特に反映されることはない。

また、字体の違いはかなり大きな問題になることもある。旧字体と新字体を丸めてどちらかに統一してしまうことは現在もよく行われるようだが、その中には、筆者が異なる文字として使い分けたはずの 2 つの字を、異体字関係であると判断して同一の文字に変更してしまうということもあるかもしれない。望ましいことを言えば、そのような使い分けはそのまま文字起こしした上で、その使い分けにどういう意味があるかを判断するのは、それを読み分析する側であるべきである。しかしながら、その資料を研究対象とする研究者であればそのような差異をきちんと記述しようとする動機づけが十分にあり得るが、たとえば、アルバイトで入力作業を依頼したり、企

業に文字起こしを発注する場合などは、そのような細かな差異に注意しながら作業してもらうことは難しい場合がある。特に後者の場合は、あらかじめ JIS の第〇水準の文字で、という風に仕様書で限定をかけることになるので、どうしても丸めざるを得ないことになる。

ここまでは漢字の話だが、仮名についても丸めたり色々なことをする場合がある。たとえば、源氏物語の研究でよく用いられてきたテキストの一つである『校異源氏物語』では字母の異なる様々な変体仮名を現代のひらがなに丸めてしまっている。一方で、近年は字母の違いを対象にした源氏物語研究も行われており、また、字母の違いを意識しながら変体仮名を Unicode で文字起こしすることがある程度可能になっている。この点は今後大きな研究課題の一つになっていくだろう。

このようなルールベースでの問題とは別に、単なる誤転記という問題もある。これは、最終的には人力でチェックするしかないので、正確性を期すなら非常に難しい。むしろ、「少し間違っていても利用可能なもの」として流通させ利用することを考えるべきかもしれない。そもそも、紙媒体でも誤植が混入することはしばしば生じるのであり、デジタルだけの問題でもないと考えることもできるかもしれない。

4．元資料との関係をどう位置づけるか

ここまでみてきたことに加えて、あらゆる面において元資料とまったく同じものをテキストデータで得ようとするのは不可能である、ということも改めて確認しておきたい。たとえば、テキストが書かれた紙や石といった元資料の任意の箇所の物質的組成を利用者が確認できるようなデジタルコンテンツを作成しようとするなら、その資料のすべての箇所の物質的組成を参照できるようにする必要がある。この場合、技術的には可能だったとしても、現在のストレージではデータ量が膨大過ぎて対応が困難であり、実際に作成し共有することはほとんど不可能である。したがって、特徴的な箇所や典型的な箇所など、なんらかの基準を設けた上で、それに沿っていずれかの箇所の物質的組成の情報を採取して利用に供することになる。また、字形の微細な違いについて、1 文字ごとの異なりをテキストデータとして利用可能な形できちんと再現しようと思ったなら、不可能ではないがかなりの労力を要することになる。さらに、Unicode に登録されていない文字を扱おうとすると、自分の PC や Web ページでは外字画像や Web フォントで表示できるように対応するとしても、サーバ上で検索できるようにしたり、あるいは他の人が作成した外字と共存させることは容易ではなく、一般に広く簡単に利用することは難しい。そのような事柄をすべて解消できたとして、その次に来る、しかしなかなか避けることが難しいのが、誤転記問題ということになる。

このようなことから、上記のような状況を斟酌しつつ、「テキストデータで何をどこまで再現したいか」ということを検討する必要がある。これは「どこまで手間暇（コスト）を費やせるか」

ということでもある。この件に関しては、その時点での技術水準や標準規格の状況による制約も大きく関わってくるため、基盤技術や関連規格の状況が流動的だった頃には 10 年単位で長く通用する具体的な対応策を立てることは難しかったが、近年は徐々に安定してきているように思われる。それについて以下にみてみよう。

5．文字が Unicode に登録されていない場合

5-1．Unicode 未収録文字という問題

まず、使いたい文字がそもそも Unicode の文字コード表に入っていない、という問題が未だに時折聞かれる。この場合、文字を表示せずに「〓」等で済ませるか、当該文字の文字画像かフォントを作って対応するか、後述する XML を使って記述するか、あるいは、Unicode で使えるようにすべくその文字を符号化提案する、という選択肢になる。

5-2．Unicode 符号化提案という選択肢

Unicode に文字が登録されれば、その文字を用いるすべてのテキストが普通のテキストデータとして記述・処理できるようになるため、その文字を扱うすべての研究分野にとって非常に有益である。これを目指す場合、上述のように国際標準化機構の当該委員会に文字の符号化提案をするという時間のかかる手続きを行うことになる。学術研究のために文字を符号化提案するためのルートはいくつか存在するが、いずれにしても、国際標準化機構の委員会での英文文書の交換に基づく議論に参加して、自らが必要とする文字を符号化する必要性を、そこでのルールに従った英文文書で提示しなければならない。これは、短くとも数年を要するプロセスであり、提案書や登録のための議論にもその都度時間をかけて対応しなければならない。

あるいは、現在は Unicode に入っていなくても、現在、符号化提案中となっている場合もある。符号化に関する議論の過程は、議論のための文書がすべて Web で公開されているため、符号化提案をする前に一度確認してみるとよいだろう。漢字に関しては IRG のサイト[2]、それ以外の文字に関しては ISO SC2/WG2 の文書リポジトリ[3]として公開されている。漢字は数年毎に数千字が提案・議論されており、それ以外の文字に関しては、提案されたもののペンディングになっているものも多く、自分が使いたい文字がそこに含まれている場合は、提案者に連絡をとって符号化に向けて議論を進めるべく協力するという方法もあるだろう。

5-3．外字として扱う場合の方法と課題

コストを勘案した結果、Unicode での文字の符号化提案は諦めて、似た文字を使ったり、あるいは「外字」として扱うという方向性も考える必要はあるだろう。ただ、そのようにした場合、その文字が「本来はどういう文字であるか」を示すことがテキストデータだけではなかなか難し

い。その状況を改善する手段の一つとして、TEI ガイドラインの gaiji モジュール[4)] がある。ここでは、文字についての様々な情報を記述した上で、本文中に記した文字に対してその文字情報を付与できる枠組みとなっている。

一方で、データとしての互換性は少し落としつつも字形をうまく表示したいという場合には、外字フォントを作成して表示させるという方法もある。フォントを作成すれば、文字を拡大・縮小した場合にもきれいに表示でき、Web ページでは Web フォントを用いることで、専用のフォントを別途ダウンロードしなくても用意したフォントを自動的に利用して表示できるため、ある程度の利便性は確保できる。フォントの作成は Glyphwiki を用いれば、やや時間はかかるものの比較的容易である。この場合、文字コードとしては、Unicode の Private Use Area (PUA) と呼ばれる文字コード領域を用いてテキストデータを記述することになる。そのため、テキストのコピー&ペーストをした際には同じフォントを用いなければ文字化けしてしまうことになり、また、同じ PUA の文字コード割り当てルールを用いたテキストデータとしか互換性を保てないため、作ったテキストデータを余所で作成されたテキストデータとあわせて幅広く利用しようとする場合にはかなり使いにくくなってしまう。つまり、独自フォントと独自文字割り当てルールを常にテキストデータと一緒に流通させなければならないということになる。そして、ここで設定した文字コード割り当てルールが失われてしまった場合、何が書いてあったのかわからなくなってしまうという難しさもある。この方法を選ぶ場合には、これらの点を踏まえた対応が必要になるだろう。

また、今昔文字鏡や GT 書体等のいわゆる多漢字フォントにおいては、複数のフォントファイルを切り替えることで同じ文字コードに複数の文字を割り当てて多数の文字の表示を実現しているため、文字に対するフォントの情報が失われるとどの文字だったわからなくなってしまうという難しさがある。この場合、テキストデータだけでは文字の情報を残すことができず、ワープロソフト等のフォントの情報を残せるソフトウェアが必須であり、他のソフトウェアにデータを移管する際にもフォントの情報が失われないようにする必要がある。今昔文字鏡は様々な漢字を必要とする研究者の間で広く使われた時期があったものの、上記のような難しさに加えて利用条件の扱いの難しさもあり、利用の際には十分な注意が必要であり、また、すでに作成されたデータを維持したり利用したりする際にも上記のような事情によく留意する必要がある。

外字フォントよりもさらに簡便な方法として、文字の形を表現するために文字画像を用いるという方法もある。この場合、データの扱い方としては外字フォントと同じであり、やはり、外字に番号を割り当て、それに対応する字形を記した表を作成しておく必要がある。Web ページで表示する場合には外字フォントよりも仕組みが簡単だが、ワープロソフト等に貼り付ける場合には不便であり、Web 公開を前提としない場合にはあまりおすすめできない。

5-4．実務的対応と長期的維持の問題

なお、「〓」の利用を除くいずれの場合にも、必要な文字に関しては、独自の文字コード割り当てルールとそれに対応する字形の対応表を作成し、維持していく必要がある。そして、それぞ

れの文字についての周辺情報を記録しておくことが望ましい。この対応表は、文字が少なければそれほどの手間はかからないが、多くなると、ある程度の用意が必要となる。たとえば、SAT 大蔵経データベース研究会では 1 万字を超える外字の対応表を作成・維持しており、これは共同作業可能な Web データベースとして運用されている。また、Unicode に登録するために符号化提案する際には、文字の意味や登場箇所、複数の利用例を撮影したデジタル画像等が求められるため、符号化提案を念頭に置いている場合には、対応表を作成する段階からその種の情報を集積しておくとよいだろう。

このような諸々の手間を省くための一つの方法として、Unicode に含まれる似た文字に置き換えてしまうという方法と、単に「〓」を置いてしまうという方法がある。前者はどれくらい意味内容に影響するかについての検討が必要だが、現実的な選択肢である。どのような置き換えを行ったか、という対応表を作成し用意しておければなおよいだろう。これについては次節も関わってくるので参照されたい。また、「〓」は最終手段ではあるが、後生に託すということで、これも一つの選択肢と考えるべきだろう。

このような、いわゆる外字の扱い方のメリット・デメリットを整理した表を表 1-3-1 としておおまかにまとめたので参照されたい。

<table>
<tr><th></th><th>一律代替文字表記</th><th>文字画像表示</th><th>外字フォント利用</th><th>XML 注記</th><th>Unicode 符号化</th></tr>
<tr><td rowspan="3">メリット</td><td rowspan="2">入力時には時間も手間もかからない</td><td rowspan="2">Web ブラウザ等では比較的うまく表示できる</td><td rowspan="2">対応する環境が整えられればきれいに表示できる</td><td>文字に関する情報を詳細に記述できる</td><td>通常のテキストデータとしてのあらゆる恩恵を受けられる</td></tr>
<tr><td colspan="2">テキストデータとしての持続可能性が高い</td></tr>
<tr><td colspan="4">字形が具体的にわかる</td></tr>
<tr><td colspan="5">そこに何らかの文字があったことはわかる</td></tr>
<tr><td rowspan="4">デメリット</td><td rowspan="4">どんな文字かわからない。
検索ができない。</td><td>画像表示機能が必要。</td><td>他の外字フォントとの共存が困難。</td><td rowspan="2">XML を処理する必要がある。</td><td rowspan="3">手続きに数年間の大きな労力を要する。</td></tr>
<tr><td colspan="2">テキストデータ単体では文字情報を伝えられない</td></tr>
<tr><td colspan="3">検索が難しい</td></tr>
<tr><td colspan="4">独自の文字割り当てルールを策定し独自に配布し続ける必要がある</td></tr>
<tr><td rowspan="2">具体的な手法の例</td><td rowspan="2">「〓」等で済ませる</td><td rowspan="2">文字画像を貼り込む</td><td rowspan="2">外字フォントを作成・利用（多漢字フォントの場合も同様）</td><td rowspan="2">TEI ガイドライン等に準拠して記述</td><td rowspan="2">Unicode に文字として登録</td></tr>
<tr><td>用意に手間がかかる</td></tr>
</table>

使いたい文字が Unicode のコード表に入っていない場合の対応の例（表 1-3-1）

6．字形・字体の相違をいかに扱うか

外字の取扱いについて一定のルールを定めたならば、次に、文字全体をいかなる方針のもとでテキスト化するかを決定する必要がある。あるいは、あらかじめ全体方針を定めた上で、その方針に従って外字の取扱いを決めるという進め方も考えられるであろう。

6-1．字形差への技術的対応

字形の微細な相違については、現在では技術的に、Unicode が提供する IVS（Ideographic Variation Sequence）の仕組みを用いることにより、相当程度まで対応することが可能である。IVS は、Unicode に登録された文字に対して、枝番号を付すことで字形の差異を識別する仕組みであり、この方式に基づいて字形と枝番号を登録したデータベースとして IVD（Ideographic Variation Database）が整備されている。場合によっては、当該データベースの中に、目的とする字形がすでに登録されていることもある。IVS は、macOS や近年の Microsoft Windows において標準的にサポートされており、IVD のリストを確認する手間のみで利用できる点で、比較的現実的な選択肢であると言える。

一方、IVD の中に適切な字形を見出すことができない場合においても、Unicode の枠組みのもとで字形の可用性を確保することを重視するのであれば、IVD への新規登録を Unicode に提案し、自らが必要とする字形を登録するという選択肢も存在する。この場合、対応するフォントを自ら作成する必要があるが、Glyphwiki を用いれば、その作業自体は比較的容易である。ただし、IVD への登録に際しては、提案書の作成をはじめとする一定の手続きが不可欠であり、Unicode 本体への文字登録と比べれば負担は軽いとはいえ、相応の時間と労力を要することに変わりはない。したがって、このようなコストを負担すべきか否かについては、慎重な検討が求められる。

6-2．運用コストと方針決定の問題

このように、技術的にも制度的にも多様な選択肢が用意されているとはいえ、最終的には「どこまでの手間と労力を費やすことが可能か」という問題に帰着する。字形の微細な相違を厳密に反映しようとする場合には、「標準的な字形とは異なるこの字形が、当該資料に出現するすべての同字に共通するものなのか」を確認しなければ、その作業の意義は大きく損なわれてしまう。この確認を人手で行うのか、あるいは画像処理やプログラミングの知識を活用して、原資料をデジタル撮影・スキャンした上で文字画像を切り出し、半自動的に分類・確認するのかによって、作業の性格は大きく異なる。人手による場合には、文字入力や OCR 結果の誤字チェックと並行して確認作業を行うことになるであろう。いずれにしても、容易な作業でないことは明らかである。さらに、その字形が既存のフォントでは表現できない場合には、前述のように IVD への登録を行うか否かについて、改めて判断する必要が生じる。

以上の点を踏まえ、字形の微細な差異に十分な労力を割くことが困難であると判断した場合には、その差異を区別することを断念するという選択も、現実的にはあり得る。ただし、その場合であっても、「当該字形をどの文字と同一視するか」といった対応方針を明示的に定めておく必要が生じることは少なくない。この作業を簡素化するためには、既存の文字コード体系を積極的に活用することが有効である。たとえば、「Unicode の範囲内で」「JIS 第 3 水準までを用いて」「新字体を基本とする」といったように、使用する文字の範囲をあらかじめ定め、その範囲から外れる文字については、対応表を作成した上でデータ入力を行う、という運用が考えられる。

この対応表が大きくなると、いちいち探す羽目になってかえって入力作業が大変になってしまうこともあり、対応表など使わずに Unicode で直接探した方がはやい、という入力者・企業もいるので、元資料の字体・字形の状況や入力担当社のスキル等の案配にも注意が必要である。

なお、検索に際しては、異体字を同時に検索する仕組みを作成することも可能であり、また、後からどちらかの字体に変換することも容易にできるため、旧字体・新字体くらいの違いであれば、元の資料に即した字体で入力しておくと、後々、手戻りの可能性を少なくすることはできる。

以上のようなことを踏まえつつ、テキストデータを作成する際の文字の範囲についてのおおまかな目安を表 1-3-2 としてまとめておいたので参照されたい。

文字の範囲	メリット	デメリット	便利な道具立て
IVS/IVD での対応	・既存の文字と同様に使える。 ・文字の意味を考えず字形で探せば済む。 ・対応可能な字形は非常に増える。	・対応文字を探すのに少し手間が増える。 ・対応フォントが必要。 ・それでも字形がみつからない場合がある。	・異体字セレクタセレクタ、CHISE、Web font、
Unicode での対応	・既存の文字と同様に使える	・対応フォントが必要な場合もある	・CHISE、Unihan データベース、花園明朝フォント
JIS 第 n 水準（n は、通常は 2 ～ 4 が多い）	・入力文字種を減らせる。 ・簡易な文字列検索には便利。	・入力時に対応字を探しにくい場合がある。 ・入力できない文字が生じる可能性が少し高まる。	・MJ 縮退マップ、CHISE、異体字データベース
新字体を用いつつ何らかの範囲で	・入力文字種を減らせる。 ・簡易な文字列検索には便利。 ・入力できる文字は表示にも困らない。	・入力時に対応字を探しにくい場合がある。 ・入力できない文字が生じる可能性がやや高まる。 ・元の資料の字とかなり変わってしまうことがある。	・MJ 縮退マップ、CHISE、異体字データベース

テキスト化する文字の範囲と意義のおおまかな目安（表 1-3-2）

7．文字の扱い方を記録しておく

文字の取扱いに関する一定のルールを定め、それに基づいてテキストデータを作成した場合には、その内容を可能な限り文書として明示的に残しておくことが重要である。すなわち、「なぜ当該文字が出現するのか」「この文字と別の文字とはいかなる関係にあるのか」といった点について、単にテキストデータのみが提供され、原資料からどのような方針や規則に基づいて文字起こしが行われたのかが明示されていない場合、第三者がそのデータを利用する際に、十分な理解を伴わないまま使用せざるを得なくなるおそれがある。

とりわけ、誤字であるのか、あるいは意図的な表記であるのか判断に迷うような箇所については、あらかじめルールが提示されていることによって、読者・利用者は適切な解釈と利用を行うことが可能となる。逆に言えば、そのような情報が付与されていない場合には、データの信頼性が低いと判断され、結果として利用されなくなることも十分に考えられる。

データ作成者以外の者が閲覧した場合であっても、その内容を理解し、適切に利用できるようにデータを整備しておくことは、デジタルデータの長期保存に関する枠組みの国際標準規格である OAIS 参照モデル (ISO 14721) においても示されている、いわば基本的な原則である。こうした情報は、テキストファイルの冒頭に記載する方法のほか、テキストデータを ZIP 形式等で配布する際に、説明文書として同梱することによっても提供し得る。

TEI ガイドライン は、この種の情報を体系的に記述するための要素群を備えており、これを利用することで、情報管理および再利用の利便性を高めることができる。関心のある読者は、特に The TEI Header の章[5] を参照されたい。同章については日本語訳も公開されており[6]、比較的参照しやすい資料となっている。

8．誤転記を含むテキストの扱い

8-1．誤転記の不可避性と分析上の問題

文字をデジタルに転記する方法については、前節までに述べたような手順により進めることができるが、次に問題となるのが、いわゆる誤転記の存在である。人手による入力は、一般にコンピュータによる自動処理よりも正確である場合が多いとはいえ、入力ミスを完全に排除することは困難である。また、OCR や HTR を用いた自動文字認識においても、常に完全な結果が得られるとは限らず、その後に人の目による修正を加えたとしても、誤りが残存する可能性は否定できない。

このような誤転記を含むテキストデータを用いることで、果たしてどの程度意味のある分析や処理が可能なのか、という問題が生じる。この点については、テキストデータの量や処理の目的、

すなわち、統計的分析を行うのか、単純なテキスト検索が可能であれば足りるのか、あるいは検索結果をコピー・アンド・ペーストしてそのまま利用したいのか、といった条件によって判断が大きく異なる。一般に、作業に多くの時間と労力を投入すればするほど、正確性の高いデータを得られることは確かであるが、十分な手間をかけられない場合には、どの点で、どの程度の妥協を許容するかを検討せざるを得ない。

8-2．正確性を重視する場合と分析可能性を重視する場合

この局面では判断が容易でない場合も多いが、基本的には二つの方向性を想定しておくと理解しやすい。一つは、可能な限り正確なテキストデータの作成を目指す方向であり、もう一つは、完全な正確性を必ずしも前提とせず、分析が成立するように分析手法そのものを工夫する方向である。この二つのいずれに、どの程度重きを置くかを決定することが、この段階において重要となる。

前者、すなわち正確なテキストデータの作成を目指す方向は、単に全体として完全無欠なデータを構築することを意味するわけではない。たとえば、特定の情報のみが必要である場合には、その部分に限って人手と機械処理を適切に組み合わせて重点的に確認を行う方法や、あるいは人手による確認作業を徹底するという選択肢も考えられる。

一方、正確性を一定程度犠牲にしてでも分析可能性を重視する方向は、特にデータ量が大きい場合に有効となることが多い。単純なテキスト検索であっても、多少の誤転記が含まれていたとしても、十分な量のテキストデータが存在すれば、必要とする情報に到達できる場合がある。また、統計的分析においても、統計的有意性を示すことが主たる目的である場合には、大規模なテキストデータであれば、一定程度の誤転記が含まれていても分析が成立することがある。このような場合には、対象とするテキストデータ全体の信頼度について、サンプル調査等によってあらかじめ明らかにしておくことで、分析結果の説得力を高めることができるであろう。

9．まとめ

本章では、テキストデータの構造化に先立って、文字起こし／デジタル翻刻の段階で避けて通れない「文字の扱い」を中心に検討した。デジタル環境において文字は文字コードとして表現される以上、元資料に見られる字形・字体の揺れや、Unicode 未収録文字の存在は、必然的に「同一視する／区別する」「代替する／注記する」「提案する／外字として運用する」といった選択を伴う。しかも、その選択は、単なる技術的可否によって自動的に決まるものではなく、作成するデータの目的、利用者層、将来の再利用可能性、ならびに作業に投入し得るコストといった条件に依存する。

Unicode は、国際標準として整備された公共的インフラであり、可能な限り Unicode の枠内で

記述できることは、互換性と持続可能性の観点から望ましい。しかしながら、歴史的資料を対象とする人文学研究においては、Unicode のみでは包摂しきれない字形差や未収録文字が現実に存在し、外字管理、TEI による注記、外字フォントや文字画像による提示、さらには符号化提案といった複数の手段を組み合わせざるを得ない局面が生じる。また、字形差の扱いについても、IVS ／ IVD といった枠組みにより表現可能性は拡大しているものの、実務上は、必要な字形を確認・同定するための作業負担が大きく、どの段階まで精密化を目指すかは、結局のところ運用上の判断に帰着する。

さらに、どのような方式を採用する場合であっても、作成したテキストデータが第三者にとって理解可能であるためには、文字の扱いに関する方針と具体的な運用ルールを文書として残すことが不可欠である。置換規則、外字の対応表、採用した字体・字形範囲、ならびに判断の根拠が明示されていなければ、利用者は誤字と意図的表記とを区別できず、結果としてデータの信頼性が損なわれる。TEI ヘッダ等を用いた記述は、この種の情報を体系的に保持し、再利用の利便性を高めるための有効な手段である。

最後に、誤転記の問題は、人手入力であれ OCR ／ HTR であれ、完全には回避しがたいことを確認した。したがって、精度の最大化を図るのか、あるいは誤りを一定程度含むことを前提に分析方法を工夫するのかという二つの方向性を踏まえ、データ量と目的に応じた現実的な設計が求められる。とくに大規模データを用いる分析では、サンプル調査等によってデータの信頼度を把握し、その前提を明示することが、研究の透明性と説得力を支える。

以上を総合すれば、文字起こし／デジタル翻刻とは、単に文字列を作成する作業ではなく、元資料との関係をどのように設計し、どの程度まで再現し、どのような前提のもとで利用可能性を確保するかを定める営みであると言える。したがって本章の結論は、特定の「唯一の正解」を提示することではなく、目的・制約・標準・運用の四点を見据えつつ、選択と記録を一体として行うことの重要性にある。

さらにこの点を敷衍するならば、ここでは主としてテキスト資料をデジタル媒体に翻刻することについて検討してきたが、その前提として、そもそもデジタルに移行する以前の文字資料を通じた理解がどこまで「うまく」成立してきたのかという点についても、内省的に問い直す余地があるかもしれない。この問題を理論的に捉える枠組みの一つとして、記号論的な視点を挙げることができる。記号論においては、意味が対象そのものに内在するのではなく、何らかの媒介的過程を通じて成立することが理論的に捉えられており、この視点は文字資料の読解を考えるうえでも示唆的である。その理論的基盤の一つとして記号・対象・解釈項から成る三項関係を提示したチャールズ・サンダース・パースの議論を参照するなら、文字は対象そのものではなく、それを媒介する記号であり、解釈項が生成される過程を通じて初めて意味理解が成立する。この枠組みにおいて重要なのは、文字の字形や字体、書写環境、媒体といった要素が、単なる付随的な外観ではなく、解釈項の形成条件として機能しうる点である。すなわち、文字資料の読解とは、対象に直接到達する行為ではなく、個々の歴史的・制度的・技術的条件のもとで構成された記号を介

して成立する理解にほかならない。我々が過去に行ってきた文字資料理解もまた、常に一定の抽象化や正規化を前提とした記号解釈の一形態であり、その意味で「完全な理解」であったと無反省に前提とすることはできないという点は、テキストデータの作成においても改めて考慮しておくべきだろう。そして、資料が用いられた文脈に即した理解が可能になるように、できる範囲で最善を尽くすべきであることもまた、ここから帰結することができる。

注

1 「デジタル翻刻」という表現は、クラウドソーシングの歴史史料文字起こしプロジェクト「みんなで翻刻」にあやかって用いている。

2 https://www.unicode.org/irg/

3 http://www.unicode.org/wg2/WG2-registry.html

4 https://tei-c.org/release/doc/tei-p5-doc/en/html/WD.html

5 https://tei-c.org/release/doc/tei-p5-doc/en/html/HD.html

6 https://www.dh.ku-orcas.kansai-u.ac.jp/?p=791