悉曇文字のデータ化に関する諸問題

―大蔵経テキストデータベース化に伴う悉曇文字作成をめぐって―

小峰智行／下田正弘*／元山公寿**

はじめに

平成19年７月30日、東京ガーデンパレスに於いて大蔵経データベース完成記念大会が開催された。『大正新脩大蔵経』（以下「大正蔵」）をテキストデータベース化するというこの大事業は1994年、東京大学の江島惠教教授が「大蔵経テキストデータベース研究会」（以下「SAT」）を発足させたことによって具体的な作業が開始された。その後、1999年に江島教授が急逝、さらに資金不足によって頓挫の危機に瀕したが、東京大学教授下田正弘が江島教授の意志を引き継ぎ事業を継続、2000年には「大蔵経データベース化支援募金会」が発足してこの危機を免れた。多くの協力者の力もあり、図像部を除いた85巻分について、この度ようやく完成に至ったのである。筆者はこの事業への協力を要請され、悉曇文字の作成と大正大学准教授元山公寿を中心に行われた入力済テキストの校正を担当した。尚、文中の悉曇文字のローマ字表記は、実際の入力作業で用いられた方式に従って表記した。

大蔵経テキストデータベース化の意義

『大正新脩大蔵経』は漢訳経論のみならず、中国・日本撰述の典籍をも網羅した一大叢書であり、世界中の研究者にとって欠かすことのできない重要な叢書である。これをテキストデータベース化することは、コンピューター上での検索・参照を容易にするだけではなく、新たな研究方法が生まれる可能性を秘めている。インターネット上で公開することによって、この膨大な知的財産を世界中の研究者・機関などが共有することになろう。その結果、仏教・歴史・文学・社会・民族など、幅広い学問分野に於いて研究の発展が期待されるのである。また、インターネット上での辞書・辞典などとリンクすることによる大蔵経を中心とした巨大な智慧のネットワーク構築の可能性など、新たな発展も望めると考

えている。

文字の作成について

筆者はSATから悉曇文字作成作業を依頼される数年前より、必要に迫られてパソコン上で少しずつ悉曇文字作成に着手していた。その際に考えた作成の方法に数種ある。

1、手書きの悉曇文字をスキャナーで取り込み、画像データとして保存する。
2、スキャナーで取り込んだ画像データをベクトルデータに変換する。
3、パソコン上でアウトラインを書きベクトルデータとして保存する。

1は恐らく最も一般的で単純な方法であろう。この方法は毛筆による手書きであるため、悉曇文字特有の縦に長く切継いだ文字などの調整がしやすく、また視覚的にも美しいが、書写に技術と手間を要する。また印刷時の文字の大きさを考慮して作成しなくてはならず、どのような大きさにも対応できることを考えて画像データを作成すると一文字ごとのデータが大きくなり、一万字を超える全ての文字を保存した場合、最終的に膨大な容量になることが予想された。

2の方法は1の文字の大きさと容量の問題を解決するものである。ベクトルデータは画像データに比べると格段に容量が小さく、また拡大・縮小しても品質を損ねることはない。この方法の問題点は書写・データ化・アウトライン化と工程が多く、より手間と時間がかかってしまうこと、さらに画像データをベクトルデータに変換する際の精度の問題もあり、最適な方法とはいえなかった。そこで筆者が最終的に選択した方法が3である。

3は毛筆で文字を書写することをせず、初めからアウトラインで文字を作成する方法である。工程は最も少なく、また容量も小さくて済む。パソコン上で文字の部品を作成し、組み合わせていくため、文字の統一感も得られるという利点がある。作成に用いるソフトウェアの操作技術と知識は必要なものの、書写の技術は基本的に不要で、部品さえ揃っていれば分業も可能に思えた。作成には筆者が使い慣れていたadobe社の

「Illustrator」というソフトウェアを用い、『梵字大鑑(1)』所収の「悉曇十八章(2)」と母音から作成を始めたのである。SATから依頼された文字作成についても、この方法を踏襲した。

ベクトルデータはこのような類のデータを作成する際に最も汎用性が高いものと考えている。様々な形式に出力することが容易で精度も高く、前述したように拡大・縮小の際に解像度の影響を受けない。将来的にフォントを作成することになったとしても十分対応ができるものと考えている。恐らく悉曇文字を初めからベクトルデータで作成する試みは筆者の知る限りでは前例がない。当初は分業も可能と考えていたが、現実的な問題として、悉曇文字の知識と「Illustrator」に関する知識及び操作技術という二点については必須であり、また当初は基本的に不要と考えていた悉曇文字の書写技術もやはり必要であることが判明した。文字の視覚的な統一性も重要であるため、文字作成については一人での作業となった。

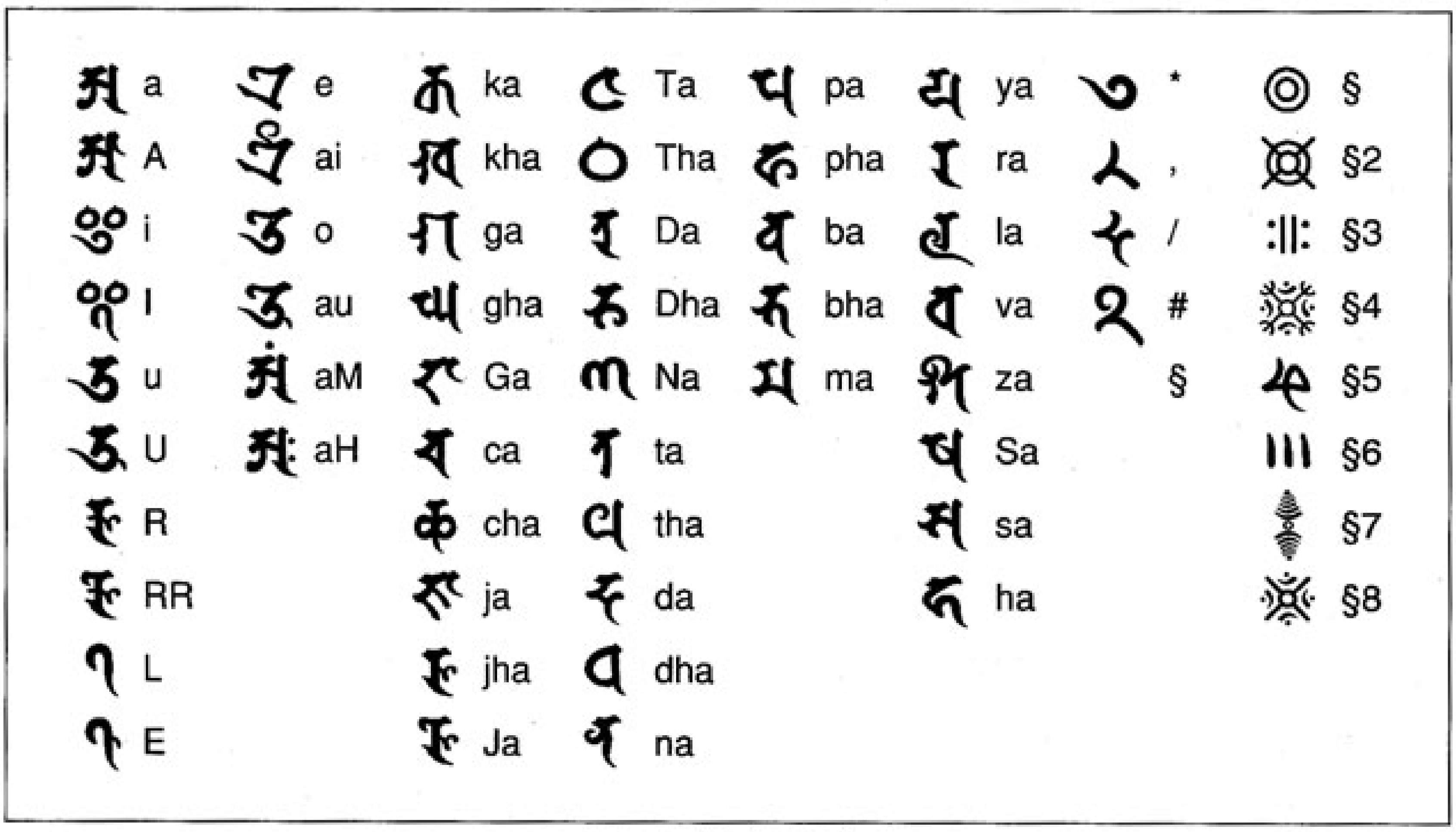
大蔵経の文字

このデータベースを作成するにあたり、様々な問題が明らかになった。その中でも特に大きな問題が、『大正蔵』中に多くの規格外文字が使用されているということである。漢字についても数千もの規格外文字が使用され、それ以外にも悉曇文字が多数含まれる。これらの文字には当然国際規格の文字コードは与えられておらず、フォントも存在しない。

漢字の表示について、SATの文字コードはJISコード（JIS X 0208：1997）が採用されたが、これで処理できない一万数千字に上る文字についてはGT書体(3)を用いるとともに、新たに作成して『諸橋大漢和辞典』などの番号を付与し、マークアップ方式で参照する方法がとられた。そもそも「JIS X 0208」はコンピューター上での情報交換用文字を示すために規定されたものであり、固有名詞などに用いられる一部の文字を除いて、基本的に異体字にそれぞれコードを割り振ることはしていない。つまり、あるコードの文字を表示した場合、そのコードに包摂された意図しない別の字形（異体字）が表示されることがある。専用のフォントを作成することにより、意図した文字を表示させる方法も考えられるが、環境に依存するためSATではこの方法を採用しなかった。これは将来的

な大規模な文字コードの公開を見据え、異体字の包摂ではなく、より厳密な分離を重要視した判断であるといえる。

悉曇文字に関してはローマナイズした上で〈siddham〉と〈/siddham〉で挟み、作成した文字データを参照する仕組みとなっている。ローマナイズに関しては一般的によく知られる「京都・ハーバード方式（以下KH方式)」を採用し、これで表記しきれない文字については新たに作成した規則に従った。よって以後本稿でも悉曇文字をローマ字表記する場合、基本的に「KH方式」を用いる。但し「KH方式」にもいくつかの種類があり、SATでは長母音を大文字で表記する方法を使用している。文字間にはスペースを挿入し、独立したものであることを示している。つまり𑖀𑖪𑖰𑖨𑖮𑖳𑖽𑖏𑖽と表示させたい場合、〈siddham〉 a vi ra hUM khaM 〈/siddham〉と入力する。そのため一つのローマ字表記につき、一つの文字データのみ表示が可能となっている。悉曇文字をローマナイズし入力する作業については、元山公寿がSATからの依頼を受け、大正大学大学院の学生など複数人を集めて行った。ここで問題となるのが同音異体字の表示、「KH方式」では表記できない文字のローマ字表記である。



KH方式（SAT）

悉曇文字の異体字について

悉曇文字には漢字同様多くの異体字が存在する。漢字においては異なる字形で同音・同義の文字について、字形の違いか字体の違いかでしばしば議論となる。漢字の異体字は一般的に「同音・同義であり、同じ文脈での交換が可能なもの」と理解されている。しかし、同音同義の複数の文字全てについて、どちらが標準字体かを定めることは難しい。悉曇文字については、日本悉曇学における字義という概念が存在するものの、そもそもが表音文字であり、種字などの特殊なものを除いては『大正蔵』の本文中では同音異体字関係にある文字自体が同義であるかどうかという問題にはならない。したがってここでは同音で表記の異なる悉曇文字を総じて異体字としている。

日本に於ける悉曇文字の伝承は師資相承によって行われ、必ずしもインド伝来の文字そのものの字体ではない。また、伝承にも様々な流派があり、それによって標準字体も異なる。そのような意味では漢字における異体字と同様の問題があるといえる。そして悉曇文字については現在でも厳密な意味での異体字の研究や認定が行われているとはいい難い。この問題については課題として今後調査していきたい。

今回は前述したように「同音で表記の異なる悉曇文字」を異体字として進めていきたいと思う。悉曇文字の異体字の多くは「ウ（ウー）点」が付加される時と切継の際に生じる。以下に『大正蔵』中に見られた異体字の類型を示す。

1、「鷲点（）」と「雲形点（）」

「ウ（ウー）点」付加の際に生じる根本的な原因は、そもそもこの点に「鷲点」と「雲形点」という二つの点画が存在することにある。悉曇文字のうち、「鷲点」を選択したことによって他の文字との混同が生じるもの（例えば「tu」）や、切継下半体に「鷲点」と「雲形点」が付加できず、別の方法で表記する文字（例えば「～ru」）などを除いては、この問題によって理論上数多くの同音異体字が発生する。実際には切継によって生じた文字よりも、「悉曇十八章」中第一章の文字に多く見られた。

2、「r～」形の文字

ローマ字表記上先頭に「r」が付く文字についても異体字が生じる。例えば「rya」は大正蔵中二種類の字体が存在する。一つは「ra」の切継上半体に「ya」を継いだもの、つまり悉曇十八章中の第八章の規則に従ったもの（  ）、もう一つは悉曇十八章中第二章の規則に従った「ra」の切継上半体に「ya」の切継下半体を継いだもの（  ）である。

3、母音の異体字

母音にも異体字が数多く存在し、その中でも特に問題になったのが「a」の異体字  である。(4) この文字については、さらに「A」「AM」「AH」などの点画が複合的に付加された文字が目立ち、出現回数も多いことから「雲形点」を「u」と見なし、後述する「+（プラス）」記号によって差別化を図った。具体的には、は「A+u」、は「AM+uH」とした。

4、その他の異体字

そもそも現在伝えられている悉曇文字は師資相承によって受け継がれてきたものであり、誤伝と思われるものや日本で変化したものもある。その顕著な例が「Dha（  ）」「Na（）」「tha（）」である。他にも「ja（）」「za（）」などにも異体字があり、『大正蔵』中にも現れる。(5)

以上の点についてこれらを使い分けようとした場合、SATのデータベース上に限れば大正蔵中の該当文字を調査し、異体字が現れなければどちらかを選択、現れた場合はどちらかに何らかの属性を付けて入力するという作業が必要となる。実際の作業では『梵字大鑑』に収められている「悉曇十八章」の字体を基本として文字作成を行うと同時に、『大正蔵』中に現れる異体字を調査し、可能な限り作成して属性を付加した。しかしながら調査対象文字数の膨大さなどから全てを網羅することは困難であり、一部の文字については現在のところ『梵字大鑑』上の標準的な字体を表示させている。

「KH方式」ではローマ字表記できない文字

『大正蔵』中には前述した異体字の他に、「KH方式」ではローマ字表記できない文字が数多く存在する。

1、空点（anusvAra）と仰月点（anunAsika）

空点と仰月点について、サンスクリット文法上の区別はともかく、悉曇ではこれらを厳密に区別はしていない。ただ種字などを書く場合、特に朴書では空点の代わりに仰月点を用いて種字を荘厳することが多く見られる。『大正蔵』中にはどちらも使用されており、種字に仰月点が付いている場合が多いものの、必ずしもそうとは限らない。これらを厳密に区別すべきとの意見もあったが、現時点では実現しておらず、文字の作成も一部種字として現れる文字のみに留まった。また「KH方式」ではアヌナーシカを「&」で表記することがあるが、「&」はSATのマークアップ方式で外字を表現する重要な記号であり、これを使用することはできない。従って「仰月点」に属性を付ける場合、SATや他の悉曇文字表記に使用していない別のASCII文字を割り当てる必要がある。

2、サンスクリット文法上ありえない文字

『大正蔵』中にはサンスクリットの文法上あり得ない文字が多数現れる。特に種字として記載のある文字に多く見られるが、それ以外にも母音である「ア」字に「イ」や「イー」などの点画が付けられているものなどもある。これらは「KH方式」ローマ字表記の規則からも外れるため、「+（プラス）」記号などを用いて入力した。この方法を用いるには「+（プラス）」によって表現される母音類をどのような順番で表記すべきかという規則を作らなければならない。例えば𑖀𑖽という文字があった場合、「hIH+R」・「hRH+I」・「hR+IH」・「hI+RH」という四種類が考えられる。このようなとき、母音の順番は伝統的な悉曇字母表に従うと同時に文法的に可能な範囲で「+（プラス）」記号を用いず、余ったものを「+（プラス）」によって付加するという規則で統一している。つまり「hRIH」は「hIH+R」ということになる。また、切継上半体・下半体についても「(+プラス)」記号で表した。例えば、「ra」の切継上半体（𑖪）は「r+」、下半体（𑖼）は「+r」とした。最大の問題はこの規則を使用する人が理解している必要があるという点にある。

3、記号

句読点や繰り返しなどの記号については「,」「/」「#」「§」などで表記

したが、『大正蔵』中には、特に繰り返しと章末記号については複数現れる。これらについては対象が限られていたため、調査のうえ「§2」「§3」などと数字を付加して差別化を行った。

問題への取り組みと今後の展望

これまで悉曇文字の異体字とローマ字表記について、その問題点を挙げてきた。これ以外にも、判別が難しい文字や明らかに誤りと思われるもの、紛らわしい文字をどう判断するかなど、多くの問題がある。しかし、今回のSATによる事業があくまでも『大正蔵』のテキストデータベース化であり、校訂は一切しないという方針である以上、主観的にこれを修正することは出来ない。いずれにしても、現時点で作業が終了した85巻全てのテキストが完全に正確に入力されているとは当然考えられない。今後、これらをどのように修正し完全な『大正蔵』のテキストデータベースに近づけていくかは、SATの大きな課題であろう。悉曇文字についてはさらに時間を費やしながら、今回作成しきれなかった異体字などの文字を追加していく作業が必要である。また、どの文字にどのような属性を付与してこれらを参照させるかということも含めて、ローマナイズされたものと悉曇文字を対応させる規則を見直し、使用者にわかりやすい形で公開する必要がある。幸いにもSATの作成したデータベースの方式では、文字作成と規則の確立がなされれば、テキスト部分の修正によって意図した悉曇文字の表示が容易であると考えている。もちろん悉曇文字の部分だけではなく、全体をより『大正蔵』そのものに近づけていくことによって初めて研究者の用いる資料として、高い評価が得られるであろう。そのためにはまだ時間が必要である。

また、今回作成されたものを基本に、『大正蔵』データベースはそれとして、別に校訂をほどこした大蔵経の改訂版を作成することも視野に入れるべきであろう。ただ、学術研究に用いる資料としての絶対的な価値を保持するためには、当然「誰がどこをどのように、そしてどのような根拠に基づいて校訂したか」が明らかでなければならない。SATの作業は一応の結実を果たしたといえるが、この大事業は今後の発展に様々な大きな可能性を秘めており、それは同時に大きな課題である。

謝辞

本論は、文部省／文部科学省／日本学術振興会・科学研究費補助金・研究成果公開促進費（データベース）の成果であり、日本印度学仏教学会の事業の一端としてSATが行った、大蔵経テキストデータベース化事業によって成り立っている。この事業はこれまでに参加した多くの方々の熱意と努力によって成されたものであり、ここに敬意と感謝を記す。

註

（1）　種智院大学密教学会編　昭和58年５月31日　名著普及会

（2）　合成字作成の規則を18の章で示したもの。日本の悉曇学では、阿闍梨が各章の頭字を書き、弟子がこれを手本に12の点画を加えることによって綴字法を習得するという伝統的な方法で相承される。

（3）　東京大学多国語処理研究会が継続的に行っている事業、「マルチメディア通信システムにおける多国語処理の研究プロジェクト」によって纏められた文字セット。

（4）　この文字を「A」の異体字とする伝承も多いが、この文字にさらに「A」点がつく文字もあり、ここでは「a」の異体字とした。SATでは異体字としてというより、「a」の「鶯点」の代わりに「雲形点」が付いた形とした。

（5）　「Na（  ）」、「za（  ）」、「ja（  ）」など。

＊　　東京大学大学院人文社会系研究科教授

＊＊　大正大学人間学部准教授