chapter.09

中世の手書き写本のOCR翻刻テスト報告

蓑輪顕量

研究協力者：ジッリオ・エマヌエーレ・ダヴィデ、余新星、田中翔悟

1. はじめに

文献のOCR翻刻の技術が進歩して、多くの資料がOCRによる翻刻が可能になってきているが、まだそれほど成果を上げていないものもある。それが手書きの崩し字で書かれた写本のOCR翻刻である。この報告では、凸版印刷が開発したOCRソフトを使用させていただき（心臓部は京都大学文学研究科の林晋教授が開発したSMART-GSと同じもの）、実際に試行してみた報告記である。取り組んでみたものは中世の資料が二つ、近世の資料が一つである。

作業手順は、一般のOCR翻刻と変わらず、画像データを作成、行を指定、一文字一文字を指定、読み込ませて辞書を作成、そして不明の字体があれば、別の箇所に探して、推測するというものであった。ソフト自体が似通った字体を並べてくれる機能を持っているので、とても助かるものであったが、実際にはなかなかうまく行かず、文字を正確には読んでくれない例がしばしばであった。以下、順を追って、実際の作業の順番に合わせて報告したい。

2. 最初の試行

最初に選んだ資料は、中世後半期の15世紀末に成立した、山梨県身延町（みのぶちょう）、身延山久遠寺（くおんじ）の身延文庫に収蔵される『諸法実相伝抄（しょほうじっそうでんしょう）』と呼ばれる手書き写本であった。大学院博士課程の学生であるジッリオ・エマヌエル氏とともに作業をしたのであるが、本写本は、講義を聴いた学生が備忘のために作成したと推

定される口訣の資料であり、書かれている文字は癖の多い、大小さまざまな字体で、しかも独特の崩し字が多く、走り書きのような悪筆であった。行は曲がっているし、文字の大きさも不揃い、一行の文字数も一定していない。

このような状況で、まず困ったものは、画像データと文字との対応で、画像は一つであるが、変換される文字は複数があり得るというケースに直面したことであった。これは仏教特有の崩し字であり、また略字と呼ぶべきであると思うが、菩薩や菩提のように、草冠の連続とその下に少ない画数を用いて、一文字分のスペースで、二文字の漢字を表すケースを、どのように辞書に記憶させるか、であった（この点に関しては、一つの画像に複数の文字が対応させられるよう、後にソフトの改善が行われた。）

また、これも仏教関連の写本にしばしば現れるのであるが、偏や旁のみで、もとの漢字を表す場合があるが、そのようなものを、どのように辞書に覚えさせ、処理するのかが問題点として浮かび上がった。

また手書きの特徴であるが、文字が次の文字と重なったり、さらには縦の行での重なりだけではなく隣の横の行の字とも重なっていたりと、変化に富む。あるいは跳ねが異常に大きく書かれていたりするものもあり、しかも文字の大きさが不揃いであることが数多く見られた。

このような場合の処理も今後の課題であろう。隣の文字と重なっている部分があったりすれば、文字の画像に大きな影響が及ぶ。そのような画像データをどのように処理するのかを考えざるを得ない場合も存在した。

また、実際に人間が手書き写本を読むときには、読めない字体はとりあえず保留にして、ほかの箇所に同じ字体が出てこないか探して、文脈の中で文字の候補を探していくという方法をとるのだが、OCR 翻刻で読む場合にも、基本的には同じ考えに基づいて、作業を行っている。結局、文脈の中で候補を考えるしかないのだが、問題は、この文脈であった。

当初、同じような候補を取り出してくれる機能は、凸版印刷の OCR ソフトに当然のごとく付いていたのだが、取り出してきてくれる前後のわずかに二文字と文字数が少ない、という難点に直面した。これは凸版印刷の開発者の方に、前後の文字の切り取り数を増やしてもらうことで解決していただいた。具体的には前後五文字前後は切り取っていただけるようにお願いした。

いずれにしても、この資料を対象にした作業のときには、ほとんど自動では読んでくれない、という事態に直面することになった。辞書を作成しても、同一の文字が同一の形で書かれているわけではない（手書きだから当然か）ので、一つの文字にたくさんの画像辞書が必要ということがよくわかった。とはいえ、読めない文字を、別の箇所の同一の形のものと比較してくれるので、作業は効率化されていることは間違いなかったが。

しかしながら、参照してくれる範囲をあらかじめ設定しておかなければならないので、その事前作業に手間がかかることも多少の問題点であった。結局、ソフトが持っている辞書データと、読もうとしているテキストの文字の筆致が大きく異なると、まったく役に立たないことがわかった。

ということは、逆に、それぞれのテキストで、文字データと画像データの対象辞書を、個別に作成することを積み重ねていけば、読めるようになる資料が増えるということである。

3. 第二の試行

第二のトライアルは高峰顕日の法語として伝わる『仏国禅師法語（ぶっこくぜんじほうご）』を対象に作業を行った。この写本【図 1】は漢字の文章ではなく仮名法語で、全部が崩し字の仮名で書かれているという資料であった。それほど大部の資料ではなかったが、全部が仮名書きであるので、必然的に字と字とがつながっている例が多く、文字同士の切れ目を判断することが困難であった。しかも文字の大きさはまちまちで、どの位のスペースが一文字に費やされているのか、簡単には推定できない。しかも、隣の行の文字と重なるような書き方もされている。

結局、OCR によって読めるようにするためには、こちらでも、一から辞書を作成しなければならず、実際に辞書を作成してみたが、思うように文字を判別してくれないことがわかったので、最終的には、人間が手作業で読むということになってしまった。ただし、読めない箇所が出てきたときには、その写本の中で、対照できる同じような字体が使われている箇所を、自動的に検索して引用し並べてくれたので、この機能は大変に助かった。とはいえ、自動で読むというレベルには到底至ってはいなかった。これは辞書に組み込む参照データ

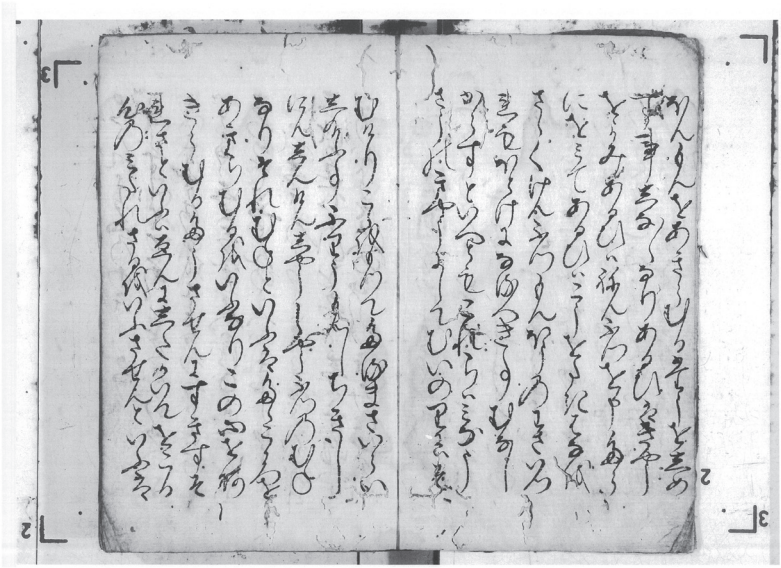

図１　仏国禅師法語（部分）

が充実していないことも一因と思われた（仏国禅師法語の例）。

このような経験から、このソフトによって、個々のわかりにくい文字については、同じような文字が書かれている箇所を探し出してもらい、それを参照して、人間が読む、という作業になった。先の事例と同じであるが、写本データベースで検索をかけると、いくつかの候補を提示してくれるので、大変参考になり、難読の文字の解決に役に立てられることが確認できた。それに、作業対象の写本をある程度まで入力したら、今回はこの写本自体が参照データとなり、後半の内容の解読に活用できる点が見込まれたので、その点は大いに歓迎すべきであるように思われた。

ところで、手書きの「かな書き写本」に関しては、複数の文字を連続書写するケースがよく見られる。参照となる写本のデータは、大抵一文字ずつ区切ってデータとして保存される。そのため、データベースで検索をかけるときに、現段階では、連続で書かれている部分を一つのまとまりとして指定しても、あるいは区切って一文字ずつ指定しても、対象イメージと参照イメージとは、マッチングする確率があまり高くはならなかったのが現実であった。

結局、文字の識別の結果はよくないのが、この仮名書き写本に対する試行に

おける結果であった。なお、ソフト上の使ってみての使い勝手の点で、提言できることがあるとすれば、文字と文字との切れ目を、後から簡単に修正できるようにしてほしい点と、領域を選ぶ際に自動で判別してくれるようになるとありがたいと思われたところであった（なお、後者の点は、後に改善され、自動で領域を判別する機能を充実していただけた）。

また、今回の試行から推定されるところは、ひらがなのような画数の少ない文字は、OCR 翻刻は、苦手なのかもしれない。その理由として考えられることは、一文字一文字の特長が少ない、文字の切れ目が判断しにくいことが原因と思われる。そして、何よりも大きな原因は、手書きであるので、同じような形で書かれていても、実際にはまったく同じものはない、ということにあるように思われた。

4. 第三の試行

本書は中世の初頭、13 世紀に東大寺尊勝院に活躍した宗性（そうしょう）の残した写本である【図 2】。11 世紀の半ばである天承 2 年よりはじまった、当時としては最も格式の高い法会（ほうえ）であったと考えられる法勝寺の御八講（みはこう）の問答すなわち論義の記録である。宗性自身が手写した部分と、他者に書写させた部分との双方からなる写本であるが、全部で 30 冊近くの分量に及ぶ大部のものである。

宗性は貴族出身の僧侶であり、尊勝院の院主も務めた当時の有力な学侶であり、また交衆（きょうしゅ）の代表的な人物でもあった。彼が残した手書きの資料は膨大な数にのぼり、平岡定海（じょうかい）氏による詳細な研究が存在する[*1]。彼の用いた書体が中世の時代の標準的なものであった可能性が高いことが指摘されているので[*2]、中世の手書きの写本を、OCR を用いて翻刻・入力する上では、最も相応しいテキストではないかと考え、本書を選んだ。実際、当時、格式の高い法会として名高い法勝寺御八講、宮中最勝講（きゅうちゅうさいしょうこう）、仙洞（せんとう）最勝講が存在したが、これらの最勝講のまとまった資料も、宗性は残している。それが『最勝講問答記』であり、同じく東大寺図書館に収蔵される。さらに彼は、華厳宗にかかわる論義資料も多く作成しており、大正蔵に収載された『華厳経香薫抄（こうくんしょう）』はその一つである。宗性には関連する写本資料が膨大に存在し、しかも翻刻も手つかずのまま

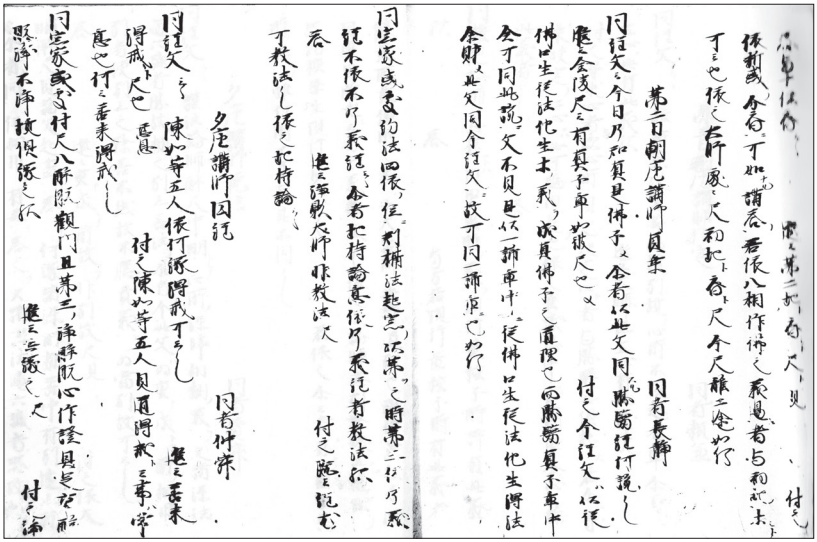

図２　法勝寺御八講問答記（部分）

のものが数多く残されているので、OCR 翻刻が可能になれば、その分野の研究を裨益（ひえき）すること、大であると考えたからでもある。

実際、『法勝寺御八講問答記』の本文はまだ、一部分しか活字になっていないので[*3]、本文がデジタルデータで提供される意義は大きい。ましてや本書は大部であるので、その全体が活字資料になり、かつデジタルデータで提供されればなおさらである。宗性の手書き資料のデータベースが完成すれば、研究は大きく広がるであろう。

宗性の手になる『最勝講問答記』（宮中や仙洞で行われたもの双方を含むと考えられる）も、東大寺図書館に残っている。すなわち、中世の時代で最も格式の高かった三つの法会、すなわち法勝寺御八講、宮中最勝講、仙洞最勝講の三講において行われた論義の内容の実際が、詳細に明らかになるであろうことが期待される。すでに、三つの講で行われた論義の内容に関する吟味は、若干であるが存在し、講は異なっても、基本的にその場で行われた論義には、共通の基盤があったことが推定されている。しかしながら、それをより確実な形で顕彰することはまだ行われていない。それが可能になることが期待されるのである。

さて、実際の作業は、前の二つと同じくまずテキストの画像データを作成し、それをOCRにかけるために行の指定などを行い、実際の文字の区切りをつけることが必要になった。本写本資料は、宗性の書き癖のおかげか、わりと一字一字がしっかりと区別されて書かれている。これが幸いし、文字の区切りを自動で処理しても（平成30年の12月の時点で凸版印刷のOCRソフトはかなり進化を遂げていた）、あまり間違うことはなかった。

自動で領域を判別し、後からその領域を訂正できるようになっており、また画像データと文字データを対応させる辞書も、単字画像を複数の文字に対応させることが可能になっていた。

この機能はとても重要である。特に仏教関係の手書き写本では、複数の文字からなる専門用語を、一つの文字で略記する場合が数多く存在するからである。典型的な例は菩薩、菩提、涅槃などであるが、これらは複数の文字が、一つの文字画像に対応している例としてあげられる。

また、次のような点も問題点として浮上した。文章の中に、「尺（しゃく）して」と読める箇所が頻出するのであるが、このような文字の場合、どのように翻刻するかが問われる。というのは、字義通りには「尺」は長さの単位であるが、この意味ではまったく文意は通じない。明らかに、「尺」は解釈の「釈」の意味であり、「釈」の代わりに用いられている。

これは、文脈から文意が明らかである場合には、偏または旁の部分のみで、あるいはその逆に、偏や旁を省略して、残りの部分のみで当該の文字を表現する例であろう。このようなことが、多く存在するのである。

このような書き方は、仏教者の手書きの場合によく見られるもので、前後の文脈からその文字が明らかであり、漢字の一部分のみでも、間違えることはほとんどないような場合に起きている。いわば、独特の省略文字が生じているのである。

ところがOCRで翻刻しようとすると、これが問題になった。『法勝寺御八講問答記』の中では「尺」の字が典型であるが、宗性自身は明らかに「尺」を「釈」の字の意味で使用している。とすれば、文脈から判断して、宗性自身は「尺」を「釈」の意味で使用しているのであるから、「尺」を「釈」として、画像辞書に登録する必要が出てくるであろう。

同様の例は、他書にも発見できる。例えば、時折、登場するものとして「泥洹」があげられる。これも文脈で明らかな場合に「さんずい」二つで表現されることがある。しかし、これを「さんずい」二つに起こしたのでは意味をなさない。なぜなら作者は「泥洹」と書こうとして「さんずい」二つを書いているからである。

このことは、次のように考えるとよいであろう。現在、われわれが認識している漢字という文字は、一文字の占める領域が、ある一定の領域を持つものと考えている。それもほぼ真四角または少し長方形の領域を暗黙裏に前提しているように思う。そして「さんずい」二つは、明らかに想定されるスペースの中に、左に偏って書かれている場合が多い。

ということは、画像データとして保存するときには、スペースを考慮して登録する必要がある文字が存在するということであろう。

また、文字の一部分をだけを書いているのであるが、スペースを考慮せずに、堂々と普通の一文字として欠いている場合も存在する。そのような典型的な例が、戒律文献の中に登場する「羊石」である。これは、文脈から推定して「羯磨」の略字であることは間違いないが、字体としてはまさに「羊」と「石」なのである。しかしながら、これも同様に、「羊石」と起こしのではまったく意味をなさない。

上記の二つの事例から推定して、OCR 翻刻のときには、仏教特有の略字を考慮しなければならないということが明らかである。しかもそれは、文脈から判断するしかない例を、どのように処理するかという問題に行き着く。偏や旁のみで表す場合には、スペースを含めた画像登録で対応できるように思われるが、明らかに普通の漢字を、まったく異なった意味合いで使用している場合は、問題が大きい。この場合は、人間と同じく、文脈を考慮するという機能を、OCR ソフトに持たせる必要が出てくるであろう。例えば、「白四」の後に「羊石」と出てきたときには「羯磨」と読ませる、などのような指示機能である。

ここでまた『法勝寺御八講問答記』の翻刻の問題に戻る。われわれが使用させていただいたソフトは領域を指定して、読めない文字があったときに、指定された範囲の中で、同じような文字画像を探し出す機能があって、それは前と変わらず便利であった。しかしながら、その指定された範囲というのが、どの

くらいの範囲までであると良好に機能するのか、まだ実際の試行例が足りないように思われた。とはいえ、この『法勝寺御八講問答記』は、以外と OCR 翻刻で、読める資料であると思われた。

5. 人間による文字の読解との比較を通じて

わたしたち人間は、文字の大きさが異なっても、また字が多少、下手であっても、文字を認識できる。さらには一部分の欠けた文字（虫食いも含めて）でも、残っている部分あるいは文脈から、正しい文字を推測することができる。ところが現状の OCR 翻刻には、それができない。なぜであろうか。おそらくそれは、文字画像と文字情報が、単に結びついただけの辞書しか、持ち合わせていないことに起因するのであろう。

人間の個人的な力量に起因する文字の上手、下手に対しても、あまり苦に感じることなく、わたしたちは文字を読んでいる。しかし、OCR 翻刻では、そのようなことはなかなかに難しいことのように感じる。

わたしたち人間が写本を読んでいるとき、わたしたちの心に起きている認識の過程が、コンピューターにできるとよい。わたしたちは、文字が読めないときにどうしているかというと、その文字の形の類似性から、候補になる文字を探し出している。そして、前後の文脈に合わせて、意味が通じるかどうかを根拠に、文字を確定している。

また、虫食いによって文字が欠損している場合でも、文脈から文字の形を推定して読んでいる。このときにも、前後関係の文脈は大きな手助けとなっている。さらには、一文字分がまったく空白になっていても、その空白の文字を、前後関係の文脈から類推することができる。

また、文字の上手、下手の場合は、一定の範囲内で、多少の形のずれがあっても、ある特定の文字であるという認識を持てている。これがなぜ、可能なのかを考えておかなければならない。

人間の認識機能は基本的には次のようなものであろう。目という感覚器官を通じて視覚という感覚機能が働き、捉えられる対象としての影像が脳の中に描かれ、それを認識しているのである。時には模様として認識することもあろう

が、文字として認識することもできる。この基本は人間も機械も変わらない。しかしながら、人間は、字体にかなりの幅を持たせて、この作業を遂行することができている。

わたしたちはすでに持っている知識と突き合わせながら、文字を読んでいるのだと思う。ということは、OCR 翻刻においても、同じような作業ができれば、効率が上がることを意味している。つまり、これは、わたしたちが普段、文字を認識するときの過程（構造と言い換えてもよいかもしれない）の解明が、必要であることを示唆してはいないだろうか。

文字の画像イメージが、既存の文字の画像イメージと一致するという、類推のときに働いている認識の仕方が詳細に解明されることによって、OCR 翻刻は、人間の認識に、より近づくことができるのかもしれない。

しかし、この発想そのものが問題なのかもしれない。まったく別の発想から、文字を認識するということを考えていかなければならないのかもしれない。それは、崩し字の読解において、よく言われることであるが、始筆と終筆の角度や方向性が、判断の材料にされることがある。これは多分に数学のベクトルの発想に近いような気がするが、そのような発想で、文字を考えることも必要なのかもしれない。

なお、この文章をまとめるに際して、別の機関でも、崩し字の OCR 翻刻が話題になっていることを知った。現在、凸版印刷の OCR システムは、さまざまな機関との共同研究に使用されているようであり、試行錯誤が行われている。ジャパンナレッジの HP によれば、立命館大学文学部チームが、崩し字の OCR 翻刻で報告を行っている（https://japanknowledge.com/personal/jknews/23920190702.html）。また、国立情報学研究所も、OCR 翻刻に実際に取り組んでいる人たちの情報を共有し、その取り組みを推進させるような試みを行っている（https://www.nii.ac.jp/news/release/2019/0710.html）。このようにさまざまな取り組みが行われていること自体が、この分野がまだ発展途上であることを示しているように思われる。

6. おわりに

OCR 翻刻による自動文字起こしは、基本的には画像データと対応する文字とを結びつける辞書機能が充実することによって、その精度を上げていく。ということは、ヒット率を上げるためには辞書をどう鍛えるかにかかっているということがよく理解できたが、それだけではだめであることも知られた。そして、現状の OCR 翻刻では、手書きの写本の場合、崩し字、特につながっているものを読むことは、まだまだ難しいことが実感された。ただ、手書きの資料でも、宗性の史料のように、一文字一文字、はっきりと区別して書いているものに対しては、威力を発揮しそうであることがわかった。

このような結果に至ったことから考えてみれば、結局、教えている人間が教師であるとすれば、機械による OCR 翻刻は教え子に相当するのであろうが、いまのレベルでは、教えている教師のレベルを超えることはなかなかに難しいように思われた。

以上、雑ぱくではあるが、試行錯誤の 3 年間の報告とさせていただく。OCR 翻刻を実際に行っている方やプログラムを書いている方々の、忌憚のないご批判やご意見を賜れれば幸いである。

注

1　平岡定海『東大寺宗性上人之研究並史料』上・中・下、日本学術振興会、1958-1960 年。

2　中世に一般に流通した書体を推定することは容易いことではないと思うが、多くの資料を見てきている名古屋大学名誉教授、阿部泰郎氏の口頭の見解による。

3　一部分であるが、翻刻が『南都仏教』77（1999 年 10 月）、拙著『日本仏教の教理形成』（大蔵出版、2009 年）などに掲載されたのみである。