COLUMN

紙資料の「データ解析」が持つ変革とコラボレーションの可能性

後藤　真

1. 質的研究と量的研究

本コラムでは、紙資料のデータ化の意味について、デジタル・ヒューマニティーズ（DH）の立場からいくつかの私見を述べるものである。人文データの計量可能性、という観点はDH研究の初期から議論されてきた。日本においても、絵画の計量分析の可能性が早くから指摘されるなど、データ化による量的な分析の試みは続けられてきた。既存の人文学の手法においては、資料を人間が読み、人間が「記述」を行うことによって、その研究成果が伝えられてきた。情報学の手法はそれらの記述という行為に加え、そのもととなるような「数」や「形式」を成果表現につなげるとしたといえるであろう。一方で、人文学のうち、とりわけモノを対象とする研究については、情報学だけではなく自然科学の手法も取り入れられるようになってきた。年輪年代分析や炭素同位体、あるいは酸素同位体による資料研究は代表的なものである。そして、本書全体を通貫する「モノとしての紙」という研究も当然そのような自然科学手法との連携によるものである。このような状況は、いわば「質的な研究であった人文学に量的な手法が取り入れられた」と表現できるであろう。

この「質的研究から（に加えて）量的研究へ」という観点は、研究をより広範な方向へと拡張するものであることはいうまでもない。しかし、量的研

究として成熟するためには、さらに次のステップが求められると考える。この後、そのステップに関して、少し検討を加えていく。

2. 量的研究のための「ステップ」

その重要なステップというのは、すなわち「量的研究には、絶対的なクライテリアが必要であり、そのクライテリアのための基礎データのさらなる蓄積が求められる」ということである。多くの自然科学・あるいは情報学における量的な研究は、その前提として「その数値が示す意味が共有」されている。量的な数字は、あくまでも数字でしかない。その数字に意味を持たせるのは「線引きする人間」であり、「数字を説明する言葉」なのである。たとえば、統計における P 値は、0.05 以下である場合に、その統計が有意であるとみなされることが標準であるとされている。これは、有意水準が 5％であることを前提とするために成立しているともいえる。このように多くの数字は、それまでの研究史にもとづいて、前提となる水準が設定されているのである。

一方で、DH ではそのような水準が設定されている例はほとんどない（無論、他分野の蓄積を用いて水準を決定することは当然行われているが、DH、ないし歴史資料の情報学的分析において取得されたデータそのもののクライテリアは存在しない）。それは、DH の研究水準の問題ではもちろんなく、純粋に研究史上データがこれまで存在しなかったか、あるいは人文学の特性上、より細かい分野設定がされているため十分な蓄積ができなかったかのいずれかが原因である。

たとえば、近年、大きく研究が進んだいわゆるくずし字の AI 翻刻の例を見てみたい。くずし字の翻刻に関わる研究が開始されたのは DH でも早い方であり、そこでさまざまな課題が発見されている。その課題をもとに、AI 活用全盛の時代となった 2020 年前後において、一挙に花開くことになったのが、AI くずし字翻刻である。本コラム執筆現在、このコンピュータによるくずし字翻刻の精度はある一定の条件下であれば、98％を超えるとも言

われている。これは、研究開始当初の精度に比すると、段違いの進歩であることはいうまでもない。この「高い」精度をもととして、ブラウザ上によるサービスや、スマートフォンアプリ、企業による展開が行われており、くずし字翻刻の技術は、「実用可能」な部分にまで至りつつあるといってよい。

しかし、ここにはこの「量的なものを評価するクライテリア」が前提として隠れている。すなわちそれは「全体の何パーセントまで読めば「(機械であれ人であれ) 読めた」といってよいのか」という前提である。仮に正解率98%という数字があったとして、逆の観点から見ればエラー率は2%である。すなわち、50文字に一文字を間違えるということになる。この数字であっても「読めている」と判断できるのかどうか、という議論を本来は行う必要があるのである。念のため付言しておくと、筆者はAIくずし字翻刻の可能性を否定するものではない。筆者自身もくずし字翻刻アプリである「miwo」を用いて、古文書の判読を行うこともあり、積極的に用いているユーザの一

khirin-a で公開している後藤家文書（CC BY4.0）より吾妻鏡1巻冒頭を miwo で解析させたもの（左。右は原文書画像）。この画像では約89%の正解率である

人である。そのような前提の上で、ここでこのことを指摘するのは、学術的な議論を今後さらに展開していくためのものである。

3. 量的基準・質的基準

このAIくずし字翻刻のクライテリアにも大きく二つの論点がある。一つは古文書を「読める」専門家のように古文書を「翻刻」することができるのかどうか。もう一つは、古文書を読むことができない、あるいは翻刻にしてもある程度までしかできない人びとが欲しい「ニーズ」に合致しているのかである。前者は純粋な研究として、後者は工学的・実用的な成果として考えるものになるであろう。前者の論点は、いわば「人間はどこまで文字の形を理解できれば、意味を理解することができるステップに移行することができるのか」という究極の問いを考えることになる。後者の論点は、「ユーザの期待値を上回ることができるのか」という、マーケティング的な観点で考えることになるであろう。現在の古文書AI翻刻は後者の点において、ユーザの期待値を上回っているであろうという判断のもとで、さまざまなサービスが展開されているといってよい。実際にこのユーザ期待値を上回っていたかどうかは、サービスに対する外的な評価によって決定されていくことになる(なお、この場合は単に精度の問題だけではなく適切なインターフェースや、典籍★か文書かなどのもとの文字種ごとの違いなどによる期待度の高低も含まれる)。一方で、前者の論点を確認することは、困難を極める。むしろ、人が文字を理解する数値に具体的なものがない以上、コンピュータの側も理解することができないのである。ただし、DHにおいてこのようなシステムの有意性を議論するためには、どこかで考える必要が、それも人文学の側で考える必要があるのではなかろうか。そしてそれは、人が文字を読む行為自体という人文学の本質的な研究にもなるし、「さらなる精度向上」を目指す情報学者にとっても有益な示唆を与えることになるだろう。

ここまで、論点を明確にするために、紙の議論ではなく具体的な数値が多く語られているAIくずし字翻刻を例としてきた。しかし、当然、このよう

な論点はモノ資料としての紙の分析にもあてはまるといえるであろう。交雑物について、どのようなものがどれぐらい入っていたら「多い」と定義できるのか、サイズ等については、いわゆる「標準」があるといえるが、科学分析による数値データについては、まだそのような「多い」「少ない」「標準的」といったことを示す共通の言葉がないように思われる。時代や文書の性質、発給者の社会的立ち位置などにおいても当然その基準は異なるであろう。このような基準を作り出すためにも、これまで以上に多くのデータが増やされていくことが求められる。多量のデータを処理し、標準的なものはどのぐらいなのか、外れ値はどのようなもので、何を根拠とすればそれを外れ値とみなすことができるのかなどを、多くの研究者のなかで議論し、決定していくことになるのであろう。そのためにも、多くのデータをさらに蓄積していくことが強く求められる。

本コラム冒頭において「質的研究から（に加えて）量的研究へ」と述べた。おそらく、この量的研究を実現するためにこそ「量的研究のための質的研究」が必要となり、その研究の共有化が求められていくことになるのではないだろうか。紙の自然科学的研究のデータ蓄積がさらに進み、それがこれまでの「質的研究」と効果的なコラボレーションを実現させるものになることを、期待するものである。