

敦煌出土漢字写本における字形分析

—— 藤枝分期法の再検討を中心として ——

高 見 美 友

はじめに

敦煌文書とは、1900年に中国敦煌市で発見された空間に収められていた古写本群の総称である。1900年に道教の道士王円籙が発見した16窟（石窟）に繋がる耳洞（第17窟）は莫高窟、藏経窟等とも呼称される。窟内の壁に山積みされた数万点の希少な古書・古文書は、後の調査により4世紀～11世紀初めのものと判明した。これらはオーレル・スタインやポール・ペリオを始めとした探検家らによって国外へ持ち出され、現在は中国・イギリス・フランスをはじめ世界各国で収蔵・保管・研究されている。日本へは大谷光瑞が組織した大谷探検隊が持ち帰った資料をはじめ、後代に購入されたものなどが京都大学や龍谷大学等に収蔵されている。

写本の内容は漢文仏教典籍が全体の8割を占め、そのうち半数は既知の經典である。そのため、当然ながら仏教学の分野からも資料の宝庫として注目されることとなり、それらを基に様々な研究・発表が行われてきた。

しかし、従来の研究では研究者の暗黙知に頼っており、年代分類法には限界があった。そこで、本稿では、書写年代が判明している15点の經典を対象に、機械学習の手法を用いて字形分析を行い、既存の年代分類法を再検討しようとするものである。

第1章 年代分類法の現状と問題点

第1節 従来の年代分類法の概要と藤枝分期法

年代研究・編年研究とは資料の作成年代等を明らかにする研究で、年代分類法は年代分類用の基準を手法として確立したもののことである。敦煌文書は、おおむね東晋末から宋代初頭にかけての600年間に書写された文書群で構成されているが、現存する敦煌文書の大半は残卷だとされている。残卷の多くは首尾・題記が無く、内容判別が困難なものが多々ある^①。それらの文書の書誌情報を研究により明らかにするために、敦煌文書の編年研究がなされるようになって

た。編年研究は敦煌文書の実態が世に披露された1909年以降、歴史学・写本学・書道史などの様々な方面から研究がなされ、藤枝晃により、敦煌写本の年代分類法が示された。藤枝自身は、年代分類法を「敦煌遺書分期示意表」「敦煌トルファン写本編年」としており、本論では、これらを含め、最終的に以下の定義で構成される敦煌写本の年代分類法を「藤枝分期法」と呼称する。

藤枝分期法は敦煌文書の年代分類を行う上で現在最も主流の手法^②で、藤枝が日本国内外の敦煌文書を比較・分析して得られた情報を基に分類されている。

藤枝分期法では敦煌文書を主に、書法・筆・紙・字形・罫線（の書き方）指標として、A～Dの4分類に分けている。

	400	600	800	1000
遺書形式	A	B／C	D	
時代	北 朝	隋 唐	吐蕃期・帰義軍期	
書 法	隸→楷	楷	楷	
筆	鹿 毫	兔 毫	木 筆	
紙	古麻紙	薄麻紙／唐麻紙	敦煌紙	
（材料）	（故麻布）	（樹皮）	（？）	
字 形	古 楷	古楷／唐楷	唐楷	
罫 線	針 孔	淡墨点／圧点	？	

図1 敦煌遺書分期示意表（藤枝晃編）藤枝[2005]p.7 より引用

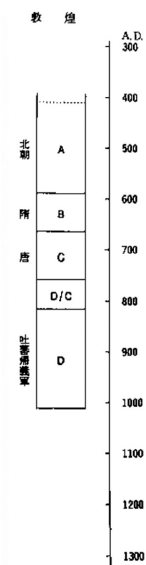


図2 「表 敦煌トルファン写本編年」藤枝[1987]p.4 より一部引用

図表掲載されている藤枝分期法は2005年時のもの。ほぼ同様の内容は英文で Fujieda[2002]にも掲載されている。上記の藤枝分期法の内容は、2005年以降に発見された見識を含めないものである。

遺書形式に示される基準とその定義は以下の通り

【時代】 時代区分についての要約が藤枝[1987]で以下のように書かれている。

唐代の拓本などで見える通常の楷書（図1のBとC）を基準とすれば判り易く、これが七一八世紀で、隋・唐の麻紙に書かれる。さういふ楷書となる以前の隸書と、隸から楷に向

ふ過渡段階（A）が北朝期（五一六世紀）で、常に北朝系の麻紙に書かれる。そして敦煌地方が中国の政治支配から離れて筆と紙とが手に入らなくなって、敦煌製の紙に木筆で書いたもの（D）が九—十世紀写本となる。CとDの中間に過渡期として古い写本（主にC）の紙背に木筆で書いてゐたD/C形式を設けた方が宜いと近ごろ考へる様になつた。^③

藤枝分期法では、概ね北朝期（A）、隋期（B）、唐期（C）、D/C期、吐蕃・帰義軍期（D）の5種類を時代区分として使用している。

【書法】 いわゆる字や文章の書き方として定義される「書法」は、その範囲の広さから厳密な定義はないが、この藤枝分期法上の「書法」は図1からも読み取れる通り、隸書や楷書等を指している。これは石塚晴通を始めとした字体史研究を行う研究者にとっては「書体」と定義されている。そのため本研究では、書き方を軸にした書写法等を「書法」とした。書体については「字体」の項目で触れる。

【筆】 用筆とも呼ばれる、書写に使用した筆の材質を指す。藤枝は、毛筆にも鹿毛や兎の毛を利用したものがあり、微細な違いがあるとしている。吐蕃・帰義軍期に入ると、毛筆から木のペン書きに主流が変化するとされる。

【紙】 紙の幅については、時代による変遷があることは池田温編[1992]の『講座敦煌5』に掲載されている呉其昱著「敦煌漢文写本概観」や坂本昭二[2016]にまとめられている。紙の幅と時代には関係があり、藤枝は「A系の紙にはA系の書体で書かれ、B系の紙にはB系の書体の字が書かれています。」^④と述べている。また、製紙時に用いる簾の跡である「簾の目」を目視によって観察し、「簾の目」の数等の傾向によって分析する手法もあるとされる。

【材料】 紙の素材は、製紙地域や製紙年代によって特徴があるとされる。この紙の材質は紙の幅（上記の「紙」という指標）とも関係があることは先ほどの藤枝の述べたもののほかにも、複数の資料により明示されている。^⑤

【字形】 漢字字体史を研究する石塚は、「字体」と関連語の定義として書体、字体、字形、字種を石塚[2012]で以下のように定義している。（以下は石塚[2012]から引用。一部筆者追記）

書体：漢字の形において存在する社会共通の様式。多くはその漢字資料の目的により決まる。（楷書・草書等）（藤枝では「書法」に相当）

字体：書体内に於いて存在する一々の漢字の社会共通の基準（藤枝が「書法」とするものに内包される）

字形：書体内において認識する一々の漢字の書写（印字）された形そのもの（藤枝が「字形」とするものに相当するが、定義が若干異なる）

字種：社会通念上同一のものと認識され、一般的に音韻と意味とが共通する相互交換可能な漢字字体の総称（藤枝分期法では扱われていない）

【罫線】 罫線は、縦の境界線で行を区切るもので、引き方は時代により変化があることが、藤枝[1990]によって示されている。藤枝[1990]では、罫線を引く方法について述べているのみだが、藤枝[1960]では罫線の幅は何らかの尺によって引かれたものであるとしたうえで、「何かの尺のラウンド・ナンバーになり得るのだと思うが、何の尺を使ったか、まだ解を得ない^⑥」と述べられている。

これらは藤枝の言う「標準形式」とされる經典群を比較した経験やその傾向から、分析されたもので、各年代における各指標の変化のまとまりを1つの分類として定義している。

なお、藤枝は、多くの研究者が研究してきた例外の形式ではなく、天地一尺の麻紙に専門の写経生が正しい楷書で、毎行十七字詰に書き、巻物に仕立てた、敦煌写本の大部分を占める正規の形式で作られたものが正式な書物として扱われていたこと、また、形式が崩されている注釈書等の個人用の書物とは異なり、宮廷図書館所属の能筆の写字生らが書写し、地方ではその書写本を書写する仕組みになっていたこと、そして、内容に関係なくこの形式が正規の書物の形式とされていたと述べている。

これは、藤枝分期法の提唱にあたり、藤枝が「敦煌出土写経で最も標準だと考えられた形式」を標準形式として定義したことを説明したもので、この形式に当てはまることを前提に、書体の違いや紙質の違いを究明し、基準として整理し、無紀年写本を分類する基準として利用していた。

藤枝の没後、石塚を含む研究者らによって、様々な研究成果が明かされた。その様相は池田[2016]によりまとめられている。

この分期法の提唱によって、敦煌学の中に「編年研究」が1つの分野として確立されたと言える。編年研究の開始にあたり、分期法の有用性について藤枝はスタイン・コレクションから標準形式のものを百年につき数点選び、複写資料を原寸大に拡大して書風の違いを確認した。この行動の目的は一時的な目録作成と無紀年写本の年代判定基準を見出すためで、結果的に、標準形式^⑦の写本のみならず、非標準形式の写本や崩し字の資料の年代推定にも有効であったと述べた。

敦煌文書全体としてある程度のグループ化傾向が年代との相関性を示すことは明らかである。この傾向と相関性については、赤尾栄慶も赤尾[1998]で調査の結果として「昨年来の知見であった時代ごとの書風・書法や料紙の紙質および大きさの変化、1行の字数が17文字になる時期が6世紀にごく近い5世紀後半であることや、7世紀から8世紀にかけての唐時代の写経の一紙あたりの行数が28行となっていることなどの点を追認することができた。」と述べている通り、研究者間の暗黙知で形成された分類が、その検証の結果、正確であることが少しずつ証明されてきた。

とはいえこれらはあくまでも計量可能な数値に基づいた結果であり、字形や風合い等の主観的基準を裏付けるには情報が不足している。次項では、藤枝分期法の課題について述べる。

第2節 藤枝分期法の現状の課題と暗黙知の限界

前述の指標に基づいて分類を行う藤枝分期法だが、近年の研究により様々な課題が浮かび上がった。

まず、書誌学的研究の中でも、書法史の観点から、書写道具・書写姿勢など書法関連の指標について、さらなる細分化の必要性と書法が字形に与える影響が史睿[2021]にまとめられた。この論文では、書写道具・書写姿勢など書法関連の指標の可能性と、史が考えた結論が示されている。史[2021]掲載内容と藤枝分期法と合わせると、用筆は、材質を穂先と軸と工法に、執筆方式は筆の持ち方と書写姿勢に、紙は紙幅・素材・製紙道具（紙漉きに使用した枠の素材・枠のサイズ）に、罫線は罫線の引き方や罫線間の幅で分けられる可能性がある。字形は藤枝分期法でしか触れられていないが、字形は用筆に影響を受けることから、字形も基準に採用できると考えられる。

また、藤枝分期法が提唱されてから少なくとも20年^⑧が経過している。

藤枝が資料を並べて比較を行った1950年代は、敦煌文書が存在が公表されてから間もない頃であり、現地での閲覧や出版された図録などを全て合わせても、閲覧可能な史料が少なかった^⑨。このことは藤枝[1999]にも述べられており、その記述から藤枝らは文書の比較研究を複写等の二次的な印刷資料^⑩を用いて行っていたことは明らかだ。近年になってデジタルアーカイブ化が進み、様々な写本をカラー且つ高解像度で確認できるようになってきたこと、杏雨書屋等の所蔵先が写本の公開を始めたこと等によって、確認可能な写本は全世界的に増加している。以前と比べて、研究環境は劇的な変化を迎えている。そのため、資料の質の向上も含めて、藤枝分期法自体の再検討を行う必要があると考えている。

また、藤枝分期法は基準が明確ではない主観的な情報から得られた結論も用いている。主観的な情報とは、上記指標における書法、筆の材質（毛先・軸）、字形、罫線の引き方に相当する。特にこの4種類は研究者間で暗黙知のような曖昧な基準で構成され、研究者間では曖昧な基準のまま利用されている。これらは書写者の個人差の影響を受けるほか、歴史的な変遷の影響も受けるため、この基準を明確で客観性の高いものにすることは非常に難しい。このことは、張も写本毎に書写者や書写目的等が異なるため、書写された字だけを見て書写年代の唯一の根拠^⑪とすることは出来ず、補助的な資料で情報を補完することが必要だとの旨を述べている。不明瞭で個人ごとに異なる基準を用いて、全員が完全に同じ回答を出すことは現状不可能である。これは、藤枝分期法最大の課題だと言える。

今までに敦煌学研究を行ってきた諸学者のあいだで、1つの写本に対して評価が大きく分かれるということはほとんどない。おおよその範囲でいつの時代のものか、どの系統に属するかを予想できる上、その予想は研究者間ではほとんど変動しない。これは、敦煌文書に限らず、日本の古写本を見る研究者にも同じことが言える。この感覚がおそらく研究者間における「暗黙知」である。

本研究の意義は、敦煌文書のみならず、トルファン文書及び日本の諸寺に伝持される古写本にも適用できる可能性を秘めていることである。

第3節 仏教学における年代研究の重要性

仏教学における敦煌文書研究について上山大峻は、新出文献やいわゆる珍本を、研究者の属する宗派的関心から涉獵して紹介するという形のものか、従来の中国仏教の史的知識の裏付けや、欠落を補うかの目的で、資料を敦煌写本中に探し求めるという性格の研究が多かったと述べた。過去の敦煌文書研究は限定的な観点から行われていたと言える。藤枝[1999]でも「従来の研究は、古文書であるとか、(中略)注釈書とか歴代大藏經には入っていない古逸書や偽經とかに重点が置かれ、普通の經典を研究する人は絶無といってよい状態でした。」と述べられている。¹³

確かにこのような、中国仏教等の史実を実証し、逸した資料の補完を行うことが可能な資料という性質は、敦煌写本の資料的意義の重要な一面である。だが、仏教学における敦煌文書の役割は、何も先に述べたような、新出文献の内容解説にとどまらない。敦煌仏教としての仏教文化の解明や、文化交流によって中国仏教に与えた影響の研究等、様々な方面の研究を行うのに大変重要かつ貴重な資料であることは明らかだ。¹⁴敦煌文書の研究と内容解明は、敦煌仏教の実像の解明につながるもので、その必要性については上山[1990]に詳しくまとめられている。¹⁵

これらの研究のためには、その資料の来歴や作成状況などを知ることも必要で、来歴や作成状況の解明のためには、資料の編年や書写者に関する情報の研究が必要だと考える。

無紀年写本の年代判定には藤枝分期法を始め、紀年写本との比較による年代推定が有効と考えられてきたが、藤枝分期法には前述した通り限界がある。

これらの現状を踏まえ、今後研究を進めるためには、新しいアプローチが必要と考えている。

第4節 新しいアプローチの提案と研究の目的

実物を利用したアプローチは、総じて文化財保護や許諾の可否等の問題があり、結果としてかなり難度が高い。そのため、非破壊な方法かつ既存の二次資料の利用をすることが望ましい。近年ではデジタルアーカイブ画像を利用した非破壊的研究が行われている。

前述の通り、紙幅や罫線幅、識語などの明文化・定量化可能な指標だけでは、年代分類の結果を確定させるための情報が足りず、それらを確定させるためには暗黙知のようなファジィな情報に頼らざるを得ない。この問題を解決するためにも、まずは既存の手法の再検討や、別角度からの検討材料を得て精度を上げる方法が効率が良いのではないかと考えた。

暗黙知のようなファジィな情報を過不足なく定量化することは難しいが、従来とは異なる手法を用いることで、何らかの新たな情報を得るのは可能であると考えている。

新たな情報を得るための新しい手法として、AIを用いた工学的手法によるアプローチを提案する。

人文学の分野において、近年ではデジタル・ヒューマニティーズ (DH) の考えが浸透しつつある。DHとは、「人文学の何らかの側面にデジタル技術を適用・応用する研究^⑩」であり、人文情報学ともされるもので、このDHの本質は方法論的なものと言える。DHの目的は人文学研究にデジタル技術を導入することで、人文学の研究方法に影響を与え、結果として新たな知識を得ることである。既存の方法とは異なるアプローチとして、DHは適していると言える。

これに加え、デジタルデータを用いた研究というのは「データ駆動型科学研究」と呼ばれ、集積してきたデータを利用する（駆動させる）ことで、新しい知見を得る研究のことを指す。

デジタルデータを起点として、複雑な人文学データを専門家と機械の分業（協力）によって解明することは、新しい手法として十分に可能性を示せると考える。

これらの理由から、本研究ではAIを用いた工学的手法によるアプローチとして、具体的には主成分分析とクラスタリングによる定量的分類を行う。この方法を利用しようと考えたのは、先ほど「おおよその範囲でいつの時代のものか、どの系統に属するかを予想できる上、その予想は研究者間でほとんど変動しない。」と述べたことに起因している。

「過不足を考慮したうえで、定量的な差を得るために新しい手法を試す」上では、前提として、何らかの定量的な差を得られる可能性が高いものを対象とすべきであり、出力結果は何らかの根拠を持って解釈する必要がある。「何らかの解釈」とは「先行研究や藤枝分期法と同様の基準を用いて解釈する」もしくは、「全く新しい基準・考えに基づいて解釈する」となる。先行研究により、分類手法の正しさは帰納法的にだが検証されている。述べた通り藤枝分期法の正しさ自体は検証が進んでいるため、機械を用いた方法でも「先行研究や藤枝分期法と同様の基準を用いて図像を解釈する」ことが出来れば、AIを用いた藤枝分期法の再検討として十分なものになると考えている。

本研究では特徴量という数値化された特徴を見つけ、クラスタリングという手法を用いてグループに分けるが、もし書体や年代、地域という指標で解釈可能な分析結果を得られれば、他の写本の分析への応用可能性を示すことが出来る。他の写本への応用の可能性は、敦煌学だけ

でなく、手書きの写本を扱うすべての分野で利用できる可能性がある事を示すものになる。仏教学においても写本研究の新しい可能性として有用であることは明白であり、強いインパクトをもたらすものと確信している。

第5節 年代分類法、年代分類研究関連の先行研究

年代分類に関する研究としては、藤枝以外に、赤尾、池田、石塚、高田時雄らが文学的なアプローチによる研究を行っており、最近では岡田至弘・江南和幸・坂本らによって工学的アプローチによる研究が行われている。

特に、岡田は紙質を写真上で判定するためのアプローチを、江南は成分分析を、坂本は簀の目等から統計的・成分分析的アプローチを行い、様々な方面から、紙という1属性に対して検証を行っていた。この部分の先行研究としては坂本・岡田[2017]や、江南[2021]がある。

中国では近年、趙聲良や張涌泉らによって敦煌文書の年代研究についての論文、著書が発表されている。趙[2000]では中国書道史、特に書道芸術の観点から敦煌写本を3つの大きな時期と16の小さなスタイルに分け、それぞれの特徴と源流¹⁷进行分析している。

張[2013]中、「敦煌文献的断代」では、先行研究を総合して、年代の判断を行うための判断基準と実際の写本を使い、自身の研究成果を並べ、年代判定の実例を示している。

ただ、日本では現状、藤枝分期法がもっともよく使われている。藤枝分期法の基準の曖昧さには疑念を持つ研究者もいる。上杉智英[2022]では、再現性がないという部分について「藤枝分期法となるべく同じ方法を用いて同じ結果が導けるのかどうか」という再検討を行っていた。

第6節 先行研究

古写本の分析にAIを用いた例としては以下のものがある。

Mladen Popović, Maruf A. Dhali, Lambert Schomaker[2019]は、古文書に対して筆跡の違いを機械学習で見出し、筆跡鑑定とその研究という観点において、機械学習の有効性を示した論文である。死海文書の「大イザヤ書巻物 (1QIsaa)」からヘブライ語の全文字種を抽出し、これらに主成分分析、ニューラルネットワーク等の複数の機械学習的アプローチを行うことで、文章の途中から文字の特徴に有意差があることを明確にし、2人の筆写者が関与していることを結論付けた。結論部分では、外的要因の関与を排除しきれないとしたものの、特徴が大きく2種類に分かれた事、伝統的な古文書学の観点から視覚的に分析しても、2種類の特徴を示す画像は書体に微妙な変化があることが、ヒートマップなどを用いて数値としても示されたこと等は、機械学習を用いることで古文書研究に新たな可能性を示したと言える。この論文は死海文書の筆写者の識別における人工知能の有用性と革新性を実証したが、これは死海文書以外に

も応用可能と推測される。

字形という観点から分析を行った研究には、雨宮健祥[2019]がある。この論文では、文字の形状、文字タイプ¹⁸先行研究で指摘されている銘文の表現等から、年代や制作者（工人）特有の特徴を分析し、最終的に鏡群や工人同士の関係性を述べている。従来の写本研究で用いられる方法と同様に、類似パターンからの推定や画像を並べて比較し関係性を探す手法だが、字形の分析に限定しても工人や字形同士の関係性を見出すことが出来るという点において、字形分析の重要性を明確にした。

第2章 AIを用いた分析方法

修士論文における機械学習を用いた分析モデル構築手順・分析モデルを利用した分析手順については、FIT（情報科学技術フォーラム）2024において発表予定である。そのため、本論文では一部割愛し、データの選定、分析手順の概要、データ処理の説明ののち、分析結果を述べる。

第1節 データセット作成とデータの選定

ここでは、本研究で使用するデータセットを作成するにあたり、前提や写本の選定基準、取得元データベース（以下DBと省略）の概要説明、実際の画像の取得手法の詳細を述べる。

写経内の文字の字形分析を行う前に、文書を構成している属性や情報について考える必要がある。中国の漢字文化の中で、使われる書体は時代とともに変遷していくが、敦煌文書の作成年代である北魏から北宋までの時期は、隸書が衰退し楷書・行書・草書に移り変わっていた過渡期に相当する。衰退した典型的な字体ではない隸書が特定の用途で使われる中で、未完成な楷行草が広く実用で使われ始めていた。これらは藤枝らの研究を始め、多くの研究者によって述べられており、本論で使用する資料は識語等の書誌事項と先行研究の成果によって、書体や書誌属性が示されている。

「楷書」について、藤枝[1999]では「楷書とは兎毫竹管の“今筆”（きんぴつ）を使って紙に書くために工夫せられた書体を、従来のものと区別するためにした呼び名のこと」と述べられている。隸書との定義を厳密にする際、藤枝は用筆と、書体そのものの用途で一応の区別をつけていたと思われる。ただ、書体の用途については、各分野で異なる内容提示・採用しており、定義としての厳密さに欠けている。

また、藤枝によれば、用筆の違いによる影響は、用筆の交代という形で年代に顕著に影響した。つまり、用筆の交代と関連事象の変遷に関係性が示されたことによって、隸書と楷書は年代で大きく区別可能になった。これにより、逆説的に楷書と隸書は用筆が不明でも年代でおおよそ区分して良いと考えられる。このため、隸書と楷書の判断が下されておらず、用筆の種類

も不明な書は、識語の書写年代を基準に隸書と楷書のどちらに分類するかを判断した。

字形の分析を行うためには、文書の書き手について検討する必要がある。敦煌文書は、中国の歴史的 policy としての写經所造営の影響を受け、写經所所属の役人と考えられる識語の書かれた經典が多く存在する。藤枝は、「敦煌の写經所の写字生によって多くの經典が書写される事業が存在した」と指摘し、書写専門の役人がおり、師弟関係を用いて書写の訓練がされていたと述べた。官営の「写經所」では、書写専門の写經生（写字生¹⁹）がいた事は藤枝[1960]を始め、他の研究者によっても認められている。しかし、写經所とは異なる寺院での書写で、専門の書写担当者がいたかどうかは疑問が残る。大屋正順[2014]には、北魏時代の後半の写經所で、専門写字生として、令狐崇哲とその弟子と考えられる「令狐」の一派が存在したとされる研究結果が示されている。

字形分析の際、文字に対し複数の字形を機械によって分類するが、それらの字形の差が書写者の個人差や時期差に起因するものであり、気分やその他の要因によって簡単に変化するものではないことを、本来であれば前提としなくてはならない。だが、本研究で使用した方法では、「字体の規範性」²⁰の度合いすらも主観的な情報として処理される。そのため、分析では、あくまでも先行研究と書誌情報を主として文字の信頼性を担保し、数値として取得可能な分散等の定量データと、先行研究で示された情報の比較によって字形の揺れや信頼性についての検討を行う。

対象の選定基準について、本研究では、分析対象として表1に示す15点の写本を、この基準で選定した。すなわち、識語が同紙内に存在すること、手書き且つ毛筆で書かれていること、標準形式であること、画像データが存在することの4点とした²¹。

画像データは、HNG・IDP データベース・Gallica の3種類のオンライン公開されているデジタルDBから取得した。

研究の最初期では漢字字体規範史データセットからデータを入手していた為、半数はデジタル写真から取得したデータではなく、HNGに収録された二次的な画像を使用している。

これらの写本は、一部を除き識語が明確な写本で、No.1~6は非写字生（主に僧侶）の書写した楷書の写經、7-12は写經所に所属する書写専門の写字生が書写した楷書とされている写經、13-15は写經所に所属する書写専門の写字生が書写した隸書とされている写經というカテゴリとした。

写字生の書写かどうかは先行研究で示されている内容に加え、識語で判断を行った。

表中の書名、(書写)年代、書写者は、池田[1990]に掲載されているものを使用した。

經生は写經所に所属する写經生（藤枝は写字生や専門写字生等と呼ぶ）を指す。写經生は訓練を受けた書写専門の役人と考えられている。No.13~15は「敦煌鎮」と呼ばれた時期の写經で

表1 用いたデータセット

クラス 番号	文書名		年代	枚数	書写者	HNG
1	『大棲炭經卷三皇后題記』	(P2413)	589	50	記載なし	○
2	『賢劫經卷一雍縣張法僧等題記』	(正倉院本)	610	57	沙門僧粲(海)校	○
3	『妙法蓮華經卷五比丘尼法藏題記』	(P2334)	617	33	記載なし	○
4	『妙法蓮華經卷八點記』	(S2577)	7C 末	35	記載なし	○
5	『妙法蓮華經卷四比丘尼善信題記』	(S2157)	691	2	比丘尼善信	
6	『魔訶摩耶經卷上彭普信題記』	(P2160)	586	54	菩薩戒弟子彭普信	○
7	『妙法蓮華經卷五使虞昶等校經列位』	(今西本)	671	44	不明	○
8	『妙法蓮華經卷三使閻玄道等校經列位』	(守屋本)	675	30	書手趙玄詳寫	○
9	『妙法蓮華經卷六使閻玄道等校經列位』	(P2195)	675	27	門下省書手袁元愆寫	○
10	『妙法蓮華經卷四使虞昶等校經列位』	(S312)	673	34	門下省羣書手封安昌寫	
11	『妙法蓮華經卷四使虞昶等校經列位』	(S3079)	671	60	經生郭德寫	
12	『妙法蓮華經卷四使虞昶等校經列位』	(S4551)	672	55	門下省羣書手劉大慈寫	
13	『華嚴經卷十六敦煌鎮經生題記』	(S2067)	514	49	敦煌鎮經(上踏黒印)	○
14	『華嚴經卷卅五敦煌經生題記』	(P2110)	513	104	敦煌鎮經生帥令狐崇哲所寫經成訖竟	
15	『華嚴經卷卅九敦煌鎮官經生題記』	(S9141)	513	2	敦煌鎮官(上踏黒印)經生帥令狐崇哲所寫經成訖竟。	

あることを示している。藤枝はこの時期の写経は、書体編年上重要であると示した。これについては大屋[2014]で事例を挙げて詳しく述べられている。

本来であれば、データの質や条件は揃える方が好ましいが、本研究ではデータ収集中のトラブルにつき、全てのデータをDB化された写真から取得することができなかった。これにより、本論文で取り扱ったデータの半数は二次的な情報源から得たものであり、ノイズの増加や情報の損失に影響があった点には留意する必要がある。

第2節 画像データの取得

一般に、画像認識ないしは画像分類は、大きく分けて機械を用いる前と後に分けられ、前者は前処理と呼ばれる。前処理においては、画像から対象を切り出す、正規化、ノイズ除去が行われる。前処理の善し悪しが全体の結果を決定するため、対象の画像に対しては最適の処理を行う必要がある。

本研究では数ある漢字の中で、「不」を用いて検証している。「不」を選択した理由としては異体字が確認されていないこと、形状的に文字が潰れにくいこと、文法上省略されにくく、出現回数が多いこと、手作業での作業時に中心を取りやすい形状であること等が挙げられる。一方、主成分分析を用いる関係上、個々人の筆致や地域毎の変化を確認するという目的に沿うには、明らかに特徴が異なる異体字を含めることは、検討事項の増加を招き、適切ではないと考えられた。そのため、計算は画像データの容量が大きくなるほど時間を要するが、画数が比較的少ないという点からも、本研究対象となる文字として「不」を選択した。

写本から得た画像は重複のない「不」636枚で、全て手作業で画像のアーカイブDBから切り抜いた。命名規則により、全ての画像がどこに位置しているかを追跡することが可能になっている。

第3節 【前処理フェイズ】

画像の取得後、前処理として、機械で効率よく分析可能なように、ノイズの減少やデータサイズの縮小などを行う。本研究ではすべての画像に対しグレースケール化と正規化を行ったあと、画像の濃淡の影響による分析結果の変化を確認するため、二値化を行った。これらの処理は全て、作成した自動化プログラムにより処理した。

本来は、すべての処理をプログラムで行うことが良いと考えられている。しかしながら、研究対象の性質上複数の問題があり、画像の切り出し、位置合わせまでを手作業で行い、グレースケール化・正規化・二値化・リサイズを自作プログラムで実装して処理した。

これらの処理をした画像を1つのフォルダにまとめ、基礎となるデータセットが完成する。次に、事前に作成した前処理用プログラムを実行して、「グレースケール化」「正規化」「二値化」「リサイズ」を行った。

グレースケール化とは、カラー画像をグレースケール画像（グレーを256階調で表現したもの）に変換することである。グレースケール化により、色情報を削減することによるデータ量減少（計算コスト低減）、テクスチャーや形状等の特徴の強調、撮影条件の違いによる影響の低減を狙った。

正規化とは、画像の画素値を一定の範囲に調整することである。正規化により各画像の撮影

環境・条件の違いによる影響を最小限に抑え、機械学習モデルが画像の本質的な特徴に焦点を当てやすくした。

二値化とは、画像を黒と白の2色に変換する処理のことで、本研究では、自動しきい値決定²⁶の手法である「大津の二値化」を使用した。

リサイズは正規化・グレースケール及び二値化された画像を特定のサイズにする処理である。画像の情報量を削減する効果があるが、情報量が削減されることによる損失には注意が必要である。

・前処理

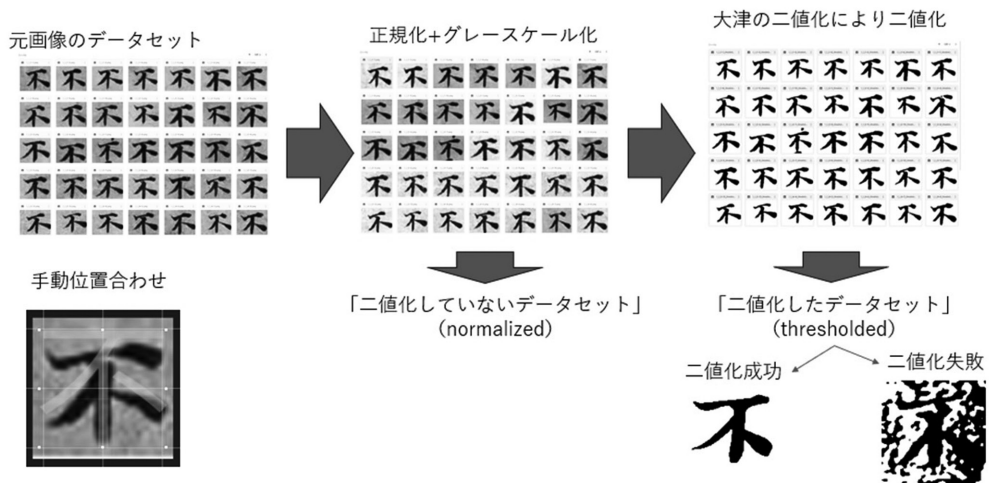


図3 グレースケール化と正規化、二値化の処理結果を示したもの

データ数ある程度保持する必要があるため、多少のノイズが残留している点については無視²⁷する。

分析・実験で使ったプログラムはすべて、Google Colaboratory を利用し、コードの記述と実行を行った。使用した PC 環境は Ubuntu 22.04.3 LTS、言語は Python 3.10.12²⁸。関連データは全て Google Drive に保存した。ライブラリは主に scikit-learn を使用している。

実験手法等の詳細については前述の理由から割愛し、以下ではフローチャートのみ掲載する。図4は、次章以降で用いる分析の流れを可視化したものである。

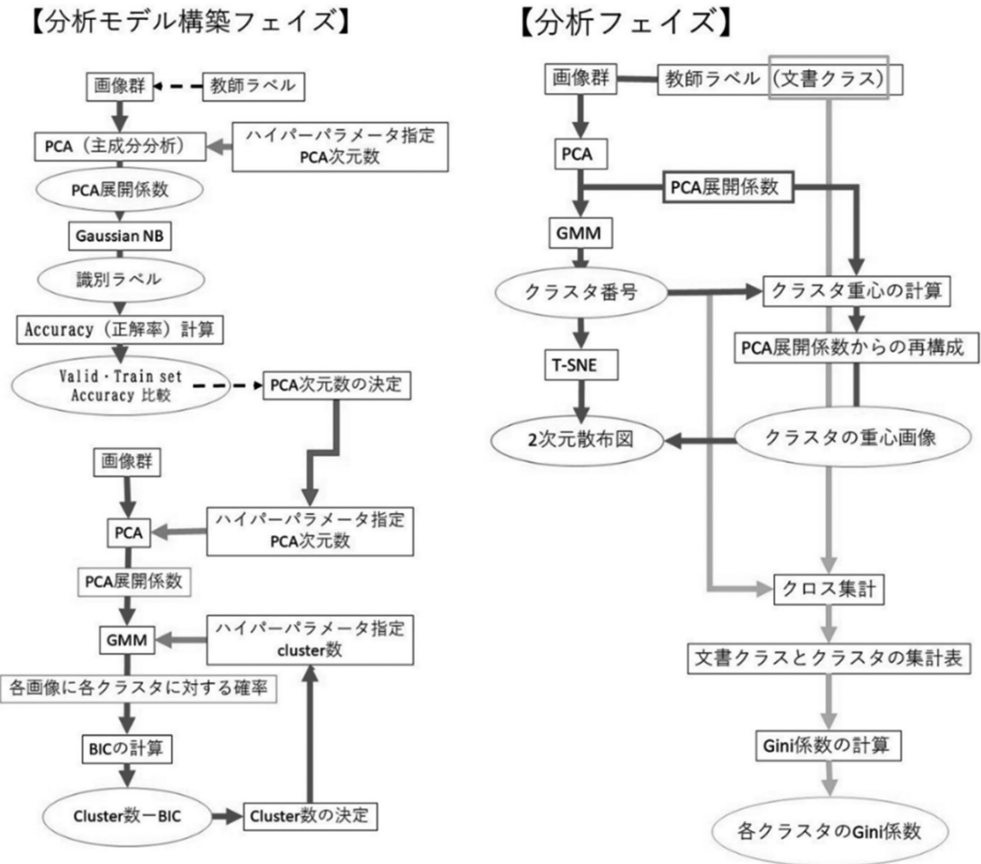


図4 処理・分析手順フローチャート

第3章 分析結果

ここでは、分析によって得られた結果について述べる。

まず前処理の影響を確かめる為、画像処理の有効性を確認するための比較を行う。

第1節 画像処理技術（前処理）の適用結果

図5では分析プログラムによって得られた散布図を並べている。左右の図は二値化処理の有無のみが異なり、それ以外は同じ条件で出力した。左図は入力画像の濃淡に影響を受けており、右図では大方、文字の形状で画像がまとまっている。この結果から、全体として「字形を見る」のであれば、二値化を行うほうが適切だと判明した。また、HNGやIDP等の各DBで分離する等の影響は見られないため、DBや撮影環境の違いによる影響は二値化によって減衰し

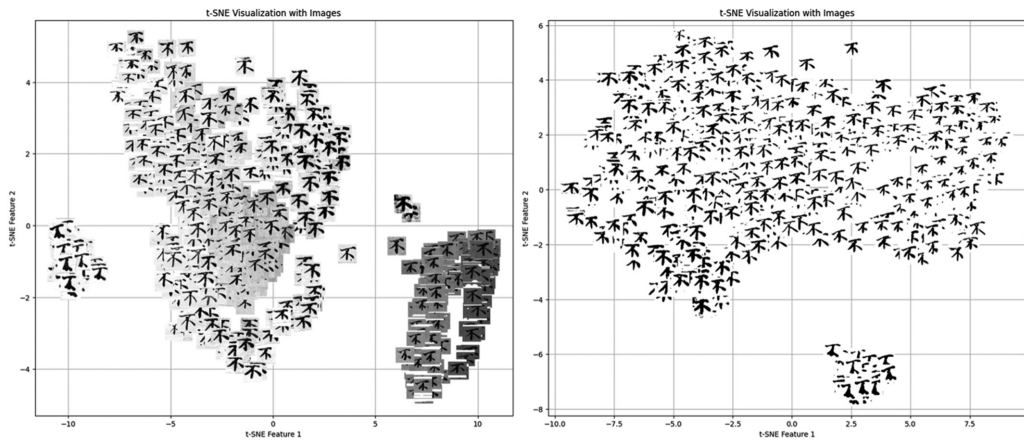


図5 PCA・t-SNEの二次元散布図 左図：非二値化データ使用・右図：二値化データ使用

ていると考えている。

前処理を適切に行うことにより、分析結果には大きな違いが生まれることを実例を用いて検証した。画像取得時に画像データが抱えていた問題は、適切な前処理によってある程度減衰させることが可能であると分かった。

第2節 PCA 展開係数の分析結果と主成分の解釈

本研究の分析で用いるパラメータの1つに、「PCA 次元数」がある。次元数の数だけ、特徴を捉える「主成分」がある。本研究では、分析をする上で効率の良い次元数を機械的に計算・探索し、次元数22をPCA次元数と決定した。

図6はPCAで分析したすべての成分の固有画像を一覧にしたもの。固有画像は各次元が捉えた特徴を可視化したもので、PCA次元数の数だけ存在する。単体ではその変化はわかりづらいが、次元数の増加は最も平均的な画像から細かな特徴を捉えた画像へ変わっていく。白や黒の色の濃い部分が特徴として注目されている箇所を示す。1・2・4画目に色が集中していることから、この例であれば1・2・4画目が特徴情報を捉えやすい部分であると分かる。

この固有画像を解釈することは難しいが、ここに書写者固有の特徴や用筆の特徴などが含まれていると考えられる。機械が捉えた特徴を示す固有画像が、人間の解釈と必ずしも合致するとは限らない。今回の方法では、人間とは異なる機械独自の観点をを用いて判断していると考えられた。

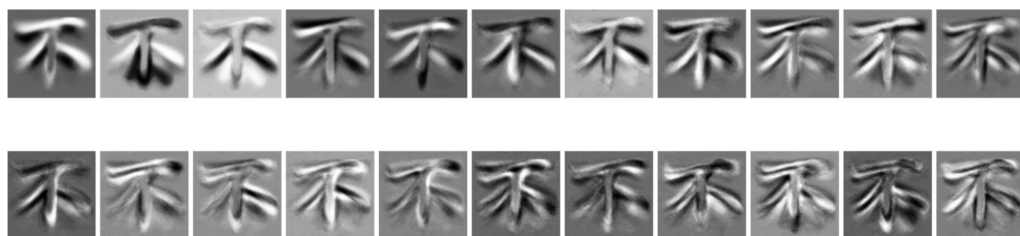


図6 固有画像22枚（次元数順）

第3節 クラスタリング結果

重心画像と BIC

ここではクラスタリングを行った結果得た散布図と、そこで使用したクラスタ数を決定する根拠、重心画像を確認する。図7の左にあるクラスタリング結果は、最適なクラスタ数で実行して得られた散布図に、それに各クラスタの重心画像を併記したもの。

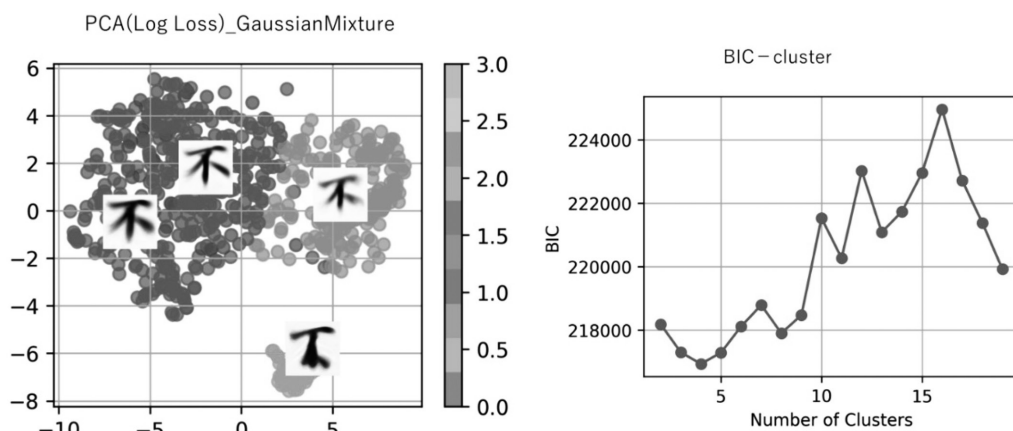


図7 GMM によるクラスタリングの結果と BIC の結果のグラフ

図7では Gaussian Mixture とその評価指標 BIC を用いて最適なクラスタ数を探した結果を示している。BIC の値は低いほどクラスタとしてデータ分離が正しくされているともいう指標になるため、筆者の検討による限り、最適なクラスタ数は4だと考えられた。

クラスタの重心画像はそのクラスタの最も平均的な特徴を捉えている。全ての重心画像には各々明らかな違いがあるが、この違いを詳細に言語化することは現状難しい。特に重心画像の上部の3種類（クラスタ0～2）は類似した字形を分類しており、何らかの特徴を捉えて、より細かく分類している。

クロス集計表

また、図8では、ファイル名・写本名・書写年代と、どのクラスにどの写本が何枚分類されているかを一覧表にしたクロス集計表を示す。

Original Class	写本名称	固有番号	書写年代 (漢語を基準とする)	枚数	藤枝分類	α β γ ϵ				α β γ				
						PCA22(Log Loss) GMM cluster = 4				PCA22(Log Loss) GMM cluster = 3				
非写経生 楷書写経	1	大樓炭經卷三 (P2413)	[隋寫本 589年]	50	B	α	49	0	1	0	α	49	0	1
	2	賢劫經卷二 (正倉院本)	[隋寫本 610年]	57	B	β	2	37	18	0	α	37	0	20
	3	妙法蓮華經卷五 (P2334)	[隋寫本 617年]	33	B	γ	0	0	33	0	γ	0	0	33
	4	妙法蓮華經卷八 (S2577)	[初唐寫本 70末]	35	C	β	13	21	1	0	α	34	0	1
	5	妙法蓮華經卷四 (S2157)	[691年]	2	C ?	γ	0	0	2	0	γ	0	0	2
	6	摩訶摩耶經卷上 (P2160)	[敦煌南北朝寫本 586年]	54	A	α	52	2	0	0	α	54	0	0
写経生 楷書写経	7	妙法蓮華經卷五 (今西本)	[初唐寫本 671年]	44	C	γ	0	20	24	0	γ	12	0	32
	8	妙法蓮華經卷三 (守屋本)	[初唐寫本 675年]	30	C	β	1	29	0	0	α	30	0	0
	9	妙法蓮華經卷六 (P2195)	[初唐寫本 675年]	27	C	α	25	2	0	0	α	27	0	0
	10	妙法蓮華經卷四 (S312)	[673年]	34	C	β	0	32	2	0	α	32	0	2
	11	妙法蓮華經卷四 (S3079)	[671年]	60	C	β	4	44	12	0	α	47	0	13
	12	妙法蓮華經卷四 (S4551)	[672年]	55	C	β	4	47	4	0	α	51	0	4
写経生 隸書写経	13	華嚴經卷十六 (S2067)	[敦煌南北朝寫本 514年]	49	A	ϵ	0	0	0	49	β	0	49	0
	14	華嚴經卷35 (P2110)	[513]	104	A	γ	2	3	99	0	γ	6	0	98
	15	華嚴經卷39 (S9141)	[513]	2	A	γ	0	0	2	0	γ	0	0	2
				枚数合計	636									

図8 クロス集計表

左から順に、オリジナルクラス番号、所蔵固有番号、写本名、書写年代、データセット入力枚数、藤枝著書に掲載されている写本とその分類、池田[1990]を始めとした先行研究に掲載された藤枝分期法の分類、クラスタ分割 (α , β , γ , ε) の順に並んでいる。図7のクラスタ0が α 、1が β 、2が γ 、3が ε に相当している。(α , β , γ , ε) はクラスタリング結果を Hard Voting で分けた時の名称として仮で使用している分類記号である。クロス集計表はPCAとクラスタリングの両方の影響を受ける。クラスタ分割の並び順自体に情報は無い。

クラスタ数4の時は β に、比較のために行ったクラスタ数3の時は α に、藤枝分期法での分類上Cに該当するものが多く分かれた。ここから、PCAと次元削減、クラスタリングを用いた分析では、藤枝分期法のCに分類される画像はまとまりやすいと分かる。

これに加え、No.14と15の敦煌鎮經生師令狐崇哲が書写した写経はクラスタ数が変化しても同じクラスタに分けられている。同一著者による写本は同じクラスタに分布される可能性はあるが、データ数が少ないため未検証である。

第4節 実験結果まとめ

- ・敦煌文書の字形分析を行う上では、適切な前処理の有無が大きな影響を与える。
- ・固有画像からPCAが特徴を可視化したが、人間とは異なった観点から分類している。
- ・使用したデータは分析とクラスタリングにより3-4つに分けることが可能。
- ・3-4つに分けられた画像のグループの平均的な特徴を可視化すると、人間から見ても異な

る特徴をとらえているため、細かな特徴を見分けられる可能性がある。

- ・クロス集計表によりから、藤枝分期法のCは他の写本とは異なる挙動を示し、定量的な違いを持っていると判明した。
- ・書写者が同一であれば、クラスタリングで同じ傾向を示す可能性がある。
- ・クロス集計表から、『華嚴經卷十六』(S2067)は、データとして独立した特徴を持つ。特異な特徴を持つ写本は定量的に分離可能である可能性が高い。

本章までで判明した分析結果は全て、機械に画像だけを読み取らせて分析・分類を行った結果得られた情報であり、奥書や識語などの書写情報を必要としない分類の実現可能性を示したものと言える。

検証データ不足のため希望的観測になるが、敦煌文書研究においては、対象に存在する無紀年の写本や断片の画像を入力するだけで、どの写本・書写者に類似しているかを定量的に判断できる可能性がある。これにより、未研究資料の同定や分類に大きな影響を与えられると思われる。手法の確立に向け、今後は他字種における傾向把握等の検証に取り組む予定である。

終わりに

本研究では、まず敦煌文書の概要を述べ、藤枝分期法の限界点を指摘した後、字形分析を行うべく、まず敦煌文書の写本にある「不」の漢字画像のデータセットを作成し、主成分分析(PCA)を始めとした機械学習を用いて分析した。次に、分析結果と先行研究の比較検討によって、各クラスがどのような漢字の字形を有しているか、各クラスにどの敦煌文書の写本が分類されるかなどについて、二次元散布図・PCAの固有画像・クロス集計表などを見て考察を行った。

その結果、特定条件下であれば、画像のみを用いて、写字生の楷書を定量的に分類できる可能性があることが分かった。また、特異な特徴を持つ写本であれば、クラスタリングの際にその写本だけを独立させ、定量的に分離できた。以上のように、機械学習を用いた工学的画像処理の手法によって字形分析に良好な結果が得られることが判明した。

今後の課題としては、検証不足やデータ不足の解消が挙げられる。研究成果として得たクラスタリングの傾向が「不」特有のものであるか否か、また、データセット作成時点で発生したデータの質の問題への対処等の機械学習関連の諸問題の解消が必要である。それに加えて、敦煌写本研究、古写本学などの写本研究や編年研究で過去の研究者が共通して持っていた暗黙知のさらなる定量化は、この研究の大きな課題である。

将来の研究への展望としては、本研究で用いた手法から進展させて、畳み込みニューラルネットワークを始めとした深層学習の手法を用いた分析や、分析対象の幅を増やしても同様の傾

向が示されるのか検証することを考えている。

このような新しいアプローチは、過去の研究を検証する上でも意義のあるものと考ええる。本研究で用いた機械学習による字形の分析手法は、今後の追加検証を必要とするものの、言語や文字に関係なく利用可能である。今後は、藤枝分期法の再検証のみならず、古写本と呼ばれる資料を扱うすべての研究に資するような、汎用的な分析手法の確立を目指していきたい。

【参考文献一覧】

Fujieda Akira

[2002] “Chronological classification of Dunhuang Buddhist manuscripts.” *Dunhuang manuscript forgeries* / edited by Susan Whitfield (The British Library studies in conservation science; 3) pp.103-114.

Mladen Popović, Maruf A. Dhali, Lambert Schomaker

[2019] “Artificial intelligence based writer identification generates new evidence for the unknown scribes of the Dead Sea Scrolls exemplified by the Great Isaiah Scroll (1QIsaa)” *PLoS ONE* 16(4)

Whitfield Susan,

[2002] *Dunhuang manuscript forgeries* / edited by Susan Whitfield (The British Library studies in conservation science; 3) British Library

赤尾栄慶

[1998] 「敦煌写本の書法と料紙に関する調査研究」1998年度 研究成果報告書概要」科学研究費助成事業データベース

[2002] 「敦煌写本における偽写本の問題について」『京都国立博物館學叢』京都国立博物館

[2003] 「書誌学的観点から見た敦煌写本と偽写本をめぐる問題」『佛教藝術（特集 敦煌学の百年）』271号 pp.33-47 佛教藝術學會編 毎日新聞出版

雨宮健祥

[2019] 「三角縁神獸鏡の銘文字形分析：編年と作成工人」『東京大学考古学研究室研究紀要32』東京大学考古学研究室 pp.35-72.

石塚晴通

[2012] 『漢字字体史研究』勉誠出版

[2016] 『漢字字体史研究 二』勉誠出版

池田温

[1990] 『中国古代写本識語集録』東京大学東洋文化研究所

池田温編

[1992] 『講座敦煌 5 敦煌漢文文献』大東出版社

池田証壽

[2016] 「漢字字体史の資料と方法：初唐の宮廷写経と日本の古辞書」北海道大学文学研究科紀要 150号 北海道大学文学研究科 pp.201-236.

上杉智英

[2022] 「徳化李氏凡将閣珍藏」本の再検討」「敦煌写本研究の現在」京都国立博物館国際シンポジウム

江南和幸

- [2021]「コディコロジーと紙の科学的分析学との邂逅：科学的コディコロジーの提案」『紙のレンズから見た古典籍—高精細デジタルマイクロスコープの世界—シンポジウムプログラム2021』pp.14-33.

榎一雄編

- [1980]『講座敦煌 1 敦煌の自然と現状』大東出版社

上山大峻

- [1990]『敦煌仏教の研究』法蔵館

大屋正順

- [2014]「北魏敦煌鎮写經の書風について」『大正大學研究紀要』巻99, pp.312-295.

北本朝展

- [2017]「IIIF Curation Viewer の開発と利用」『IIIF Japan シンポジウム～デジタルアーカイブにおける画像公開の新潮流～』

西原一幸

- [2015]『字様の研究：唐代楷書字体規範の成立と展開』勉誠出版

坂本昭二

- [2016]「敦煌漢文文書の Scientific Codicology」『研究報告人文科学とコンピュータ (CH)』pp. 1-7.

坂本昭二・岡田至弘

- [2017]「敦煌漢文文書の紙に使用されている粒状物の分析」『考古学と自然科学 = Archaeology and natural science : 日本文化財科学会誌』72号 日本文化財科学会 pp.63-75.

高田智和

- [2012]「第四部字 体史研究の方法 国語研本金剛頂經の漢字字体」『漢字字体史研究』

藤枝晃

- [1960]『墨美』第97巻 墨美社1960.

- [1961]「敦煌出土の長安宮廷写經」『仏教史学論集：塚本博士頌寿記念 通号29』塚本博士頌寿記念会, 647-667(R)

- [1985]「「德化李氏凡将閣珍藏」印について」『学叢』第7巻京都国立博物館, pp.153-173.

- [1987]「中国北朝写本の三分期」『古筆と国文学』八木書店, pp.4-5.

- [1999]『敦煌学とその周辺』ブレーンセンター

- [1999]『文字の文化史』講談社学術文庫

- [2005]『トルファン出土仏典の研究：高昌残影积録』法蔵館

永崎研宣

- [2021]「デジタル・ヒューマニティーズ (DH) の概要と人文学 DX」人文学・社会科学特別委員会 (第7回) 会議資料

趙聲良

- [2000]「早期敦煌写本书法の分期研究」『1994敦煌学国际研讨会文集』敦煌研究院

史睿 (翻訳者：波多野由美子、矢越葉子)

- [2021]「敦煌写經の年代判定と書法史」『古代学研究所紀要』第31巻 pp.1-21. 明治大学日本古代学研究所

张涌泉

- [2013]『敦煌講座書系 敦煌写本文献学』甘肅教育出版社

註

- ① 池田温編[1992]p.18を参照のこと。
- ② これ以外の分類法については、次項「現状の課題」で詳しく述べる。
- ③ 藤枝[1987]pp.4-5
- ④ 藤枝[1999]p.44
- ⑤ 池田温編[1992] (p.19)のほか、赤尾栄慶[1998] (https://kaken.nii.ac.jp/report/KAKENHI-PROJECT-08041034/080410341998kenkyu_seika_hokoku_gaiyo/ (閲覧日2024年1月17日))等
- ⑥ 藤枝晃「敦煌写経」『墨美』97号 p.3
- ⑦ 藤枝[1999]pp.39-40
- ⑧ 1960年代に提唱されたのち、最後に藤枝から言及・修正などが為されたのは、大英図書館の「The British Library studies in conservation science ; 3」が最後で、内容はFujieda[2002]に書かれている。同時期に編集されていた藤枝編著[2005]でも記載はあるが、内容としてはFujieda[2002]から殆ど変化はない。そのため、最低でも20年と考えてよいと思われる。
- ⑨ 少なくとも、藤枝は「まだジャイルズの目録が出ていないとき」に日付のない個々の写本の年代判定のための“モノサシ”を作ろうとしていた。ジャイルズの目録とは大英図書館が出版した1957年の「Descriptive Catalogue of the Chinese Manuscripts from Tunhuang in the British Museum」で、それ以前に作成された目録については、「講座敦煌 5」p.10に、1931年刊『敦煌劫余録』（北京蔵8679点）が挙げられるに留まる。他の目録の有無は不明だが、1909年のスライン・コレクションの公開から半世紀あまり時間で、それほど多くの目録は作成されていないと考えるのが自然である。
- ⑩ 原本が一次資料であることは明らかなが、写真と印刷では情報の損失具合が異なる。印刷は写真とは異なり、印刷という製法上、微妙な歪みや色ムラなどが出来ることで、多少情報を失う。写真も勿論歪みや色ムラは撮影環境上発生し得るものの、処理技術の向上によって補正を行うことが可能であり、情報の損失は印刷よりも少なくすることが可能。そのため、「印刷したかどうか」は注視すべき点だと考えている。
- ⑪ 張[2013]第二节 据书法断代の一部。著者が日本語で要約。
- ⑫ 上山[1990]p.3より著者要約。
- ⑬ 藤枝[1999]p.37
- ⑭ 榎一雄編[1980]『講座敦煌 1』大東出版社 p.392でも、敦煌を研究することの重要性について述べている。
- ⑮ 上山[1990]『敦煌仏教の研究』
- ⑯ 永崎研宣「デジタル・ヒューマニティーズ (DH) の概要と人文学 DX」科学技術・学術審議会 学術分科会 人文学・社会科学特別委員会（第7回）資料から引用。
- ⑰ 論文の注釈には「关于早期敦煌写本的分期,藤枝晃先生根据纸张的质地等方面作过精辟的研究,参见藤枝晃《中国北朝写本的三个分期》(《敦煌研究》一九九〇年二期),颇有启发性。但本文意在作书法史的考察,故未采用藤枝先生的时代分法。」と書かれており、あくまでも藤枝分期法とは異なった観点で年代を分ける方法を示しているため、藤枝分期法は使用しないとしている。
- ⑱ 【データセット作成】の部分に詳しく述べている。
- ⑲ 表1の「経生」に同じ。
- ⑳ 字体規範が存在するという石塚らの論から、字体にあるなんらかの「規範性」というのがデータ点の分散度合い等、数値として得られる可能性があると考え、石塚らの述べる「字体規範」とはあえて異なった言い回しをしている。感覚的情報として、字体に存在する「硬さ」等がこれに

当たると著者は考えているが、実際その感覚の出どころが個々の文字に由来するのか、書全体を俯瞰した結果として述べられるものなのかはわかっていない。

- ②① 7番『妙法蓮華經卷五使虞昶等校經列位』は、個人所蔵とされており、図録その他で原本を確認することができなかったが、修士論文にも掲載しているため、本論文でもそのまま掲載した。このことから、識語の確認ができず、書写者は不明とした。
- ②② 漢字字体規範史データセット (HNG) は漢字字体規範史データセット保存会が公開・保存しているデータセットのこと。公開 HP には「漢字字体規範史データベース (Hanzi Normative Glyphs)」(公開期間: 2005-2015年) の前身である「石塚漢字字体資料」の情報カードの画像、漢字字形の切り抜き画像、及びメタデータを「漢字字体規範史データセット」と称して公開するリポジトリです」とある。(漢字字体規範史データセットトップページより <https://www.hng-data.org/> (閲覧日2024年1月9日)) 元々は石塚晴通らが作成した「石塚漢字字体資料」が基で、関連研究成果は『漢字字体史研究』『漢字字体史研究2』として出版されている。
- ②③ IDP の公式 HP URL : <http://idp.bl.uk/>
- ②④ Gallica 公式 HP URL : <https://gallica.bnf.fr>
- ②⑤ 書写されている奥書の情報に誤りがいないかどうかは不明だが、「年代・先行研究で言及されている場合に、どのように分類されているか、という基準においてはどうか」という仮定を含めている。
- ②⑥ 事前に固定閾値 (Fixed Thresholding) を設定する方法と、適応閾値 (Adaptive Thresholding) 決定という自動閾値決定方法を検証したが、固定閾値・適応閾値ともに処理後のノイズがひどく、ノイズが特徴としてとらえられることが多々あった為、採用しなかった。
- ②⑦ 現状、目視確認ではノイズが顕著な画像を前処理後の画像で確認することが出来なかったためそのままにしているが、今後データを増やす際はノイズの有無や量に十分に注意する必要がある。これは、PCA の基底にデータの質が影響を与えることが判明しているからである。上の図に示される「二値化失敗」の例示画像は、あくまで一例として示したものであり、データセットには入力されていない。
- ②⑧ Google Colaboratory 上での環境データの参照日時 (main, Nov 20 2023, 15:14:05) [GCC 11.4.0]
- ②⑨ Hard Voting とは、いわゆる多数決のこと。ここではクラスタリングで分けられたデータを最も多いクラスに分類することで、その写本がどのクラスに属するのかを判断しやすくしている。