人文科学とコンピュータ　45－1
（2000．1．21）

古文書翻刻支援システム開発プロジェクト報告
（1）－プロジェクト概要－

○山田奨治(1), 加藤寧(2), 川口洋(3), 原正一郎(4), 石谷康人(5)
柴山守(6), 笠谷和比古(1), 小島正美(7), 梅田三千雄(8), 山本和彦(9)
(1) 国際日本文化研究センター, (2) 東北大学, (3) 帝塚山大学
(4) 国文学研究資料館, (5) 東芝デジタルメディア機器社, (6) 大阪市立大学
(7) 東北工業大学, (8) 大阪電気通信大学, (9) 岐阜大学

この報文では，平成11年度より3ヶ年の予定で開始した「古文書翻刻支援システム開発プロジェクト」の概要とねらい，および進行状況について報告する．このプロジェクトは，手書き日本語文字認識技術を発展的に応用して，古文書の翻刻作業を支援するシステムを開発するための諸研究を実施するものである．われわれは現在，(1) 古文書用例データベースの作成，(2) 古文書文字データベースの作成，(3) 古文書文字切り出し，(4) 古文書文字認識について研究を進めている．

Project report on development for a historical document research supporting system
(1) Project outline

○ Shoji YAMADA(1), Nei KATO(2), Hiroshi KAWAGUCHI(3)
Shouichiro HARA(4), Yasuto ISHITANI(5), Mamoru SHIBAYAMA(6)
Kazuhiko KASAYA(1), Masami KOJIMA(7)
Michio UMEDA(8), and Kazuhiko YAMAMOTO(9)
(1) International Research Center for Japanese Studies, (2) Tohoku Univ.,
(3) Tezukayama Univ., (4) National Institute of Japanese Literature,
(5) Toshiba, (6) Osaka City Univ., (7) Tohoku Institute of Technology,
(8) Osaka Electro-Communication Univ., (9) Gifu Univ.

In this article, we report on the outline, aim, and current status of a project on development for a historical document research supporting system, which has been started since April 1999. This project covers various research topics to develop the supporting system, applying advanced hand-written character recognition technology. We are continuing the following researches for historical document: (1) corpus database, (2) character database, (3) character segmentation, and (4) character recognition.

1　プロジェクトの経緯

歴史学研究においては，古文書の翻刻が研究プロセスの重要な基礎的作業である．古文書翻刻作業は高度に知的な作業で，歴史の基礎知識，文書の種類やレイアウトに関する知識，定型文言・慣用表現の知識，文字の異体字やくずし方に関する知識と翻刻経験の蓄積が必要であり，人間が古文書翻刻作業をひととおりこなせるようになるまでには，相当の訓練期間を必要とする．古文書翻刻の知的プロセスを解明し，その知見にもとづいて古文書翻刻作業の一部を支援するシステムがあれば，歴史学研究の有効なツールとして活用しうるかもしれない．

古文書の文字認識をにらんだ研究は，文献[1, 2, 3]など，これまでごくわずかしか発表されていない．これらの先行研究はいずれも，古文書文字認識の可能性を検証したにすぎないもので，本質的な技術的課題について解答を示したものではない．古文書翻刻支援システム実現のための，基本的かつ特殊な技術的課題に以下のものがある．

1. 古文書文字認識の技術　－　古文書特有の毛筆くずし字，つづけ字の辞書と認識．
2. 文書形式・定型文言の認識技術　－　近世文書に特有の文書類型，「恐々謹言」「仍而如件」などの頻出熟語の考慮．
3. システムと人間のインタラクション技術　－　古文書文字認識において人間が与える情報の範囲，認識結果の提示法，誤り修正方法など．

これらは従来の日本語手書き文字認識研究では未開拓の内容で，あらたな技術開発が必要な分野である．

上記の個別技術課題に関しては，共著者のひとりである柴山が，科学研究費基盤研究「東洋学における大量マルチメディア情報の提供方式の研究」（平成7～8年）で基礎的検討をおこなった．そこでは，歴史史料を対象にした画像資料の入力とデータベース化，ネットワークによる文字テキストや画像資料の提供方式についての研究の一部として，(1) ビデオ撮影による古文書の効率的画像入力法とコンピュータ上での史料復元，(2) 古文書画像の文字切り出しと文字認識に関する基礎的検討をおこなった．

また科学研究費補助金特定領域研究「人文科学とコンピュータ」（平成7～10年）のイメージ処理計画研究，公募研究において，山田，原，小島，川口が，劣化した古文書の画像処理，古文書のひらがな・漢数字に関する文字認識研究を実施し，文書を限定したひらがなにおいて65.8％，漢数字において92％の文字認識率を得ている．

以上のような個別的な古文書認識技術に関する研究成果をもとにして，平成10年8月5～6日に国際日本文化研究センターにおいて「第1回古文書ＯＣＲ（自動読み取り）シンポジウム」が開催された[4]．同シンポジウムでは共著者等が研究発表をおこない，日本史・古文書学研究者，手書き文字認識研究者ら約60名が参加し，(1) 歴史研究者からみた古文書ＯＣＲへの期待，(2) 古文書ＯＣＲ研究の現況，(3) 日本語手書き文字認識の最先端技術の3つのテーマについて討議をおこなった．このシンポジウムの結果，当面の研究方略として以下の4点推進することで，参加者の意見の一致をみた．

1. 対象の選択において，書体の安定した公文書であり歴史的価値のたかいものを対象にする．
2. 文字認識のための辞書構築を進めるために，標準文字データベースを作成する．
3. 古文書読解に関する専門知識を整理し，システム化する．
4. 人間と機械の作業分担を明確化し，両者を円滑につなぐ知的ユーザインタフェースを構築する．

日本語手書き文字認識の最新技術を展開的に応用しつつ，上記課題の (1) ～ (3) を達成し、課題 (4) であげられた知的ユーザインタフェースを備えた，古文書翻刻支援システムの開発をめざした研究の必要性が認識されている．

2　目的と概要

2.1　プロジェクトの目的

本プロジェクトの目的は，古文書翻刻支援システム開発に向けて，文字データベースなどの必要な研究環境の整備とシステム実現のための基礎的な検討を実施することにある．システム実現のための技術的なアプローチは，つぎの3点にある．

1. 古文書学の専門家が持つ古文書認識における認識過程をモデル化し，古文書解読のメカニズムを実証的にあきらかにする．
2. 日本語手書き文字認識技術を古文書に対して展開的に応用する．
3. 古文書翻刻支援に真に有効なマン・マシンインタフェースを検討する．

専門家の古文書解読プロセスをモデル化することは，知能情報学研究として興味深いテーマであるばかりでなく，その知見を利用することにより，古文書解読訓練方法の開発や支援ツールの開発にもつながる．古文書文字認識は，すでに性能向上の限界点に達している日本語手書き文字認識技術研究に，あらたな展開を与えうるものでもある．人文科学研究の現場で使用するコンピュータという観点からは，人間とコンピュータの作業分担のありかたを具現化する部分として，インタフェース研究が重要である．

本プロジェクトは，文字のくずしのはなはだしい文書を含むすべての古文書の解読や，古文書解読の完全自動化を目指すものではない．古文書解読プロセスのモデル化とシステムへの実装を通して，古文書解読という高度な知識処理過程を実証的に解明することと，同一文型・書体の文書が大量にあるような古文書の翻刻において，人間の作業負荷軽減に有効なシステム，人間が得意とする作業は人間が，機械が得意とする作業は機械がおこない，両者の円滑なインタラクションが確保できるシステムの開発が狙いである．

2.2　プロジェクトの概要

本プロジェクトの眼目は，つぎの3点にあるといえる．

1. 古文書専門家がもつ古文書解読の専門知識を構造化し，モデル化する．
2. 30年来の研究の蓄積を有する文字認識技術，なかでも日本語手書き文字認識に関する近年の飛躍的研究成果をもとに，文字認識の範囲を近世（江戸時代）古文書にまで展開して適用する．
3. 文字認識機能と古文書読解の専門的知識を内蔵した知的インタフェースを構築し，翻刻作業に関する習熟度のひくい作業者であっても，短時間によりおおくの翻刻作業がおこなえるシステムを開発する．

そのために当面必要となる作業には，つぎのようなものがある．

- ○ 古文書解読のための専門的知識の抽出と構造化．
- ○ 文字認識に必要な古文書文字認識用辞書の作成．
- ○ 古文書文字認識のアルゴリズム検討．
- ○ これらの作業を実施するための，基本的ツール群の開発．

具体的には，行書体および一部草書体を含み，語彙や文言が限られた証文・触書を中心とした近世文書を対象に，古文書文字認識のための辞書の作成，近世文書のレイアウト・頻出慣用表現などに関する専門的知識の構造化，古文書文字認識エンジンの開発，知的インタフェースの開発をおこなう．当面対象とする近世文書は，以下の文書である．

- ○ 「伏見屋善兵衛文書」（以下「伏見屋文書」と略す）（大阪市立大学所蔵）
- ○ 陸奥国会津郡小松川村「宗門改人別家別書上帳」（以下「宗門改帳」（しゅうもんあらためちょう）と略す）（個人蔵，年齢表記部分）

○ 『柳営日次記』(りゅうえいひなみき)(国立公文書館内閣文庫所蔵，一部)

本プロジェクトの意義は，以下の点にある．

1. 古文書解読のための専門知識をモデル化することで，人間の知能情報処理を解明できる．
2. 日本語手書き文字認識の手法を，古文書に拡大適用するための方法論を確立できる．
3. 知識処理と文字認識を統合した，知的インタフェースのプロトタイプが作成できる．

本プロジェクトを実行するための予算として，日本学術振興会科学研究費補助金・基盤研究 (B)(1) 一般研究「古文書解読プロセスの知能情報学的解明」(平成 11 ～ 13 年度，研究代表者：山田奨治)，同展開研究「手書き文字ＯＣＲ技術を援用した古文書翻刻支援システムの開発」(平成 11 ～ 13 年度，研究代表者：山田奨治) 同一般研究「古文書ＯＣＲの試論的研究」(平成 11 ～ 13 年度，研究代表者：柴山守) を幸いにも得ることができた．本報告共著者らをプロジェクトメンバーとして，現在，鋭意研究を遂行中である．

3 進捗状況

平成 11 年 12 月現在の研究進捗状況を，①古文書解読のための専門的知識の抽出と構造化にむけた「古文書用例データベース」の作成，②文字認識に必要な「古文書文字データベース」，③文字認識の前処理であると同時にデータベース作成の基本ツールでもある「古文書文字切り出し」，そして④「古文書文字認識」の各項目について報告する．

3.1 古文書用例データベース

古文書にみられる用例データとして，定型文言が頻出するタイプの文書に焦点をあて，多数の文書の翻刻された全文をテキストデータとして作成し，用例を蓄積している．具体的には，「伏見屋文書」(図 1) の翻刻作業と全文入力作業を並行して進めている．「伏見屋文書」は，大阪の元伏見坂町 (現在の大阪市南区坂町) の茶屋，伏見屋善兵衛家に伝わった文書である．伏見屋善兵衛は，遊興の地である伏見坂町のなかでも最大の茶屋として栄えた．また町年寄をつとめ，芝居興業にも関係し，何軒かの貸家をもち，金融業を営んだ．

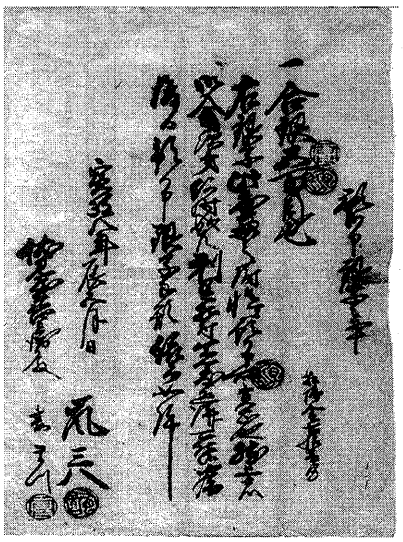


図 1: 「伏見屋文書」

本文書は，文化から慶応年間にいたる各種の証文類である．芝居関係では，天保年間を中心に歌舞伎役者の芝翫，我童らの手附証文がある．伏見屋の金融・借家，同家内部の親族関係に関する諸証文・議定等も含まれている．文書の総数は，証文類が約 1,300 である．

平成 11 年 12 月現在，全文書の見出しはテキストデータとして利用可能である．「伏見屋文書」の見出しに登場する字種は 552，字数は 9,665 である．サンプル数が 20 以上ある字種は 77 で，それで全体の 80.3 %をしめている．頻出字種の上位 20 位までを，表 1 に示す．借金証文の決まり文句である「預り申金子之事」の 7 文字が，上位 7 位に相当していることがわかる．本文の翻刻・電子化は平成 11 年 10 月から作業を開始し，現在全体の約 20 %が完了している．

表 1: 「伏見屋文書」見出しの頻出文字

字種	出現頻度	累積%
之	653	6.8
事	645	13.4
申	348	17.0
り	307	20.2
子	306	23.4
預	290	26.4
金	274	29.2
覚	270	32.0
文	256	34.6
證	229	37.0
一	225	39.3
請	211	41.5
屋	183	43.4
札	182	45.3
銀	156	46.9
月	154	48.5
年	136	49.9
家	125	51.2
状	124	52.5
借	105	53.6
通	103	54.6

3.2 古文書文字データベース

古文書文字認識の試験データとなる文字データベースは，以下の観点から作成している．

1. 字種が限られているが，さまざまな筆跡のサンプルが多数得られるもの．
2. 用例データとともに文字データが提供でき，知識処理を加えた文字認識の開発に供せられるもの．
3. 歴史研究上の汎用性のたかい文書からの文字．

1. の観点からは，共著者の川口が「宗門改帳」に記載された数字表現 16 字種 3,066 サンプルを採字し，2 値画像データとして作成した．当データベースを HCD1 (Historical Character Database 1) と名付け，後述する古文書文字認識の基礎実験に供している．HCD1 に収録されている字種とサンプル数は，表 2 のとおりである．

2. の観点からは，前記の「伏見屋文書」の全文字を切り出してデータベース化する予定である．「伏見屋文書」全文書は，すでに画像データベース化され，大阪市立大学学術情報総合センターから公開されている．その画像総数は 1,989 画像，総文字数は 20 万文字を越えると予測されている．

表 2: HCD1 収録の字種とサンプル数

字種	サンプル数	字種	サンプル数
ツ	200	八	200
一	200	九	200
二	200	十	200
三	200	壱	200
四	200	弐	200
五	200	年	200
六	200	拾	200
七	200	廿	66

3. の観点からは，『柳営日次記』（図 2）を選択している．「柳営」は幕府もしくは将軍を意味する言葉で，『柳営日次記』は江戸幕府の公日記になる．江戸時代研究の一級資料である『徳川実記』は，『柳営日次記』をもとに作成されている．『柳営日次記』は，江戸時代における将軍の動静，幕府における各種の行事，叙任，法令等の研究上，非常に重要な資料で，『徳川実記』の典拠として江戸幕政史のみならず，江戸時代史研究にも欠かすことのできない基本資料とされている．しかしながら，『徳川実記』がすでに翻刻されていることと，その量の膨大さがわざわいして，『柳営日次記』にはいまだに翻刻の手がつけられていない．

『柳営日次記』は，江戸幕府の書記官である右筆（ゆうひつ）によって書写されたものであるため，書体がいわゆる「御家流」に統一されているという，文字認識上有利な点がある．とはいえ，筆記者による字形の差はそれでもおおきく，文字認識が困難であることには変わりはない．

平成 11 年 12 月現在，『柳営日次記』を撮影した 35 ミリマイクロフィルム全 130 巻のなかか

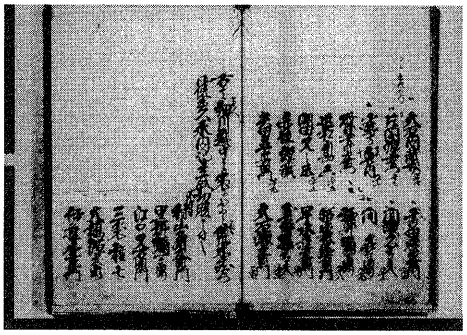


図 2: 『柳営日次記』(元禄 16 年 2 月 4 日の記録の一部)

ら，元禄 15 ～ 16 年部分にあたるリール 1 巻を選択し，文字データベース作成のための全頁画像ディジタル化を完了している．ディジタル化は，マイクロフィルムスキャナを使って 400dpi で取り込むことによっておこなった．元禄 15 ～ 16 年の 2 年分の総コマ数は，664 コマである．現在，その文字データベース化のための翻刻作業を進めている．

3.3 古文書文字切り出し

古文書によくみられるようなつづけ字から 1 文字を切り出す作業は，文字データベース作成のために，また古文書文字認識における前処理として，きわめて重要である．われわれは，文字データベース作成用の文字切り出しツールを開発するとともに，文字の自動切り出しをにらんだ基礎研究を進めている．

古文書文字データベース作成用の文字切り出しツール「切」は，古文書画像をグラフィカル・ユーザ・インタフェース上に呼び出し，文字の切り出しと同時に翻刻文字を入力することで，文字データベース作成の作業効率をたかめることを目的とするツールである．文字の切り出しは，ユーザがマニュアルで，文字を囲む矩形を切り出すか自由多角形で文字を囲んで切り出す．切り出し後の 2 値化，ごま塩状ノイズの除去，上下左右からの侵入ノイズの除去は，自動的におこなえる．翻刻文字は，SJIS コードのほかに大漢和のコードでも入れられるようになっている．切り出された文字が，もとの文書のどの位置にあったかも確認できる．「切」のインタフェースを図 3 に示す．

文字切り出しの自動化をめざした基礎研究として，ピラミッド型によるレイアウト画像の抽出とヒストグラムを用いて，「伏見屋文書」の標題を自動抽出する方法を検討した．その結果，標題のある古文書画像に関して全体の約 3/4 の割合で標題を抽出することができた (図 4) [5].

3.4 古文書文字認識

古文書文字認識エンジンの開発は，本プロジェクトのなかでも技術的にもっとも挑戦的な部分である．われわれは，これまでの手書き日本語文字認識研究の技術的ベースのうえに，古文書文字の認識に必要な要素技術の開発と評価を実施している．これまでの成果の詳細については，文献 [6] に発表しているので，そちらを参照されたい．

4 今後の予定

古文書翻刻支援システム開発プロジェクトは，今年度スタートしたばかりである．当面は，研究遂行に必要なツール群の整備と文字データベースの作成，文字認識の基礎研究を中心に進めることになる．

古文書用例データベースに関しては，「伏見屋文書」全文の入力を急ぎたい．それが完成すれば，借金証文類の用例データとしては，じゅうぶんな内容をもったものとなるだろう．

古文書文字データベースに関しては，「伏見屋文書」の標題文字の切り出しと文字データの対応付けを平成 12 年度のなるべく早い段階で完了させたい．全文字切り出しと用例データベースとの対応づけは，平成 12 年度中の完成が当面の目標である．宗門改帳からの文字データは，さらに字種を増やすして文字認識実験に供する．『柳営日次記』の文字データベース化

図 3: 古文書文字ＤＢ作成用文字切り出しツール「切」

は，将来をにらんだアドバンスドなデータベースとして，整備を継続したい．さしあたり，その一部もしくは人名のみについての文字データベース化が目標である．また研究の進展とともに，現在の古文書字典の検索性がわるく，これを電子化して検索性をたかめることが，古文書翻刻支援におおいに役立つであろうことが認識されてきた．電子化古文書字典の研究も，余力のあるかぎり同時に進行させたい．

古文書文字切り出しに関しては，現在のツールを改良・安定させることと，文字自動切り出し方法の研究を，平成12年度以降も継続する．

古文書文字認識に関しては，できあがったデータベースをもとに平成12年度以降，本格的に遂行できるであろう．古文書文字認識は，いわば究極の手書き日本語文字認識ともいえるもので，3年間で完全なものが得られるとは思えない．データベースとともに，今後の研究環境を徐々に整備することが，3年間の目標である．

謝辞

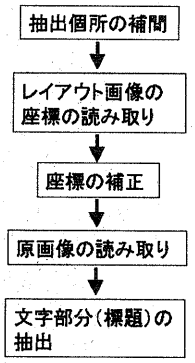
本研究は，日本学術振興会科学研究費補助金・基盤研究 (B)(1) 一般研究「古文書解読プロセスの知能情報学的解明」（平成 11 ～ 13 年度，研究代表者：山田奨治），同展開研究「手書き文字ＯＣＲ技術を援用した古文書翻刻支援システムの開発」（平成 11 ～ 13 年度，研究代表者：山田奨治），同一般研究「古文書ＯＣＲの試論的研究」（平成 11 年～ 13 年度，研究代表者：柴山守）の支援を得て実施しているものである．

参考文献

[1] 山田奨治: 高次局所自己相関特徴による古文書かな文字認識, 情報処理学会研究報告, Vol.95-CH-25, pp.21-30, 1995.

[2] 山田奨治: 変体かなの認識実験とその応用, 人文学と情報処理, No.18, pp.71-75, 1998.

[3] 日置慎治, 上原邦彦, 川口洋: 年齢を表記した古文書文字の認識－「宗門改帳」古文書画像データベースを用いた実験－, 情報処



(a) 補間前　(b) 補間後

図3 文字列の抽出個所

図4 抽出した標題

図 4: 「伏見屋文書」標題抽出

理学会研究報告, Vol.98-CH-37, pp.35-42, 1998.

[4] 「挑戦　古文書ＯＣＲ」特集号, 人文学と情報処理, No.18, 1998.

[5] 尾崎浩司, 柴山守, 荒木義彦: 古文書レイアウト画像のピラミッド型抽象化と標題の自動抽出, 平成 11 年電気関係学会関西支部連合大会発表論文, 1999.

[6] 和泉勇治, 加藤寧, 根元義章, 山田奨治, 柴山守, 川口洋: ニューラルネットワークを用いた古文書個別文字認識に関する一検討, 情報処理学会研究報告, Vol.99-CH-45, 1999.