ヒューマンインタフェース 53－9
（1994．3．10）

古文書を分析するためのＨＣＩ

北村 啓子
国文学研究資料館 研究情報部
keiko@nijl.ac.jp

コンピュータ上で古文書を読むことに主眼を置き、国文学の研究活動をコンピュータによって支援する研究を行なっている。HCIとして、古文書の写真（イメージデータ）をモニター上で読む、テキストを縦入出力を行なうなどの環境整備を行なってきた。究極の姿としては、古文書に描かれている文字情報を処理する時のインタフェースとしてイメージデータを使い、テキストはコンピュータ処理のための内部表現であると考えている。それを実現するために古文書のイメージデータとそれを翻刻したテキストとの間で、文書中の同じ所を示す対応関係をコンピュータで利用する。この例として、あたかもイメージデータ上で文字列を検索しているような手書き文字検索を試作した。その中で使用するツールとして、イメージデータを扱うために開発した文字の認定に充分な品質で古文書を表示するviewerとテキストデータを縦書きするためにttermを開発した。

HCI for Studying Handwritten Classical Books

Keiko Kitamura

Information Processing Section
National Institute of Japanese Literature
1-16-10 Yotaka-cho, Shinagawa-ku, Tokyo 142, Japan

It is important to read handwritten classical books on computer for supporting research activity by computer. It is discribed that the idea of HCI is image data and text is internal representation of computer when human scientists study classical text on computer. For implementing of this idea, it is used the relation express the same point between image data and text data. It is developed the prototype system to show search results on image data. It is also developed VIEWER to display image data and TTERM is a terminal emulator to input/output texts virticaly.

1 はじめに

文化資産である古文書を保存し、解読、調査、分析などの研究活動を支援する計算機環境の研究に取り組んでいる[5]。アプローチとしては、データベースとして蓄積、管理、検索というよりも、人文科学者個人の研究を支援することを目指している。そのために、軽装なコンピュータで自分の必要なデータを集めれば利用できるツール群のようなアプリケーションを考えている。その際のHCIとして、古文書を扱うのに古文書のイメージを直接見ながらコンピュータ処理をするのが当然であり、その究極の姿としてイメージをHCIとしテキストはコンピュータの内部表現であるという考え方について述べる。

2 古文書の特徴

国文学をはじめ多くの人文科学系の分野で研究の対象となる古文書は、草書体の手書き文字で書かれている。そこには文字情報が描かれているのだが、コンピュータで扱うためにはイメージデータとして表現することになる。特に国文学では、文字情報と同時にその形状も重要な情報である。

写本は数百年を渡って書き写し継がれてきたものであるため写し間違いが潜在的に存在する。また写す際に写す人の裁量で、写し間違いと思われるものを直したりその時代に応じた表現に書き直すなど行なわれた可能性もある。

殆んどの場合、著者が直接書いたオリジナル（元本）は残っていないので、現存する多くの他の古文書（異本）を比較することによってオリジナルを推察し、翻刻するのが国文学の重要な研究である。

まず翻字の段階で、文字の認定の困難さがある。変体仮名文字の種類は多くその認定は難しい。漢字ー変体仮名ーかなのくずれ方は連続的なので無数にあると言ってよく、その区別は困難を伴う。手書きである上、草書体であるため、字のくずし方・つなぎ方が無数に存在する。

さらに翻刻の段階で、意図的に次のような修正を加える。

・写本時の写し間違いと思われるものを修正する

・古くは濁点が存在しなかったので濁点を補う

・読みの正解がない（不明）ため推察で補う

・句読点は使わないか使っても使い分けをしないので、句読点を補ったり句点と読点の区別を判断する

・現在の古文文法（歴史的仮名使い、送り仮名、尊敬謙譲語など）に合わせる

・当時の正しい表現で書かれているが、長い年月で見れば流行であり相対的に時代ごとに異なる

・現代的表現やこなれた表現に直す

・習慣的に通用体という小さい文字セットに縮退する

・多くの異表記の中からどの字が使われたのか、どの字を使うのが最適かを判断する（漢字／変体仮名／かな、異体字、同義文字、正字／俗字、新字／旧字）

・多くの異本の差から正解一つを判断する

逆に言うと一つの翻刻したテキストがあった時にそのオリジナルは異なる表現である可能性が高く、潜在的に不確定性を多く含んでいると言える。直接イメージデータを見ることで不確定性は国文学研究者個人に判断を委ねることができる。

コンピュータで文字情報を扱う時には、これらの問題によりバラつきが発生し、精度を落す原因となっている。このため、翻刻と言っても単純に一つのテキストにはならず、より複雑な構造が必要になる。潜在的に存在する不確定性をも包含した処理が期待される。

3 国文学研究へのコンピュータのかかわり

古文書を対象とした国文学研究の現状を、コンピュータとのかかわりという見地から分析したのが図1である。研究活動と扱う対象の表現形態によって、次の三つに分類

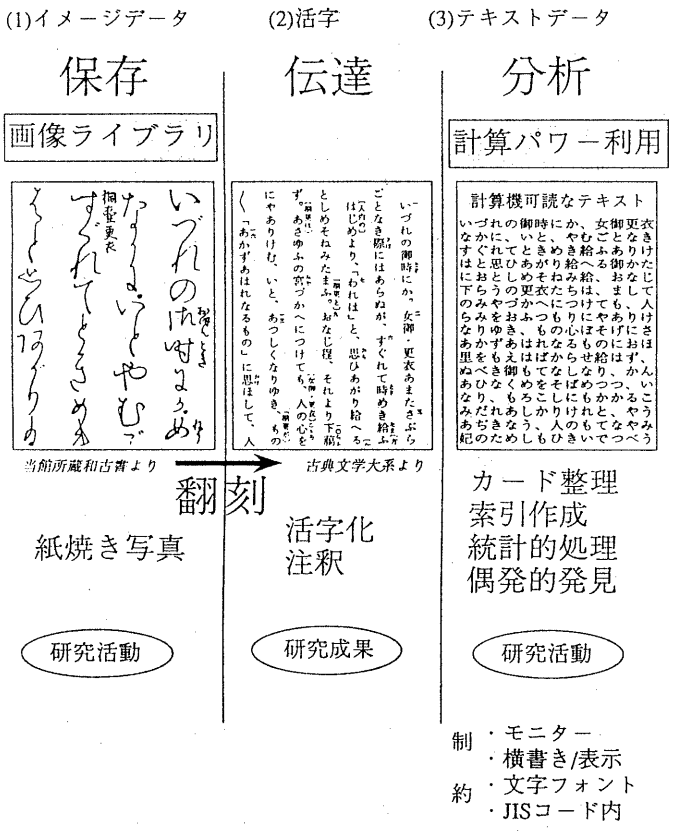

図 1: 国文学研究へのコンピュータのかかわり

できる。

（1）イメージデータ

古文書の写真をコンピュータで扱うためには、字形情報をそのまま表現できるイメージデータで表現することになる。最近、画像ライブラリとして入力されており、主な目的は保存である。デジタルデータ化することによりデータの劣化を防ぐことができる。また、イメージデータをプリントすることにより比較的容易に紙焼き写真を入手できるというメリットがあり、配布目的にも利用できる。

（2）活字

研究成果を第三者に伝達することを目的に、翻刻したものや作成した索引、論文を活字化する（出版による研究成果の公開）。現在、コンピュータは殆ど介在していない。最近では、出版社へワープロ原稿や計算機処理した索引データを入稿する例が増えつつある。

（3）テキストデータ

従来研究者自身がマンパワーで行ってきたカード整理、索引作成、統計的処理などの分析作業をコンピュータパワーを利用して行うために新しくできた世界である。モニターを見なければならない、横書きを強要される、使える文字コードが制約される、使える文字フォント（字形）が制約される等々の制約を強いられるなどの困難にもかかわらず利用する研究者が増加している。

4 HCIの要件

国文学研究をコンピュータ支援する際のHCIとして求められる要件を述べる。まずコンピュータ支援のアウトラインを図2に示す。（2）活字メディアは電子出版の言葉通り活字本とテキストの差異はないのでテキストメディ

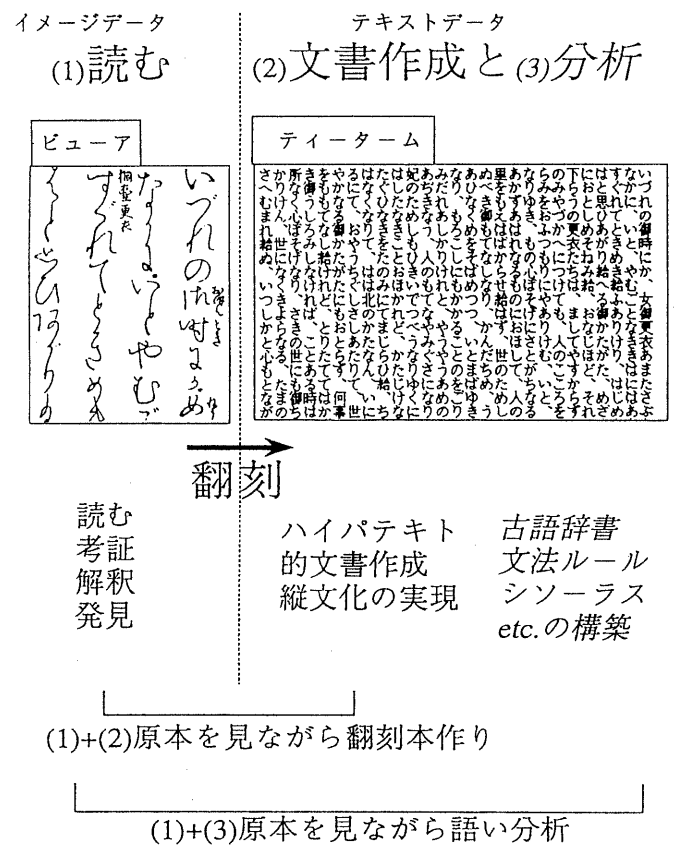

図 2: コンピュータ支援のアウトラン

アに入れ、目的で分類し（2）文書作成（3）分析とする。（1）、（2）、（3）、（1）＋（3）の各々におけるHCIの要件を述べる。

（1）読む（イメージデータ）

1. 高精度で表示

現物と遜色ない高精度で見られると、写真に比べ複製が容易に作れ入手もし易くなり写真を代行できる。文化財である古文書の劣化を防ぐことにも役立つ。読むことを単独で支援しても紙めくりの使い勝手など紙焼きにはかなわない。しかし、検索などコンピュータの得意な分野の機能をより効果的に使った研究活動の全体を統合した環境が実現すると、基本的な機能としてコンピュータ上でも「読める」ことが大変重要となる。また、テキストをコンピュータ処理する時のインタフェースとしてテキストデータを強要されることなく、イメージを見る（読む）ことも意味しており、国文学者にとって格段に親近感のあるコンピュータとなるであろう。

2. 紙の使い勝手

コンピュータで書物を扱おうとする場合まず問題になるのが紙めくりの使い勝手である。国文学では書物に頻繁に接しているため、ザッと目を通す、概略を把握する、当りをつける、特定の目的なく眺めることによる発見など紙としての使い心地を捨ててまでコンピュータに移行することは難しい。HI研究の一つとしてブックメタファー [1][2] の研究がなされており、この応用が期待される。古文書では、縦書き、右開きや巻物などの考慮が必要となる。

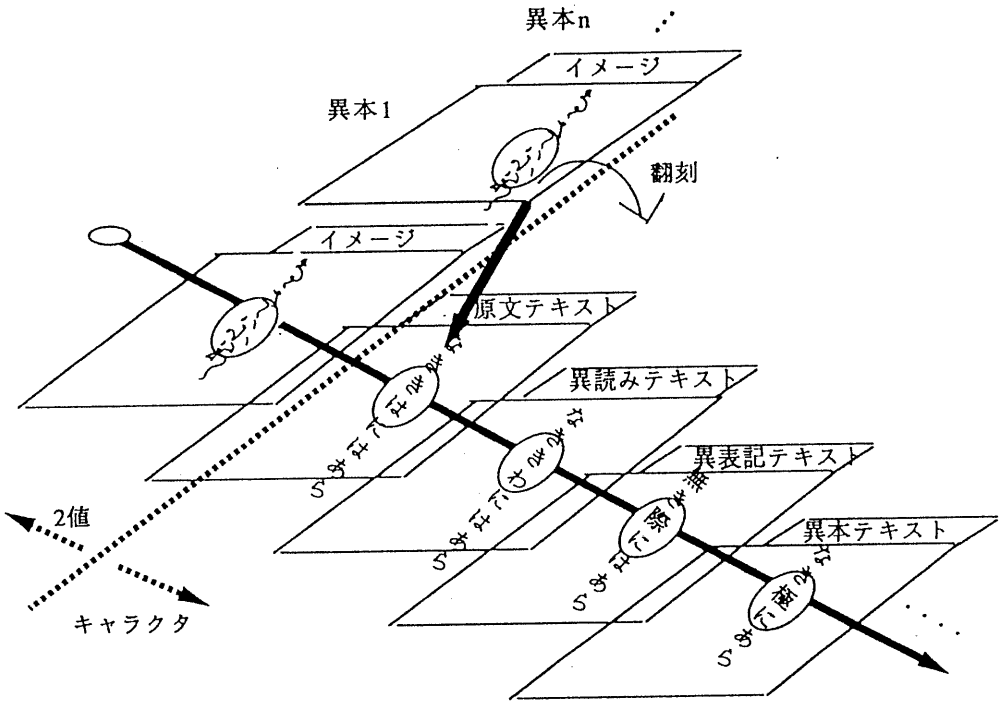

図 3: マルチレイヤ構造のコンセプト

（2）文書作成（テキストデータ）

1. 縦文化の導入

図1を見れば明らかなように、縦書きの原本に対して、テキストになった途端横書きになるのは不合理である。翻刻本も縦に書くのは当然であり、縦書き文化を計算機に持ち込むべきである。

（3）分析（テキストデータ）

1. 個人用データとそれを処理するツール群

研究段階で必要なデータは未整理のばらばらなデータの寄せ集めである。個人用データとそれらを処理するツール群を整備することが重要と考えている。

2. 個人の解釈を重要視

国文学は他の分野に比べて個人の解釈が非常に重要であるため、その解釈や解釈に基づいた翻刻のデータを個人データとして持つことができ、それを他のデータと一緒に処理できることが重要な点である。

（1）読む＋（3）分析

1. イメージを見ながらテキスト処理

モニター画面上で直接イメージを見ながらそこに描かれている文字情報を処理できるのが理想的である。コンピュータが古文書の手書き文字を認識できない現在、その疑似的実現方法としてイメージとテキストをペアで一つのデータとして扱うことを考える。両者には同じ文章が書かれいるので、文書中の同じ所を示す対応情報を持つ。イメージデータ上で、この対応情報をたどってテキストの持つ文字情報やテキスト処理のための付加情報を利用するという方法

である。

ここで重要なマルチレイヤの構造の考え方を図3に示し述べておく。先に述べた通り古文書自身に潜在的に不確定性を多く含んでいる。それらも包含した処理をするために、一字一句忠実なデータ、文法的に正しいデータ、（異なりも含めた）読みのデータ、異表記のデータなど学術的な判断で必要なデータをマルチレイヤ化する。イメージ-テキスト間、テキスト間相互に対応情報をたどってそれらのデータを利用するという考え方である。図3中イメージデータと複数のテキストすべての文書中の同じ所を串座しして見られる点が重要である。

（1）＋（2）、（1）＋（3）のアプリケーション例としては、前者はイメージを見ながらの翻刻本作り、後者はイメージを見ながらの語彙分析があげられる。後者については次章で文字列検索機能について報告する。

5 ツールとアプリケーション例

5.1 表示用ツール

HCIとして古文書のイメージデータを使うために開発した表示用ツールviewerとテキストを縦に入出力するためのツールttermを紹介する。

・viewer

パソコンやワークステーションで手軽に利用できるスキャナで入力したイメージデータでも、読める（文字認定ができる）程度の精度でモニター画面に表示する。図4左と右下のウィンドウ。古文書の写真を400dpiでスキャナ入力した2値データをモニターの精度に合わせて間引く比率によって階調を付けて表示している。1.280x1.024dotのモニターに原寸大で表示すると16階調で、筆文字のかすれや朱書きもグレー濃淡で識別でき、原本の字形情報をほぼ再現できる。X11R4, XToolkitを使って開発した。

・ttterm

マルチウィンドウ環境で、入出力の全てを縦に表示するターミナルエミュレータ。ktermを右に90度寝かして文字の入出力が左→右、上→下を上→下、右→左の順で行うように改造した。UNIX上の全てのソフトウェアを利用できる汎用性がある。図4右上は、nemacs上でwnnを使った日本語入力をしている。

縦書きを実現する時問題になるのが、縦書き用の文字フォントと、全角／半角の混在、アルファベットをどう書くかである。ttermでは、全角文字のフォントの殆どはそのまま横書きのものを使用している。例えば「」（）、。カーソルマークのように横と縦で変換の必要なものだけ縦書き用を作成している。半角／全角が混在する場合、文字数と見た目の長さを合わせる必要があるので、半角文字は、横に寝かすか、縦方向に半角サイズの横長フォントを作成するかになる。アルファベットの場合、どちらも読み易くはない。古文を扱う中でアルファベットの出現は非常に低いと考えて、あまり労力はかけないことにし、横書き用のフォントを90度回転して横に寝たまま使っている。

5.2 連動ツール

イメージデータの内部表現としてテキストを使う例を紹介する（厳密にはイメージとテキストの両方を使う例である）。先に紹介したviewerとttern上のnemacs間で対応関係を渡すことにより連動して使えるようにした。ここではnemacsの文字列検索機能を使い、検索した結果から文字列マッチした行番号をviewerに渡し、同じ行のイメージを先頭行として表示する。対応関係の単位は行を採用し、行番号を受渡ししている。nemacsのgnu-lispでプロセス間通信により実現した。現在は行のX座標のデータを人手で作成している。

[1] ktermは日本語を扱えるようにxtermを籠谷氏（東工大）が改造したもの。

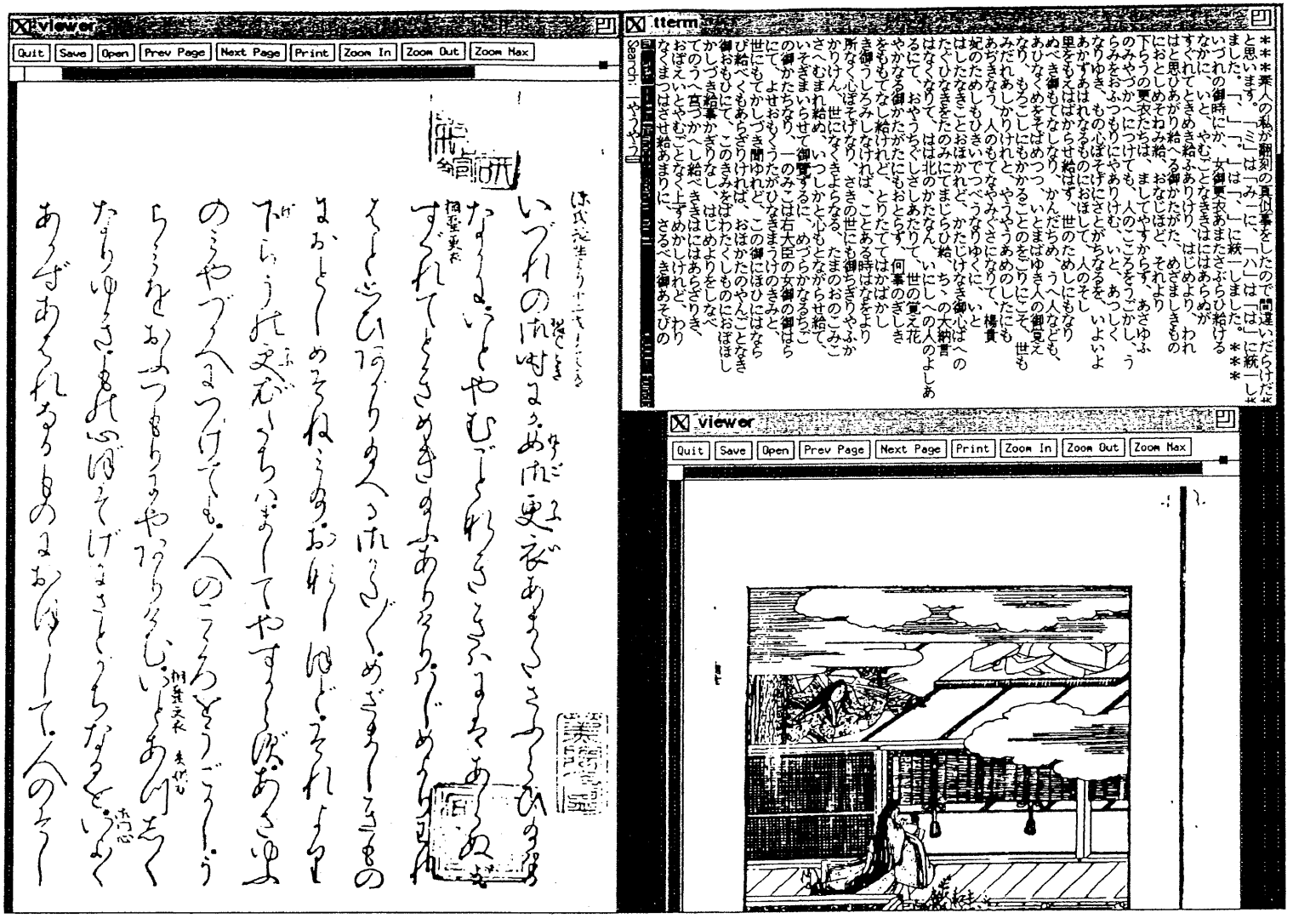

図 4: 表示用ツール viewer と tterm

nemacs の持つすべての文字列処理機能の処理結果をこの連動を利用してイメージで見られる。先に述べたマルチレイヤのテキストデータを使った国文学特有のさらに高度なテキスト処理 [3][4] もこの連動を利用して同様にイメージデータ上で結果を見ることができる。

ユーザは、イメージのインタフェースを通して文字列検索の結果をイメージ上で見ることができる。またＫＷＩＣ（KeyWord In Context）リストも、同様に文字列検索でマッチした文字列の前後のイメージデータを集めることにより、手書き文字ＫＷＩＣリストの作成も可能である。これによって、検索した文字列を含む行の手書きイメージを見ることができることになる。

ほんとうの意味でのテキストを内部表現として使うためにはイメージ上で検索文字列を指示できなければならない。イメージデータ上の該当する文字列をマウスで選択するなどの入力のインタフェースが必要となる。

6 おわりに

国文学研究にコンピュータを使うための究極の姿として、HCI は古文書の手書きイメージでありテキストはコンピュータ処理のためのデータ形式（計算機の内部表現）であるという考え方について述べた。字形を表現するイメージデータと文字情報を表現するテキストの対応をコンピュータが知っていることは重要なことである。その具体的な実現方法として、イメージデータとテキストの対応関係を利用する方法について述べた。

具体的に試作したシステムとして、viewer, tterm, イメージ - テキスト間の対応情報を使った連動ツールを紹介した。現在は nemacs の文字列検索機能だけを使っているが、国文学特有のさらに高度なテキスト処理も組み込んで行きたい。

現在の viewer はグレー階調を使っているだけなので写真を紙に焼付けた紙焼き写真に相当する精度である。（現在当館では白黒マイクロフィルムで収集しているので研究に供するのはこの紙焼き写真であるという意味では充分かもしれない。）現物の紙の質感や虫喰い、裏写りなどを考えるとカラーデータ入力、表示が望まれる。

最後に本稿で紹介した tterm はフリーソフトウェアとしての配布を計画している。

謝辞：本研究を行うに当たり、当館所蔵和古書の電子複写コピーサービスを利用させて頂いた。日頃国文学の立場からアドバイス頂いている当館当部新井教授、中村助教授に感謝する。また、本研究は稲盛財団の研究助成を受けている。

参考文献

[1] 工藤正人 岡田謙一 松下温: 人間の空間情報処理能力を活用したユーザインタフェース：BookWindow, 情処 HI 研究会, 93-HI-48(1993)

[2] 荒井恭一 佐藤敬行 木下薫 横山光男 松下温: ページめくり機能を持ったウィンドウインタフェース：BookWindow, 情処 HI 研究会, 91-HI-36(1991)

[3] 北村啓子: 古文書を表現するためのマルチメディアデータモデル, Advanced Database Syatem Symposium'93,pp.235-244,(1993)

[4] 北村啓子: 古い著作物を分析するための計算機環境－和歌文学を題材に－情処全大 47 回,2C-1,vol.4,pp.95-96,(1993)

[5] 北村啓子: 計算機で写本版本を読む－写本版本を計算機で扱うためのマルチメディアデータモデル－、国文学研究資料館紀要、第 19 号, pp.1-21(1993)

[6] 北村啓子: 古文書を表現するためのマルチメディアデータモデルの構想, 情処全大 45 回,1R-1,vol.4,pp.99-100,(1992)

[7] 北村啓子: 縦書きテキスト編集機能の検討と X Window 上での試作、情処全大 43 回, 7L-1,vol.4,pp.81-82,(1991)

[8] K.Kitamura: Data Base Delivery for Japanese Literature by CD-ROM, Proc. of ACH/ALLC'91,(1991)