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——

北　村　啓　子

要　旨　計算機上で写本版本を読むことに主眼を置き，国文学の研究活動を計算機によって支援する研究を行なっている。写本版本のイメージデータとそれを翻刻したテキストデータとの間で，文書中の同じ所を示す対応関係を計算機で利用することを考えている。この対応関係を利用することにより，両者を連動しながら調査／分析したり，双方向に参照することや，テキストデータを計算機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うことが可能になる。後者の例として，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ので紹介する。また，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写本版本を表示するビューアとテキストデータを縦書きするために開発したティータームの紹介を行なう。

1．はじめに

タイトルの「計算機で写本版本を読む」というのは，「計算機が写本版本の手書き文字を認識する」のではなく，翻刻，考証，解釈を行うために「計算機上で研究者が写本版本を読む」ことを意味する。もちろん，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，検索など計算機の得意な機能をより効果的に使った研究活動全体を統合した環境が実現すると，基本的な機能として計算機上でも「読める」ことが大変重要となる。また，テキストを計算機処理する時のインタフェースとして写本版本を見る（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計算機となるであろう。このように計算機で支援する様々な研究活動を代表して「読む」を使ったのであって，研究活動全体を意味している。

最近イメージデータの入出力が容易に行なえるようになり，写本版本の写真を保存するために計算機に入力することが実際に行なわれている。一方，文字情報を使った計算機処理をするためには，イメージデータとは別に，文字コードになったテキストデータが必要であり，研究者が翻刻した活字を電子化してこれを作成している。このテキストデータを処理するのに，横書きを強要され，使える日本語コードや文字フォントなどの制約を受けながら計算機を利用している。これまでこれらの制約を緩めて，より国文学に近い計算機の構築を目指し，縦書き環境の整備などの研究を行ってきた[1][2]。これまでの研究を通して生じて来た，「写本版本を扱うのに写本版本を直接見ながら計算機処理をするのが当然である」という発想から本研究を始めた。

目標は，計算機で次のことを可能にし，翻刻，考証，解釈を行うのに利用に耐える計算機環境を構築することである。

（1）文字認定に充分な品質でイメージデータをモニターに表示する

（2）縦書き文化を持ち込み，縦に入出力を行う編集や古文を解する日本語入力を行う

（3）イメージデータとテキストデータを連動しながら調査／分析したり，双方向に参照する

（4）イメージデータ上で，テキストデータの持つ文字情報，構文情報を利用する

2．国文学研究への計算機のかかわり

写本版本を対象とした研究活動の現状を，計算機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，次の三つの世界に分類できるであろう。

（1）イメージデータの世界

写本版本の原本またはその写真である。計算機で扱うためには，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。

(1)イメージデータの世界 (2)活字の世界 (3)テキストデータの世界

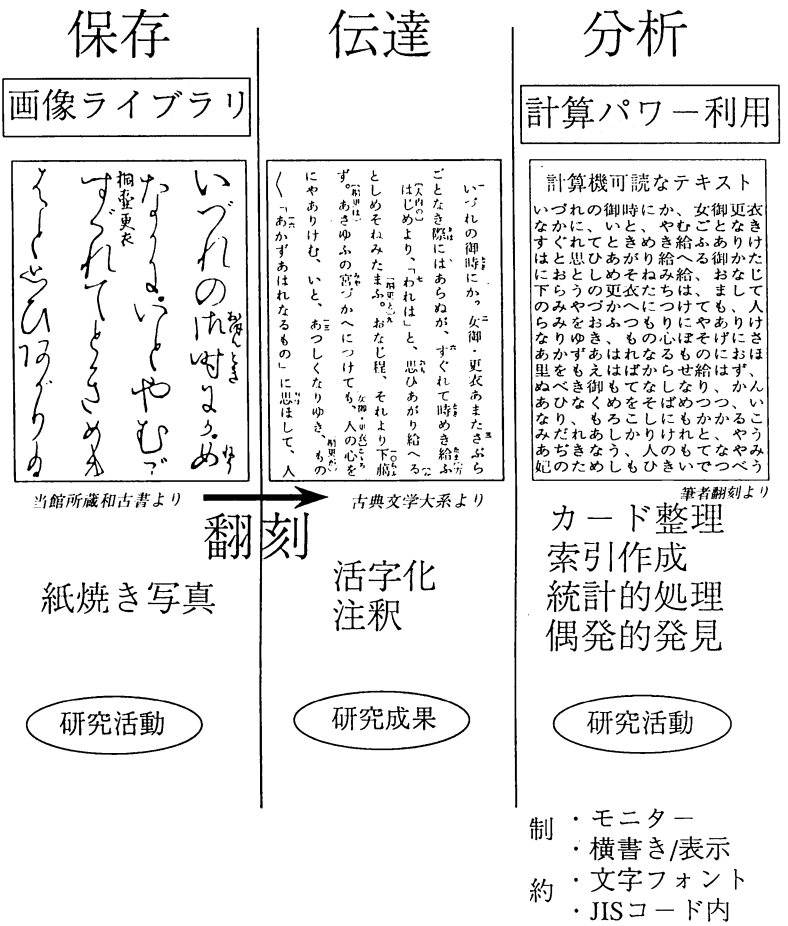


図1　国文学研究への計算機のかかわり

（2）活字の世界

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，計算機は殆ど介在していないようである。（厳密には，翻刻者の直接関与しない所でＣＴＳ出版など計算機と関わりはある。）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ようである。

（3）テキストデータの世界

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計算機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困難にもかかわらず利用する研究者が増加している。機械的な作業については，強力な威力を発揮する計算機パワーが評価されているからであろう。

3．国文学研究に計算機を利用するために

現在の計算機技術は，主に理工系，ビジネス分野などでの利用を目的に構築されてきた計算機科学の上に成り立っている。徐々に日本語を扱える環境，文字列処理を始めとするテキスト処理の技術も研究されてきているが，写本版本は未だその範疇には入っていないと言ってよかろう。現存する技術を応用してみて，可能不可能を明らかにするのも意義あることではあるが，写本版本を扱うための新しい計算機科学の基礎作りこそが急務であると考える。この考え方は，国文学側から新井教授の提唱された「文芸工学」の必要性[3]に通ずるものがあるかもしれない。

写本版本を計算機で扱うための計算機科学の一つの基礎となるであろうマルチメディアのデータモデルが本研究の重要テーマである。字形情報を表現しているイメージデータと文字情報を表現しているテキストデータには，表現形態は違っても同じ文書が書かれている。それにもかかわらず，それぞれが無関係なデータとして入力されているので，文書中の同じ所（対応関係）を同定することすらできない。この対応関係を表現できるモデルを構築し，両者の情報を計算機上で補完しあって扱えるようにすることが目的である。

ここで提案するマルチメディアデータモデルの概略は，図2に示すようにマルチレイヤーのデータ構造をしている。ここで重要な点は，それぞれのレイヤー間で文書中の同じ所を同定することができることである。つまり，図に矢印で示した通りすべてのレイヤーの同じ箇所を串刺しして見ることができるということである。これによって，イメージデータとテキストデータを連動しながら読み／分析したり，双方向に参照することが可能となる。また，イメージデータ上で，テキストデータの持つ文字情報，構文情報を利用することもできる。これは言い換えると，両者の同じ所を見ながら，お互いに参照しあって有機的に利用することもできるし，あくまでもイメージデータを見ながら，必要な文字情報はイメージデータを通してテキストデータから得るという使い方もあるということである。

ここで，テキストデータが複数のレイヤーになっている点に触れておく。新井教授が日頃提唱されている「国文学を扱うには，目的によって数種類のテキストを用意する必要がある」という考え方[4]を支持し，ここで提案しているマルチメディアデータモデルでもテキストデータをマルチレイヤー化している。翻刻された活字の多くは漢字仮名混じり文であり，その他仮名だけの仮名文，

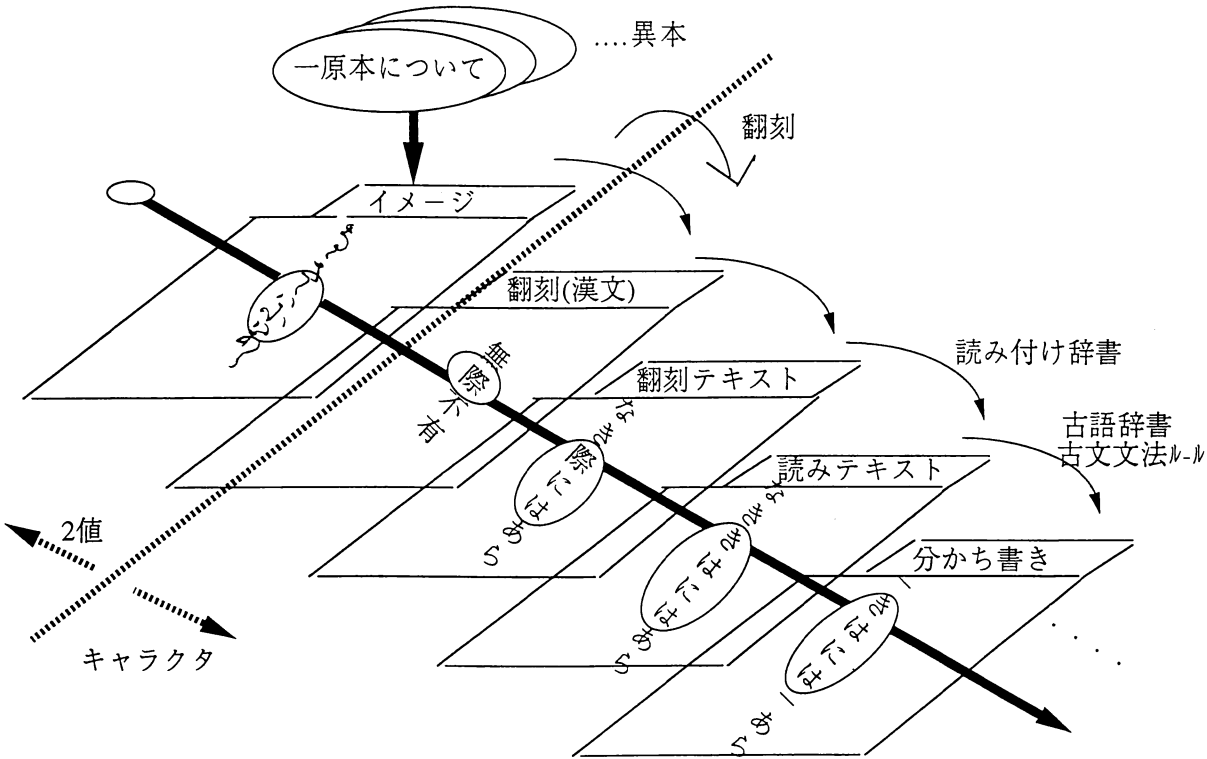


図2　マルチレイヤー構造のコンセプト

漢字だけの漢文／万葉仮名文などである。この文字情報をそのまま電子化（テキストデータ化）しても認識できるのは日本語文字コードだけである。現在の計算機は，表記のヨミ，新字旧字の対応，異体字，同義語も知らない。古文の語彙，古文文法，語彙の意味などももちろん知らない。語彙分析をするためには，これらの情報を教えなければならない。これらの情報には，辞書データとして計算機に実装した方がよいものと，作品ごとに翻刻されたテキストデータと同様にテキストデータとして準備した方がよいものとがあると考えている。例えば，漢文であれば翻刻した漢文以外に，それを書き下したテキスト，ヨミのテキスト，分かち書きしたテキスト，...等々を目的の処理によって準備する方がよいであろう。もちろん，古語辞書を始めヨミ付け辞書などが整備されれば，目的に応じたテキストデータを計算機で作成するのに有効である。

この複数のテキストデータは，高度な語彙分析を実現する目的でマルチレイヤーのデータ構造の中で提案したものだが，２章で言う（３）テキストデータの世界に閉じた話しである[5]。本稿では，イメージデータとテキストデータとを連動して扱うことが第一の目的であるので，イメージデータとテキストデータ間の対応（（１）＋（２），（１）＋（３）の組合せで利用する情報）に限定して述べる。

４．何ができるのか？

では一体計算機で何ができるのであろうか？２章で分析した研究活動に基づいて議論する。使用するデータの表現形態によって分類すると，図３に示すように読むためのイメージデータの世界と文書作成や分析のためのテキストデータの世界になる。（１）イメージデータの世界では，モニター画面で写本版本

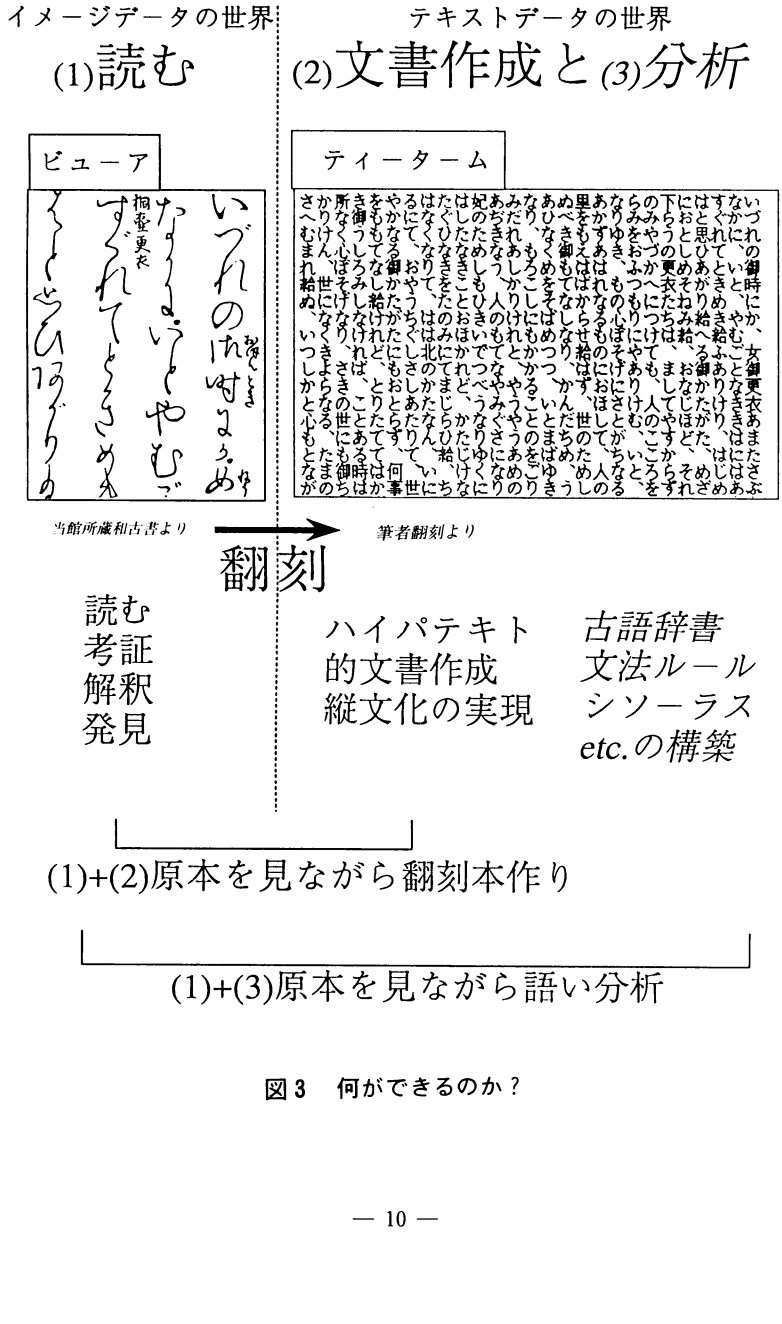

イメージデータの世界
テキストデータの世界
(1)読む
(2)文書作成と(3)分析
ビューア
ティーターム
当館所蔵和古書より
筆者翻刻より
翻刻
読む
考証
解釈
発見
ハイパテキスト
的文書作成
縦文化の実現
古語辞書
文法ルール
シソーラス
etc.の構築
(1)+(2)原本を見ながら翻刻本作り
(1)+(3)原本を見ながら語い分析

図3　何ができるのか？

を読むこと，（2）テキストデータの世界ではワープロ感覚で翻刻本を作ること，（1）＋（2）の世界では原本を見ながら翻刻本を作ること，（1）＋（3）の世界では原本を見ながら語彙分析を行なうことについて述べる。

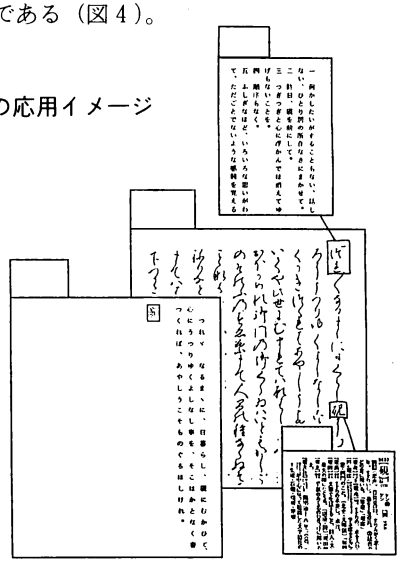
（1）モニター画面で写本版本を読む

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，筆文字のかすれ，朱書きも読める程度の精度でモニター画面に表示する。

（2）ワープロ感覚で翻刻本を作る

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。また，翻刻本で重要な注釈は，ハイパテキストの考え方に基づいて，関連情報として書き込みや参照を支援することも計算機の得意とするところである（図4）。

図4　ハイパーテキストの応用イメージ



また，日本語入力で古文を対象にすることも重要な点である。一般の日本語変換用の辞書では，殆んどの古語は入っていない（例えば，「おもふ」→「主婦」「尾も負」などに変換される）。現在古語辞書の電子化にも取り組んでおり，古語変換用の辞書にして特定の日本語変換機能に組み込むことを計画している。現状では，日本語入力機能が提供しているユーザ辞書に個人で登録しているようであるが，数量的な制限や古文／現代文での活用形など文法の差などの問題がある。

（1）＋（2）原本を見ながら翻刻本作り

国文学研究では，何種類もの異本の紙焼きを並べて注釈の書き込みをしながら翻刻本を作成をすると聞いている。紙焼きの代用に足るだけの精度で計算機上でイメージを読めれば，何種類もの異本に注釈を書き込み，翻刻原稿作成時にそれを参照（流用）するのは，まさにハイパテキストの本来的な使い方である「文書作成」の支援である。

ここでは（1）で述べたモニター上でも充分読める表示手段と（2）で述べた縦文化を持ち込んだテキストデータ編集機能を利用して，翻刻本作りを支援する。ハイパテキスト的に文書の構造に沿った（語彙→注釈）ビュー（見方）や書式に従って本としてのビューなど，目的に合わせた見方ができる。

（1）＋（3）原本を見ながら語彙分析

イメージデータを直接見ながら語彙分析を行う。利用者はあくまでもイメージデータを見，文字情報を使った語彙分析は（イメージデータと対応のとれた）テキストデータを裏で使うという方法である。つまり，イメージデータ上でテキストデータの持つ文字情報，構文情報を利用する訳である。語彙分析の技術が高くなれば（それをそのまま使えるので），それだけ高度な処理を行える。現

在のテキスト処理技術で可能なものとして，手書き文字列検索（完全一致／部分一致，あいまい検索），手書きＫＷＩＣ（特定のキーワード（文字列）の前後を抜き出して出現する全ての箇所のそれを並べてリストにしたもの），手書き文字の用例集，手書き文字の索引（実用的な電子古語辞書を作成するかまたはデータを分かち書く必要がある）などが挙げられる。どれもテキストを使って処理した結果をイメージ上で見たり，それを保存／印刷して研究ノートを作ったりできる。

このように，研究者は計算機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，イメージデータ化した写本版本を直接見，処理の指示をし，その結果をイメージデータ上で見ることができる。

5．手書き文字列検索の試作

ここではまず，（1）（2）それぞれを支援するために開発したツールを紹介する。次に，これらを利用して，（1）＋（3）の例として開発した手書き文字列検索を紹介する。

これらのツール，システムはUNIX（ユニックス）ワークステーション上で，UNIX上のウィンドウシステムの標準であるX Window（エックスウィンドウ）の環境で開発を行った。X Widowがあれば利用可能である。計算機としては，UNIXワークステーション，X専用端末，最近ではパソコンでも利用できる。（X Windowをご存知でない方は，Macintosh（マッキントッシュ）のマルチファインダーやMS-WINDOWS（エムエスウィンドウズ）を，または画面上に沢山のウィンドウがあって，同時に沢山の仕事ができる様子をイメージして欲しい。）

（1）ビューア（viewer）

手軽に利用できるスキャナで入力した写本版本（またはその紙焼き写真）を原本の字形情報を正確に再現できるよう表示するツール。白黒2色だが4階調を付けているので，筆文字のかすれや朱書きもグレー色で識別できる。拡大／縮小をして，さらに詳細に／全体的に画面上で見ることができる（図5左のウィンドウ）。

イメージデータを扱う場合はデータサイズの制約が現実的な問題として上がってくる。データサイズは，A4版1枚を400dpi（1インチあたり400ドット）で入力して約1Mバイトである。これをUNIXの標準データ圧縮であるcompress（コンプレス）をすると，十分の一の約100kバイトになる。これは，光磁気ディスクやCD-ROMに約5千～6千枚分入る計算になる。他にJPEG（ジェイペグ）（厳密には近似であって100%情報量を保存していない）などさらに高い圧縮率が得られる画像データ圧縮方式もある。

（2）ティーターム（tterm：Tate TERMinal emulator）

マルチウィンドウ環境で，入出力の全てを縦に書き／表示できるウィンドウである。ティータームのウィンドウ内ではコマンドからアプリケーションまで何を動かしても縦書きで利用できる。図5右下2つのウィンドウでは，UNIXのmore（モア）コマンドで古今和歌集を表示している。図5右上のウィンドウでは，エディタ（nemacs（エヌイーマックス））で源氏物語を編集している。日本語入力はWnn（ウンヌ）を利用している。）

特定のワープロやエディタを縦書きに改造すれば目的は達成したのだが，日本語入力のフロントエンドプロセッサにも改造が及ぶのを避けるのと，他のアプリケーションでも使える汎用性を考慮して，文字の入出力（とカーソル・マウスの制御）だけを行っているウィンドウを改造することにした。概略を説明

図5　ティータームの例

すると，X Window 上の一つのウィンドウである kterm（ケータァーム）を右に90度寝かせて文字の入出力が左→右，上→下であるのを上→下，右→左の順で行うように改造した。ここで，kterm は X Window 上で日本語を表示するために xterm（エックスタァーム）を東京工業大学籠谷氏が改造したもので，X Window の開発元であるマサチューセッツ工科大学（MIT）からフリーソフトウェアとして配布される中に含まれている（日本語フォントも）。従って，海外の日本語対応でない計算機でも，X Window を使えば日本語の表示ができる。

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「 」（ ）、。｜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た。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，横に寝かせるか，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度回転させて横に寝たまま使っている。

また汎用性の高さの例として，このウィンドウから大型計算機に接続すると，検索システムも縦書きで利用できる。図 6 に国文学論文データベースを検索している例を載せる。

（1）＋（3）手書き文字列検索システム

語彙分析の中でも一番単純な文字列検索を写本版本を見ながら行うシステムを試作した（図 7）。viewer と tterm を利用し，tterm 上のエディタ（nemacs）で文字列検索を行い，その結果（文字列マッチした行）を nemacs から viewer

図６　ティータームを使った国文学論文目録データベースの検索例

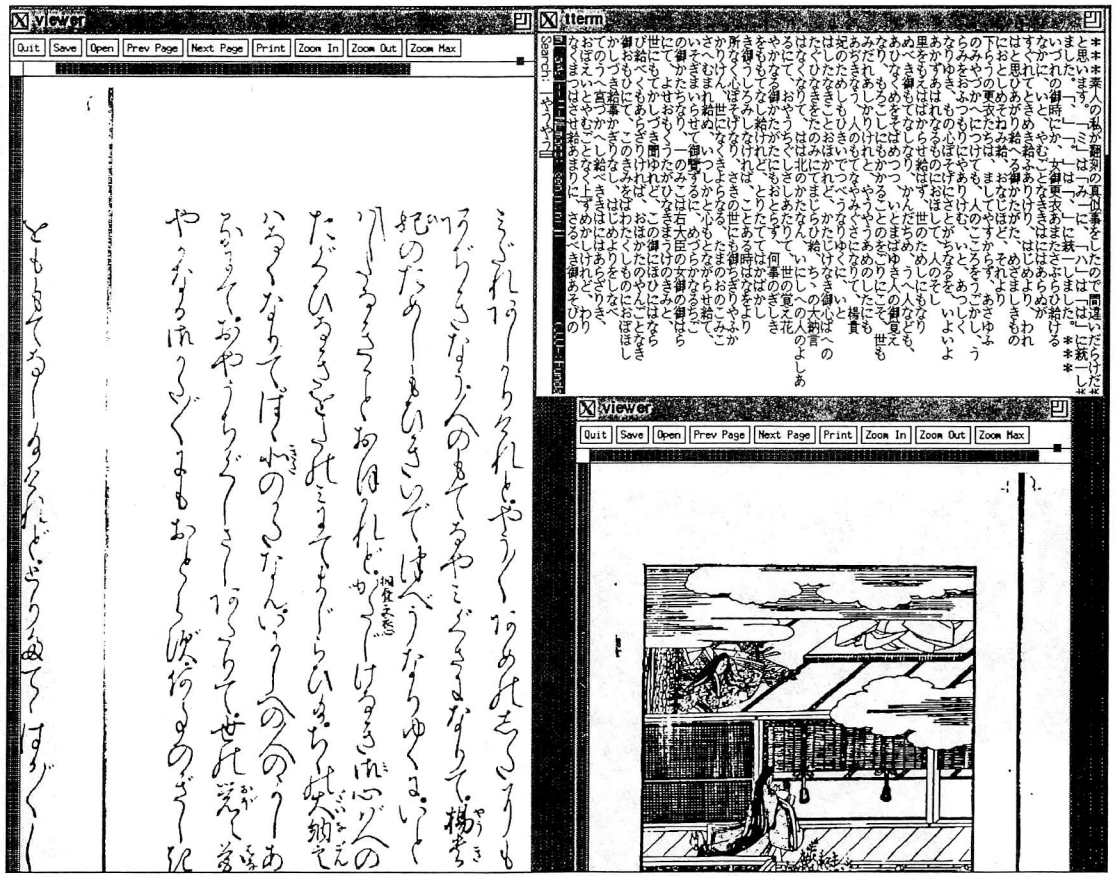


図7　手書き文字列検索システムの検索例（「やうやう」）

に渡して，viewerは文字列マッチした行を先頭行として表示する。図7では，「やうやう」を検索し（右上のウィンドウ），viewerはマッチした行を先頭にして表示している（左のウィンドウ）。

内部的な話に少し触れておくと，イメージデータとテキストデータ間の同じ箇所を示す単位を，簡単のために行単位とした。行単位の対応は，例えば翻刻する時にテキストの行をイメージデータに合わせて改行することで，容易に採取できる。既にテキストデータが存在する場合にも，イメージデータの行に合わせて切っていく作業は比較的負担が軽い。実際問題として，この例の文字列検索のように行単位の対応だけ，またはもっと粗くページ単位の対応だけでも有効な使い方もあると思われる。対応の単位については，テキストデータの構文情報などの利用を考えるとある程度文法的に明確な単位が好ましく，ほぼ文節程度が妥当と考えている。作品ごとに文節単位の対応を人手で付けるのは現実的ではないので，平仮名と典型的な漢字だけを拾って手書き文字認識を行い，認識できた文字だけを頼りにテキストとの関係を自動的に対応付ける研究を計画している。

6．おわりに

イメージデータとテキストデータの両者の持つ情報を補完しながら利用する，また計算機のインタフェースとしてイメージデータを使うという考え方を述べた。この考え方で具体的に何ができるのかを紹介した。さらに，インタフェースとしてイメージデータを使う例として，単純な文字列検索を試作した報告を行った。これは，考え方の有効性を確認するための単純な試作である。研究に供することのできるものとして，次に手書きKWICの開発に取り組んでいる。

イメージデータとテキストデータの両方を使うといっても，完璧なテキストが不可決な前提とは考えていない（完璧なテキストがあるということは研究は終わっているという説もあるようだ)。例えば次のような状況での利用を考えている。

- 未決定の文字はテキストの方にワイルドカードを埋め込んでおくなど，あいまい検索の技術により（研究途中の）不完全なテキストでも役立つ
- 一文字誤り／二文字誤りなど検索時のあいまい検索の技術により他の人が翻刻したテキストデータでも役立つ

一方，自分用のテキストデータの作成を動機付ける次のような意味も持ち得るであろう。

- テキスト作成の支援に魅力があれば翻刻というプロセスで個人の研究活動の一部としてテキストデータは自然に作られる
- 分析で得られる恩恵が大きければ，分析だけを目的にしてもテキストデータは作られる

計算機のインタフェースとして，テキストデータを強制されることなく，直接イメージデータを使う話をした。これは見方を変えると，テキストデータはあくまでも計算機処理のためのデータ形式（計算機の内部表現）に過ぎないという考え方も可能であり，計算機の分野でも新しい発想である。この考え方は，国文学だけに限らず写本版本を研究する分野で共通に使えるものであり，新しい計算機科学の基礎作りに役立つことを期待している。

本稿の中でも触れたが，平仮名と典型的な漢字の手書き文字認識（これが，まさに「計算機が写本版本を読む」である)，古語辞書を使った自動分かち書きの研究，実験を計画している。

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，当館所蔵和古書源氏物語を電子複写コピーサービスにより利用させて頂いた。常日頃，国文学の立場からアドバイス頂いている研究情報部長新井教授，データベース室長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] 北村啓子，“CD-ROMによる国文学研究材料データベースの配布”，国文学研究資料館紀要，第17号，(1991)
[2] 北村啓子，“古典テキストCD-ROMの開発”他4編，科学研究費補助金（試験研究（1））研究成果報告書課題番号：63810007，pp.17-60，(1991)
[3] 新井栄蔵，パネル討論より，『国文学とコンピュータ』シンポジウム（第1回）講演集，pp.59-73，(1990)
[4] 新井栄蔵，“古典本文データベースとその検索をめぐって”，『国文学とコンピュータ』シンポジウム（第2回）講演集，pp.22-35，(1990)
[5] 北村啓子，“古典テキストCD-ROMシステム”，『国文学とコンピュータ』シンポジウム（第2回）講演集，pp.123-146，(1990)