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

1R−1

北村 啓子
国文学研究資料館 研究情報部

1.はじめに

古文書をイメージデータとして計算機に格納したり、古文書を翻刻した文書をテキストデータとして計算機に蓄積されてきている。古文書の特徴には、書かれている文字情報はもちろんその形状も重要な情報であり、研究者が分析・研究の対象として扱うのはイメージであるという点がある。しかし、古文書には多くの異本が存在し、数多くのイメージを扱わなくてはならない。テキスト化されたデータを計算機処理するにも、古語辞書を始め古文を処理するための蓄積が殆んどない。またハイパメディアへの期待も大きいが、書かれた文字情報の利用を考えると充分とは言えない。

2.古文書の表現モデル

古文書を計算機で扱うための表現モデルを考える。研究者が分析・研究するために、古文書をマルチレイヤーの構造でとらえている（図1）。筆記体の形状をそのまま保持するイメージデータ（2値データ）と、その文字情報を文字として計算機で処理するために翻刻したテキストが必要である。さらに、テキストデータも使用目的によって数種類を利用することが必要となる。例えば、漢字／かな表記、新字／旧字の問題を回避するための読みのテキスト、索引作成や用語単位の処理をするための分かち書きしたテキストなどが考えられる。これらは、古文書の性格やその利用方法によって異なってくるであろう。このようなマルチレイヤーのデータ構造の上で、各レイヤーの同じ箇所を串刺しにして見られることが、重要なポイントである。これによって計算機処理の様々な可能性を広げることができる。

ここでは、レイヤー間のマッピングの中でもイメージデーターテキストデータ間のマッピングについて述べる。

3.語彙の対応関係の考え方

ハイパテキスト、ハイパメディアの研究が盛んに行われ、人文科学での利用が期待されている。しかしこのデータモデルでは、メディアの異なるデータも含め任意の情報の塊を同等に扱える柔軟性の反面、情報の塊とそれらの関連情報をゼロから作り上げなければならない。さらに、それらは意図的に指定されたものに限定され、意図されていない情報の塊やそれらの関連を見ることはできない。

古文書の多くは筆記体で書かれた写本・版本である。その形状はイメージデータで表現することになるが、そこに書かれているのは文字情報である。イメージデータの形状を判読し、活字化（テキスト化）することを翻刻といい、この作業自身が研究活動である。イメージデータとテキストデータと表現形態が違っても、書かれている文字情報としては同じであり（必ずしも文字列としては一致しない）、語彙レベルでの対応関係が潜在的に存在する。この語彙が、文章に対する人間にとって自然な情報の塊であり、データを行き来する時の該当箇所（同じ所）を示す手がかりである。この語彙の対応関係を利用することにより、イメージデータのインタフェースを通して他のレイヤーの持つ情報を利用することができる。また人文科学で研究対象として古文書を使う場合、文章中のキーワード的なものだけでなく、全ての語彙を扱えることという要請がある。それら語彙の対応関係を利用することにより、どの語彙ももれなく情報の塊として利用することができる。ハイパメディアの考え方に、この語彙の対応関係という考え方を導入する。

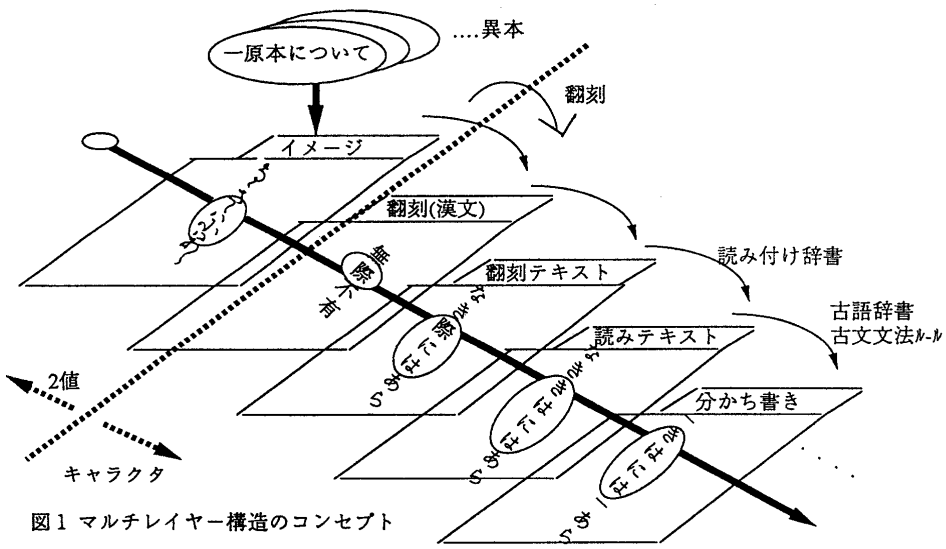

図1 マルチレイヤー構造のコンセプト

Concept of Multimedia Data Model for Classical Books
Keiko KITAMURA
National Institute of Japanese Literature

この表現モデルによって可能となる利用例を説明する。

○イメージ上での文字列検索、語彙分析

テキストデータを使って文字列検索した結果を語彙の対応関係を利用してイメージ上に提示する。利用者は、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。

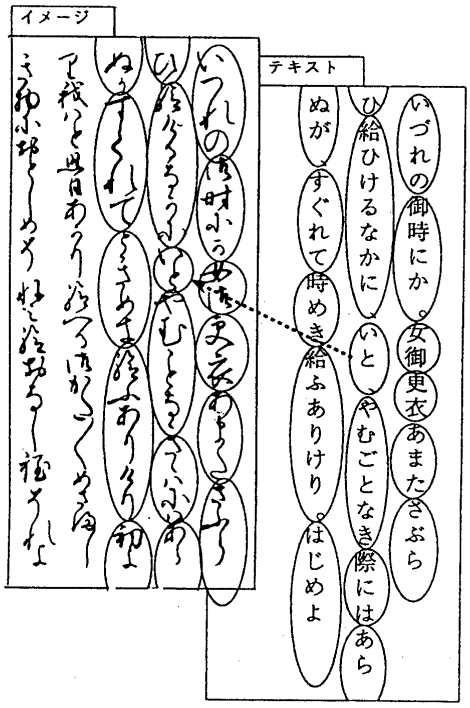

図２ イメージ上での文字列検索、語彙分析

○異本間の差異比較

異本の差異の比較は重要な研究である。異本Ａのイメージデータと異本Ｂのイメージデータを比較する時に、異本Ａのテキストデータと異本Ｂのテキストデータとの文字列上の差分を取って、それを異本ＡＢのイメージデータにそれぞれマッピングすることにより、文字情報としての差をイメージ上で見ることができる。

４.語彙の単位と対応関係の作成

語彙の定義を厳密に（例えば文節単位、語単位のように）行ない、研究者が人手でその語彙の対応関係を設定するのは、利用価値は高いが現実的ではない。（ハイパメディアで全ての用語をノードとし、その関連情報をゼロから作成する困難と同じである。）実際に利用する場面として、a.イメージデータがあり翻刻作業に計算機を使ってテキストを入力する場合、b.イメージデータとすでに翻刻されたテキストがある場合が考えられる。a.の場合は、翻刻しながらテキストに何らかのイメージとの対応関係を入力することが可能である。入力の負担にならず価値のある情報として、行単位の対応関係が考えられる。例えばイメージの行に合わせてテキストを改行するのである。b.の場合でも、イメージに合わせてテキストを行分割するのは可能である。このようにして少なくとも行単位の対応関係は抽出できる。

さらに細かい対応関係を計算機で抽出するために、語彙の対応関係も単なる内部表現と考え、語彙の単位は文節程度で意味的なものを厳密には定義しないことにする。そして語彙の対応関係を、読める文字だけ読む手書き文字認識によって抽出することを考えてみる。古文の手書きを読めない素人がテキストとつき合わせて簡単に読める文字だけ拾ってマッピングしながら読んで行けるのと同じ戦略である。筆記体でも何種類かに限られるひらがなやパターンマッチングしやすい特徴的な漢字を拾って、テキストとつき合わせて行く。認識できない文字列（unknown）は、テキスト上のつき合わせで残った文字列と対応させる。先に述べた行単位の対応情報を使えば突き合わせる範囲が行内に絞られるので、マッピングの取れる確率は上がるであろう。さらにテキストデータに分かち書きの情報があれば、テキスト上の厳密な語彙に対してイメージ上の語彙を、１対多、多対１の対応を取ることもできる。イメージ上の厳密に定義していない語彙は、利用者が使う時の用語の単位とは一致しないが、対応関係をたどる時に双方向に語彙の部分マッチング／複数にまたがるマッチングを考慮すれば正しく対応は取れるであろう。

５.手書き古文書上での文字列検索の試作

３章で説明した手書き古文書のイメージ上での文字列検索を試作する。４章で述べた語彙の単位として、最少は文字単位から最大は行単位まで考えられるであろう。まず対応関係を使うことの有効性を評価するために、対応の取り易い行単位の対応関係を使うことにする。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。翻刻作業でイメージを見ながらテキストを入力するという状況を想定し、テキストにイメージの行に合わせて改行コードを入れ、それから対応関係を抽出することにする。この試作は現在インプリメント中である。

６.おわりに

古文書を計算機で扱うために、語彙の対応情報という考え方を導入し、古文書の表現モデルについて検討を行なった。応用例として、手書き古文書のイメージ上での文字列検索の試作に取り組んでいる。文字列検索機能の高度化も重要であるが、文字情報が書かれた形状が重要である古文書を研究する分野では、潜在的に存在する語彙の対応を計算機が知っていることは重要なことである。ユーザインタフェースはイメージであり、テキストは計算機の内部表現であるという考え方も可能である。表現モデル自身まだ基本的なアイデアだけで、充分練れているとは言えない。様々な応用が期待できると思うが、それに耐え得る表現力を持つかどうかも検証せねばならない。また、語彙の対応関係の作成については、実際どの位抽出できるか実験による評価が必要である。

謝辞：日頃国文学の観点からご教授頂いている、当館当部新井教授、松村教授、中村助教授に感謝します。
また、本研究は稲盛財団の研究助成を受けています。

＜参考文献＞

[1] 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)
[2] K.Kitamura, Data Base Delivery fot Japanese Literature by CD-ROM, Proc. of ACH/ALLC'91,(1991)
[3] 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, pp.1-27,(1991)
[4] 北村啓子、Hypertext技術を応用した横断的利用技術の提案、科研#63450059報告書、pp.27-50,(1990)