2C－1

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－

北村 啓子
国文学研究資料館 研究情報部

1.はじめに

古い著作物を計算機で扱うためにマルチレイヤ構造のマルチメディアデータモデルを提案している。ユーザインタフェースとしてのイメージデータ（筆記体の形状をそのまま保持）と計算機処理のための内部表現としてのテキストデータ（筆記体を活字に翻刻したテキスト；複数種類のテキストがレイヤーを成す）の間で各レイヤの同じ箇所を同定しお互いに情報を補完しあうことを特長とする。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は[1]で述べた。本論文では、古い著作物固有の特徴を考慮したマルチレイヤ構造のテキストについて述べる。また実際に和歌文学を題材として、テキストの表現方法とテキストを扱うために開発した基本的な処理ライブラリとそれらを利用した分析の例を紹介する。この環境はjperlで構築している。

2.マルチレイヤ構造のテキスト

国文学をはじめ古い著作物を対象とする分野一般の特徴として文字の認定、読み方の正解が無いという難しさがある。その要因として、古い著作物に本質的に存在するもの、古い著作物→翻刻→マシンリーダブルテキストのプロセスで発生するものなどがある。具体的に挙げると、

・読みが不明
・異表記が多い：漢字／かな、異体字、同義文字、正字／俗字、新字／旧字
・異本による差
・句読点を使わない／使い分けしない

が大きな差を生んでいる。これらはテキスト検索をはじめあらゆるテキスト処理の精度を落す原因となる。これらの差を包含し表現する方法として図１に示すマルチレイヤ構造のテキストを利用することによりこの問題を解決する。

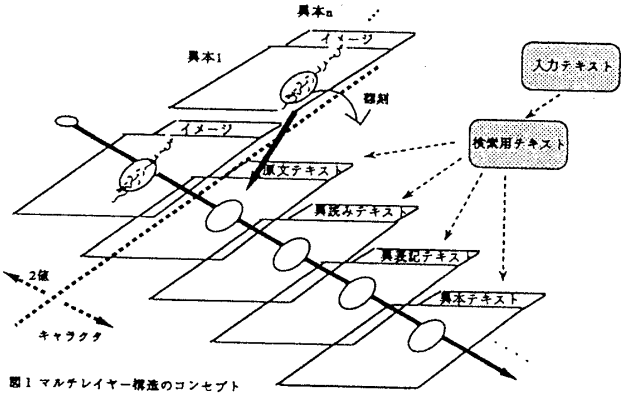


図１ マルチレイヤー構造のコンセプト

3.古い著作物の構造

国文学が対象にしている古い著作物をテキスト処理するという観点で、和歌、散文、漢文／万葉仮名文、その他に大別する。それぞれの特徴とその構造を分析する。

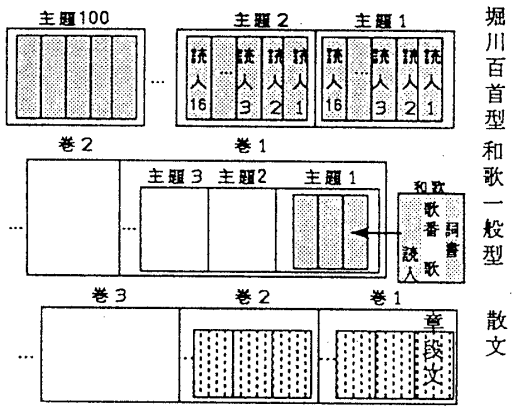


図2 古い著作物の構造

○和歌：

図２に示すように物理構造が論理構造を表現している。特に単純な構造で扱いやすい堀川百首型と和歌一般型に分類する。和歌のように論理構造が明解である場合、次の２つのアプローチが考えられる。

a. 物理構造から論理構造を抽出

論理構造を処理プログラムに埋め込んでおくことにより物理構造から論理構造を抽出する。データ作成が楽で標準化の煩わしさもなく、見慣れた物理構造をしているので違和感がない。論理構造を示す物理構造の区切り（空行、改行、字下げ、下合わせ、etc.）を定義可能とすることにより、自由度を高くできる。

b.論理構造のマークアップを定義

論理構造を明示するマークを決めてマーク付けを行う。標準化に伴う煩わしさがあるが、マークが標準化されると処理プログラムも共有できデータ互換性もある。標準化が進むとその分自由度は低くなる。 マーク自身を定義可能とすることにより自由度を高くできる。

両者は逆のアプローチであるが、b.で定義したマークを物理構造の区切りに縮退したものと考えると、両者を包含した高い一般化のテキスト表現が可能である。簡単な論理構造のものはa.のみで、複雑な論理構造のものは大まかな構造はa.を、細かな構造はb.を使った折衷が妥当と考える。

○散文：

図２に示すように、物理構造は文、段、章、巻で構成される。文については、句読点がないものが多く、句点と読点を区別していないものもある。その場合、句読点の正解は存在しない。散文の中の多くは物語であり、What（何を書いているのか）を明解には言えない。書誌情報（書名、著者名、etc.）以上の論理構造はなく、物語を分析するために意味ある論理構造のマーキングは困難であろう。

散文を分析するには、テーマに応じた機能語を使うことを考えている。例えば、

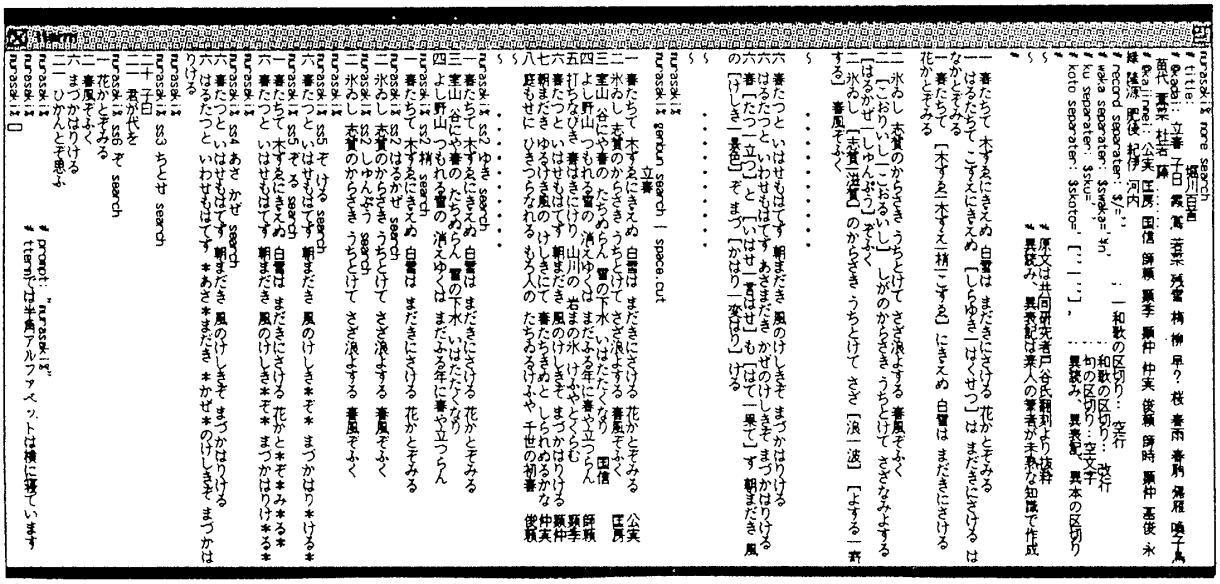
軍記物語：～に～参じる、～に充満す、～に著く、etc...を拾って登場人物の移動経路を辿っていく。または、

会話物語：～曰く「～ 」、～と申す、～と仰る、etc...を拾って登場人物の会話集を作成するなどである。

○漢文、万葉仮名文：漢字仮名混じり文に書き下して散文として扱う。

○その他：謡曲（楽譜）、台詞（シナリオ）、絵文字や造字を使った物の一般化は困難。個々の対応を考える。

Text Stracture and its Text Processing Libraries for Classical Books
Keiko KITAMURA
National Institute of Japanese Literature



4.和歌向けのテキスト表現

前章で述べた堀川百首型と和歌一般型に分けて、論理構造を示す基準を定義可能とすることにより一般化したテキスト表現ならびにそれらを処理するライブラリを開発した。特に、異なる読み、異表記、異本による差のバラつきを吸収したテキスト処理ができること、ならびに和歌特有の五七五七七を活かした処理や句切れを利用して検索ノイズを減少できることを考慮した。

入力テキスト、検索用テキストは、1レコード目にヘッダ情報として物理構造の区切り、論理構造を明示するマークの定義、ならびに巻名や歌題、歌人名の設定を書く。2レコード目から和歌テキストが始まる。図1に点線で示したように、入力テキストからヘッダ情報を基に特殊化や省略の補完等を行う検索用テキストへの変換、ならびに検索用テキストから各レイヤのテキスト作成はライブラリを使って行う。

堀川百首型のテキスト表現に従って物理構造の区切りと歌題、歌人名を設定した入力テキスト（この型は検索用テキストと縮退）のヘッダレコードを以下に示す。

```
#####1st record: header: このデータのマクロ値、セパレート記号の指定
# title:                        ; 歌集名
#
### マクロ値
# 歌題: @kadai:
# 歌人名 or 選者名: @kajinmei:
#
### セパレート記号
# file: <record>                ; 和歌集：<和歌>
# record separater: $/=''       ; 一和歌の区切り： 空行
#
 record: ($genbun, $yomi, $ihyouki)    ; 和歌：<原文、異読み、異表記>
 waka separater: $swaka='Yn'           ; 和歌の区切り： 改行
#
 $genbun: ($n, $ku1, $ku2, $ku3, $ku4, $ku5);原文：<歌番号 五 七 五 七 七>
 $yomi: ($n, $ku1, $ku2, $ku3, $ku4, $ku5);異読み：<歌番号 五七 七五 七七>
 $ihyouki: ($n, $ku1, $ku2, $ku3, $ku4, $ku5);異表記：<歌番号 五 七五 七 七
 ku separater: $sku=' '                ; 句の区切り：空文字
#
 koto separater: $skoto='<''|''>'      ; 異読み、異表記、異本の区切り
#
#####after 1st record: 和歌テキスト
# - - - - - - - - - - - - - - - -
#
                                ### 実際の和歌テキストは上の画面を参照
```

5.和歌向けのライブラリ

国文学者をはじめ人文科学系研究者の数少ない計算機利用者の多くはパソコンユーザであることの考慮と、以下の理由でjperlを利用することにした。ライブラリはすべてjperl scriptである。各ライブラリはＵＮＩＸ／ＭＳ－ＤＯＳのコマンドとして実行でき、パイプ機能を利用してコマンドをつないで新しいテキスト処理を実現できる。

・移植性：ＵＮＩＸ、ＭＳ－ＤＯＳマシンのほとんどにインプリメントされている、日本語コード変換のみで移植可能、言語方言がない
・文字列処理能力：非常に強力である
・カスタマイズ容易性：リーダビリティが高い、記述性が高い、コンパクトに書ける、数多くの実現方法がある、インタプリタなので即実行できる

以下に和歌用に用意したライブラリを列挙し、その実行例を上に掲載する。tterm[2]を使って縦表示をしている。

```
○検索用データへの変換
  gene:# 入力データのヘッダ情報に基づいて検索用データに変換する
○原文抽出
  genbun:# 検索用テキストから原文だけのテキストファイルを作成する
○レイアウティング
  space.cut:# 空行を取る
  layout:# XX本のレイアウトに編集する
○文字列検索（単純、異なる読み／異表記／異本）
  ss: # 単純文字列マッチ
  ss2:# 原文、異読み、異表記全てを検索してマッチした歌の原文を表示する
○２ワード検索（係り受け、枕詞）
  ss5:# 係受けの言葉を検索してマッチした歌の原文を表示する
      # 各句の最後にマッチのみ（〜ぞ 〜ける）
      # 原文以外にマッチした時は原文とマッチした読み／異表記を表示する
  ss4:# 各句の途中もマッチ（ss5のマッチング条件を#に変更）
      # 原文以外にマッチした時は原文とマッチした読み／異表記を表示する
○２ワードの相手捜し検索（係り受け、枕詞）
  ss3:# 枕詞捜し
      # ss2と同じ検索をしてマッチした歌の最初の句を表示する
  ss6:# 係り受け捜し
      # ss2と同じ検索をしてマッチした歌の最後の句を表示する
      # (ss3 の $ku[ -> $ku5)
```

6.おわりに

テキストのマルチレイヤ構造の考え方について述べ、和歌文学を題材にテキスト表現とその処理プログラムのライブラリを紹介した。ライブラリは整理してより一般的なものにリメイクする予定である。ユーザの習熟度に合わせて、デフォルトのテキスト表現を利用→自分のテキストに合わせてカスタマイズ→処理プログラムの拡張と段階的な使い分けが可能である。このテキスト表現とライブラリが人文科学系研究者のperlへの入門になることを期待する。散文については、機能語を使った処理に現在取り組んでいる。さらに文字の補助集合の取り込みの検討と[2]で報告したttermのリリースの準備をしている。

なお、本研究は稲盛財団の研究助成を頂いている研究課題の一環として行っている。また本研究中、堀川百首を対象にした部分は当館共同研究により研究の機会を得た。

＜参考文献＞

[1] 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,'92
[2] 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)