

留学生の日本語作文データベース(2)

— 書き直された作文の活用 —

中川 正弘

0. はじめに

先年、留学生の書いた日本語作文の手書き原稿と電子テキスト版、添削済み作文をデータベース化し、これを日本語研究に活用することについて書いた¹⁾。その際、添削後の作文よりも間違いを直される前の留学生の書いたそのままの作文を特に日本語研究に有用なデータと考えた。日本語学習者の日本語能力の現状を分析、理解することが、日本語の教材、教授法の改善に結びつくと考えたからである。そのデータベースに添削後のものを含めたのは、外国人の使う日本語が日本人の使う日本語とどのように違うかを浮かび上がらせるためであった²⁾。

今回データベース化したものはそれとは少し違う。広島大学留学生センターで現在行っている6ヶ月集中予備教育(日本語研修コース)の1期(1985年10月～1986年3月)から27期(1998年10月～1999年3月)までの修了スピーチ作文の文集である。

このような文集は長期保存など特に考えられていなかったため、散逸したものも多く、何人もの関係者に保存してあるものを調べていただいて、何とか26期分集めることができたのだが、どうしても見つからないものも一期分あった。残っていたものもコピーをスケーパーで留めただけの簡易製本とも言えないほどの作りであるため、傷みんでもいた。

その中で記憶に残っている作文をもう一度読もうと思っても、記憶だけが頼りではなかなか探し出せない。このコースで日本語を学んだ修了生にとって記念碑的なものであるだけでなく、コースを担当した者にとっても思い出が詰まったものでもあり、適当な保存方法が求められていた。

そこで、上述の作文データベースに倣い、コンピュータで管理、検索が容易な画像と電子テキストに変換することにした。

1. 手書き原稿から画像データ、電子テキストへ

27冊の文集だと、1期平均20名として、各自3枚、27冊で1620枚もの手書き原稿があることになる。しかし、これをスキャンして画像データにするだけなら、それほど時間はかからない。分類、整理に必要な時間を別にして、1枚のスキャンに30秒もかからない高速スキャナーを使えば、13.5時間ほどということになる。

しかし、画像データにするだけでは保存しやすいだけで、読みたいものをすぐ探し出せ

ない。データとしての活用を考えるなら電子テキスト化し、さまざまな検索ができる形にしなければならない。その基礎データになるワープロ入力は学生アルバイトに依頼した。その際、手書き原稿に含まれる様々な情報は可能な限り残してもらった。ひらがなとカタカナの混乱、漢字の間違い（他の漢字との混乱）、句読点とコンマ、ピリオドの混用、字下げ、スペースの使い方もである。ただし、漢字の部首の間違いで日本語で使われない漢字となっている場合には仕方なく正しい漢字に置き換えた。該当する画像がすぐ参照できるようにその箇所に何かマークを付けることも考えられたが、作業が煩雑になるため今回は見送った。そのようなものは必要になった時点で付け加えることもできる。

スキャナーは紙を差し込めば自動的にスキャンし始める小型のもの（Visioneer PaperPortVX）を使用した³⁾。既に製造・販売を終了している旧式のものだが、データベースとして1ビット、300dpi程度の画質で十分なのであるため、新しい高精度のものを使う必要はない。



図1 スキャナーの設定

2. ウィンドウデザインとレイアウト

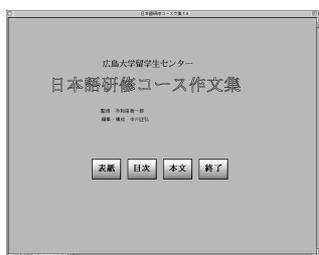
今回のデータベースにもファイルメーカーProを使用した。バージョンが3.0から4.0に上がったが、デザイン、レイアウトなど、基本的な部分で変更はない。3.0で作成したものを4.0で開くことも、その逆も可能だ。

- カバーページ-

各文集の表紙画像、目次、一冊分の本文ページへのリンクボタンと終了ボタンのみ。

- 表紙画像のページ-

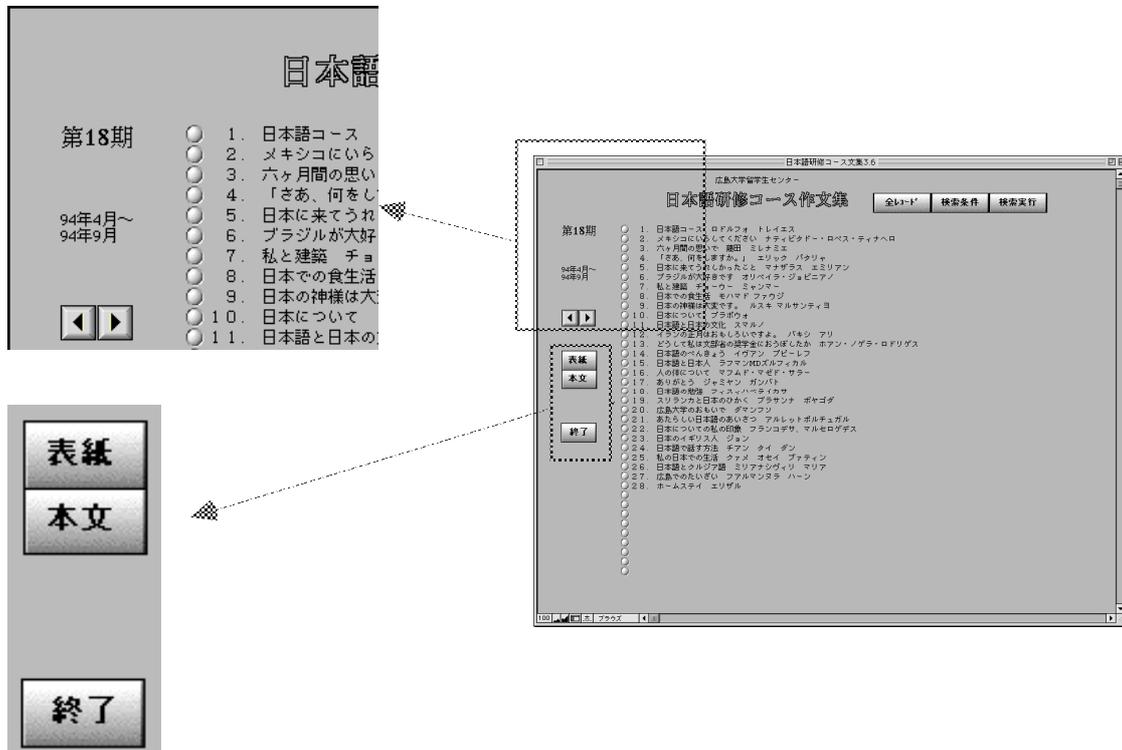
一つ前の期／一つ後の期へのリンクボタン、その期の目次へのリンクボタン、文集一冊分を一つのフィールドに流し込んである本文ページへのリンクボタンと終了ボタンのみ。



- 目次のページ-

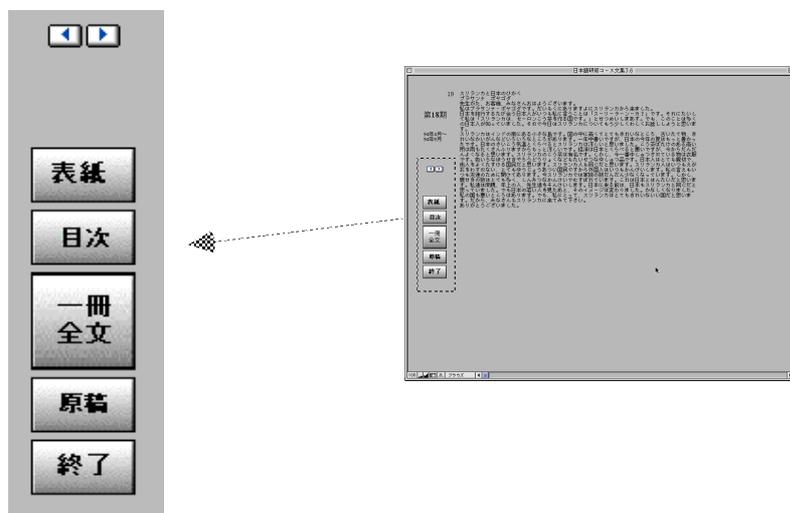
個別の作文へのリンクボタン、一つ前の期／一つ後の期へのリンクボタン、表紙画像のページと一冊分の本文ページへのリンクボタン、終了ボタンに加え、全レコード／検索条件／検索実行のボタンがある。検索条件ボタンをクリックすると、全フィールドが空白に

なるので、年度、期、学生名、作文題目、作文中の言葉のどれでも必要なところへ書き込み、検索実行ボタンをクリックする。検索すると表示は検索したレコードだけになるので、再び文集の全データを対象とするには全レコードボタンをクリックしなければならない。



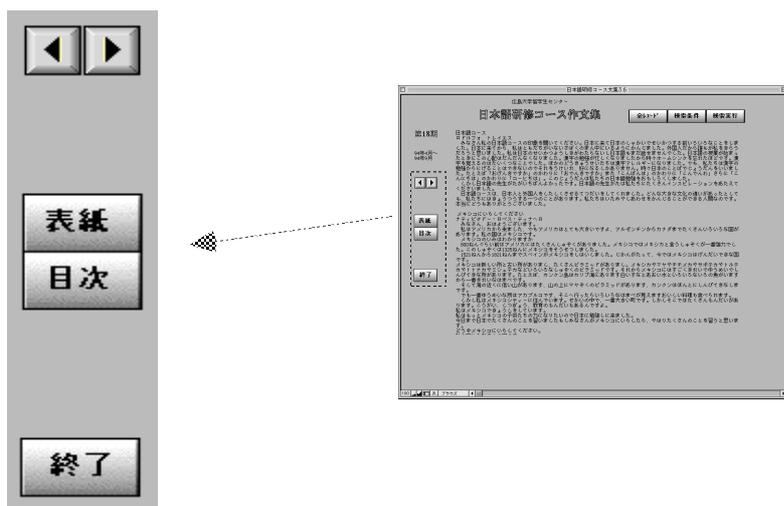
－ 単独作文のページ－

目次の一つ前／一つ後へのリンクボタン、表紙画像のページへのリンクボタン、目次、一冊全文、原稿画像へのリンクボタン、終了ボタンがある。



一冊分の本文ページ

作文が一つずつ読めるページの他に、文集一冊分を一つのフィールドで一括表示するページを作った。これは何期の何番か分からない学生名、あるいは作文中の言葉を検索するためであるが、このデータベースの全作分を他のアプリケーションにコピーしてさまざまな検索、分析を行うことを想定し、27期なら27回のコピーで写し取れるようにした。



原稿画像のページ

実物の50パーセントぐらいの縮尺の原稿画像だが、コンピュータ画面で書き癖まで判別できる。また、少しかすれた程度で十分読めるぐらいの精度で実寸で印刷できる。ここも目次の一つ前／一つ後へのリンクボタン、表紙画像のページへのリンクボタン、目次、一冊全文、終了ボタンがある。

現時点では手書き原稿の1ページ目だけにしてある。画像でページめくりができるようにも、同じページに2ページ目、3ページ目を並べて縮小表示し、どれも実寸で印刷が可能構成にもできなくはないが、そのような使用が必要になった時点で作り替えることにした。



3. 検索

一人名一

目次のページ、もしくは文集全文ページで検索条件をクリックすると、目次のフィールド、文集全文のフィールドが空っぽの検索モードに変わるので、そこに探したい名前を書き込み、検索実行をクリックする。

該当するページが出ると、目次の場合は一覧表示なのですぐ分かる。文集全文のページはテキストがフィールドにごく一部しか見えていないため、その名前がそのフィールドのどこかにあるということしか分からない（何期、何年何月―何年何月の者であるかはどのページでも左上角に表示される）。その人の書いた作文へ行くにはボタンをクリックして目次ページへ行き、そこで作文題目の頭にあるボタンを押す。

―作文中の語彙―

作文中の語彙は文集一冊分の本文ページからしか検索できない。また、該当するページが表れても、現在のバージョンではその語彙がどこにあるか反転などで表示させることができない。ファイルメーカーPro はデータベースソフトウェアとしては非常に便利なものなのだが、フィールド内テキストのある語彙を特別な形で表示するのは難しいようだ。

研究のためにさまざまな検索をしたい場合には、全文集のテキストをワープロソフトなどに移植して（冊数と同じ回数のコピー・アンド・ペーストで一文書にできる）、それぞれのソフトの得意な検索機能を使うほうがいだろう。

単純な語彙の検索以外に単文中で他の文字列をまたいだ複数の文字列の組み合わせを探するような特殊な検索までしたい場合には、「強力検索」機能が売り物のマーキュリーソフトウェアのNisusWriterのようなものもあり、検索式を工夫すればさまざまに使える。

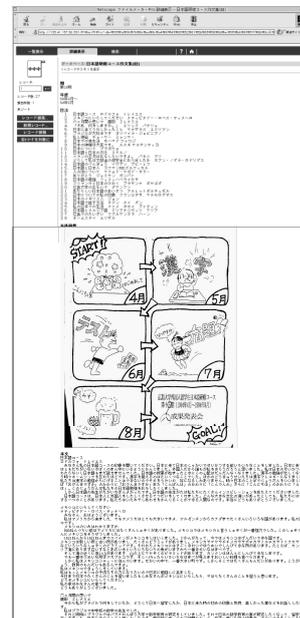
4. 配布と閲覧

上述のデータ構成で40MB規模になるため、配布しようとするフロッピーディスクでは無理である。Zip、MOを使うことも考えられるが、メディア単価がもっとも安く、データが壊れる心配もないのでCD-ROMに焼き付けることにした。ファイルメーカーProはWindows版と互換性があるため、ハイブリッド版を作ることも難しくないようだが、現在、テスト版の段階では、Macフォーマットにしてある。また、CD-ROMにはデータベースのファイルだけではなく、ファイルメーカーProを使っていない方も使えるようにこのソフトの評価版も入れてある⁴⁾。



Mac フォーマットの CD-ROM であるが、Windows コンピュータでも MacDrive98 のような Mac フォーマットを使用可能にするソフトをインストールすれば、Windows 版のファイルメーカーPro で開くことができる。

また、ファイルメーカーPro はバージョン 4.0 から Web で公開し、ブラウザで表示することができるようになったので、広島大学留学生センターのホームページで閲覧できるようにする予定でいる⁵⁾。ただ、ファイルメーカーPro のバージョン 4.0 のインスタント Web 機能ではデザイン、レイアウトまで反映させることは困難で、右図のように、一つのページにテキストデータも画像データも一繋がりでは表示できず、使い勝手がいいとは言えない。



2000年1月発売のバージョン 5.0 は Web 公開でレイアウトまでほとんど表示できる (Mac 版なら、CSS (Cascading Style Sheets) に対応した Internet Explorer 4.0 以上) はずなのだが、こちらのデザイン、レイアウトの構成方法に問題があるためか、うまく表示できない。バージョン 4.0 のカスタム Web を使うほうが表示も正確で軽快に動くそうなのだが、この html 変換は易しくはない。

5. データベースの活用

この文集に入っている作文は留学生が書いたそのままの日本語ではなく、コースの担当者などが手を入れた後、清書したものである。従って、外国人の使う日本語がどんなものか、文法や語彙の間違いを考察するためには使えない。しかし、添削を受けていると言っても、完全な日本人の日本語になっているわけではない。短い単位で見ると文法が間違っているとは言えないが、ここかしこに違和感を覚える。ここには添削した日本人の文法を扱うスタイルの違いが表れているのである。そう考えると、単なる文法、語法の間違いの水準とは違う水準で日本語の問題を考えることができる。

また、言葉ではなく内容の水準での利用も可能だ。日本語教材の大部分は日本人の考えを日本語で表したものでいいのだが、一部、外国人の考えを日本語で表したものも必要ではないだろうか。

予備教育の初級日本語に限らず、中級、上級で必要な発話、作文能力を考えると、日本人の知らないこと、日本人の考えないことを日本語で表せるようにしなければならないはずで、そのためには日本語教材も少し違ったものも使うほうがいいだろう。

このデータベースにあるような、ある程度書き直された作文を完全な日本人の日本語に更に書き直せばどうだろうか。内容は外国人の考えたもの、日本語は日本人のスタイルと

のである。

中身となる作文データの入力、スキャナーによる画像の取得は単純作業なので外注もしやすいだろうし、その張り込み作業もテキストと画像のコピー・アンド・ペーストだけなのだから、これなら自分でもできるという方は多いだろう。余裕があれば、背景やボタン、フィールド枠のデザインを変えればいい。

そのような利用を考え、データレコードなしのレイアウト、デザインのテンプレートを留学生センターのホームページからダウンロードできるようにする予定でいる。(了)

-
- 1) 中川正弘、**留学生の日本語作文データベース**、広島大学留学生センター紀要、第9号、1999年
 - 2) 参考 中川正弘、「作文」を「読む」／「書く」技能の位置づけと展開、広島大学留学生日本語教育、第4号、1992年
中川正弘、**作文の誤りと文体**、広島大学留学生センター紀要、第3号、1993年
中川正弘、**作文と解釈行為**、広島大学留学生センター紀要、第4号、1994年
中川正弘、**外国人の日本語、日本人の日本語 - 言葉の問題から教授法の問題へ-**、広島大学留学生日本語教育、第6号、1994年
中川正弘、**作文の添削と文体差**、広島大学留学生日本語教育、第7号、1995年
中川正弘、「は／が」と助詞選択の零度、広島大学留学生日本語教育、第8号、1996年
中川正弘、**添削文が語る日本語のスタイル(1)完了表現と時制**、広島大学留学生教育、第1号、1997年
中川正弘、**添削文が語る日本語のスタイル(2)言葉の選択と序列**、広島大学留学生教育、第2号、1998年
中川正弘、**スタイルから見る日本語文法**、広島大学留学生センター紀要、第10号、2000年
 - 3) カラー対応の後継機種 PaperPortStrobe が現在発売されている。
 - 4) ファイルメーカーを使った市販データベースならどれにでも添付されているようだが、今回はアクシスパブリッシング社の『万国辞典』のCD-ROMに含められていたバージョン3.0のものをコピーさせていただいた。評価版は機能制限があるだけで、インストールすれば普通に使える。また、ファイルメーカー社のホームページ (<http://www.filemaker.co.jp/downloads/index.html>)から最新のものがダウンロードできる。
 - 5) 現在、留学生センターのホームページにリンクを張らずにテスト中である。閲覧を希望される方は、その旨連絡いただければ、テスト公開のURLをお知らせする。E-mail: masahiro@hiroshima-u.ac.jp