機械可読テキストと〈索引〉

高木　元（愛知県立大学文学部）

情報処理語学文学研究会第 17 回大会（1995 年 7 月 22 日　共立女子大学）

概 要

最近まで、機械可読テキストには索引の機能が内在化しているものと考えていた。ところが、実際に一冊の本を作成し、その索引を作ってみると話は単純にはいかないことが分かった。そこで、フロッピー入稿から校正そして索引作成まで、実際の本造りの作業過程を辿りながら、直面した諸問題について報告したい。さらに、機械可読テキストの特質と、モノとしての本との差異についても考えてみたい。

1　論文集という本

1.1　出版企画

およそ売れる見込みがない学術論文集の出版企画は、一種投機的な要素の濃い他の分野の出版とは、少しく性格を異にしているはずである。何しろ対象読者数が限られているので、捌ける部数を予測しやすい。となれば、編集出版流通に掛かる経費の計算から容易に付けるべき単価がはじき出せる[1]。

有能な編集者は、世に出すべき優れた著者を発掘すると、その著書の出版企画を立案するに際して、書型や頁数や装幀などについて検討する。さらに、所属する出版社（ブランド）の編集会議を通して経営者や営業サイドの了解を得るために、少しでも生産コストを下げるように努力するはずである。

そこで著者が要求されるのは、完全原稿での入稿と、時期を逸しない迅速な出版へ向けての校正作業であろう[2]。

1.2　電算写植

そこで近年一般化した電算写植による組版は、完全原稿の機械可読テキストさえ用意出来れば、編集校正に費やされる労力が低減できる劃期的な技術のはずであった。ところが、実際には念校あたりで活版やタイプ印刷の時代だったら我慢して諦めてしまうところであっても、その修正が技術的に容易なことを知っているために、気軽に書き直してしまうから、結局は労力の低減にはならない場合が多い（らしい）。

もっとも、漢文が多い場合や字体（字母）にこだわる場合には、いまだに活版に依存している出版も残っているようだが、時代の流れからいっても風前の灯であろう[3]。

[1]当然のこと需要と供給の相関関係で決まるわけであるが、やはり一冊が数万円という値段設定は、売ることをまったく考慮していないとしか思えない。単に経費を部数で割っただけとしか思えないからである。斯様な出版は、取り敢えずその本が流通に乗ったということだけでしかない。そもそも、公費はもちろんのこと、私費でもとても買う余裕はないし……。

[2]誰しもが予想できるように、これがそうは簡単にはいかないのである。それと、やはり二月と八月の出版は営業側から注文で差し控えられることが多いらしい。斯くいう拙著も九月出来予定に延期（になって助かった）。

[3]あの活字印刷の持つ重厚な迫力や、印面を撫でると凸凹していてそれとわかる物理的な刻印の痕跡には、いわく謂いがたい魅力があるのだが……。

ワープロなどの普及によって、科研の報告書や資料集などは、自分で板下を作成してしまってから、オフセット印刷に回されることも多い。また、御承知の通り本会の会報累積版のように LaTeX による板下作成に至っては、印刷品位として申し分ない仕上がりになる。まだ余り一般的にはなっていないが TeX でコンパイルした結果得られる dvi ファイルをそのまま印刷所に持ち込んで、高品位のプリンタで出力して板下にすれば、電算写植並みの品位を得ることも可能になっている。実際、LaTeX に関する本などはこの方式で作られるものが多いし、ある学会では投稿論文はすべて LaTeX のソースでなければ受け付けないというところもあると聞く。

いずれにしても、このような校正を完全に済ませた後の板下での入稿であれば、確実に合理化が可能になりコストダウンも見込める。

1.3 原稿整理

長期間にわたって書き溜めてきた論文の一つひとつを、全体のテーマにそくして構成し直して一冊の本として纏める時に問題になるのが、時間的経過による文体の変化や、掲載誌による制限などに起因する表記の揺れである[4]。

たとえば、邪悪な機種依存文字や利用者定義文字[5] や 1byte 片仮名[6] などは `sed` を用いて一括して正規化しておいた[7]。

また、表記の凡例は、しっかりと立てておく必要がある。たとえば、平仮名に開くことにした副詞や形式名詞などの語句を書き出してみると「良い、全く、拠る、所、行く、来る、分かる、何故、窺う、様に、如何、又、為、於て、所謂、〜達、出来る、此、居る、成る、他、無し、程、最も」などである。また、送りがなの基準も統一する必要があるし、漢数字は原則として羅列式（百・十を入れない）などということも決めておく必要がある。さらに、「後印本（後摺本）」のような用語の統一も大切だし、いわゆる書式に関ることで、箇条書最後に「。」を付ける、刊年に「刊」を付けない（文化二年）、強調は◇（「」は雑誌等のタイトルか引用のみ）を始めとして、書誌の記述様式も「『書名』（叢書シリーズ名（巻数）、出版社、刊年）、「‥引用‥」（「論文名」、［発行所、］刊年月）に統一した。

これらは、正規表現に拠って一気に変換するのではなく、原稿のチェックを兼ねて、丁寧に文脈をたどりながらエディタでの検索と置き換えとを繰り返した[8]。その結果、基準そのものを何度も変更することになったが、この作業は原稿自体が機械可読化されていたからできたことで、手書き原稿であったらほとんど気が遠くなるような作業だったと思われる。

1.4 フロッピー入稿

今回は LaTeX で板下を作製したわけではなく、実際の入稿は MS-DOS のプレインテキストファイルで行うことに成っていたため、最終的には LaTeX のコマンドを剥ぎとってルビや角書のタグを入れた文字コードだけのテキストファイルに直す必要があった。

そもそも、いわゆるワープロソフトではなく LaTeX を使った最大のメリットは、ルビや割注および後注の処理にあったのであるが、 `dvi2tty` や `dvistrip` などを使ってもルビがうまく取り出せないという難点がある。また、ソースからコマンドを剥ぎ取る `detex` も macro などの部分でど

[4] もっとも、それ以前の問題として、パソコン導入以前に書いた論文の打ち込みから始めなければならなかったが……。
[5] いわゆる外字のこと。
[6] 清書に LaTeX を使ったため 1byte 片仮名があるとコンパイルできない。
[7] regular.sed というスクリプトが公開されている。
[8] ただし出典書誌事項の年号などは西暦に統一したので、これはテーブルを参照させ、 `AWK` で一括して書き換えた。

うしても手作業が不可欠であるし、大体これでは注が後ろに回らない。仕方ないので、一端 LaTeX のコマンドをテキストフォーマッタ fin 用のコマンドに変換し、さらに注を番号を振って最後に付すための前処理してくれる notes.awk [9] を使用して入稿用のテキストファイルを作成した。

以下、LaTeX のソースと入稿用のテキストファイルに変換したイメージとを例示しておく。

LaTeX のソース

たとえば、当時すでに次のような観点からの区別が存在していた。
\begin{quote}
\RUBY{御伽這子}{おとぎばふこ}の書は漢土の小説を\RUBY{皇國}{みくに}の
事に\RUBY{摸}{うつし}たる\RUBY{鏑矢}{はじめ}にて。文体いにしへにちかく
猶物語の\RUBY{余波}{なごり}あり。\RUBY{繁}{しげ／＼}\RUBY{英}{はなぶさ}
の二書はこれを\RUBY{襲}{つぎ}\RUBY{間}{まゝ}皇國の事を飜案して古に非今
に非。文章の奇絶國字小説第一といはんに論なし。

入稿用テキストファイル

たとえば、当時すでに次のような観点からの区別が存在していた。
★御伽這子▲おとぎばふこ▼の書は漢土の小説を★皇國▲みくに▼の事に★摸▲うつし▼たる★鏑矢▲はじめ▼にて。文体いにしへにちかく猶物語の★余波▲なごり▼あり。★繁▲しげ＋＋▼★英▲はなぶさ▼の二書はこれを★襲▲つぎ▼★間▲まゝ▼皇國の事を飜案して古に非今に非。文章の奇絶國字小説第一といはんに論なし。

この入稿用のファイルに使用したタグは次の五種類である。[10]。

1. ルビは　★漢字▲かんじ▼　　漢字（かんじ）
2. 左ルビ　★稿▲こう▼＄シタカキ＃　　稿（こう）シタカキ
3. 割注は　{右｜左}　　右 左
4. 踊字は　＋＋
5. 濁点踊字は　＃＃

このテキストファイルに付したタグを、印刷所で CTS 用のファンクションに置き換えて板下を作成するのであるが、LaTeX に拠る清書をレイアウト指定紙として用いたので、今回はこれで問題なく校正刷が得られた。

[9] 杉本武氏製のフリーソフトウエア。

[10] 具体的には LaTeX のソースの「¥RUBY {節亭琴驢}{せつていきんろ}」という部分を「★節亭琴驢▲せつていきんろ▼」というように置き換える sed のスクリプトを用意して一括変換した。

1.5 校正あるいは書換え

本の校正としては念校まで取るのが一般的であるようだが、入稿からの時間が経つほど新しい情報が増えるし、また間違いにも気付くことになる。拙著の場合は書誌データや書目などが多いこともあって校正に手間取り、結局入稿から校了までに約一年を費やしてしまった。この間に新しい論文が発表されたり、新たに入手した原本があったり、あるいは引用した論文が増補訂正されて論文集に納められたりと、様々な補訂が生じ、校正と云うよりは書き直しに近い部分も出てきて、部分的には四校まで取る羽目になってしまった。

さらに、一番問題なのは、これらの修正をこまめに入稿テキストに反映させておくことである。が、実際には赤を入れたゲラを見ながら入力していく時間的な余裕が無く、その結果後で索引を作るときに大変な苦労をすることになった。

2 索引の作成

以下は、具体的な索引作りの実際を見て行きたい。これは、試行錯誤の結果として小生が試みた方法であり、もっとスマートで間違いのない方法が別にあるのではないかと思う[11]。

2.1 ページ情報の添加

まず、最終校の校正を見ながらファイルに頁の区切りを入力しておく。具体的には改頁の部分で改行し、行頭にタグを付して「==53」のように頁数を入れておいた。

> おそらく、この説を敷衍した江戸読本の成立を説き続ける限り、京伝馬琴以外の読本作家たちとその作品、および読本の刊行をめぐる板元の演出など、大きな枠組みとしての出板界の様子が覆い隠されてしまう危険がある。とくに京伝読本を評価する場合には、この対立抗争説という文学史の呪縛から自由になった勝ち負けとは別の
> ==53
> 新たな視座が必要になるはずである▼8。
> 三　京伝馬琴不和説の検討
> ところで、藤村作『國文學史總説』▼9を見ると、

2.2 固有名詞の切出し

次に、このファイルから固有名詞以外の部分を削除する。これは本文を読みながらの手作業であるが、エディタのマクロで文字種が変わるところまでジャンプするという擬似的な単語認識をさせると比較的楽に作業が可能となる。

[11] LaTeX の ¥indexx を使うのも一つの方法であるが、折角のテキストファイルの通読性をひどく妨げるし、固有名詞すべてを拾うにはあまり合理的な手段だとは思えない。

```
==422
==423
鈴木重三
==424
鈴木牧之
曲亭馬琴
『西遊記』
『金毘羅舩利生纜』
曲亭馬琴
渓斎英泉
『修紫田舎源氏』
```

書名には『』を、作品名には「」を付ける方針をたてたために、ここで入力しておく。

また、次の作業のために、このファイルに出てくる字体を旧漢字に統一しておいた[12]。

また、ソートするための準備として、それぞれの単語にページ情報のフィールドを与えておいた。

```
111,曲亭馬琴
111,『化競丑満鐘』
111,柳亭種彦
111,『勢田橋龍女本地』
111,『曲亭伝奇花釵児』
111,『女敵討記念文箱』
111,曲亭馬琴
112,『繪本加々見山列女功』
112,川関楼主人
112,山崎平八
112,『敵討連理橘』
113,振鷺亭
113,『水滸傳』
113,『いろは酔故傳』
113,『源氏物語』
113,曲亭馬琴
113,『水滸傳』
113,山東京伝
```

同一ページの同一単語はソートした後に `unique` で処理して重複をなくす[13]。

なおこの書式への変換は、前に付けておいたページのタグがあれば、以下の簡単な AWK のスクリプト[14] で可能である。

[12] これは、基本的には旧漢字の出現頻度が高かったためである。近世の用字法はいい加減であり、原則的に字体には手を加えなかったので表記が混在している。あとで元に戻すためには `diff` でも使って変更箇所を記録しておくほうが良かった。なお、拙編 `qkan.sed` を用いた。

[13] 実際には `sortf` の `-U` オプションで可能である。

[14] もちろん `perl` でも書けるが、今回は新しいツールの修得に逃避している時間的な余裕がなかった……。

```
BEGIN {
    FS = ",";
}
/^[^#]/ {

    if (/^=/) {
        num = (substr($0, 3));
        next;
    }

    print num "," $0;
}
```

2.3 名寄せ –同定作業–

今度は、上の固有名詞に注目してソートして、表記の揺れや、原稿のミスを捜して本文の訂正をする。もっとも、まだ訓みを入力していないので単純にコード順に並ぶだけである。しかし、この作業は一冊の本全体を吟味し直すに当たって、**かなり有効性を持つ手段**であった。

また、複数の呼称を持つ固有名詞については名寄せをする必要があるが、その過程では厄介な考証が必要となることが多く、下手するとそれで一本論文が書けてしまいそうなものもあった[15]。

なお、ここで名寄せをして呼称を統一したものには、空見出しを付けておく。

```
生田芳晴→芳春
一榮斎→芳春
一梅斎→芳春
朝香樓→芳春
242,芳春(芳晴・一梅斎・一榮斎・朝香樓・歌川・生田)
243,芳春(芳晴・一梅斎・一榮斎・朝香樓・歌川・生田)
244,芳春(芳晴・一梅斎・一榮斎・朝香樓・歌川・生田)
256,芳春(芳晴・一梅斎・一榮斎・朝香樓・歌川・生田)
(中略)
281,芳春(芳晴・一梅斎・一榮斎・朝香樓・歌川・生田)
332,芳春(芳晴・一梅斎・一榮斎・朝香樓・歌川・生田)
395,芳春(芳晴・一梅斎・一榮斎・朝香樓・歌川・生田)
435,芳春(芳晴・一梅斎・一榮斎・朝香樓・歌川・生田)
```

2.4 よみの入力

今度は、コード順に並べたファイルに訓みのフィールドを付加して入力していく。これも予想外に大変な作業であった。原本に訓みが記されているものは問題ないのであるが、精確な呼称が不明

[15]具体例を一つだけ挙げれば、「松亭」で名寄せをしたときに種彦の読本『綟手摺昔木偶』の序者「松亭陳人」も拾ってしまったが、これは「松亭金水」ではなく「鳥海松亭」である。人名の別号には実体不明のものがかなりあり、おそらく多くの誤謬を含んでいるものと危惧している。また、初代と二代の区別が難しい笠亭仙果などの例も多く、曖昧な判断を残したままになっている箇所がある。

のものもかなりあり、最終的には音読みでしか入力出来ないものも出てきた[16]。

```
すずきけいいち,46,鈴木圭一
すずきけいいち,210,鈴木圭一
（中略）
すずきじゅうぞう,54,鈴木重三
すずきじゅうぞう,59,鈴木重三
（中略）
すずきとしゆき,153,鈴木俊幸
すずきとしゆき,196,鈴木俊幸
（中略）
すずきぶしゅん,45,鈴木武筍
すずきぶしゅん,78,鈴木武筍
（中略）
すずきぼくし,59,鈴木牧之
すずきぼくし,424,鈴木牧之
（中略）
すずきぼくしぜんしゅう,59,『鈴木牧之全集』
すすきよはち,359,鈴木與八
```

2.5 索引書式

最後に訓みのフィールドに注目して「あいうえお順」にソートして、再び不合理がないかチェックし、良ければ索引の書式に直す[17]。

[16] 人名録などが部首順の索引を備えている理由はここにある。

[17] ここでは、新字体のものを元に戻したり、括弧の付かない見出し語を半角左にずらしたり、三頁以上連続する部分をハイフォンで繋いだり、頁の区切りにコンマを使うために、データのデリミタを変更したりと云う処理が加わるが、今は触れない。

『観音靈應譚』　195
『巖柳嶋』　475
『{かたき|う　ち}岸柳縞手染色揚』　447,453
『函嶺復讐談』　129,168,497
感和亭鬼武（曼亭・蛭牙亭・難淺簾・前野曼七・萬七・満治郎・倉橋羅一郎）　25,68,72,83,
125-130,145,146,153,161,168,172,175,195,203,229,247,340-342,
344,349,367,396,398,404,478-512

■き
『奇異雑談集』　148
『{江戸|時代}戯曲小説通史』　511
菊地茂兵衛（晴雲堂）　48,65,78,85,86,88,91,92,174
菊亭文里　267
『聞説女自来也』　340

これも以下の AWK のスクリプトでできる。

```
BEGIN{
    FS = ",";
    key = " ";
}
{
    if ($2 != key) printf("\n");

    if ($2 == key) {
        printf( ", %s", $1);
        next;
    }
    printf("%s  %s", $2,$1);
    key = $2;
}
```

これで、何とか索引原稿の入稿用ファイルができたわけであるが、角書などには前述したのと同様のタグをつけ、「あいうえお」の見出しも付けておいた。

最後に索引の凡例を示しておこう。

■凡例
1) この索引は、本書中の書名、作者名・画工名・序跋者名・筆耕名・彫工名・書肆名などの固有名詞を抜き出し、五十音順に排列したものである。ただし、地名や機関名、資料の所蔵者名、また作品中に登場する人物名は省いた。
2) 人名と区別するために、書名作品名には『』「」を付した。
3) 一般的な呼称や異称別号などには空見出しを施し、可能な限りの名寄せを試みた。
4) 統一書名には原則として内題 (草双紙は見返題、その他は巻首) を採用した。また、角書を持つ場合は付したが読みには含めなかった。
5) 画工名を除いて、人名には可能な限り氏名を採用した。ただし、戯号堂号屋号などしか分からない場合はこの限りではない。
6) 原則として、表記が相違していても同一のものを指示していると判断できる場合は、表記の相違を適宜吸収した。逆に、まったく同様の表記でも別のものである場合は分けて表示した。
7) 読みが決定できなかった場合は、原則的に音読みによって排列した。
8) この索引の末尾に、切附本の柱刻 (ノド) から「切附本書目年表稿」の整理番号を検索するための索引を付した。

このほかにも細かい問題は多いのであるが、作業のあらましは紹介できたと思われるので、ここに記した以外のことについては御質問頂ければ答えられる範囲でお話できると思います。また、色々と不備な点も多いと思いますので、御教示頂ければ幸甚と存じます。

2.6 後日譚

じつは、三校の後に赤の多い部分だけ四校を取ることにしたのであるが、これが戻ってきてからが大変であった。かなりページが動いたために、索引の総当たり（逆引き）をしなくては成らない羽目に陥ったのである。約三千項目で、のべ一万件を越える固有名詞を点検するのに一ヶ月以上も消耗な作業を強いられてしまった。

索引は最終校を見てから作成すべきである。以下は、泣きながら作った索引書式をページごとの一覧に戻す AWK のスクリプト[18] と、その結果である。

[18]ただし、何故か「竹」など 0x7c を含むコードでこける。当時は直す気力も無かったのでそのままである。

```
BEGIN {    FS = ",";}
{
    if(/→/) next;
    if(/【/) next;

    count1 = index($1, "｜");
    yomi = substr($1, 1, count1-1);
    data0 = substr($1, count1+1);
    count2 = index(data0, "｜");
    record = substr(data0, 1, count2-1);
    $1 = substr(data0, count2+1);

    for(i = 1; i < NF+1; i++){
        if(i > 1){page[i] = substr($i, 2);}
            else {page[i] = substr($1, 1);}
        print page[i] "｜" record "｜" yomi;
    }
}
```

```
# 『江戸読本の研究』索引チェックリスト (頁順)
6｜上田秋成｜うえだあきなり
6｜『雨月物語』｜うげつものがたり
6｜曲亭馬琴 (馬琴・曲亭主人・曲亭蟬史・曲亭陳人・玉亭・著作堂・瀧澤清右衞門)｜きょくていばきん
6｜山東京傳 (醒々斎・洛橋老店主人・山東庵・山東軒)｜さんとうきょうでん
6｜『古今奇談英草紙』｜ここんきだんはなぶさぞうし
6｜『忠臣水滸傳』｜ちゅうしんすいこでん
6｜都賀庭鐘｜つがていしょう
6｜『南總里見八犬傳』｜なんそうさとみはっけんでん

7｜羅貫中｜らかんちゅう
7｜『剪燈新話』｜せんとうしんわ
7｜『席上奇観垣根草』｜せきじょうきかんかきねぐさ
7｜『新斎夜語』｜しんさいやご
7｜『古今奇談繁野話』｜ここんきだんしげしげやわ
7｜『後篇古實今物語』｜こうへんこじついまものがたり
7｜『古今奇談英草紙』｜ここんきだんはなぶさぞうし
7｜『古今奇談莠句册』｜ここんきだんひつじぐさ
7｜『童唄・古實今物語』｜こじついまものがたり
7｜曲亭馬琴 (馬琴・曲亭主人・曲亭蟬史・曲亭陳人・玉亭・著作堂・瀧澤清右衞門)｜きょくていばきん
7｜『怪談前席夜話』｜かいだんぜんせきやわ
7｜『伽婢子』｜おとぎぼうこ
```

3 蛇足

3.1 機械可読テキスト

「本」を底本としている場合は、基本的に位置情報を如何に持たせるかと云う問題がある。しかし、用例の検索にしても文脈の中で見たいわけで、cgrep のようにターゲットの前後を任意に表示してくれるものが必要となる[19]。

[19]いわゆる kwic 索引なども斯様な要求から生まれたものであろう。擬似的な kwic 風の出力をする grep もある。

また、所与の機械可読テキスト全体の凡例を熟知していないと、シソーラスを思い描くことすらできない。これは検索時には決定的な問題となるはずである。つまり、データの作成者が一番上手く検索利用ができるのは当然のことである。したがって、凡例をどれだけきちんと書けるかという点が重要になるだろう。

また、その折りにデータに加えた改変、例えば新字体に統一したとか、利用者定義文字をどう処理したかなどと云う点も不可欠な情報になるだろう。

これは当然のことであるが、物理的大きさでは機械可読テキストに優るものはない。ノートパソコンの内蔵ハードディスクが大きくなった現在、様々な機械可読テキストを持ち歩くことが可能になった[20]。

機械可読テキストの最大の弱点は斜め読みができないことである。パラパラと捲って全体を知ることは、作業能率上不可欠な準備であるが、これができない。

あとはテキストの精度、文字種の限定などの問題もあるが、この辺は割り切ってしまっても余りある利用価値が存すると思われる。

3.2 本というモノ

通読するテキストとしての本というモノは、やはり大切である。装幀をも含めて「読む」対象としてのテキストなのであるから。

一方、検索するツールとしての本のメリットは、視認性の良さと寄り道の可能性に在ると思われる。頁全体を一覧しつつ、こちらの問題意識に応じて自然と目に入ってくることは誰しも経験することだろう。これは、能動的に検索しなければならない機械可読テキストとは本質的に相違する点であろう。

もっとも、そのツールの編集方針にもよるわけで、恣意的判断を停止して、悪く云えば何も判断しない索引など、逆に使いようが無いかもしれない。

あとは、置き場所と云う物理的な問題と、腱鞘炎になった時に重い大きな本を使うのが辛いと云う欠点もある[21]。

[20]普段使っている ThinkPad220 は内蔵ハードディスクを 320M に換装して、国文学研究資料館のデータベースをはじめとして、EB 版の『広辞苑』など種々のデータを詰め込んであり、これさえあれば授業の予習から、簡単な原稿書きまで出来てしまう。

[21]今春、突然右手首の腱鞘炎に悩まされた時の実感。『国書総目録』や『大漢和』などは使うのが億劫になってしまったが、機械可読テキスト化されていないために、使わないわけにはいかず、結果的に完治が遅くなってしまった。

4 目次

目 次