国立国語研究所学術情報リポジトリ

Taiyo Corpus : Language database of the journal Taiyo published from 1895 to 1928, Research on the formative era of contemporary Japanese based on the Taiyo Corpus

メタデータ	言語: jpn 出版者: 公開日: 2019-03-25 キーワード (Ja): キーワード (En): 作成者: 岡島, 昭浩, OKAJIMA, Akihiro メールアドレス: 所属:
URL	https://doi.org/10.15084/00002165

国立国語研究所資料集15
太陽コーパス　雑誌『太陽』日本語データベース

国立国語研究所編　2005年3月31日　博文館新社

国立国語研究所報告122
雑誌『太陽』による確立期現代語の研究
—『太陽コーパス』研究論文集—

国立国語研究所編　2005年3月31日　博文館新社

岡島　昭浩
(大阪大学)

1. はじめに

以前から公開が期待されていた「太陽コーパス」が，CD-ROMの形で公刊され，また，それを利用した研究（利用するためのモデルとなる研究）などを採録した報告も刊行された（以下，報告『雑誌「太陽」による確立期現代語の研究』を『研究』と呼ぶ)。

「太陽」のCD-ROMといえば，八木書店の出したものがあるが（1999年)，それは画像であり，今回の「太陽コーパス」は電子テキストである。また，『太陽』全点ではなく，創刊の1895年と，1901年，1909年，1917年，1925年の5年分（増刊号は除く毎年12号)，60冊を抜き出したものである。『太陽』は1928年の休刊まで531冊出ており，その中の60冊ということである。

この「太陽コーパス」は，国立国語研究所の「日本大語誌」構想の流れを継ぐものだが，用例採集からコーパス構築へと発展したものである。明治以降の現代に至る日本語の姿を跡づけようというものとして，国定読本の次に，総合雑誌『太陽』が選ばれたものである。

明治から継続して刊行されているものとしては，新聞や他の雑誌類もあるが，『太陽』には次のような評価がある。

> 硬い漢文くずし訓読体の演説調から，その軟かい調子のもの，また「だ，である」体，「です，ます」体，あるいは文芸欄に限られるが「和文体」まで，さながら「文体」のデパートの観もあった。欄の種類によって一定の規制は働くものの，これは文体の選択を基本的に執筆者の自由にまかせていたからのことであり，これもまた，当時の各種新聞にも，民友

社の機関誌である『国民の友』などにも見られない『太陽』の大きな特徴である。
鈴木貞美「総合雑誌『太陽』と博文館」(『彷書月刊』1999年11月号)

このような点に着目して『太陽』が選ばれたわけであるが,『国定読本用語総覧』のように,プレーンテキストと語彙表のデータを配布するのみには留まっていない。未整備といわれる日本語コーパスの先駆けとして,「構造化テキストタグ付きコーパス」の雛形となるものを目指しているという。

2. CD-ROM

「太陽コーパス」CD-ROMをドライブに載せると自動的にInternet Explorerなどのブラウザが立ち上がるが,そこで「CD-ROMの中身を見る」をクリックするか,ファイルエクスプローラーなどでCD-ROMを見ると,index.htmlやlicense.txtというファイルの他に,Himawariというフォルダと,XMLというフォルダが見える。

Himawariには,全文検索システム「ひまわり」に関する諸ファイルが入っており,XMLフォルダに「太陽コーパス」の本体が収められている。拡張子xmlを持つ60のファイルがそれで,1冊につき1ファイルとなっている。ファイルサイズの合計は約60MBである。総文字数は約1450万字というから,本文の文字だけで30MB弱ほどあり,残りがタグの部分ということになる。

『国定読本用語総覧』(三省堂1997年)がプレーンテキストで3MB弱であったのと比べると,「太陽コーパス」の規模の大きさがうかがえる。

ちなみに,インターネット上の電子テキスト図書館として知られる青空文庫(http://www.aozora.gr.jp/)は,野口英司編著『インターネット図書館 青空文庫』(はる書房2005年)刊行の時点で,これもプレーンテキストでの容量であるが160MB程度(若干重複があるので実際にはもう少し少なくなる),『新潮文庫の100冊 CD-ROM版』(新潮社1995年)から翻訳物を除いたサイズが26MB程度,『新潮文庫の絶版100冊 CD-ROM版』(新潮社2000年)の,同じく翻訳物を除いたサイズが,40MB程度である[1]。

国文学研究資料館がインターネット上で公開している「日本古典文学大系本文データベース」[2]は,岩波の旧大系に東京堂の噺本大系が加わり,現在,909作品で,「文字数」が約3650万字[3]あるから,「太陽コーパス」と比較すると,2.5倍ほどの量ということになる。

「太陽コーパス」の記事の数は約3400で[4],著者は1000人程度である(八木書店によれば総執筆者は6500人)。無署名記事が1052件あって,これは博文館編集部内の人物によるものであろうし,「ＸＹＺ生」「△生」といった変名の署名には同一人物の異名もあろうが,それを差し引いても多数の人の文が収められていることがわかる。コーパス化の対象となるに際して見込まれていた,多様な著者によって書かれているということを,数字の上で見せてくれるものである。たとえば,青空文庫には5000ほどの作品があるが,その著者の数はわずかに300人程度であり,これと比べて『太陽』の多彩さがはっきりする[5]。また,『新潮文庫の100冊』等でも,著者のバラエティという点では物足りないものである。

なお，多彩といっても，「太陽コーパス」の場合は，同時代で同媒体という場を踏まえた多彩性であり，青空文庫のように土佐日記から昭和20年代までの作品があるという多彩性とは異なる。

『研究』の巻末には，「『太陽コーパス』に含めなかった記事の一覧」があり，310件の記事が載せられている。「著者の著作権が未処理」（著作権が残っていて，継承者からの掲載の同意が得られなかったもの）である。

独立行政法人国立国語研究所が文化庁から出た独立行政法人であることもあり，著作権処理は慎重である。民間の出版社に往々にして見られるような，外側からはわかりにくい著作権処理ではなく[6]，それを明示して，著作権者からの掲載許諾の得られなかったものについては，連絡の付かなかった人も含めて掲載しないという方針で臨まれたのは，慎重で公正な判断であったろうと思う。

本書発行時点で著作権の保護期間が終わっていたのは，1954年以前に亡くなった人だが，本書に収録されたうち，著作権が残っていた方が88名ある。著作権の保護期間が残っていた1955年以降に亡くなった方々であるが，この88名は，国立国語研究所の掲載の許可を求める依頼に対し，諾とされた方々である。諾とされたおかげで162件の記事が採録されたことになり，ありがたいことである。著作権継承者の方に感謝すると同時に，少しでも多くの作品を採録しようと著作権処理にあたられた国立国語研究所にも感謝するものである。

ただ，残念なのは，事前（2004年頃）にネット上で行われた著作権者に関する情報探しのことである。人名のリストを掲げるだけで，国立国語研究所側で判明していない情報が，著作権継承者だけであるのか，没年についても不明であるのかが分からなかった点である。2003年に行われた国立国会図書館の近代デジタルライブラリーにおける著作権情報調べの際には，そのあたりのことも分かるようにしてあったので，それを参考にした調査にして欲しかったものである。

たとえば，

1925年４号　「無線電話の真空管に就て」佐野昌一

の佐野昌一は逓信省に勤めていて後の筆名が海野十三である人物なら1949年没であり，著作権保護期間は終了している。同名異人の可能性もあるが，こうした情報が寄せられなかったとすれば，調査が情報を集めにくかったものであった，ということになるであろう。

また，

1925年３号　「『理研酒』から『理研水』へ　山本敬三氏の理研酒の研究をよむ」根本正

の根本正は，コーパスに収められている1895年12号の「帝王国の政変」の著者（1933年没）とは別人なのであろうか。著者情報にも「禁酒運動家」とあり，題名からすると同一人物でありそうに思えるのだが，どうなのであろう。佐野昌一にせよ，根本正にせよ，既知の人物と違うことが明らかなのであれば，注記をしておいて欲しかったところである。

さて，「太陽コーパス」に収められた記事の著者にどのような人々が居るかというのは，『研究』p77にみえる「ひまわり」の著者一覧機能で見ることが出来るが，

¥Himawari¥Corpora¥Zassi¥Taiyo¥authors.xml

というファイルが，その本体である。上記の通り，1000人ほどの人の名がある。所属・分野・生年・没年の記載があるが，第二欄の「所属」については，『太陽』刊行当時のものもあれば，後のものもある（『研究』p22)。両者は出来れば，区別しておいて欲しかった。また，「著者一覧機能」が，「ひまわり」では，「ツール」→「一覧」から見られることを明記して欲しかった。

3. ひまわり

ここで，Himawari フォルダの中を見ておくと[7]，himawari.exe が，全文検索システム「ひまわり」の本体であり，他に関連ファイルがあるが，Corpora フォルダと，j2re1.4.2_06 フォルダ，Manual フォルダがある。

j2re1.4.2_06 フォルダは，「ひまわりを動かしたり結果を表示するのに必要な JavaScript を動かすためのファイル群が入っている。

Corpora フォルダに入っているのは，「ひまわり」による検索のためのインデックスファイルである。その中に Zassi フォルダがあり，さらにその中に Taiyo フォルダがあり，そこに「太陽コーパス」検索のためのインデックスファイル群がある。ここには全文が 1 ファイルとなっている corpus.xml（85MB 程度）も収められている。

xml ファイルを，InternetExplorer 等で表記させるための設定ファイルが，Taiyo フォルダの下の xslt フォルダに置かれている（検索の際の結果表示のための一時ファイルもここに作られる)。更にその下の Gaiji フォルダに，JIS の第一・第二水準以外の字を表示するための画像ファイルが収められている。その下の Large フォルダには，外字の拡大表記のための画像ファイルが収められている。「今昔文字鏡」にある文字はそれを使い（数字 6 桁の gif ファイル)，ないものは原文の画像を切り出して用いている（T で始まり数字 3 桁の続く png ファイル)。同内容の Gaiji フォルダは，XML フォルダの下にもある。一つ一つはそれほど大きなファイルではないが，1500以上の文字の大小ふたとおりが二箇所にあるのは，もったいない気がする。

「ひまわり」は〈太陽を追うもの〉を意味するのであろうが，「太陽コーパス」にしか使えないわけではない。「太陽コーパス」と同様の構造を持つテキストであれば，「ひまわり」自身によってインデックスを作って検索することが出来る。

検索すべきファイルを追加した場合には，Corpora フォルダの下に複数のフォルダが入ることになる。その方法については，国立国語研究所のサイト「言語データベースとソフトウェア」(http://www.kokken.go.jp/lrc/）参照のこと。テキストファイル群を「ひまわり」で検索可能な形式にするツール「えだまめ」や，青空文庫の xhtml ファイルを検索可能な形式にする「あおまめ」等も配布されている。

インデックスファイルを作成しておいて素早く検索してくれるソフトは，たとえば，namazu などが知られ，また最近では，Google デスクトップサーチなどもあるが，namazu は初心者には扱いにくく，Google デスクトップサーチはブラックボックス的で心配を感じる人もいるだろう。この「ひまわり」のインデックスファイル作りを試してみるのもよいだろう。

なお，「ひまわり」は，「太陽コーパス」に同梱のものは再配布できないが，国立国語研究所の

サイトには，一定の条件を満たせば再配布可能なものがある。また，「太陽コーパス」用のひまわりをバージョンアップしたものもある。

4. タグ

前述のように，「太陽コーパス」は，太陽をテキスト化し，タグを付けたものである。単語や形態素のレベルまで構造化した「形態素解析タグ付きコーパス」までは行かず，「文章や文のレベルまで構造化し，ジャンル，文体，著者，引用箇所，話者などの情報をタグ付けしたもの」という「構造化テキストタグ付きコーパス」ということであるが，どの程度のタグが付されているかは，『研究』p16以下に示されている。

記事について，題名・著者等の他に，「文体」「ジャンル」も示される。文体は，口語・文語の二種のみだが，引用文についても文体を示し，口語文中に文語の引用のある場合等にも対応している。

「ジャンル」は，図書館で使われる分類のNDCによって示されているが，たとえば，NDC800の言語にどのような記事があるのかを見ようという検索のツールは準備されていない。

〈注〉タグは，誤りを訂正したり，揺れを統一しようとしたものであるが，下記の通り，大量に付与されている。

A 誤字通用	9587
B 衍字	592
C 脱字	226
D 転倒	44
E 欠損	197
F 濁点脱落	12185
G 仮名遣	28177
H 正誤表	33

このうち，「仮字遣」は，「仮名遣いが歴史的仮名遣いの規範と異なる使われ方をしているもの」ということであるが，『研究』の小木曽智信「仮名遣いについて」にも，「「誤用」と言うよりは『太陽コーパス』の仮名遣い規準との「不一致」と表現する方が正確」とされるとおり，さまざまなものが含まれる。

「酔・睡・衰・推・水・追」等の「すゐ」「つゐ」を「すい」と訂正したものなどは，『研究』p376でも指摘されるとおり，当時は，「すゐ」が規範的なものであった。「太陽コーパス」において「水すゐ」を「すい」と訂したものが230例あるが，それに対して原文で「水すい」となっているものが9例であり，当時の規範が比較的よく守られていることが分かる。他の字音仮字遣についても，「衆・従・中・忠」等の「しう・ちう・じう」表記を「しゆう・じゆう・ちゆう」と訂しているものなども，何を規範としているのかについて見極めねばならないところである。

〈外字〉タグは，JISの第一水準・第二水準にないものを示し，「今昔文字鏡」にあるものはその番号を付し，ないものは独自の番号を付してある。表示の際に，「今昔文字鏡」のフォントを使うようになっており（必要な分がCD-ROMに収められている），「今昔文字鏡」にないものは，原本の画像を使っている。

〈小書〉タグでは，カタカナの小書きが示される。「キコネハフヘホムルワヰヱヲ」にあるとのことだが，187例中，「ヰ」が105と飛び抜けて多い。カタカナ以外で，「王覇」に「忘八」を〈小書〉でタグ付けしてあるものが一箇所あり[8]，本来はカタカナと思われる伏字部分に付けられているものもある（一箇所五文字）。

なお，JIS規格で小書きのある「ァィゥェォヤユヨワッ」は，〈小書〉タグを使わずに入力されており，これらは2145例ある。また，厳密にはカタカナではないが，「ヶ」が1713例ある[9]。また，平仮名でも，「っゃ」については小書がある。

他に，533個の〈敬意欠字〉タグもある。

なお，〈注〉〈外字〉〈引用〉タグの抽出は，後述のプリズムの「注一覧」で行うことが出来るが，他のタグについては検索を行うツールが用意されていないようである。プリズムのスタイルシートを自作することによっても可能だが，評者はgrep類のツールと，エディターによる検索を用いた。

5. プリズム

XMLフォルダの中に，冊ごとの「太陽コーパス」本体ファイルと同じ場所にある「プリズム」Prism.htaは，「太陽コーパス」形式のxmlファイルから様々なスタイルの文書へ変換することが出来るツールである。用意されている変換を行う場合には，操作は至って簡単である。

大きな画像入りの外字の一覧を出力するように変換するものなど，これを出来上がった形で公開するのではなく，ユーザーの手で変換させて生成するようにしたのは，ユーザーを導くものとして有意義なことであると思う。記事の一覧も，プリズムによって生成されるものであり，これを既製のファイルとして提供しなかったのも面白い試みである。

ルビもないプレーンテキストが生成されるものがある。この形式を求める人も多いであろう。評者もこれで生成したファイルを，青空文庫などのファイルとともに，grep検索用に置いている。およそ30MB弱になる。

また，用意されてはいないが，ルビ付きテキストを出力するためのスタイルシートの作り方が『研究』p100に書いてあり，これを参考にして青空文庫形式のルビに変更することも可能である。青空文庫形式のルビは，それを読みやすく表示してくれるビューワーも多数あり[10]，そうしたビューワーで読むことが可能になる。シンプルなhtmlファイルへの変換も，ルビタグを残したものにして，ブラウザで読んだり，そこからさらに，他のデバイス用に変換して読むことも可能である。

しかし，これらのファイルは，あくまでもコーパスから機械的に作られたものであり，読むのには適した形にはなっていない。まず，これらの形式では，漢文・欧文・図表などが省略された

部分であることを示すタグまでもが削除されてしまうからである。もう少し読むことを目的としたファイルを作りたい人は，これらのタグを削除しないスタイルシートを作る必要がある。とはいえ，元のxmlファイルでも，「非入力対象」があることが分かるだけで，その中身を見ることは出来ないのである。779の「《図表》」，16の「欧文」，7つの「漢文」，1つの「ハングル文」が，「非入力対象」とされ，実際の姿は参照することが出来ず，そこを読むには，元の雑誌か複製かにあたらねばならない。

コーパスは，内容を見ずに数量的に処理するものだからこれらは参照出来なくてもよい，という考え方もあろうが，孤例のようなものが出た際に，元の姿に戻って確かめる必要が生じることがある。その意味からも，これらについては画像ファイルによって提供があってもよかったのではないかと惜しまれる。欲をいえば，収録された号については，『研究』p13にいう「対象外」の部分も含めて，著作権処理ができないもの以外は，画像が参照できると有り難かった。CD-ROMの容量を考えると，全冊の画像を収録するのは無理なようだが，一年に一号ずつでよいから，画像が付されていれば，と思う。雑誌『太陽』の姿を示すことで，雑誌『太陽』と「太陽コーパス」の間，「太陽コーパス」が雑誌『太陽』からどのような情報を電子化したものであったのかが，より分かったものと思われる。しかし，それは欲として，せめて図表などテキストで省略したものだけについてでもよいから，入れておいて欲しかったというのが，現実的な希望である。

さて，プリズムで用意されている変換にはTeX形式に変換するものもある。これをpLaTeXを使って印刷することが出来る。「詳しくは生成されるtexファイルの冒頭」を見よとあるが，XMLフォルダにあるtx2latex.xslをエディタなどで開くと分かる。このTeXファイルをコンパイルして印刷可能なdviファイルにするのに必要なのは，

藤田眞作氏作成のkyakuchu.sty（ルビ／割書／脚注用）。

堀田耕作氏作成のmjfonts.sty（今昔文字鏡フォント用）。

である。どちらもネット上に公開されているものであり，入手自体は容易なのだが，今昔文字鏡を導入し，これらのスタイルファイルを使って生成されるdviファイルが，どのようなものであるのか，見本のdviファイルを添付して欲しかった。解説書p41，『研究』p88-89に画像があるが，小さくて分かりづらい。「今昔文字鏡」を入れることに抵抗を感じている人もおり，見本dviを示すことで，どの程度のものが出来るのかを見せて欲しかったのである。

dviを変換したPDF形式などもあるとなおよい。本CD-ROMの刊行にあたり，今昔文字鏡の使用許可を得ているので，フォントを埋め込んだPDFファイルの添付も可能だったであろう。

なお，このプリズムによる変換の出力先は，ブラウザかファイルかであるが，ファイルに出力する場合，ファイルをそのまま変換するか，いくつかのファイルを一つにまとめるしかできないのだが，これを記事ごとに切り分けたファイルにすることも示していただければ有り難かった。読むためのファイルにせよ，印刷するためのTeXファイルにせよ，記事ごとに切り分けられていた方が，扱いやすいように思うからである。

6. たんぽぽ

「たんぽぽ」は，「ひまわり」をハードディスクにインストール出来ないような人のためのツールで，CD-ROMからも実行出来るものである。「ひまわり」のようにインデックスファイルを作って検索するのではないので，残念ながらその検索はかなり遅い。振り仮名部分の検索は「ひまわり」には出来ないが，「たんぽぽ」では〈フリガナを本文と見なした〉検索も可能であるし[11]，結果の表示もコンパクトで見やすいので，使わないのはもったいない気もするが，検索の遅さで，使い勝手のよいものではない。

振り仮名部分を検索したいという要求に対しては，振り仮名を本文に開いて，漢字を振り漢字のようにしたファイルを作っておいて[12]，それを「ひまわり」に検索させるような手段を示してもよかったのではなかろうか。

また，ハードディスクにインストール出来ない人のことを考えるのであれば，CD-ROMに，プリズムによって生成されるプレーンテキストと，振り仮名を開いたテキストとを収録しておく，という方法もあったろう。ルビタグが付されたのは小説の類だけなので，それらだけについてルビを本文に開いて収録しておけばよいと思うのである。

7. おわりに

さて，この「太陽コーパス」は，国立国語研究所において，現在，「明治から現代にいたる日本語の全貌を把握するための言語コーパスKOTONOHA」の中に位置づけられ，多くのコーパスが提供されようとしている，という。たとえば，「太陽コーパス」に比べると小規模のものであるが，「近代女性雑誌コーパス」の提供が始まった。

国立国語研究所の外でも，古代語の資料を，「太陽コーパス」のタグ付けを拡張しながら，コーパス化しようとしているものの存在も聞いている。

今後，さまざまな日本語コーパスが作られてゆくことになるだろうが，死蔵されることなく，多くの人が使えるものになって欲しいものである。今回の「太陽コーパス」のように，比較的安価で提供されて欲しいと思う。

以上，書評というよりも，追加説明のようになってしまった感がある。また，勝手な要望ばかりを書き並べた感もある。評者は，タグの有効性を知りつつも，これまでプレーンテキストを検索することを中心に行ってきたので，的はずれなことを言っているおそれもある。

『研究』の個々の論文には触れることが出来なかった。「太陽コーパス」そのものを見てゆくことを中心としたかったためであるが，『研究』は論文集の側面と，CD-ROMの解説書の側面とを持っている。「太陽コーパス」を使おうとする人にとっては，CD-ROMに付された解説書だけではなく，『研究』も参照せねばならないものであることを言い添えて，書評の終わりとする。

注

1　新潮文庫のCD-ROMは，エキスパンドブック形式のものであるが，テキスト相当の容量を計

算したものである。

2 国文学研究資料館 http://www.nijl.ac.jp/ の「電子資料館」にある。利用登録が必要。

3 「一覧表示」で得られる各作品の「文字数」を合計した。

4 『太陽』を全点収録した八木書店のCD-ROM付載のデータによれば，総記事数は四万点以上あるが，八木書店のデータは，複数著者の記事の場合に，記事数が著者数だけあることや，表紙なども記事数に含まれることなどから，実際の総記事数はそれよりも少ない。しかし，それでも，実際に『太陽』に載せられていた記事は，「太陽コーパス」の十倍以上の量であろうことは推測できる。

5 青空文庫はコーパスとして設計されたものでなく，ボランティアの人が入力したいと希望したものを入れてゆくものであるから，そのような傾きが出るのは仕方がないことである。

6 「一部，著作権継承者の分からない人があるのでご存じの方は連絡を」とだけ書き，その「一部」に誰が該当するのかを示さないなど。

7 CD-ROMに付されている『解説書』にも『研究』にも，CD-ROM内のファイルリストのようなものが存在せず，そのことで「太陽コーパス」が，「ひまわり」等のツールを介したブラックボックス的なものになってしまっているかに見える人が多いであろうことは残念なことである。

8 1901年14号。「王忘覇八」となっており，「王覇」「忘八」では検索できない。

9 「ゃ」は，1907年11号，中村星湖「歌さんの幻影」にみえる32例のみ。すべて「あびゃびゃ」のような形である。

10 http://www.sky.sannet.ne.jp/at-sushi/aozora/viewer.html などを参照。

11 「〈r rt=“なごり”〉名残〈/r〉〈r rt=“な”〉無〈/r〉く」という形式から「なごりなく」を検索できる。

12 「〈r rt=“名残”〉なごり〈/r〉〈r rt=“無”〉な〈/r〉く」というような形式。

岡島　昭浩（おかじま　あきひろ）
大阪大学大学院文学研究科
560-8532　大阪府豊中市待兼山町1-5
okajima@let.osaka-u.ac.jp