総合雑誌『太陽』の本文の様態と電子化テキスト

著者	田中 牧郎, 小木曽 智信

雑誌名	日本語科学
巻	8
ページ	141-152
発行年	2000-10
URL	http://doi.org/10.15084/00002048

『日本語科学』8（2000年10月）141-152　〔研究ノート〕

総合雑誌『太陽』の本文の様態と電子化テキスト

田中　牧郎
（国立国語研究所）
小木曽智信
（東京大学大学院）

キーワード

「太陽コーパス」，明治大正期文献，電子化テキスト，ＸＭＬ形式

要　旨

国立国語研究所では総合雑誌『太陽』（1895-1928）のコーパス（「太陽コーパス」）作成を進めている。その一部をなす『太陽』1901年12冊分の本文の電子化テキストを試験的に公開し，批判を得たい。『太陽』の本文は，現代の総合雑誌とは異なる様態を種々示しており，それらは，電子化に際して相応の問題を生じる。そうした雑誌本文の様態を整理し，その様態に応じて必要になる仕様を策定し，その仕様にしたがって電子化テキストを作成する。基本となるデータには，ＸＭＬ形式を採用し，この形式を通して，データの保守，変換，検索等も行う。

1．「太陽コーパス」の作成

総合雑誌『太陽』（1895-1928,博文館刊）は，19世紀末から20世紀初めの日本語の実態をよくうつしだした資料として，特に記事ジャンルや読者層が広い点で，一定の価値をもつものである[1]。国立国語研究所では，大規模コーパス構築の一部をなすものとして，『太陽』のコーパス作成をすすめている。国語研が作成する「太陽コーパス」は，1895，1901，1909，1917，1925の各年，計60冊に，終刊年の1928年（２号で終刊）の２冊を加えた62冊の本文のうち，原著者の著作権上の問題がある記事を除いた全文を対象とし（英文・広告等は除く），本文テキストと単語タグから構成される予定で，2002年度の完成・公開を目指している[2]。

コーパスを構築するには，作成段階で研究を要することがらが色々あるが，この研究ノートでは，とくに，明治大正期の文献の本文を電子化する際の方法論的な問題を整理し，その問題に対応できる仕様を策定し，その仕様に基づいて電子化した結果の一部を試験公開し，批判を得たい。得られた批判を参考にしつつ，完成版の公開時の最終的な仕様を固めていきたい。今回の具体的な検討材料には『太陽』1901年12冊分の本文テキストを用い，その試用版として「太陽コーパス Ver 0.3（1901本文テキスト）」を作成した[3]。

なお，「太陽コーパス」の単語タグには，スカウト式によって採集された語について，読み・品詞・語種等のインデックスを付す予定であるが，その仕様策定の問題については別に整理・検討する必要がある[4]。今回は，単語タグの問題は扱わず，「太陽コーパス Ver 0.3」にもそのデータはおさめない。

2.『太陽』本文の様態

2.1. 特徴

『太陽』は，刊行間隔や，判型，総ページ数，組版等を何度か変更しているが，1901年については，月刊（第6号と第11号は臨時増刊。これはコーパスの対象から除外），四六倍判，各冊本文230ページ余，縦書き。5号活字2段組みを原則として一部に6号活字や3段組みをとることがある。

1901年本文の様態で，現代の総合雑誌とは異なる特徴として，次の諸点をあげることができる。

・句読点の付け方は記事によってまちまちで，句点を用いない記事も多い。
・振り仮名を豊富にもつ。記事によって総ルビ，ぱらルビ，無ルビの，さまざまな様相を示す。
・誤植あるいは誤用ではないかと疑われる箇所が目立つ。
・誤植・誤用ではないが規範からは逸脱しているのではないかと考えられる表現が少なくない。
・濁点が脱落しているのではないかと思われる場合が多い。
・仮名遣いは，歴史的仮名遣いによることが一般的だが，徹底しているとは言い難い。
・字体の変異が大きい。

本文の様態のこうした特徴は，電子化テキストの作成に際してそれぞれに問題を生じてくる。

2.2. 句読法

次は，1901年1号の巻頭記事「明治三十四年」の冒頭を，原文のままの行取りで掲げたものである（一部中略）。

人生には限りありて冀望には際涯なし、有限の生を以て、無
限の望を懐く、其發して理想となる者、千萬年後の光景を豫
（中略）
可からず、此信仰を懐きて社會を觀る者、誰か無限の快感を
生ぜざらんや。
明治の世となつて以來、三十四回の年を迎へり、個人より之

文の末尾は「、」で示され，読点と区別がない。そして，段落の末尾が「。」で示され，段落の最初は一字下げを行わない。こうした句読法については，土屋(1966)が指摘する通り，他に，「。」を一切用いないものや，「、」を文末にのみ用いるもの，読点にあたるところにも「。」を用いるものなど，さまざまなタイプがあり，一定していない。同一の記事のなかでも句読法の基準はゆれることがあり，どうやら，編集に際して統一をはかることはなかったと見え，著者の表記法をそのまま掲載しているものと思われる。

2.3. 振り仮名

総ルビ，ぱらルビ，無ルビの，三種類の文章がある。

牛込(うしごめ)の辨天町(べんてんちやう)から早稲田町(わせだまち)へだら～～と二間(けん)ばかりの阪(さか)を下(お)り（01081A05）
吾人が農界(のうかい)に瞥(ほつ)めんと欲する所のもの一にして足らず、（01161A04）
昨年六月以來の清國事變に因り露國は其の準備の未だ整はざる前に（01003A05）

どの形式をとるかは，記事のジャンルによって決められている模様であり，句読法の場合と異なり，一定の編集方針に基づいていたものと考えられる。すなわち，小説は総ルビ，歴史地理・商業・工業・農業・家庭・科学・社会・海外等はばらルビ，論説・評論・彙報等は無ルビである。

2.4. 誤植

誤植と判断される箇所は，現代の総合雑誌に比べてかなり多い。

①誤字　秒糖税可ならん、酒税可ならん、絹布税可ならん。〔「砂」の誤字〕(01008A23)

②衍字　努めたるたるを以て　〔「たる」衍字〕(01065A20)

③脱字　太え事を抜すない。　〔「か」の脱字〕(01102A15)

④転倒　喰つてかりゝしは　〔「ゝり」の転倒〕(01036A10)

⑤欠損　廣■豫備病院の近状　〔「廣」の後一字分欠損〕(01228A06)

上記のような，誤字・衍字・脱字・転倒・欠損が，たとえば1901年１号だけで78箇所もある。

2.5. 通用

規範から外れる形や用法と疑われながら，誤用と見ることにも抵抗がある場合がある。

①　語形

a 成る様にしきやならなからうからね〔しか〕(01096A08)／涙を浮めた〔浮べ〕(01084A09)／缺ぐべからざるものなるべし〔缺く〕(05038B02)

b 竹を骨とし籐を皮として〔とう〕(01114B18)／喜みて〔喜び〕(01151A09)／丑満頃を〔うしみつごろ〕(13107A09)

②　語法

a 禁じ能はざりき〔禁ずる〕(07041A23)／小さながらも〔小さい〕(09078B13)／顫上るばかり吃驚したのは〔ふるへあがる〕(01089B05)

b 心持を察しるとね〔察する〕(01105B05)／一班を伺ふに足るべし〔伺ふ〕(01123B12)／取引しられて居るのに〔取引せられて〕(13077B02)

③　漢字表記

a 人々の利害干係を〔関係〕(09026A03)／陸兵と更代せしむるに〔交代〕(12227B12)／直段高く〔値段〕(07151B02)

b 毒焔を延上せしめたり〔炎上〕(08030A25)／遊客の氣嫌を取るを〔機嫌〕(08038A19)／拂戻すべき體度を示したれば〔態度〕(05060B18)

上掲いずれも，〔　〕中に記した規範的な形式に対して，異例と考えられるものである。しかし，2.4に掲げた誤植の類とは一線を引きうるものである。このうち，aとしたものは，『日本国語大辞典』（小学館）や『新潮現代国語辞典第二版』（新潮社）等，明治期の言語の実態を記述している辞書類に，何らかの形で記されているもので，通用として認められる。ところが，bは，そうした記述を簡単に見出すことはできず，当時の通用と認められることの証拠が簡単には得られない。

2.6. 清濁

例えば，次は，「ど」「だ」「ぎ」の濁点が付されなかった例であると考えられる。

殆と (02182A15)／甚た (05192B11)／黴菌の傳播を防き (0107224)

下に示すように，３語いずれも，濁音表示が一般的であるが，わずかながら清音表示の例も見られる。この数値は，右欄の「検索文字列」を，「太陽コーパスVer0.3」の全号で検索し，「注」により原文の形に復した場合の結果を数えたものである（したがって，他の表記法による当該語の例は数えていない)。清音表示の用例が，特定の号や記事に偏ることはなく，清音表示に有意な分布は指摘できない。濁音形に濁点を付けることが徹底していなかったことの反映と見るべきものであろう。

	濁音表示	清音表示	計	検索文字列
殆ど	675 (97.3%)	19 (2.7%)	694	殆ど／殆んど／幾ど／幾んど
甚だ	759 (99.1%)	7 (0.9%)	766	甚だ／太だ
防ぐ	93 (93.9%)	6 (6.1%)	99	防が／防ぎ／防ぐ／防げ／防ご
免がる	13 (25.0%)	39 (75.0%)	52	免がる

ただし，「免がる」のように，清音表示の例が多い場合もあり，これは，清濁の間で語形がゆれていた語であると考えられるものである。

2.7. 仮名遣い

次にあげるものは，歴史的仮名遣いに合致しない例である。

或(ある)ひは〔或いは〕(01102B05)／手中(しゅちう)の〔中(ちゆう)〕(01102A21)／堪ゆる〔堪ふる〕(01014B19)／堪えざらしむ〔堪へ〕(01023B03)／教ゆるは〔教ふる〕(02014A20)／教えない〔教へ〕(03087B19)

歴史的仮名遣いの普及は，明治後期から大正期へと進展していくと言われるが(築島(1986))，1901年の『太陽』において，仮名遣いは，語や語形により複雑な様相を示す。上記の例について，2.6で行ったのと同様の検索を行い，原文の形に復元して数えると，次のようになる。

	歴史的仮名遣いに合致する			合致しない			計	検索文字列
或いは	或いは	0	0.0%	或ひは	101	100.0%	101	或いは，或(あるい)は
中(ちゆう)	中(ちゆう)	17	7.9%	中(ちう)	199	92.1%	216	中(ちゆう)
堪へ	堪へ	166	79.8%	堪え・堪ゑ	42	20.2%	208	堪へ／耐へ
堪ふ	堪ふ	13	29.5%	堪ゆ・堪う	31	70.5%	44	堪ふ／耐ふ
教へ	教へ	75	92.6%	教え・教ゑ	6	7.4%	81	教へ
教ふ	教ふ	31	88.6%	教ゆ・教う	4	11.4%	35	教ふ

このように，「或いは」「中(ちゆう)」の９割以上が，歴史的仮名遣いに合致しない。また，「堪ふ（へ)」と「教ふ（へ)」の仮名遣いの状況をみると，「教ふ（へ)」よりも「堪ふ（へ)」の方が合致しない率が高く，また，ｅ段音よりｕ段音の場合に合致しない場合が多い。清濁の場合以上に，複雑な様相を呈していることがわかる。

2.8. 漢字の字体

現段階でもっとも一般的な電子媒体の日本語文字セットは，JIS X 0208:1997（以下，JIS）である。もっと大規模なものもあるが，現時点の日本における普及状況を考えるとJISによるのが，現実的である。ところが，『太陽』には，次に示すように，JISに含まれない字体が多く見られる。

a 羽ぶり〔羽18〕(01075A12)／身に浸む〔浸54〕(01152A07)／青年の〔青146〕(01018B22)／幽玄神秘を〔神161〕(01017B17)

b 沿海交通の〔沿〕(01065A07)／開港場〔港〕(01054A10)／信憑すべき〔憑〕(01005A01)／煮沸するか〔煮〕(01168A06)

上記のうち，aは，JISの包摂規準によって，〔 〕内の字体と同一視できるものである（数字は包摂規準の連番）。ところが，bは，この規準に含まれていない字体の変異で，〔 〕内の字体と無条件に同一視することはできないものである。

3．電子化の方法

3.1. 基本原則

本文電子化にあたっては，言語研究の観点から検索性と再現性に一定の満足が得られることと，データを作成する側にも利用する側にも負担があまりかからないこと，を基本的な原則とする。

検索性とは，求める用例を洩れなく確実に引き出すことができる性質のことで，決まった方針で均質化した本文を作成することと，いくつかの検索ツールを用意することによって，実現させる。再現性とは，原資料の姿に立ち戻って確認できる性質のことで，誌面通りの配置の本文を提示することと，均質化本文で修正された例にも原文の形を添えることによって，実現させる。

負担の軽減については，作成する側では，入力・校正・管理に至る一連の作業で，コストがかからず，誤りが起こりにくくし，利用の側では検索に便利で読みやすいものとする。

3.2. テキストの構造

号，記事，行を階層的に構造化したテキストとする。文を単位として階層化することも考えられるが，2.2で述べたとおり，形式上その切れ目が明確でない文章も多いため，その方式はとらない。データの保守・管理に秀で，検索や表示機能も満足できる，XML形式で構造化を行う。

3.3. 振り仮名の埋め込み

原文にある振り仮名は，それがないと読めない場合など有用な情報を提供する場合が多いが，振り仮名を入れることでテキストが煩雑になることも否めない。そこで，情報価値が特に高い振り仮名は入れ，それ以外は入れない方針をとることにする。振り仮名の採否を個々の語や漢字ごとに判断するのは大変手間がかかり，現実的でないので，記事ジャンルによって採否を決定する。具体的には，小説記事は，熟字訓や特殊な宛字等が多く，振り仮名の情報価値が特に高いと判断して，振り仮名を入れるが，他のジャンルの記事では，振り仮名がなくても読める語がほとんどで，その情報価値はそれほど高くないと考えて，振り仮名は入れないことにする。

3.4. 誤植の修正

2.4に示した類の誤植は，電子化にあたって修正する。ただし，再現性の観点から，原文の形を注で示す。なお，注は本文に埋め込む形式とし，随時参照できるようにする。この「修正」注には，2.4の①②③④⑤にしたがって，「A誤字」「B衍字」「C脱字」「D転倒」「E欠損」の類別を付す。

3.5. 通用の処理

2.5の通用は，原文のまま修正しないこととする。ただし，場合によっては，誤りではないかと利用者に疑いをもたれることもありえよう。特に，bの辞書類に指摘がない現象についてはその恐れが大きいので，古典の校訂などでとられる「ママ」と同じ趣旨から，注を付して規範形を示し，注意を喚起することにする。ママ注には，2.5の①②③にしたがって，「1語形」「2語法」「3漢字表記」の類別を付す。

3.6. 清濁の処理

2.6に示した「殆ど」「甚だ」「防ぐ」等の濁点が脱落した例は，修正して濁点を付す。3.4の場合に準じて注を付ける（類別は「F濁点脱落」）。清濁がゆれている語（「免がる・免かる」など）については通用の扱いとして，特に注は施さない。

3.7. 仮名遣いの処理

2.7に見た通り複雑な様相を示す仮名遣いを原文のままとすると，本文の均質化の観点から問題である。また，当時一般的であった仮名遣いに統一する方法も考えられるが，一般的な仮名遣いを語ごとに判定するのは手間がかかる。歴史的仮名遣いに統一するように修正するのがもっとも現実的であり，原文の仮名遣いがわかるように「修正」注を残す（類別は「G仮名遣」）。

3.8. 字体の処理

2.8に述べた字体の変異のうち，aは包摂して示すことで問題ないであろう。問題はbであるが，検索性の観点からは，JISの字体で示すことが望ましい。JISの包摂規準には一部不備と思われる部分もあり，明治大正期の字体に適用することに限界がある。包摂規準をいくつか補足することで，JISに同一視できる字体が求められる場合があるが，こうした場合は包摂を行うこととする（例：沿・港）。また，字体のゆれとして包摂する規準を明示できないが，音訓や意味など機能的に等価と見てよいものもある。これについては，包摂に準じた扱いとし（以下，「準包摂字」と呼ぶことがある），必要に応じて原文の字体を画像で参照できるように注を付した（例：凉・煖）。包摂規準を補足したり，包摂に準じる扱いをしても，JISに同一視できる字体がない場合は，外字（JIS外字）となる。この場合は，画像を参照することになる。なお，こうした字体の処理方法の詳細は，木村・田中・飯島・笹原(1999)に述べた。

4.「太陽コーパスVer 0.3（1901本文テキスト）」について

「太陽コーパス」のうち，本研究ノートで述べてきた仕様にしたがって作成した試用版として，「太陽コーパス Ver 0.3（1901本文テキスト）」を試験的に公開する。以下には，「太陽コーパス Ver 0.3」の説明を簡潔に記す。詳細については，「太陽コーパス Ver 0.3」の README ファイルを参照されたい。

4.1.「太陽コーパス Ver 0.3（1901本文テキスト）」の概要

「太陽コーパス Ver 0.3」は，「太陽コーパス」として構想しているもののうち，1901年12冊分（1～14号，6・11は除く）の本文テキストを，各冊単位にまとめたものである。所在コードは，年（2桁：01）・号（2桁：01～14），頁（数字3桁）・段（英字1字：ＡＢ～）・行（数字2桁）を，数字と記号で示す。おさめる文章は，著者の著作権保護期間（死後50年）を過ぎたものに限り，保護期間内の可能性のある記事は，除外している。

テキスト内の，特殊書式や特殊文字のうち，代用の規則を定めたものがある。

〈 〉　片仮名小書き，宣命書き，アクサン記号，二の字点などの場合，これで囲む

～～　～゛　くの字点　　■　本文の欠損等による判読不能箇所

＄　JIS 漢字で代用した字　　＝　入力不能文字（外字）　　＿　意識的空白

などである。

「太陽コーパス Ver 0.3」は，データファイルとして XML ファイルと TXT ファイル，HTML ファイルを収録する。このうち基本となるデータファイルは XML ファイルであり，このほかのデータファイルは XML ファイルから機械的に生成したものである。XML[5]とは W3C（World Wide Web Consortium，インターネット関連技術の標準化組織）によって勧告された規格であり，今後インターネット上での標準的なマークアップ言語となることが期待されている。XML 形式には，1．各種の情報を一つのファイルの下に構造化して統合することができる，2．データの妥当性の検証機能によってデータの保守・管理が容易になる，3．データ形式を変換したり必要な情報だけを抜き出したりすることが容易である，など利点が多いことから，この形式を採用した。

これらデータファイルのほかに TXT ファイルを検索するための perl スクリプトと XML ファイルを検索・形式変換するための Internet Explorer 5 用スクリプト（HTA ファイル，XSL ファイル）を収録する。

4.2. データファイル

4.2.1. XML ファイル

すべての情報を含むデータファイルである。W3C 勧告 XML Version 1.0 に準拠し，文字符号化形式（encoding）はエディタ類の普及状況に鑑み Shift_JIS を採用した。今回の XML 化は原文のページ・行割りなどの体裁と文字表記に着目した簡略なものであり，文・単語を単位としたマークアップは行っていない。文書に用いたタグと要素間の親子関係は次の通りである（文書定義ファイル TAIYO. DTD で規定している）。行内要素はタグを取り去ってテキスト値だけを取り出した場

合にもプレーンテキストとして可読であるように定義した[6]。

	タグ	属　　性	子 要 素
ブロック要素	太陽	年・号・Version	複数の「記事」
	記事	題名・著者・著作権・保護期限・ジャンル	複数の「行」
	行	id（原文での位置を示す）	テキストと「外字」「r」「注」「br」
行内要素	外字	文字鏡・記号	テキスト（漢字または＝1文字）
	r	rt（振り仮名）	テキスト（原則的に漢字1文字）
	注	原文・規範形・分類	
	br	なし	

◎「太陽」タグ　『太陽』1冊分の情報を要素とするルートタグ。

「年」属性――『太陽』の刊行年（西暦）。

「号」属性――『太陽』の号数。

「Version」属性――『太陽』XML文書のバージョン。

◎「記事」タグ　『太陽』に含まれる記事の一つを要素とする。記事の範囲は目次にとられているかどうかを目安として決定した。

「題名」属性――記事の題名。

「著者」属性――記事の著者。ペンネームなどはその著者の代表的な署名で示した。

「著作権」属性――記事が著作権保護期限内であるかどうかを示す。

「保護期限」属性――著作権保護期限を年（西暦）で示す。

「ジャンル」属性――記事のジャンル。原文の欄名（論説・小説等）を指標にした。

◎「行」タグ　原テキストでの1行を要素とする。

「id」属性――原文における頁・段・行を「001A01」（1頁上段1行目）の形式で示す。

◎「外字」タグ　3.8に述べたJIS外字や準包摂字の画像や関連する情報を表すためのタグ。『今昔文字鏡』の番号を用いる[7]。

「文字鏡」属性――原文の字形にもっとも近い『今昔文字鏡』収録字の文字番号。対応する文字が見つからない場合（文字鏡外字）は空とした。

「記号」属性――文字鏡属性を補足する情報。文字鏡外字には独自の記号を入力した。また原文の字形と文字鏡の字形との間に違いがあると判断した文字（文字鏡異体字）には「？」を入力した。

対　象	方 針 と 用 例	
準包摂字	同一視した字を外字タグで囲む。 例：露清密約として傳へらるゝ條<外字 文字鏡="16085">款</外字>の内容は	欵
文字鏡外字	記号属性に独自の記号を入力。 例：運輸交通の利便を<外字 文字鏡="" 記号="T23">缺</外字>くに至り，	缺

JIS外字		=を外字タグで囲む。 例：人の社會を化して<外字 文字鏡="40001">=</外字>より醇に赴かしむ	醪
	文字鏡 異体字	記号属性に「?」を入力。 例：化學的に風化<外字 文字鏡="42302" 記号="?">=</外字>爛するときは,	霉 (母一毋)
	文字鏡 外字	記号属性に独自の記号を入力。 例：爾も目<外字 文字鏡="" 記号="T 01">=</外字>のない廣い處で	碭

◎「r」(ルビ) タグ　振り仮名の情報を示すタグ。原則的に漢字一字ごとに付与した。ただし熟字訓については語に対して付与した。

「rt」属性――振り仮名。全ての「r」タグがもつ。

例：<r rt="かみ">髪</r><r rt="ゆひ">結</r><r rt="どこ">床</r>の<r rt="かど">角</r>を

例（熟字訓）：<r rt="ひとりごと">獨語</r>，<r rt="うなづ">首肯</r>いて

◎「注」タグ　「注」情報を含む空要素タグ。誤植・誤用等につき原文に手を加えた「修正」注と，規範形からは外れるが原文に手を加えなかった「ママ」注との２種類があり，それぞれ原文属性または規範形属性のいずれか１つと種類属性をもつ。振り仮名に対して注記をつける場合には振り仮名付きの漢字全体につけることとした。属性中の振り仮名は漢字の直後の［　］に入れた。

「原文」属性――「修正」注がもつ。原文の形を示す。

「規範形」属性――「ママ」注がもつ。規範形を示す。

「分類」属性――３に記した通り，「修正」注（Ａ誤字・Ｂ衍字・Ｃ脱字・Ｄ転倒・Ｅ欠損・Ｆ濁点脱落・Ｇ仮名遣）と，「ママ」注（１語形・２語法・３漢字表記）のそれぞれの種類を示した。

例（修正）：長さ<注 原文="長き" 規範形="" 分類="Ａ誤字"/>百六十メートル

例（ママ）：裁判所搆成法は<注 原文="" 規範形="裁判所構成法は" 分類="３漢字表記"/>

◎「br」(改行) タグ　原文での意識的な改行（段落末・箇条書きなど）を示す空要素タグ。原文での改行位置にこだわらない形式の文書に変換した場合にも改行として示される必要がある箇所に挿入した。原文で行末が文末と一致しており，段落ごとの字下げが行われていない場合などには挿入すべきかどうかを個々に判断した。

4.2.2. TXT ファイル

原文の行にあわせて改行し，行頭に原文位置を示す出典コードを付したテキストファイル。ルビは漢字の直後の［　］内に入れた。同梱の perl スクリプトによって検索することができる。

4.2.3. HTML ファイル

WEBブラウザで本文を閲覧するためのファイル。冒頭にリンク付きの目次を設け，外字を文字鏡研究会のサーバ上の画像で表示するほか，ルビも表示する。注記は「注」記号をクリックすることで表示される。

4.3. 検索・形式変換ツール

4.3.1. PLファイル

■TTFIND. PL　TXT ファイルを検索するための perl スクリプト。ファイル中の改行や出典コードを除去した全文検索が可能で，正規表現を用いることができる。結果は記事情報付きの KWIC 形式で出力される。詳しくは ttfind.pl ファイルの冒頭を参照。

4.3.2. HTA ファイル

XML ファイルを直接利用するためのアプリケーション。いずれも Windows 用 Internet Explorer 5 が必要。詳しくは各アプリケーションの「使い方」を参照。

■TX_FIND. HTA（本文検索ツール）　XML ファイルを全文検索し，用例を表示する HTML アプリケーション。正規表現を用いた検索ができ，ジャンル別の検索対象指定も可能である。また，検索対象として，ルビなし・ルビ付き・ルビを開いたテキストのいずれかを選ぶことができる。結果は指定した文脈長の KWIC 形式で表示され，ファイルへの出力やクリップボードへのコピーを行うことができる。

■TX_CONV. HTA（XSL（T）変換ツール）　XML ファイルと次に述べる XSL ファイルを組み合わせて別の形式のファイルを生成するための HTML アプリケーション。変換結果をブラウザで表示したり，ファイルに出力したりすることができる。

4.3.3. XSL ファイル

XML ファイルを各種の形式に変換したり，XML ファイルから特定の情報を抜き出したりするためのファイル。Internet Explorer 5 に実装されている XSL に準拠した。変換結果の詳細は README ファイルを参照。

	ファイル名	概　　要
形式変換	hon.xsl	原文通りに改行した行番号付きのテキストファイルを生成する。ルビは削除する。
	rub.xsl	原文通りに改行した行番号付きのテキストファイルを生成する。ルビは漢字の後の［　］に入れる。（TXT ファイル）
	plain.xsl	プレーンテキストファイルを生成する。
	htm _ rich.xsl	本文を閲覧するための HTML ファイルを生成する。（HTML ファイル）
	htm _ hon.xsl	原文通りに改行した行番号付きの HTML ファイルを生成する。
	htm _ smpl.xsl	外字画像などを使わないシンプルな HTML ファイルを生成する。
	tex.xsl	TeX（pLaTeX 2e）形式のファイルを生成する。
情報抽出	gaiji.xsl	外字の一覧（出現順）を生成する。
	gaijic.xsl	外字の一覧（コード順）を生成する。
	stj.xsl	記事に関する情報（題名・著者・著作権・保護期限・ジャンル・開始位置・終了位置）の一覧を生成する。

	tyu.xsl	注の一覧（種類順）を生成する。

4.4. 試験公開について

- 「太陽コーパス Ver 0.3（1901本文テキスト）」（約35MB）は，国立国語研究所が公開準備中の「太陽コーパス」の一部について，研究を目的とする利用希望者（個人）に対して，試験的に公開するものである。
- 試験公開の目的は，利用者の意見を完成時の最終版に生かすことにある。本コーパスへの意見，要望，誤りの指摘等を寄せてほしい。
- 本コーパスを利用して研究成果を発表する場合は，利用した旨を明記すること。また，抜刷等を国立国語研究所国語辞典編集室に送付すること。
- 本コーパスの全体あるいは部分について，複写物や加工物を流通させてはならない。
- 以上の条件に同意し，本コーパスの利用を希望する場合は，氏名，所属，連絡先（住所，電話番号，E-mailアドレス），具体的な研究目的を明記し，mtanaka@kokken.go.jp（田中牧郎）に申し込んでほしい。本コーパスを無料で提供する。

注

1 『太陽』が，時代の潮流のなかでどのような意味をもっているのかについては，国際日本文化研究センター(1996-1999)，永嶺(1997)が詳しい。『太陽』がどのような層の日本語を反映しているのかを考える場合にも，この二つの研究は参考になる。

2 もともとは国語辞典編集のための用例採集事業の一環として，『太陽』からスカウト式によって用例を採集していたものを，コーパス構築を目的とする研究事業へと変更した。その間の経緯は，田中(1998)，木村・加藤・田中(1999)に述べた。

3 「太陽コーパス」は，Ver 0.2までは内部資料。試験公開は本バージョンが最初である。

4 単位認定，品詞体系，検索システム等，検討を要することがらが色々あるので，別の機会にまとめたい。スカウト式用例採集方法とこの方法で得られた語の性格については，木村・加藤・田中(1999)に概略を述べた。

5 Extensible Markup Language（拡張可能なマーク付け言語）。XMLについてはW3C（http://www.w3.org/XML）の他，日本XMLユーザーグループ（http://www.xml.gr.jp/），XML FAQ（http://www.fxis.co.jp/DMS/sgml/xml/xmlfaq.html）等が参考になる。

6 要素とはそれぞれの開始タグ（<タグ名>）と終了タグ（</タグ名>）によって囲まれる情報，空要素タグとは開始タグと終了タグとの間に何の要素ももたないタグ（多く<タグ名/>の形式をとる），属性とは開始タグまたは空要素タグ内に<タグ名 属性名="属性値">の形で付与される情報，ルート要素とはこれ以外のすべての要素を含む最上位の要素をさす（用語の正確な定義はW3C勧告 XML 1.0（http://www.w3.org/TR/1998/REC-xml-19980210・日本語訳http://www.fxis.co.jp/DMS/sgml/xml/rec-xml.html）を参照）。また，本稿でいうブロック要素とは原文における一定の領域を構成する要素，行内要素とは主にテキストを修飾するための要素をさす。

7 『今昔文字鏡』Ver2.00によった。その詳細は，文字鏡研究会（http://www.mojikyo.gr.jp/）等を参照。文字鏡番号1～49964までは，諸橋轍次著『大漢和辞典』（大修館書店）の番号と共通。なお，

『今昔文字鏡』に未収録の文字については今後，文字鏡研究会に登録を申請する予定である。

参考文献

木村 睦子・加藤 安彦・田中 牧郎（1999）「国語辞典編集のための用例データベース」『日本語科学』5，109-128

木村 睦子・田中 牧郎・飯島 満・笹原 宏之（1999）『『太陽』コーパスの漢字処理—『太陽』1901の漢字調査—』新プロ「日本語」研究成果刊行物

国際日本文化研究センター（1996-1999）「共同研究報告：総合雑誌『太陽』の総合的研究」『日本研究　国際日本文化研究センター紀要』13-19

田中 牧郎（1998）「『太陽』コーパスの作成」『国立国語研究所創立50周年記念研究発表会資料集』39-46

築島 裕（1986）『歴史的仮名遣い　その成立と特徴』中公新書

土屋 信一（1966）「雑誌「太陽」（明治28—昭和３）に見る表記の変遷」『言語生活』182，36-42

土屋 信一（1967）「雑誌「太陽」の用字の変遷」『言語生活』193，34-43・87

永嶺 重敏（1997）『雑誌と読者の近代』日本エディタースクール出版部

付　記

『太陽』本文の電子化は，国立国語研究所国語辞典編集室の事業課題「国語辞典の編集」（1988-2000年度），および，科学研究費創成的基礎研究「国際社会における日本語についての総合的研究」（研究代表者：水谷修，1994-1998年度）研究班４の研究課題「情報発信のための言語資源整備に関する研究」のなかで進めてきた。本文電子化に着手した1995年度以後の担当者は次の通り（所属は担当した当時のもの）。小椋秀樹・加藤安彦・木村睦子・笹原宏之・田中牧郎・中山典子・藤原浩史・山口昌也・山田貞雄（国語研），飯島満・石山順子・乾とね・大塚みさ・緒方典裕・小木曽智信・奥村大志・貝美代子・小島聡子・中尾比早子・服部紀子・本多久美子・吉川明日香（国語研非常勤）。また，所外から次の方の協力も得た。大木一夫（埼玉大学），柴田雅生（明星大学），島田泰子（香川大学），服部隆（上智大学），馬場俊臣（北海道教育大学），平澤啓（和歌山大学），湯浅茂雄（実践女子大学）。

（投稿受理日：2000年７月４日）

田中　牧郎（たなか　まきろう）

国立国語研究所　115-8620　東京都北区西が丘3-9-14

mtanaka@kokken.go.jp

小木曽　智信（おぎそ　としのぶ）

東京大学大学院人文社会系研究科日本文化研究専攻日本語日本文学専門分野

165-0022　東京都中野区江古田4-33-4-101

ogiso@eva.hi-ho.ne.jp