国立国語研究所学術情報リポジトリ

New XML-tagging program for Japanese linguistic study : Its function and application

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: 小木曽, 智信, 近藤, 明日子, OGISO, Toshinobu, KONDO, Asuko メールアドレス: 所属:
URL	https://doi.org/10.15084/00002187

日本語研究のためのXMLタグ付けプログラム
——その開発と活用例——

小木曽　智信
(国立国語研究所)

近藤　明日子
(国立国語研究所)

キーワード
XML，構造化文書，用例検索，タグ付け，太陽コーパス

要　旨

現在XMLで作られた日本語の言語資料が普及しつつある。これを日本語研究で有効に活用するため，資料が持つ情報を十分に引き出した調査を行い，その調査結果を再利用可能な形で保存，時々の研究場面に応じて参照する研究手法を提案する。まず，その手法を実現するために必要なXMLタグ付けプログラム「たんぽぽタガー」の開発について，プログラムの概要とともに報告する。次に，このプログラムを使ったタグ付けの方法と，タグ付けしたXML文書にXSLTを適用して研究に有用なリストに変換する方法を，研究手順に沿って具体的に紹介する。

1. XML文書のタグを利用した日本語研究の手法

1.1. 日本語研究と構造化文書

従来，日本語研究のための電子化された資料は単純なテキストデータで公開されることが多かった。しかし，今日ではテキスト本文だけでなく言語研究に必要な情報を付与するために，マークアップ言語を用いた構造化文書として公開されることが多い。不定型な部分を持つ言語資料の格納に適していることから，XMLを用いて作成されることが普通である。これまでに国語研究所の『太陽コーパス』などの大規模なデータがXMLで作られてきた。現在開発中の「現代日本語書き言葉均衡コーパス」[1]をはじめ，今後も多くの言語資源が同形式で作られていくことは間違いない。また，インターネット上で公開されている資料には，「青空文庫」[2]をはじめ，XHTML形式で作成されたものが多いが，これもXML形式の一種である。はじめから日本語研究を目的として作られたコーパスだけでなく，こうした周辺的なデータを含めれば，XMLで作成された言語資源は膨大な量に上る。

1.2. XML文書利用の問題点

このように多くの資料が存在するにもかかわらず，現状では日本語研究——ここでは文献資料を対象とした表記・語彙・文法等の研究を念頭においている——においてXMLで作られたデータ（XML文書）が十分に活用されているとは言いがたい。XML文書によるコーパスの特長として，言語研究にも活用できるさまざまな情報がタグによって付与されている点があげられる。し

かし，文系研究者の多くは，XML 文書からタグを取り去って単なるテキストデータとして利用したり，コーパス付属の検索ツールが出力するテキストデータを表形式で利用したりしているようである。

このような従来の手法でも十分に研究に役立つ場合もあるが，次のような問題がある。

(1) **タグを取り去ったテキストだけを使用する場合，データの誤った利用につながることがある**

XML で作られたデータは，一般にタグを含めた全体で十分な情報になるよう設計されている。そのため，タグを取り去ってテキストだけを利用すると，誤った結果を引き出す可能性がある。たとえば，原文を検索しやすい本文に訂正して，タグを用いて原文情報を記録している XML 文書の場合，そのテキスト部分だけで判断すると，原文では出現していない語をカウントしたり，出現していた語を見落としたりといった誤りを招くことになる。

(2) **元のデータが持つ情報のごく一部しか利用できない**

タグを取り去ってしまった場合にその情報が失われるのはいうまでもないが，コーパス付属のツールを使った場合であっても，そのツールが対応している範囲でしか情報を得ることができない。また，検索時・用例処理時に必要であると考えていなかった情報を後から引き出すことが困難である。

(3) **調査結果が元のデータに反映されないため，その場限りの使い捨てになり，別の調査結果との複合的な分析ができない**

検索結果に対し，用例を取捨選択したり分類したりといった編集を行う場合，その結果はその場限りの使い捨てになってしまい再利用できないことが多い。

たとえば，助動詞 X を検索した結果を編集して X の用法分類リストを作った場合，そのリストは X について論じる場合にしか利用できない。動詞 Y について同じような処理を行ったとしても，X と Y の関係についての情報（X が Y に接続するかどうか，X の用法と Y への接続の関係はどうか，X が Y に接続する文の属性に偏りはあるか…といった情報）は簡単には得られない。

こうした問題を乗り越え，情報をより活用するには，XML 文書を XML 文書として処理する必要がある。これまで XML 文書はコーパスの格納形式としてその有効性が注目されてきたが，コーパスの利用においてもその利便性を活用することが望まれるのである[3]。

1.3. XML 文書の活用

XML 文書をより高度に活用するために“研究者自身が XML 文書に情報をタグとして埋め込み，XML 関連技術を積極的に利用してタグの情報を引き出す”という方法を提案したい。

XML 文書中の用例が持つ情報をすべて引き出すには，文書中の用例そのものに情報を埋め込んでおくことが有効である。調査結果をタグとして直接埋め込んでおくことにより，書き込みをしたり付箋を貼ったりした本のように，自分だけのカスタマイズされた資料とすることが可能になる。そのうえで，検索やデータの抽出に XPath，XSLT などの技術を用いることで，コーパスに元から含まれる情報とともに自分で埋め込んだ情報を自由に利用することができる。

これにより，調査結果を使い捨てにすることなく再利用することが可能になるだけでなく，検索時に想定していなかった要因を後で調査することも可能になる。情報を埋め込んでおけばXML文書につけられている属性などの情報がすべて参照できるほか，記事中の出現位置などの文書構造情報も取り出すことができる。このような方法は，多くの国語資料に対して有効な研究手法となりうるが，特に国立国語研究所の雑誌コーパス[4]のような，構造化され言語資料としての情報が付加されたXML文書に対しては利用価値が高い。

こうした研究手法を誰もが利用できるようにするためには，プログラムが書けなくても使用できるツールを提供するとともに，実際にそれを使ってどのようなことができるのか実践して示すことが必要であろう。本稿では，2.でそうしたツールの開発について報告し，3.でそのツールを研究に利用した実践例を紹介する。

2. タグ付けプログラム「たんぽぽタガー」の開発

2.1. 設計方針

1. で示した研究手法を可能にするために，XML文書中の指定した文字列に対して任意の情報をタグによって埋め込む（タグ付けする）ことのできるプログラムを開発した。“日本語研究のための使いやすいツール”とすることを目標に，次のような点に配慮しつつ設計・開発を行った。プログラムはフリーソフトとして公開する予定である[5]。

日本語研究のためのツールとして必要な次の機能を実装した。

【正規表現に対応した検索・タグ付け機能】 漢字や送り仮名などの表記の揺れ，活用語尾などに対応するため，正規表現を使って検索を行いタグ付けすることを可能にした。

【「ふりがな」「踊り字」に対応した検索・タグ付け機能】 ふりがながタグ付けされている場合に，たとえば「走(はし)る」という本文を「はしる」でも「走る」でも検索可能とした。また，踊り字に対応し，「こゝろ」や「心(こゝろ)」を「こころ」で検索してタグ付けすることができるようにした。

【タグに情報がいくつもつけられる機能】 必要に応じて自由に情報を記述し，あとで容易に組み合わせて取り出すことができるよう，タグに複数の属性を埋め込めるようにした。

【タグ付けした情報を簡単に取り出す機能】 日本語研究で必要となるXSLTスタイルシートのサンプルを同梱し，変換ツールと連携して簡単に利用できる機能を用意した。

【各種のXML形式への対応】『太陽コーパス』のXML形式のほか，XHTML形式等，日本語研究で用いられる他の形式にも対応した。

また，使いやすいツールをめざし，次の点に配慮した。

【ガイド付きのインターフェイス】 見た目がわかりやすく，その都度使い方を案内するユーザインターフェイスを用意した。

【標準的なアプリケーションとの連携機能】 データの編集に際して，テキストエディタや表計算ソフト等，ユーザが使い慣れた標準的なアプリケーションソフトを呼び出して利用する機能を用意した。

【特別なソフトウェアに依存しないこと】 別途特別なソフトウェアをインストールすることなく，一般的なパソコン環境で動作することを前提として設計した。

プログラムは一般的な Windows 環境で動作する[6]。ユーザインターフェイス部は HTML アプリケーション，テキスト処理部は Perl で作成して PAR[7]により実行形式化したものである。

特徴である「ふりがな」「踊り字」に対応した検索機能は『太陽コーパス』付属の検索プログラム「たんぽぽ」の機能を引き継ぐものであるため，名称を「たんぽぽタガー」とした。「たんぽぽタガー」の実行画面を図1に示す。

2.2.「たんぽぽタガー」の機能

「たんぽぽタガー」の基本的な機能は，XML 文書中の任意の文字列に対し任意のテキスト情報をタグ付けすることである。単純なテキストの置換とは異なり，検索対象の文字列がタグをまたいでいる場合にもタグ付けを行うことができる。

利用の手順としては，画面上に表示されるメッセージに従って(1)から(4)のステップに移ってゆくことでタグ付けが完了する流れとなっている。

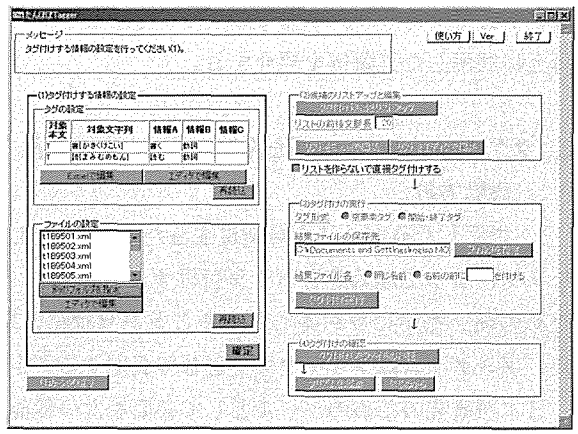


図1 「たんぽぽタガー」の実行画面

以下，この流れに沿って「たんぽぽタガー」の機能について説明する。ここでは機能・仕様の説明にとどめ，実際の利用方法については具体例とともに3. で示すこととする。実際にタグ付けされた本文の例も3.4. で示した。

(1) タグ付けする情報の設定

ここでタグ付け対象のファイルと，タグ付け対象の文字列，本文の種類，タグに付与する情報を設定する。

タグ付け対象のファイルは，日本語で書かれた一般的な XML 文書であればどのような形式でもよい。また，一部の形式については，ふりがなや踊り字に配慮した検索・タグ付けが可能となっている。ふりがなは，雑誌コーパスの形式のほか，XHTML のルビ形式[8]に対応している。踊り字については，すべての XML 形式で「ゝゞヽヾ」を展開した本文を検索・タグ付けできるほか，雑誌コーパスの踊り字タグを展開した本文にも対応している。

タグ付け対象の文字列やそこに埋め込む情報は設定ファイルに記述し，それを読み込んで使用する形を取っている。これは一度に複数の対象文字列に対してタグを埋め込めるようにするとともに，タグ付け条件を保存しておくことを可能にするためである。

タグの設定ファイルは，タブ区切りのテキストファイルで，次の形式による。

［対象本文］［対象文字列］［情報 A］［情報 B］［情報 C］

［対象文字列］はXML文書中のタグ付け対象となる文字列で，正規表現が利用できる。［情報A～C］はタグに属性として付与する情報で，任意の文字列を指定することができる[9]。［対象本文］は表1に示す略号T・R・Aで指定する。Oは踊り字を展開するオプションである。

表1　対象本文の種類

略号	本文の種類	踊り字の展開	例：其（そ）のまゝ
T	通常のテキスト（「ルビなし本文」形式）	しない	其のまゝ
TO		する	其のまま
R	「ルビを開いたテキスト」形式	しない	そのまゝ
RO		する	そのまま
A	「ルビ入りテキスト」形式	しない	其［そ］のまゝ
AO		する	其［そ］のまま

(2) 候補のリストアップと編集

「たんぽぽタガー」では，検索した文字列に直接タグ付けするのではなく，いったんタグ付け候補のリストを作成し，それを編集した後に実際にタグ付けするという手順を踏む。これは，検索結果の中に含まれている意図しない用例を，タグ付け前の段階で除去するためである。

タグ付け候補のリストは，タブ区切りのテキストファイルで，形式は次の通りである。

［ファイル名］［対象文字列の開始位置］［対象文字列の終了位置］［埋め込むタグ］［KWIC］

［対象文字列の開始位置］と［対象文字列の終了位置］はファイル先頭からの文字数で記録している。［KWIC］は用例の要不要を判断するための文脈で，その長さは画面上で指定できる。

(3) タグ付けの実行

続いて実際にタグ付けを行う。タグには，(1)で指定した情報がA，B，C属性として付与されるほか，対象として指定した文字列（正規表現）と本文の種類が属性として埋め込まれる。

埋め込むタグの形式は，「空要素タグ形式」と「開始・終了タグ形式」の二種類から選択できる。前者は，タグ付け対象文字列の開始位置と終了位置にそれぞれ空要素のタグを挿入するもので，後者は，タグ付け対象文字列を開始タグと終了タグで囲むものである。空要素タグ形式の場合には，候補リストの開始・終了位置に基づくidが二つのタグに属性として付与され，これによって対応するタグが確認できる。

空要素タグ形式

いとしい〈t:tag text="通常" search="貴方" A="二人称代名詞" B="あなた" C="" id="100:102"/〉あなた〈t:endtag id="100:102"/〉は今どこに

開始・終了タグ形式

いとしい〈t:tag text="通常" search="貴方" A="二人称代名詞" B="あなた" C=""〉あなた〈/t:tag〉は今どこに

前者はどのようなXML文書であっても整形式でタグ付けすることができるが，後者は不正なXMLになる可能性がある。例えば，次のような場合には，ふりがなのrタグと新たに挿入されたタグが入れ子にならないためエラーとなる。

〈r rt="さみだれ"〉五月〈t:tag〉雨〈/r〉が〈/t:tag〉降る

空要素タグによる形式ではこの問題は起きない一方，情報を抽出するXSLTスタイルシートの記述が複雑になるという問題がある[10]。

(4) タグ付けの確認

開始・終了タグ形式でタグ付けした場合のエラーに対処するため，ここでタグ付け後のXML文書を検証することができる。エラー発生時にはタグ付けしたXML文書を修正する必要がある。

2.3.「プリズム」と付属XSLTスタイルシート

タグ付けしたXML文書を利用するために，研究で役立つと考えられるスタイルシートを用意し，これを適用するためのソフトウェア「プリズム」[11]を同梱している。

付属のXSLTスタイルシートは表2の通りである。いずれも，空要素タグ形式と開始・終了タグ形式の両方に対応している。スタイルシートは，雑誌コーパスXMLとXHTML形式に対応したものを用意したが，これ以外の形式ではそれに合わせたものを用意する必要がある。

表2　附属XSLTスタイルシート

スタイルシートファイル名	スタイルシートの用途
タグリスト.xsl	t:tagタグのリストの作成
表記リスト.xsl	表記のリストの作成
kwic.xsl	KWICの作成
集計表.xsl	タグ数の集計表の作成（集計対象のタグは3.3で紹介する研究例に対応）

3. タグ付けを利用した研究例

ここでは，「たんぽぽタガー」とXSLTスタイルシートを利用して，資料とするXML文書中の研究対象とする語にタグ付けし，日本語研究に有用な情報とともに収集，リストや集計表に整形する例を紹介する。

とりあげる研究例は，例1から例3までの3つの部分に分かれる。例1では，資料とするXML文書での研究対象語の表記リストを作成する（3.1.）。例2では，例1で作成した表記リストをもとに，研究対象語にタグ付けを行う。例1で作成した表記リストを利用することで，現代語の表記法からは推定困難な表記も含めた検索が可能となり，検索漏れを最小限に抑えることができる（3.2.）。例3では，例2でタグ付けをした研究対象語と他の語との共起関係を調査する場合を想定し，タグ付けした語数の集計表を作成する（3.3.）。

なお，以下，図中で具体例としてあがっているのは，二人称代名詞「あなた」「おまへ」を研究対象語とし，文体を敬体と決定する要素である「ございます（ござります）」「です」「ます」との共起関係を考察する場合を想定したものである。資料としたXML文書は，『太陽コーパス』を構成するXML文書の一つ（t190901.xml）である[12]。

3.1.〔例１〕研究対象語の表記リストの作成

〔手順１〕タグ付けする情報の設定

まず，「たんぽぽタガー」の「(1)タグ付けする情報の設定」にある「Excelで編集」ボタンまたは「エディタで編集」ボタンをクリックし，Excelまたはテキストエディタでタグを設定するファイルを表示・編集する。編集後，ファイルを上書保存し終了する。「(1)タグ付けする情報の設定」の「再読込」ボタンをクリックすると，編集後のタグの設定が表示される（図２）。

図２　表記リスト作成のためのタグ付け設定

この例では，研究対象語と語形（よみ）が同一であることが確実な文字列の表記リストの作成を目標とするので，仮名表記かふりがなの振られた表記の文字列をタグ付けの対象とする。そのために，「対象本文」に「RO（ルビを開いたテキスト／踊り字を展開する)」，「対象文字列」に研究対象語を仮名表記で入力する。

次に，「ファイルの設定」でタグ付けの対象とするXML文書を指定，「確定」ボタンをクリックすると，タグ付けする情報の設定が完了する。

〔手順２〕候補のリストアップと編集

「たんぽぽタガー」の「(2)候補のリストアップと編集」にある「リストを作らないで直接タグ付けする」にチェックを入れる（この例ではタグ付け候補の編集は行わない)。

〔手順３〕タグ付けの実行

「たんぽぽタガー」の「(3)タグ付けの実行」にある「タグ形式」で，「空要素タグ」にチェックを入れる。この例では，タグ付け対象のXML文書を上書きしないようにするために，「結果ファイルの保存先」でタグ付け対象のXML文書のあるフォルダとは別のフォルダを指定するか，「結果ファイル名」で「名前の前に□を付ける」にチェックを入れ，□に適宜文字列を入力するかする。

「タグ付け実行」ボタンをクリックすると，タグ付けが開始される。「メッセージ」にタグ付け完了のメッセージが表示されることを確認する。

〔手順４〕XSLTによる表記リストの作成

「たんぽぽタガー」の「(4)タグ付けの確認」にある「タグ付けしたファイルの検証」ボタンをクリック，タグ付けしたXML文書が正しい旨のメッセージが表示されることを確認する。「プリズムを起動」ボタンをクリックし，XSLTスタイルシートを適用するためのアプリケーション「プリズム」を起動する。

「プリズム」の「入力XMLファイル」でタグ付けを終えたXML文書を指定，「適用するスタ

イル」で表記一覧を作成するためのスタイルシート「表記リスト.xsl」を指定，「変換（ブラウザで表示）」ボタンをクリックすると，表記リストが表示される（図3）。リストの左から1列目が本文中での表記（次の例2で利用するためルビは削除），2列目がタグ付け対象文字列である。これで，研究対象語と語形（よみ）が同一の文字列の表記リストが完成した。

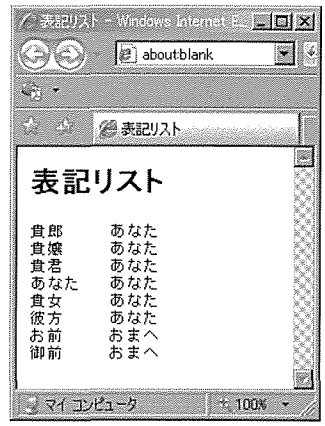


図3　表記リスト

3.2.〔例2〕研究対象語へのタグ付けとKWICの作成

〔手順1〕タグ付けする情報の設定

3.1.の〔手順1〕同様に，「たんぽぽタガー」の「(1)タグ付けする情報の設定」でタグの設定を表示させる（図4）。

この例では，語形（よみ）が確定できなくとも，研究対象語である可能性がある文字列すべてをタグ付け候補とすることで，検索漏れを最小限にすることを目標とする。そのため，「対象本文」に「TO（通常のテキスト／踊り字を展開する）」，「対象文字列」に3.1.で作成した表記リスト（図3）を参照し，ルビのない表記を入力する。これにより，ルビのない表記の用例があったとしても，タグ付けの候補として拾い上げることができる。また，「情報A」以降の列は，タグ付けの目的に応じて適宜入力することになるが，この例では，「情報A」に研究対象語の分類，「情報B」に研究対象語の基本語形を入力する。以上の作業は，図3の画面上でリストを選択・コピーしたものを利用すれば，手間を省くことができるであろう。

図4　研究対象語へのタグ付け設定

なお，「ファイルの設定」で指定するのは，3.1.で作成した，表記リスト作成のためにタグを付けたXML文書ではなく，元のXML文書とする。

〔手順2〕候補のリストアップと編集

「たんぽぽタガー」の「(2)候補のリストアップと編集」にある「リストの前後文脈長」を入力後，「タグ付け候補をリストアップ」ボタンをクリックすると，タグ付け候補のリストアップが開始される。「メッセージ」にリストアップ完了のメッセージが表示されたことを確認した後，「リストをExcelで表示」ボタンまたは「リストをエディタで表示」ボタンをクリックし，候補リストを表示する（図5）。

リストの最右列に，タグ付け候補の文字列が【　】に括られ前後文脈とともに表示されるので，これを手がかりに【　】内が研究対象語か否かを判断し，研究対象語ではないと判断される候補は，その行ごと削除する。なお，候補リストでは判断が困難なものは，いったん候補として残しておき，タグ付け後，XML文書をテキストエディタ等で開いて，改めて削除することも可能である。編集完了後，リストを上書保存し終了する。

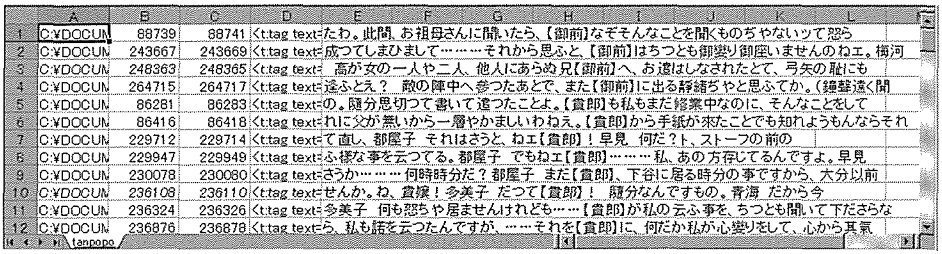


図5　タグ付け候補のリスト

〔手順3〕タグ付けの実行

3.1.の〔手順3〕同様に行う。

〔手順4〕XSLT による KWIC の作成

「プリズム」を起動するまでは3.1.の〔手順4〕と同様である。「プリズム」の「適用するスタイル」で，XML 文書をタブ区切りテキストファイル形式の KWIC（文脈付き索引）に変換するスタイルシート「kwic.xsl」を指定する。「ファイル出力オプション」を適宜設定後，「変換（ファイルに出力）」ボタンをクリックすると，ファイルへの出力が開始される。出力完了のメッセージを確認後，出力されたファイルを Excel で読み込むと，表形式で表示される（図6）。

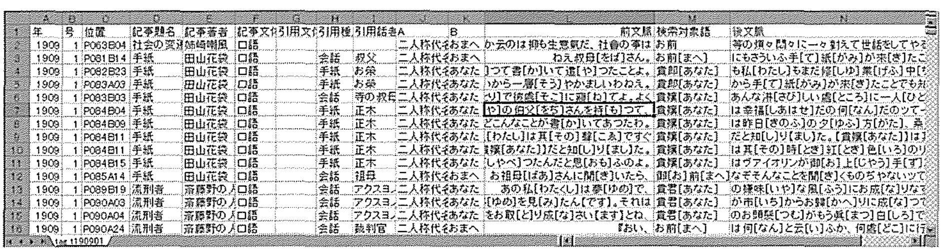


図6　KWIC

これで，研究対象語の KWIC が完成した。『太陽コーパス』付属のアプリケーション「ひまわり」[13]「たんぽぽ」でも KWIC の作成は可能であるが，これらは，「検索対象語と前後文脈とを区切る」「検索対象語と前後文脈をルビ付きで表示する」「検索対象語に自分で付与した情報を表示する」といった機能の一部にしか対応していない。

3.3.〔例3〕研究対象語と共起する語の集計表の作成

〔手順1〕共起する語へのタグ付け

研究対象語と共起する語へのタグ付けは，3.1.〔手順1〕から3.2.〔手順3〕までと同様に行う。ただし，「ファイルの設定」においては，3.2.で研究対象語へのタグを埋め込んだ XML ファイルを指定する。これは，研究対象語へのタグと

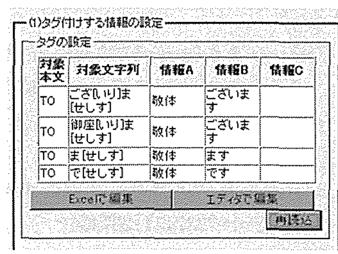
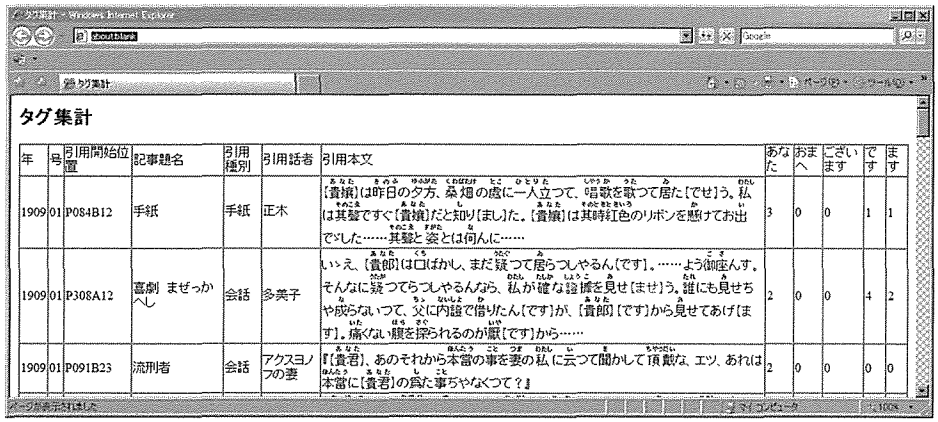


図7　共起する語のタグ付け設定

共起する語へのタグとを同一のXMLファイルに共存させるためである。なお,「(1)タグ付けする情報の設定」の「タグの設定」の「対象文字列」では,「ござ［いり］ま［せしす]」のように正規表現を使用できる（図7）。

〔手順4〕XSLTによる語数集計表の作成

「プリズム」を起動するまでの手順は3.1.の〔手順4〕と同様である。「プリズム」の「適用するスタイル」で,XMLファイルをタグ集計表に変換する「タグ集計.xsl」を指定する。「変換（ブラウザで表示）する」ボタンをクリックすると,集計表が表示される（図8）。これで,タグの集計表が完成した。



タグ集計

年	号	引用開始位置	記事題名	引用種別	引用話者	引用本文	あなた	おまへ	ございます	です	ます
1909	01	P084B12	手紙	手紙	正木	〔貴嬢〕は昨日の夕方、桑畑の處に一人立つて、唱歌を歌つて居た〔でせ〕う。私は其聲ですぐ〔貴嬢〕だと知り〔まし〕た。〔貴嬢〕は其時紅色のリボンを懸けてお出でした……其聲と姿とは何んに……	3	0	0	1	1
1909	01	P308A12	喜劇 まぜっかへし	会話	多美子	いゝえ、〔貴郎〕は口ばかし、まだ疑つて居らつしやるん〔です〕。……よう御座んす。そんなに疑つてらつしやるんなら、私が確な證據を見せ〔ませ〕う。誰にも見せちや成らないつて、父に内證で借りたん〔です〕が、〔貴郎〕〔です〕から見せてあげ〔ます〕。痛くない腹を探られるのが厭〔です〕から……	2	0	0	4	2
1909	01	P091B23	流刑者	会話	アクスヨノフの妻	『〔貴君〕、あのそれから本當の事を妻の私に云つて聞かして頂戴な、エツ、あれは本當に〔貴君〕の爲に事ぢやなくつて？』	2	0	0	0	0

図8　語数集計表

3.4. 付けたタグの活用

以上,例1から例3で「たんぽぽタガー」を使った研究例を紹介したが,実際のXMLファイルには次のようにタグが付けられている。太字体の部分が新たに付けられたタグである。

〈s〉いゝえ,〈/s〉
〈s〉〈t:tag text=" ルビなし " search=" 貴郎 " A=" 二人称代名詞 " B=" あなた " id="237393:237395"/〉〈r rt=" あなた "〉貴郎〈/r〉〈t:endtag id="237393:237395"/〉は〈r rt=" くち "〉口〈/r〉ばかし,〈/s〉
〈s〉まだ〈r rt=" うたぐ "〉疑〈/r〉つて〈r rt=" ゐ "〉居〈/r〉らつしやるん〈l位置 ="P308A07" /〉〈t:tag text="ルビなし" search="で[せしす]" A="敬体" B="です" id="237414:237416"/〉です〈t:endtag id="237414:237416"/〉。〈/s〉
〈s〉……よう〈r rt=" ご "〉御〈/r〉〈r rt=" ざ "〉座〈/r〉んす。〈/s〉
〈s〉そんなに〈r rt=" うたが "〉疑〈/r〉つてらつしやるんな〈l位置 ="P308A08" /〉ら,〈/s〉
〈s〉〈r rt=" わたし "〉私〈/r〉が〈r rt=" たしか "〉確〈/r〉な〈注 原文 =" 證[しやう]" 分類 ="G 仮名遣 "〉〈r rt=" しよう "〉證〈/r〉〈/注〉〈r rt=" こ "〉據〈/r〉を〈r rt=" み "〉見〈/r〉せ〈注 原文 =" ましや " 分類 ="G 仮名遣 "〉〈t:tag text=" ルビなし " search=" ま[せしす]" A=" 敬体 " B=" ます " id="237454:237456"/〉ませ〈/注〉〈t:endtag id="237454:237456" /〉う。〈/s〉

このように調査した結果をタグの形でXML文書内に残しておいて，後から活用できるのが本稿で提案する研究手法の特長である。タグをさらに追加することも可能であるから，例えば常体の文末辞「だ」「である」を調査対象に加えて共起関係を再調査する必要が生じた場合も，同じXML文書に追加してタグ付け作業を行い，XSLTスタイルシートで集計し直せば，比較的簡単に再調査が終了する。このように，一度たんねんに調査した結果を次の調査の際に活かすことができれば，新たな発見につながることもあろうし，また，単独では役に立たなかった調査結果が重要な意味を持つことになるかもしれない。

さらに，研究例では「kwic.xsl」を使ったKWIC（図6）や「タグ集計.xsl」を使った集計表（図8）を紹介したが，XSLTスタイルシートを研究者が新たに作成することで，XML文書内の別の情報を参照したリスト作成や集計も可能である。適用するスタイルシートを変えるだけで新たな観点に立った分析ができることも，この研究手法の特長の一つである。しかし，XSLTに不慣れな研究者にとっては，白紙の状態からのスタイルシートの自作は敷居が高く感じられるかもしれない。そのような場合には，まずは「kwic.xsl」「タグ集計.xsl」等の付属スタイルシートに改造を施して利用するところから始めるのも一つの方法である[14]。

4. おわりに

現在，コンピュータを利用した日本語研究はますます広がりを見せ，XML文書をはじめとするコンピュータ上で扱うことのできる資料の量は膨大なものになりつつある。こうした流れの中で，個人の研究者が，膨大な資料の中から用例を効率よく過不足なく収集し，再利用しやすい形で保存，時々の研究場面に応じて参照できる手法を開発することは，重要な課題の一つと考えられる。本稿ではその一例として，XMLタグ付けプログラムとXSLTスタイルシートを活用した研究手法を紹介した。このような新たな資料と新たな研究手法が，従来とは異なる視点を日本語研究者にもたらし，より深化した研究へとつながることが期待される。

注

1 国立国語研究所の言語コーパス整備計画 KOTONOHA http://www2.kokken.go.jp/kotonoha/

2 青空文庫　http://www.aozora.gr.jp/

3 言語研究におけるXML活用の有効性について日本語で平易に解説したものとして千葉(2006)がある。

4 『太陽コーパス』のほかに『近代女性雑誌コーパス』が公開されている。文書定義は両者共通。http://www.kokken.go.jp/lrc/index.php? 近代女性雑誌コーパス

5 公開場所は国立国語研究所「言語データベースとソフトウェア」http://www.kokken.go.jp/lrc/

6 Internet Explorer 6以降が動作するWindows環境に対応する。

7 Perlスクリプトから実行形式のファイルを生成するパッケージ。http://par.perl.org/

8 W3C Ruby Annotation http://www.w3.org/TR/ruby/

9 インターフェイス・付属XSLTスタイルシートは，A～Cの3属性までの対応であるが，プログラムは最大26属性（Zまで）の埋め込みに対応している。

10 たとえば，タグ付けされた部分をXPathで取得する場合に，開始・終了タグ形式であれば単に「t:tag」と指定できるところを，「t:tag/following-sibling::node()[following-sibling::t:endtag/@id = current()/@id]」などとしてid属性を用いて指定する必要がある。

11 「プリズム」は国立国語研究所「言語データベースとソフトウェア」のページで公開中。

12 本稿で紹介する手法を用いた『太陽コーパス』の二人称代名詞に関する論考は近藤(2007)を参照のこと。

13 「ひまわり」は国立国語研究所「言語データベースとソフトウェア」のページで公開中。

14 小木曽(2005)で『太陽コーパス』対応のXSLTスタイルシートの改造例を紹介している。

参考文献

小木曽智信(2005)「構造化テキストを直接利用するアプリケーション—『プリズム』と『たんぽぽ』—」『雑誌「太陽」における確立期現代語の研究—「太陽コーパス」研究論文集—』，83-113，博文館新社

国立国語研究所(2005)『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』，博文館新社

近藤明日子(2007)「明治末期の二人称代名詞—『太陽コーパス』を資料として—」『日本語日本文学論集』，笠間書院

田中牧郎(2005)「言語資料としての雑誌『太陽』の考察と『太陽コーパス』の設計」『雑誌「太陽」における確立期現代語の研究—「太陽コーパス」研究論文集—』，1-48，博文館新社

千葉庄寿(2006)「構造化された言語データが言語研究にもたらすもの—コーパスを利用する言語研究者の知識基盤としてのXML—」『麗澤大学紀要』82，43-65，麗澤大学

付　記

本稿は，日本学術振興会科学研究費補助金・基盤研究(C)「コーパス言語学の方法に基づく言文一致現象の解析」(2006～2007年度　研究代表者：田中牧郎　研究分担者：岡島昭浩・岡部嘉幸・小木曽智信・近藤明日子）による成果の一部を含む。

（投稿受理日：2007年1月31日）
（最終原稿受理日：2007年6月21日）

小木曽　智信（おぎそ　としのぶ）
国立国語研究所研究開発部門
190-8561　東京都立川市緑町10-2
togiso@kokken.go.jp

近藤　明日子（こんどう　あすこ）
国立国語研究所研究開発部門
190-8561　東京都立川市緑町10-2
kondo@kokken.go.jp

New XML-tagging program for Japanese linguistic study:
Its function and application

OGISO Toshinobu
The National Institute for Japanese Language

KONDO Asuko
The National Institute for Japanese Language

Keywords
XML, structured document, example retrieval, tagging, *Taiyō Corpus*

Abstract

At present, Japanese linguistic resources in XML format are becoming common. It is required to use these resources efficiently for Japanese linguistic studies.

Therefore, we have developed a XML-tagging program *Tanpopo Tagger* which provides new methods for linguistic research using XML documents. This program enables linguists to mark up text strings in any XML document with original tags which have useful attributes. With this program, linguists can fully extract necessary information from the resources for their research. And also, they can save the results of the research as XML tags to reuse them in other studies.

In this paper, we first describe the function and usage of this program and its usage. Next, we show some examples of study using this program, and XSLT style sheets we made for linguistic research. By applying these style sheets to the XML documents tagged by this program, linguists can easily create their original lists or tables of the strings.